



**HAL**  
open science

# LELA: an LLM-based Entity Linking Approach with Zero-Shot Domain Adaptation

Samy Haffoudhi, Fabian M Suchanek, Nils Holzenberger

## ► To cite this version:

Samy Haffoudhi, Fabian M Suchanek, Nils Holzenberger. LELA: an LLM-based Entity Linking Approach with Zero-Shot Domain Adaptation. 2026. <hal-05445830>

**HAL Id: hal-05445830**

**<https://hal.science/hal-05445830v1>**

Preprint submitted on 7 Jan 2026

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization

# LELA: an LLM-based Entity Linking Approach with Zero-Shot Domain Adaptation

Samy Haffoudhi, Fabian M. Suchanek, Nils Holzenberger

Télécom Paris

Institut Polytechnique de Paris, France

{samy.haffoudhi, fabian.suchanek, nils.holzenberger}@telecom-paris.fr

## Abstract

Entity linking (mapping ambiguous mentions in text to entities in a knowledge base) is a foundational step in tasks such as knowledge graph construction, question-answering, and information extraction. Our method, LELA, is a modular coarse-to-fine approach that leverages the capabilities of large language models (LLMs), and works with different target domains, knowledge bases and LLMs, without any fine-tuning phase. Our experiments across various entity linking settings show that LELA is highly competitive with fine-tuned approaches, and substantially outperforms the non-fine-tuned ones.

## 1 Introduction

Entity linking, also known as Entity Disambiguation, is the task of mapping ambiguous mentions of entities in natural language text to reference entities in a knowledge base (KB). Typically, the input is a sentence in which the mention of an entity has already been identified, as in this example:

France hosted the Olympics in Paris.

The knowledge base contains each entity, associated with a short textual description, e.g.:

Paris (city)	Capital city of France
Paris (novel)	1897 novel by Emile Zola
France	Country in Europe

The goal is to map the mention of Paris to the correct entity in the KB (here, the first one).

Entity Linking is an important preprocessing step in tasks such as information extraction (Martinez-Rodriguez et al., 2020), knowledge-based question answering (Welty et al., 2012), and knowledge graph completion (Ji and Grishman, 2011). Most of the work on entity linking focuses on linking to general knowledge graphs such as Wikidata (Vrandečić, 2012), DBpedia (Auer et al., 2007) or Yago (Suchanek et al., 2007). However,

	Domain	Knowledge Base	Documents
Classical	test=train	test=train	test≠train
Zero-shot	test≈train	test≠train	test≠train
True zero-shot	test≠train	test≠train	test≠train

Figure 1: Classification of entity linking settings (“training” refers to fine-tuning after initial pre-training). LELA is in the *true zero-shot* setting.

in real-world applications, the knowledge base is often proprietary or domain-specific, as in the legal or biomedical domain, or inside a company.

Therefore, several entity linking methods use training. In **classical entity linking** (Figure 1), the models are trained on the same domain and same knowledge base as the one used during testing. In **zero-shot entity linking** (Logeswaran et al., 2019), the set of entities seen during training is different from the one seen during testing. However, even this setting, as defined in Logeswaran et al. (2019), requires labeled entity linking training data for fine-tuning from related domains and document sources. Fine-tuning is not just computationally expensive, but also requires high-quality labeled data. Thus, to apply entity linking to a new domain (say, a company wants to disambiguate chemical products), one would first have to label hundreds of mentions. This task cannot be outsourced easily, as it requires navigating specialized knowledge bases and involves privacy restrictions. Hence, the quest that we tackle in this paper is **true zero-shot entity linking**, i.e., entity linking that works out of the box for any domain without the need for fine-tuning.

Several approaches target such true zero-shot scenarios (Ding et al., 2024; Liu et al., 2024c; Zhou et al., 2024; Ye and Mitchell, 2025). However, even these require hand-chosen Chain-of-Thought examples and/or strongly domain-specific candidate generator approaches, limiting their generalizabil-

ity. Furthermore, their performance lags behind the specialized approaches.

In this paper, we propose LELA, a *coarse-to-fine* fine-tuning-free method for entity linking. LELA (1) generates an initial entity-candidate set, (2) selects top-k candidates through pointwise reranking, and (3) passes them to the LLM for final candidate selection through reasoning over the context and entity descriptions. This approach is modular, LLM-, reranker-, and retriever-agnostic, and can hence be used with both open-source and proprietary models. As LLMs become ever more of a commodity, the performance of LELA thus only stands to gain. Through extensive experiments, we show that LELA systematically outperforms the true zero-shot methods, and even beats state-of-the-art classical and zero-shot approaches on multiple datasets. We attach our code for reviewing, which will be made available as an open-source library.

## 2 Related Work

### 2.1 Entity Linking

Before the blossoming of machine learning and deep learning, many entity linking approaches (such as AIDA (Hoffart et al., 2011) and SOFIE (Suchanek et al., 2009)) were training-free, but were bound to a given knowledge base.

Today, methods that work on any input KB are usually embedding-based: BLINK (Wu et al., 2020), AcroBERT (Chen et al., 2023), ExtEnD (Barba et al., 2022), ReFinED (Ayoola et al., 2022), VERBALIZED (Rücker and Akbik, 2025), the approaches in Févry et al. (2020); Yamada et al. (2022), as well as the different zero-shot entity linking methods evaluated on the ZESHEL benchmark (Logeswaran et al., 2019): ReS (Xu et al., 2023), Uni-MPR and Bi-MPR (Tang et al., 2021), and E-repeat (Yao et al., 2020). GENRE (De Cao et al., 2021a) is a generative approach and mGENRE (De Cao et al., 2022) is its multilingual extension. Although these methods can technically operate over different knowledge bases, they are not true zero-shot approaches, because they require supervised training on the target KB or closely aligned KBs to achieve good performance. LELA obviates this requirement, and we show in our experiments that it still performs competitively with these approaches.

Several LLM-based approaches have also been proposed. The approaches in Vollmers et al.; Xin et al. (2024) use LLMs for contextual augmentation.

Li et al. (2025) categorizes mentions and routes them to either a low-computational entity linker or a reasoning-LLM. Wang et al. (2026); Shi et al. (2024); Liu et al. (2024b) focus on the multimodal setting. GenDecider (Zhou et al., 2024) uses an LLM fine-tuned on the reranking task on ZESHEL using LoRA (Hu et al., 2021). OneNet (Liu et al., 2024c) is a pipeline approach of LLM-based candidate filtering, entity linking, and hallucination tackling. EntGPT (Ding et al., 2024) prompts the model to augment the mention before selecting the best candidate entity and Ye and Mitchell (2025) uses in-context learning (Brown et al., 2020) for candidate selection.

All of these LLM-based approaches leverage task- or domain-specific candidate retrieval, fine-tuning, or complex prompt engineering, which limits their applicability in true zero-shot settings. In contrast, LELA requires no fine-tuning or domain-specific components, and our experiments show that it outperforms these approaches across multiple datasets, in a true zero-shot setting.

### 2.2 LLM-based Reranking

LLMs have recently been explored as general-purpose rerankers in information retrieval: pointwise approaches compute a score for each query-document pair (Pipitone et al., 2025; Zhang et al., 2025); pairwise approaches compare the relevance of a pair of documents (Qin et al., 2024; Luo et al., 2024); and listwise approaches process a list of documents and output them ranked by their relevance (Wang et al., 2025; Liu et al., 2025b). Other methods are setwise (Zhuang et al., 2024), tournament-inspired (Chen et al., 2025b), efficiency-centered (Dedov, 2025), as well as attention-pattern-based (Chen et al., 2025a) or conditional perplexity-based (Shi et al., 2025). Some approaches make use of reasoning token generation (Abdallah et al., 2025; Liu et al., 2025a; Zhuang et al., 2025), and study its impact on performance (Jedidi et al., 2025).

While this body of work primarily targets document and passage retrieval, we demonstrate that pointwise LLM-based reranking, followed by reasoning-driven selection, can be effectively adapted to the entity linking setting without task-specific fine-tuning.

### 3 Methodology

**Task Definition.** Let the KB  $\mathcal{E}$  be a set of entities, each with a textual description. Given a mention  $m$  in a text  $S$ , the goal of entity linking is to find the entity  $e \in \mathcal{E}$  that  $m$  refers to in the context of  $S$ , or, optionally, return a special entity  $e_{\text{none}}$ , which indicates that no entity is relevant.

LELA uses a three-stage *coarse-to-fine* pipeline, as illustrated in Figure 2: Candidate entities are first generated (Sevgili et al., 2022; Wu et al., 2020), then ranked, and then selected. Beyond the text, mention, and KB, LELA requires only a parameter  $k \in \mathbb{N}^+$  as input.

**Candidate generation.** The goal of the candidate generation step is to identify the entities  $C_m \subseteq \mathcal{E}$  that could potentially be relevant to  $m$ . For this step, LELA is retriever-agnostic, and can work with any method that is capable of pre-selecting a subset of relevant entities. In practice, any generation method that requires domain-specific training would jeopardize LELA’s goal of being fine-tuning-free. Hence, we primarily consider BM25 (Robertson and Sparck Jones, 1988) and dense retrieval, and also experiment with prior-based dictionaries.

**Candidate reranking.** Each candidate  $c \in C_m$  is then given a score  $s(c) \in \mathbb{R}$  by pointwise reranking. While any off-the-shelf reranker can be used, we opt for instructed rerankers, as the candidate entity reranking task is not typically covered during reranker training. Rerankers based on generative language models typically take as input a query  $q$ , a document  $d$ , and a reranking-instruction  $I$ . They then compute a score as:

$$\begin{aligned} \text{score}(q, d) &= \frac{e^{\text{logit}(\text{yes}|I,q,d)}}{e^{\text{logit}(\text{yes}|I,q,d)} + e^{\text{logit}(\text{no}|I,q,d)}} \\ &= \text{softmax} \left( \frac{\text{logit}(\text{yes}|I,q,d)}{\text{logit}(\text{no}|I,q,d)} \right)_0 \end{aligned}$$

The score can also be expressed as a single-logit binary classifier<sup>1</sup>. In our case, we want to obtain a score for each candidate  $c \in C_m$ , and hence, following the formalism of information retrieval, we set  $q = S$  (marking the mention  $m$ ),  $I$  as an entity disambiguation instruction (see Appendix A), and  $d$  as the name of  $c$  followed by its description. We then identify the top- $k$  scored candidates,  $C_m^{(k)}$ .

<sup>1</sup><https://huggingface.co/Qwen/Qwen3-Reranker-0.6B/discussions/3>

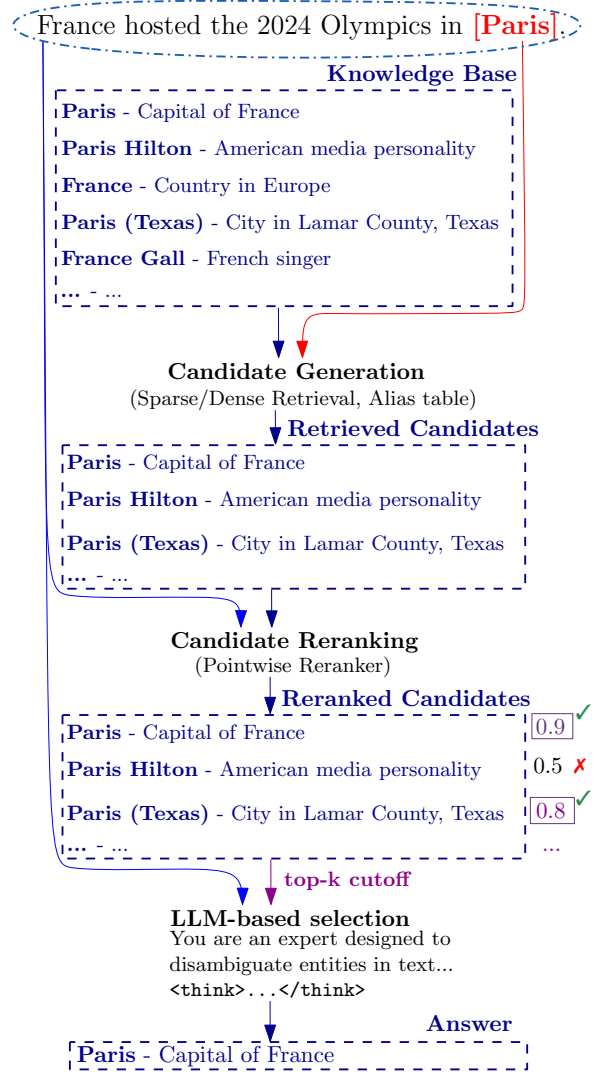


Figure 2: Overview of the LELA approach.

**Candidate selection.** To select the best candidate from  $C_m^{(k)}$ , we construct an LLM prompt that consists of (1) a system instruction that asks the model to return the index of the entity that is most likely mentioned in the text (shown in Appendix A), (2) input context  $S$  with the mention  $m$  surrounded in square brackets, and (3) the candidates as a newline-separated list of triples, each consisting of a zero-based index, an entity name, and an entity description. Optionally, the first list item is “0. None of the candidates”. We use self-consistency (Wang et al., 2023), sampling different candidate selection reasoning paths, with majority voting to determine the final candidate  $c^* \in C_m^{(k)}$ .

**Implementation** LELA uses vLLM<sup>2</sup> (Kwon et al., 2023) for LLM-inference. It is implemented as a Python library and can be used easily in spaCy

<sup>2</sup><https://docs.vllm.ai/>

pipelines as an entity-linking component<sup>3</sup>.

## 4 Experiments

We evaluate LELA on various entity linking settings, using five different benchmark datasets. Our evaluation spans four complementary entity linking settings: (i) domain-transfer zero-shot, (ii) domain-specific with None-linking, (iii) acronym disambiguation without entity descriptions across specialized domains, and (iv) classical Wikipedia-based benchmarks with strong supervised baselines. Each considered benchmark emphasizes a distinct challenge, and together they demonstrate LELA’s robustness beyond any single scenario.

**LELA configuration.** We report results for LELA using the open-source Magistral-Small-2509 (Mistral-AI et al., 2025) reasoning LLM. For all datasets and models, we use  $k = 10$  and sample 10 outputs for self-consistency. We use the Qwen3-Reranker-4B model (Zhang et al., 2025) for pointwise reranking. We run our experiments on two A100 40GB GPUs. As we are interested in true zero-shot entity linking, we do not perform any hyperparameter tuning and discuss other values of  $k$ , self-consistency samples and rerankers in Section 4.5 and 4.6. For dense retrieval, we use Qwen3-Embedding-4B (Zhang et al., 2025) along with flat FAISS indexes (Douze et al., 2025). BM25 retrieval is implemented using the bm25s Python library<sup>4</sup>. In addition, we also report results for the mixture-of-experts (MoE) (Shazeer et al., 2017) Qwen3-30B-A3B, and the small Qwen3-4B (Yang et al., 2025) reasoning LLMs.

**Competitors.** We compare LELA to all state-of-the-art entity linking methods that have been previously evaluated on the considered benchmarks. The same method can be zero-shot for one dataset and true zero-shot for another one.

### 4.1 Zero-Shot Scenario

**Benchmark.** ZESHEL (Logeswaran et al., 2019) is a benchmark for evaluating entity linking methods in a zero-shot setting, built with documents from Wikia<sup>5</sup>, a Wikipedia-like Web site for specific subjects such as fictional universes or sports. ZESHEL consists of 8 Wikia domains for training,

4 for validation, and 4 for testing. The test has a total of 10,000 mentions. For each mention, the dataset provides 64 candidates retrieved by BM25. Entities correspond to Wikia pages, where the entity name is the title of the page and the entity description is its content. The BM25-based candidate retrieval has a top-64 recall of 68% on the test set. Following previous work (Logeswaran et al., 2019; Xu et al., 2023; Li et al., 2022), we consider the *normalized accuracy* as the evaluation metric, which is defined as the accuracy on the subset of mentions for which the correct entity is among the top-64 candidates. For LELA, we use the provided BM25 candidates and use the first three sentences of the entity page as the entity description.

**Results.** As Table 1 shows, LELA ranks first on the benchmark. It achieves the highest performance across all test domains in the true zero-shot setting, improving the previous best reported macro-averaged normalized accuracy by 8.84 percentage points. On half of the domains, it outperforms even the non-true zero shot approaches, improving the normalized accuracy by 1.62 and 5.06 percentage points respectively. Category-specific results are given in Appendix B.

### 4.2 Domain-specific and None Scenarios

**Benchmark.** The ESCO benchmark (le Vrang et al., 2014) is designed for the task of mapping occupational skills to the ESCO taxonomy, evaluated on the manually annotated dataset from Decorte et al. (2022). Some mentions have to be mapped to the special entity UNK, because they have no counterpart in the taxonomy (813 out of the 1824 test mentions).

**LELA configuration.** We generate our own candidates either with BM25 (retrieving 200 candidates per mention) or with dense retrieval (retrieving 64 candidates per mention). BM25 yields a recall of 80.8%, and dense retrieval yields 95.6%. We use the provided entity descriptions.

**Results.** Table 2 shows that, for the accuracy on actual entities, LELA performs on par with the classical approaches. In the overall accuracy, the baseline that always predicts UNK has the highest accuracy, which is due to the high ratio of UNK mentions. LELA still outperforms the other true zero-shot competitors, as well as the classical approaches. Dense retrieval of candidates allows for higher non-UNK accuracy across the models, while

<sup>3</sup><https://spacy.io/>

<sup>4</sup><https://bm25s.github.io/>

<sup>5</sup><https://www.wikia.com/>

Method	Type	Forgotten Realms	Lego	Star Trek	YuGiOh	Macro Acc.
Baseline** (Logeswaran et al., 2019)	zero-shot	-	-	-	-	77.05
BLINK* (Wu et al., 2020)	zero-shot	86.80	76.39	77.95	67.46	77.15
E-repeat** (Yao et al., 2020)	zero-shot	-	-	-	-	79.64
Uni-MPR** (Tang et al., 2021)	zero-shot	87.25	78.57	80.56	67.31	78.42
Bi-MPR* (Tang et al., 2021)	zero-shot	89.60	80.50	81.04	68.74	79.97
ReS* (Xu et al., 2023)	zero-shot	88.10	78.44	81.69	75.84	81.02
GenDecider* (Zhou et al., 2024)	zero-shot	<u>91.10</u>	<b>81.42</b>	<b>82.48</b>	<b>75.99</b>	<b>82.75</b>
BLINK*** (Wu et al., 2020)	true zero-shot	84.11	74.45	72.22	66.30	74.27
ExtEnD* (Barba et al., 2022)	true zero-shot	79.62	65.20	73.21	60.01	69.51
GENRE** (De Cao et al., 2021b)	true zero-shot	55.20	42.71	55.76	34.68	47.09
OneNet (Liu et al., 2024c)	true zero-shot	55.80	53.80	53.90	40.80	51.08
<b>LELA (ours)</b>	true zero-shot					
w/ Magistral-Small-2509		<b>88.69</b>	<b>83.04</b>	<b>79.64</b>	<b>81.05</b>	<b>83.11</b>
w/ Qwen3-30B-A3B		85.69	79.14	72.68	69.65	76.79
w/ Qwen3-4B		80.08	67.93	65.28	58.69	68.00
w/o LLM selection		31.23	22.10	30.56	24.70	27.15

Table 1: Normalized accuracy on the test set of ZESHEL. \* means reported in (Zhou et al., 2024), \*\* means reported in Xu et al. (2023), \*\*\* means reported in Li et al. (2022), - means not reported. Best results within each group are bold, best overall results underlined. Ablations of LELA in gray (see Section 4.5).

BM25 results in higher UNK accuracy, and thus in higher general accuracy given the high UNK proportion, but overall, the choice of the retrieval method does not lead to statistically significant differences in accuracy.

### 4.3 Scenario of Acronym Disambiguation

**Benchmark.** GLADIS (Chen et al., 2023) is a benchmark for the task of mapping acronyms to their inferred long form, spanning three domains: the general, scientific and the biomedical domain, with, respectively, 7125, 14066 and 2909 test mentions. The average number of candidates per acronym is 200.

**LELA configuration.** We use the short-form to long-form mapping for candidate generation, and use the long-forms as entity descriptions.

**Competitors.** Classical entity linking approaches have not been applied to the acronym disambiguation task. Rather, the original paper proposed AcroBERT, a method that can be run both in a classical and in a zero-shot fashion. Since it outperformed alternative approaches by a wide margin (usually by a factor of two), we compare only to AcroBERT and the baselines in these categories. For the true zero shot group, we use GENRE (De Cao et al., 2021b) and the BLINK cross-encoder (Wu et al., 2020) (not trained on the dataset).

**Results.** Table 3 shows that LELA outperforms all true zero-shot competitors by a huge margin of up to 54 percentage points. It also outperforms all other competitors, whether or not they were

fine-tuned on the training data, on all domains except the biomedical one. LELA improves the best-reported macro-accuracy by 4.6 percentage points, and the scientific domain accuracy by 13.9 percentage points.

### 4.4 Wikipedia-based Scenario

**Benchmark.** ZELDA (Milich and Akbik, 2023) groups several established Wikipedia-based classical entity linking benchmarks. It aims at allowing for a fair comparison of supervised methods by mapping all mentions to the same Wikidata knowledge base and by providing a common Wikipedia-based train set and candidate dictionary to be used across all benchmarks and methods. We also evaluate LELA on the WikilinksNED Unseen Mentions (Onoe and Durrett, 2020) benchmark and report the results in Appendix C.

**LELA configuration.** We use the provided candidates and entity descriptions.

**Competitors.** We compare LELA to all entity linking methods that have been trained on ZELDA (Rücker and Akbik, 2025; Milich and Akbik, 2023). These supervised approaches benefit from strong alignment with Wikipedia-based training, which favors memorization-based linking. This makes ZELDA a stress-test for LELA.

**Results.** Table 4 shows that the best-performing methods are FusionED and VERBALIZED<sub>it</sub><sup>6</sup>. In this setting, LELA cannot achieve the performance

<sup>6</sup>VERBALIZED does not use the provided candidate dictionary.

Method	Type	Accuracy	UNK Accuracy	Non-UNK Accuracy
Predict-UNK baseline	baseline	<b><u>44.57</u></b>	<b><u>100.0</u></b>	0.0
Random baseline	baseline	0.22	-	-
TF-IDF baseline	baseline	2.25	-	-
BLINK Wiki+ESCO (Wu et al., 2020)	classical	<b>23.55</b> $\pm 1.95$	1.38 $\pm 0.81$	<b>41.38</b> $\pm 3.03$
GENRE Wiki+ESCO (De Cao et al., 2021b)	classical	11.48 $\pm 1.46$	<b>1.65</b> $\pm 0.88$	19.39 $\pm 2.43$
BLINK Wiki (Wu et al., 2020)	true zero-shot	23.30 $\pm 1.94$	-	-
GENRE Wiki (De Cao et al., 2021b)	true zero-shot	6.91 $\pm 1.16$	-	-
<b>LELA (ours)</b>	true zero-shot			
<b>w/ Magistral-Small-2509 &amp; Dense Retrieval</b>		<b>26.75</b> $\pm 2.03$	5.41 $\pm 1.56$	<b>43.92</b> $\pm 3.06$
<b>w/ Magistral-Small-2509 &amp; BM25</b>		26.21 $\pm 2.02$	<b>7.26</b> $\pm 1.79$	41.44 $\pm 3.04$
w/ Qwen3-30B-A3B & Dense Retrieval		36.62 $\pm 2.21$	28.04 $\pm 3.09$	43.52 $\pm 3.06$
w/ Qwen3-30B-A3B & BM25		39.36 $\pm 2.24$	39.11 $\pm 3.36$	39.56 $\pm 3.02$
w/ Qwen3-4B & Dense Retrieval		26.59 $\pm 2.03$	8.36 $\pm 1.91$	41.25 $\pm 3.04$
w/ Qwen3-4B & BM25		29.44 $\pm 2.09$	14.88 $\pm 2.45$	41.15 $\pm 3.04$
w/o LLM selection & Dense Retrieval		-	-	35.71 $\pm 2.96$
w/o LLM selection & BM25		-	-	34.92 $\pm 2.94$

Table 2: Accuracy on the ESCO entity linking benchmark. Accuracy and UNK accuracy from Zhang et al. (2024) (Non-UNK accuracy is derived), best per category in bold, globally best underlined. We report 95% confidence intervals to account for the smaller test set size. Ablations of LELA in gray (see Section 4.5).

Method	Type	General	Scientific	Biomedical	Macro Acc.
BM25 baseline (Chen et al., 2023)	baseline	25.8	10.7	14.3	16.2
Popularity baseline (Chen et al., 2023)	baseline	53.2	12.3	54.0	38.8
AcroBERT (fine-tuned) (Chen et al., 2023)	classical	<b>76.1</b>	<b>27.6</b>	<b>69.2</b>	<b>57.7</b>
AcroBERT (Chen et al., 2023)	zero-shot	<b>72.0</b>	<b>27.4</b>	<b>61.4</b>	<b>53.6</b>
GENRE (De Cao et al., 2021b) (our implementation)	true zero-shot	63.9	3.9	7.1	25.0
BLINK (Wu et al., 2020) (our implementation)	true zero-shot	15.8	1.5	1.8	6.4
<b>LELA (ours)</b>	true zero-shot				
<b>w/ Magistral-Small-2509</b>		<b>78.9</b>	<b>46.9</b>	<b>61.0</b>	<b>62.3</b>
w/ Qwen3-30B-A3B		80.1	45.5	61.5	62.4
w/ Qwen3-4B		73.8	37.7	54.3	55.3
w/o LLM selection		44.2	22.7	22.6	29.8

Table 3: Accuracy on the GLADIS benchmark. Best in each group in bold (AcroBERT was previously shown to outperform classical and fine-tuned methods), globally best underlined. Ablations of LELA in gray (see Section 4.5).

of the classical methods. However, it remains competitive, performing on par with GENRE, despite being the only method that operates without supervised training signals.

#### 4.5 Ablation studies

Tables 1, 2, 3, and 4 show that our default model Magistral-Small-2509 stands out on ZESHEL, where differences in performance between the models are the largest. On the other datasets, Qwen3-30B-A3B catches up. It also handles None prediction better. On the more general domains, the smaller Qwen3-4B also performs well. We discuss these differences in more detail in Section 4.6. Overall, our analysis shows that the performance of LELA is robust to the choice of the underlying LLM.

The tables show also that omitting the selection LLM (“w/o LLM selection”) lowers the results

drastically on domain-specific datasets, indicating that LELA’s joint consideration of candidates is vital. We study several other ablations on the ZESHEL benchmark:

**No Reranker:** We pass all retrieved candidates to the LLM without reranking nor cutoff.

**No Self-Consistency:** We sample only one LLM answer.

**No Reasoning:** We skip thinking token generation and prompt the selection LLM to directly output an answer.

**No Descriptions:** We pass empty entity descriptions to the selection LLM.

Table 5 shows that **removing the reranker** results in a drop in performance (in addition to a longer runtime and increased cost given the longer contexts). While the drop in performance is significant for Qwen3-30B-A3B and Qwen3-4B, it is however negligible for Magistral-Small-2509, which is

	AIDA	Tweeki	Reddit		WNED		Slinks			Avg
			posts	comments	CWEB	Wiki	Tail	Shadow	Top	
Upper bound* (Milich and Akbik, 2023)	91.1	94.0	98.4	98.3	92.4	98.8	98.8	56.7	73.1	89.1
MFS baseline* (Milich and Akbik, 2023)	63.5	72.3	83.4	81.0	61.2	65.1	99.4	14.9	41.3	64.7
<b>Classical (with training data)</b>										
FEVRY <sub>CL</sub> * (Févy et al., 2020)	79.5	76.9	89.0	86.5	70.3	84.5	87.6	31.9	47.7	72.7
LUKE <sub>FT</sub> * (Yamada et al., 2022)	81.2	77.9	81.5	78.5	70.3	76.5	98.0	22.5	51.8	71.0
GENRE <sub>CL</sub> * (De Cao et al., 2021b)	78.6	80.1	92.8	91.5	<b>73.6</b>	88.4	<b>99.6</b>	37.3	52.8	77.2
FusionED (Wang et al., 2024)	80.1	<b>81.4</b>	<b>93.9</b>	<b>92.3</b>	<b>73.6</b>	89.0	98.3	41.5	57.9	78.7
BLINK <sub>bi</sub> (Rücker and Akbik, 2025)	65.5	72.1	83.1	79.1	58.1	73.1	96.3	41.8	42.6	68.0
VERBALIZED <sub>IT</sub> (Rücker and Akbik, 2025)	<b>88.2</b>	78.9	92.2	88.4	71.5	<b>90.8</b>	98.2	<b>66.3</b>	<b>65.9</b>	<b>82.3</b>
<b>True Zero-Shot (without training data)</b>										
<b>LELA (ours)</b>										
w/ Magistral-Small-2509	<b>83.3</b>	<b>79.1</b>	<b>80.7</b>	<b>79.5</b>	<b>66.7</b>	<b>87.5</b>	<b>97.2</b>	<b>48.9</b>	<b>66.7</b>	<b>76.6</b>
w/ Qwen3-30B-A3B	84.2	79.1	83.9	84.6	64.8	87.5	97.3	48.8	67.0	77.5
w/ Qwen3-4B	84.0	78.0	84.1	82.6	66.3	85.1	97.1	48.2	65.0	76.7
w/o LLM selection	76.4	75.1	79.0	79.9	65.0	76.2	97.0	43.6	60.6	72.5

Table 4: Accuracy on the ZELDA benchmark. \* means reported in (Milich and Akbik, 2023). Best in each group in bold, globally best underlined. Ablations of LELA in gray (see Section 4.5).

due to the fact that this larger model is better at handling a large number of candidates, resulting in a slower drop in selection accuracy as  $k$  increases, which can thus be offset by the increased candidate retrieval recall (see Appendix D). The **No Self-Consistency** variant, too, systematically results in a drop in performance (more details in Appendix E). **Disabling reasoning** results in the next biggest drop in performance across the three models. This shows that simply considering the candidates together is not sufficient and that context-based reasoning is needed. Finally, the **No description variant** provides the selection LLM with less precise entities, and it also performs worse in all configurations. Together these results show that all components of LELA contribute to its performance. **Examples** of candidate selection outputs for different variants are given in Appendix J.

#### 4.6 Further Analyses

##### The Recall vs Selection Accuracy Trade-off.

The value of  $k$  results in a trade-off between recall and selection accuracy (Appendix D): lower values allow for a high selection accuracy, but may result in the correct entity not being among the candidates, while higher values reduce the selection accuracy but improve the candidate recall. The impact of  $k$  on these two components depends on the reranker and selection LLM in use. Overall,  $k = 10$  is a safe choice, with lower values leading to a substantial

degradation in recall (see Appendix I for a detailed visualization of recall retention across stages).

**Self-consistency.** Given the impact of self-consistency highlighted in Section 4.5, we further study its impact over different numbers of sampled answers (see Appendix E). Notably, we observe that while performance increases with the number of samples, an accuracy plateau is reached quickly across models, suggesting that as little as three samples are optimal.

**Positional Bias.** LLMs have been shown to be subject to positional bias (Liu et al., 2024a). In Appendix F, we shuffle the list of candidates given to the LLM. Larger models appear robust to this bias, while smaller models are more sensitive to it. Overall, the results justify our approach of sorting the candidates according to their reranker-assigned score in LELA.

**Runtime and Scalability.** In Appendix G, we observe that selection accuracy decreases as the input context length increases, with a model-dependent rate of decline. We also notice traces of *underthinking* and *overthinking* behaviors in reasoning LLMs (Su et al., 2025).

In addition, we report the average mention processing time across benchmarks. Our measurements indicate that while LELA is naturally more computationally intensive than shallow retrieval methods, it remains within a respectable runtime

Method	Forgotten Realms	Lego	Star Trek	YuGiOh	Macro Acc.
<b>LELA</b>					
w/ Magistral-Small-2509	<b>88.69</b>	83.04	79.64	81.05	<b>83.11</b>
w/o reranker	87.79	<b>83.45</b>	<b>80.00</b>	<b>81.15</b>	83.10
w/o self-consistency	81.38	76.98	73.29	74.28	76.48
w/o descriptions	76.58	71.22	72.53	71.24	72.89
w/ Qwen3-30B-A3B	<b>85.69</b>	<b>79.14</b>	<b>72.68</b>	<b>69.65</b>	<b>76.79</b>
w/o reranker	83.58	74.31	67.07	64.44	72.35
w/o self-consistency	82.08	73.48	65.39	63.18	71.03
w/o descriptions	71.67	62.28	65.92	60.64	65.13
w/o reasoning	65.97	53.24	52.06	45.64	54.23
w/ Qwen3-4B	<b>80.08</b>	<b>67.93</b>	<b>65.28</b>	<b>58.69</b>	<b>68.00</b>
w/o reranker	75.78	60.43	61.01	51.14	62.09
w/o self-consistency	77.48	62.49	60.25	53.82	63.51
w/o descriptions	66.87	54.16	61.08	51.57	58.42
w/o reasoning	51.55	36.79	36.45	35.75	40.14

Table 5: Normalized accuracy on the test set of ZESHEL across LELA variants. The results for the *w/o reasoning* variant are reported for the two Qwen models which explicitly support switching between reasoning and non-reasoning mode.

envelope for offline and batch processing. For example, on ZESHEL, we observe a processing time of 44 seconds per mention on average<sup>7</sup>. The MoE Qwen3-30B-A3B offers the most compelling trade-off of speed vs. performance, achieving inference speeds faster than the smaller Qwen3-4B model (26 vs 31 seconds on ZESHEL) while significantly outperforming it and often matching the performance of Magistral-Small-2509.

**Reranker performance.** We compare the performance of different rerankers in Appendix H. Other models can achieve a recall@k similar to the one of our default model Qwen3-Reranker-4B, showing that LELA is not tied to a specific reranker.

#### 4.7 Discussion

Our results suggest that LLM-based candidate reranking and selection can effectively compensate for the absence of labeled training data, and even outperform the fine-tuned approaches. Our ablation studies suggest that these gains stem primarily from the ability of LLMs to jointly reason over multiple candidates rather than from candidate ranking alone. The fact that the candidate ranking alone does not perform well also mitigates concerns regarding data contamination: If the models were simply recalling memorized gold labels from pre-training, the reranking alone would nearly always find the best candidate – which it does not.

Our work points to a fundamental shift in how entity linking can be approached in resource-constrained environments. While traditional methods optimize for inference efficiency, front-loading

costs onto human annotation, LELA demonstrates that allocating computational budget to test-time reasoning can effectively substitute for domain-specific supervision. This suggests that the ability for disambiguation is a general capability of reasoning models that transfers across settings, unlike the distributional patterns learned during fine-tuning. Consequently, the usual barrier to entry for entity linking in specialized domains (viz. the availability of labeled data) could be mitigated with additional inference compute.

## 5 Conclusion

We have proposed LELA, a simple and modular *coarse-to-fine* approach to entity linking that leverages the capabilities of LLMs. LELA is *true zero-shot*, in the sense that it can be applied to a given knowledge base and domain without requiring annotated data or fine-tuning. Our experiments show that LELA outperforms other true zero-shot approaches by a large margin. It performs on par, and often beats, fine-tuned and classical approaches, even though these require training data.

Our results suggest a blind spot in entity linking, where inference compute can make up for training data. We thus hope that LELA opens the door for a wider use of LLMs in entity linking, and for new approaches that perform the task without any need for fine-tuning. In particular, we encourage LLMs to be included as baselines in entity linking studies. Future work could extend LELA to end-to-end entity linking, where mentions have to be detected as well.

<sup>7</sup>With 10 samples for self-consistency.

## 6 Limitations

Being LLM-based and model-agnostic, LELA directly benefits from evolving model capabilities, but also inherits their limitations. In particular, the generation of reasoning tokens makes LELA slower than specialized entity linking models and thus more suitable for offline processing. Additionally, LELA inherits LLMs' limited interpretability.

While our evaluation spans a broad range of benchmarks, it cannot cover every possible knowledge base or domain. Although context-based reasoning improves generalization, further research is required to assess LELA's performance across additional settings. In particular, our evaluation focuses on English benchmarks, and evaluating LELA in multilingual or low-resource settings remains an important direction for future work.

## References

- Abdelrahman Abdallah, Jamshid Mozafari, Bhawna Piryani, and Adam Jatowt. 2025. [DeAR: Dual-Stage Document Reranking with Reasoning Agents via LLM Distillation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5710–5723, Suzhou, China. Association for Computational Linguistics.
- Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. DBpedia: A nucleus for a web of open data. In *Proceedings of the 6th International The Semantic Web and 2nd Asian Conference on Asian Semantic Web Conference, ISWC'07/ASWC'07*, pages 722–735, Berlin, Heidelberg. Springer-Verlag.
- 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, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.
- Edoardo Barba, Luigi Procopio, and Roberto Navigli. 2022. [ExtEnD: Extractive Entity Disambiguation](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2478–2488, Dublin, Ireland. Association for Computational Linguistics.
- Hannah Bast, Matthias Hertel, and Natalie Prange. 2022. [ELEVANT: A Fully Automatic Fine-Grained Entity Linking Evaluation and Analysis Tool](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 72–79, Abu Dhabi, UAE. Association for Computational Linguistics.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20*, pages 1877–1901, Red Hook, NY, USA. Curran Associates Inc.
- Lihu Chen, Gael Varoquaux, and Fabian M. Suchanek. 2023. [GLADIS: A General and Large Acronym Disambiguation Benchmark](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2073–2088, Dubrovnik, Croatia. Association for Computational Linguistics.
- Shijie Chen, Bernal Jiménez Gutiérrez, and Yu Su. 2025a. [Attention in Large Language Models Yields Efficient Zero-Shot Re-Rankers](#). *Preprint*, arXiv:2410.02642.
- Yiqun Chen, Qi Liu, Yi Zhang, Weiwei Sun, Xinyu Ma, Wei Yang, Daiting Shi, Jiabin Mao, and Dawei Yin. 2025b. [TourRank: Utilizing Large Language Models for Documents Ranking with a Tournament-Inspired Strategy](#). In *Proceedings of the ACM on Web Conference 2025, WWW '25*, pages 1638–1652, New York, NY, USA. Association for Computing Machinery.
- Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021a. [Highly Parallel Autoregressive Entity Linking with Discriminative Correction](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7662–7669, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021b. [Autoregressive Entity Retrieval](#). *Preprint*, arXiv:2010.00904.
- 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.
- Jens-Joris Decorte, Jeroen Van Hautte, Johannes Deleu, Chris Develder, and Thomas Demeester. 2022. [Design of Negative Sampling Strategies for Distantly Supervised Skill Extraction](#). *Preprint*, arXiv:2209.05987.
- Evgeny Dedov. 2025. [JointRank: Rank Large Set with Single Pass](#). In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*, pages 208–217.

- Yifan Ding, Amrit Poudel, Qingkai Zeng, Tim Weneringer, Balaji Veeramani, and Sanmitra Bhat-tacharya. 2024. [EntGPT: Linking Generative Large Language Models with Knowledge Bases](#). *Preprint*, arXiv:2402.06738.
- Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvassy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2025. [THE FAISS LIBRARY](#). *IEEE Transactions on Big Data*, pages 1–17.
- Yotam Eshel, Noam Cohen, Kira Radinsky, Shaul Markovitch, Ikuya Yamada, and Omer Levy. 2017. [Named Entity Disambiguation for Noisy Text](#). In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 58–68, Vancouver, Canada. Association for Computational Linguistics.
- Thibault Févry, Nicholas FitzGerald, Livio Baldini Soares, and Tom Kwiatkowski. 2020. [Empirical Evaluation of Pretraining Strategies for Supervised Entity Linking](#). *Preprint*, arXiv:2005.14253.
- Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenauf, 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.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. [LoRA: Low-Rank Adaptation of Large Language Models](#). *Preprint*, arXiv:2106.09685.
- Nour Jedidi, Yung-Sung Chuang, James Glass, and Jimmy Lin. 2025. [Don't "Overthink" Passage Reranking: Is Reasoning Truly Necessary?](#) *Preprint*, arXiv:2505.16886.
- Heng Ji and Ralph Grishman. 2011. [Knowledge base population: Successful approaches and challenges](#). In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 1148–1158.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. [Efficient Memory Management for Large Language Model Serving with PagedAttention](#). In *Proceedings of the 29th Symposium on Operating Systems Principles, SOSP '23*, pages 611–626, New York, NY, USA. Association for Computing Machinery.
- Martin le Vrang, Agis Papanioniou, Erika Pauwels, Pieter Fannes, Dominique Vandesteene, and Johan De Smedt. 2014. [ESCO: Boosting Job Matching in Europe with Semantic Interoperability](#). *Computer*, 47(10):57–64.
- Xiuxing Li, Zhenyu Li, Zhengyan Zhang, Ning Liu, Haitao Yuan, Wei Zhang, Zhiyuan Liu, and Jianyong Wang. 2022. [Effective Few-Shot Named Entity Linking by Meta-Learning](#). In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, pages 178–191. IEEE Computer Society.
- Yajie Li, Albert Galimov, Mitra Datta Ganapaneni, Pujitha Thejaswi, De Meng, Priyanshu Kumar, and Saroni 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, Suzhou (China). Association for Computational Linguistics.
- Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024a. [Lost in the Middle: How Language Models Use Long Contexts](#). *Transactions of the Association for Computational Linguistics*, 12:157–173.
- Qi Liu, Yongyi He, Tong Xu, Defu Lian, Che Liu, Zhi Zheng, and Enhong Chen. 2024b. [UniMEL: A Unified Framework for Multimodal Entity Linking with Large Language Models](#). In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM '24*, pages 1909–1919, New York, NY, USA. Association for Computing Machinery.
- Wenhan Liu, Xinyu Ma, Weiwei Sun, Yutao Zhu, Yuchen Li, Dawei Yin, and Zhicheng Dou. 2025a. [ReasonRank: Empowering Passage Ranking with Strong Reasoning Ability](#). *Preprint*, arXiv:2508.07050.
- Wenhan Liu, Xinyu Ma, Yutao Zhu, Lixin Su, Shuaiqiang Wang, Dawei Yin, and Zhicheng Dou. 2025b. [CoRanking: Collaborative Ranking with Small and Large Ranking Agents](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5098–5110, Suzhou, China. Association for Computational Linguistics.
- Xukai Liu, Ye Liu, Kai Zhang, Kehang Wang, Qi Liu, and Enhong Chen. 2024c. [OneNet: A Fine-Tuning Free Framework for Few-Shot Entity Linking via Large Language Model Prompting](#). *Preprint*, arXiv:2410.07549.
- Lajanugen Logeswaran, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Jacob Devlin, and Honglak Lee. 2019. [Zero-Shot Entity Linking by Reading Entity Descriptions](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3449–3460, Florence, Italy. Association for Computational Linguistics.
- Jian Luo, Xuanang Chen, Ben He, and Le Sun. 2024. [PRP-Graph: Pairwise Ranking Prompting to LLMs with Graph Aggregation for Effective Text Re-ranking](#). In *Proceedings of the 62nd Annual Meeting*

- of the Association for Computational Linguistics (*Volume 1: Long Papers*), pages 5766–5776, Bangkok, Thailand. Association for Computational Linguistics.
- Jose L. Martinez-Rodriguez, Aidan Hogan, and Ivan Lopez-Arevalo. 2020. [Information extraction meets the Semantic Web: A survey](#). *Semantic Web*, 11(2):255–335.
- Marcel Milich and Alan Akbik. 2023. [ZELDA: A Comprehensive Benchmark for Supervised Entity Disambiguation](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2061–2072, Dubrovnik, Croatia. Association for Computational Linguistics.
- Mistral-AI, Abhinav Rastogi, Albert Q. Jiang, Andy Lo, Gabrielle Berrada, Guillaume Lample, Jason Rute, Joep Barmmentlo, Karmesh Yadav, Kartik Khandelwal, Khyathi Raghavi Chandu, Léonard Blier, Lucile Saulnier, Matthieu Dinot, Maxime Darrin, Neha Gupta, Roman Soletskyi, Sagar Vaze, Teven Le Scao, and 81 others. 2025. [Magistral](#). *Preprint*, arXiv:2506.10910.
- Yasumasa Onoe and Greg Durrett. 2020. [Fine-Grained Entity Typing for Domain Independent Entity Linking](#). *Preprint*, arXiv:1909.05780.
- Nicholas Pipitone, Ghita Hour Alami, Advait Avadhanam, Anton Kaminskyi, and Ashley Khoo. 2025. [zELO: ELO-inspired Training Method for Rerankers and Embedding Models](#). *Preprint*, arXiv:2509.12541.
- Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Le Yan, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, and Michael Bendersky. 2024. [Large Language Models are Effective Text Rankers with Pairwise Ranking Prompting](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1504–1518, Mexico City, Mexico. Association for Computational Linguistics.
- Stephen E. Robertson and Karen Sparck Jones. 1988. Relevance weighting of search terms. In *Document Retrieval Systems*, pages 143–160. Taylor Graham Publishing, GBR.
- Susanna Rücker and Alan Akbik. 2025. [Evaluating design decisions for dual encoder-based entity disambiguation](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 15685–15701. Association for Computational Linguistics.
- Özge Sevgili, Artem Shelmanov, Mikhail Y. Arkipov, Alexander Panchenko, and Chris Biemann. 2022. [Neural entity linking: A survey of models based on deep learning](#). *Semantic Web*, 13(3):527–570.
- Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. [Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer](#). *Preprint*, arXiv:1701.06538.
- Senbao Shi, Zhenran Xu, Baotian Hu, and Min Zhang. 2024. [Generative Multimodal Entity Linking](#). *Preprint*, arXiv:2306.12725.
- Yuling Shi, Yichun Qian, Hongyu Zhang, Beijun Shen, and Xiaodong Gu. 2025. [LongCodeZip: Compress Long Context for Code Language Models](#). *Preprint*, arXiv:2510.00446.
- Jinyan Su, Jennifer Healey, Preslav Nakov, and Claire Cardie. 2025. [Between Underthinking and Overthinking: An Empirical Study of Reasoning Length and correctness in LLMs](#). *Preprint*, arXiv:2505.00127.
- Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. [Yago: A core of semantic knowledge](#). In *Proceedings of the 16th International Conference on World Wide Web, WWW '07*, pages 697–706, New York, NY, USA. Association for Computing Machinery.
- Fabian M. Suchanek, Mauro Sozio, and Gerhard Weikum. 2009. [SOFIE: a self-organizing framework for information extraction](#). In *Proceedings of the 18th International Conference on World Wide Web, WWW 2009, Madrid, Spain, April 20-24, 2009*, pages 631–640. ACM.
- Hongyin Tang, Xingwu Sun, Beihong Jin, and Fuzheng Zhang. 2021. [A Bidirectional Multi-paragraph Reading Model for Zero-shot Entity Linking](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13889–13897.
- Daniel Vollmers, Hamada Zahera, Diego Moussallem, and Axel-Cyrille Ngonga Ngomo. Contextual Augmentation for Entity Linking using Large Language Models.
- Denny Vrandečić. 2012. [Wikidata: A new platform for collaborative data collection](#). In *Proceedings of the 21st International Conference on World Wide Web, WWW '12 Companion*, pages 1063–1064, New York, NY, USA. Association for Computing Machinery.
- Fang Wang, Tianwei Yan, Zonghao Yang, Minghao Hu, Jun Zhang, Zhunchen Luo, and Xiaoying Bai. 2026. [DeepMEL: A multi-agent collaboration framework for multimodal entity linking](#). *Information Processing & Management*, 63(3):104507.
- Feng Wang, Yuqing Li, and Han Xiao. 2025. [Jina-reranker-v3: Last but Not Late Interaction for Listwise Document Reranking](#). *Preprint*, arXiv:2509.25085.
- Junxiong Wang, Ali Mousavi, Omar Attia, Ronak Pradeep, Saloni Potdar, Alexander M. Rush, Umar Farooq Minhas, and Yunyao Li. 2024. [Entity disambiguation via fusion entity decoding](#). In *Proceedings of the 2024 Conference of the North*

- American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 6524–6536. Association for Computational Linguistics.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. **Self-Consistency Improves Chain of Thought Reasoning in Language Models**. *Preprint*, arXiv:2203.11171.
- Chris Welty, J. William Murdock, Aditya Kalyanpur, and James Fan. 2012. *A Comparison of Hard Filters and Soft Evidence for Answer Typing in Watson*, pages 243–256. Springer Berlin Heidelberg, Berlin, Heidelberg.
- 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 2020, Online, November 16-20, 2020*, pages 6397–6407. Association for Computational Linguistics.
- Amy Xin, Yunjia Qi, Zijun Yao, Fangwei Zhu, Kaisheng Zeng, Xu Bin, Lei Hou, and Juanzi Li. 2024. **LL-MAEL: Large Language Models are Good Context Augmenters for Entity Linking**. *Preprint*, arXiv:2407.04020.
- Zhenran Xu, Yulin Chen, Baotian Hu, and Min Zhang. 2023. **A Read-and-Select Framework for Zero-shot Entity Linking**. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13657–13666, Singapore. Association for Computational Linguistics.
- Ikuya Yamada, Koki Washio, Hiroyuki Shindo, and Yuji Matsumoto. 2022. **Global Entity Disambiguation with BERT**. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3264–3271, Seattle, United States. Association for Computational Linguistics.
- An Yang, Anpeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. **Qwen3 Technical Report**. *Preprint*, arXiv:2505.09388.
- Zonghai Yao, Liangliang Cao, and Huapu Pan. 2020. **Zero-shot Entity Linking with Efficient Long Range Sequence Modeling**. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2517–2522, Online. Association for Computational Linguistics.
- Christophe Ye and Cassie S. Mitchell. 2025. **LLM as Entity Disambiguator for Biomedical Entity-Linking**. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 301–312, Vienna, Austria. Association for Computational Linguistics.
- Mike Zhang, Rob van der Goot, and Barbara Plank. 2024. **Entity linking in the job market domain**. In *Findings of the Association for Computational Linguistics: EACL 2024, St. Julian’s, Malta, March 17-22, 2024*, pages 410–419. Association for Computational Linguistics.
- Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. **Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models**. *Preprint*, arXiv:2506.05176.
- Kang Zhou, Yuepei Li, Qing Wang, Qiao Qiao, and Qi Li. 2024. **GenDecider: Integrating “None of the Candidates” Judgments in Zero-Shot Entity Linking Re-ranking**. 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 239–245, Mexico City, Mexico. Association for Computational Linguistics.
- Shengyao Zhuang, Xueguang Ma, Bevan Koopman, Jimmy Lin, and Guido Zuccon. 2025. **Rank-R1: Enhancing Reasoning in LLM-based Document Rerankers via Reinforcement Learning**. *Preprint*, arXiv:2503.06034.
- Shengyao Zhuang, Honglei Zhuang, Bevan Koopman, and Guido Zuccon. 2024. **A Setwise Approach for Effective and Highly Efficient Zero-shot Ranking with Large Language Models**. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 38–47.

## A Prompts

The reranker task, retriever task and selection LLM prompt are given in Figure 3.

## B Category-specific results on ZESHEL

In ZESHEL, mentions are split into four categories, based on token overlap with the entity name:

**High Overlap (HO):** the entity name is identical to the mention text

**Multiple Categories (MC):** the entity name is the mention text followed by a disambiguation phrase (e.g. mention string: “Batman”, title: “Batman (Lego)”)

**Ambiguous Substrings (AS):** the mention text is a substring of the entity name

**Low Overlap (LO):** all other cases

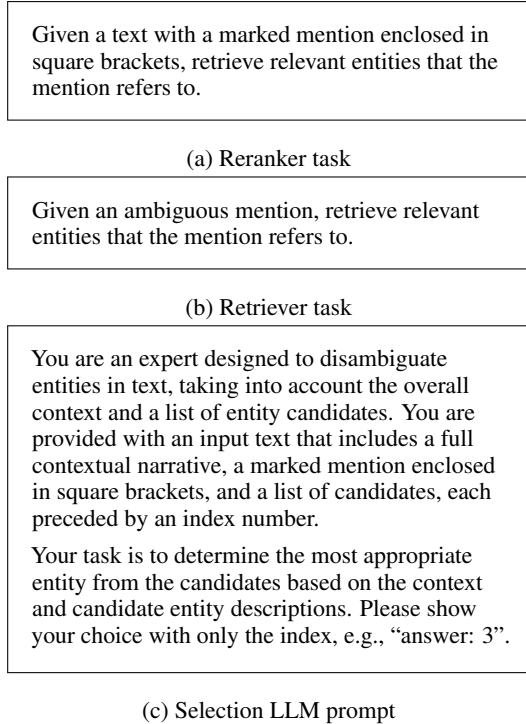


Figure 3: Reranker, retriever and LLM selection prompts.

These categories respectively make up 50%, 28%, 8% and 5% of the mentions in the dataset. The category-specific normalized accuracy in Table 6 shows that using `Magistral-Small-2509`, LELA ranks first on the four categories, and shines in the AS and MC cases. We conjecture that performance on the AS case can be explained by the high exposure of LLMs to such cases during pre-training, which occur in all kinds of text corpora, while performance on the MC case can be attributed to the need for understanding fine-grained distinctions between particularly similar candidates, and which can benefit from reasoning. Regarding the LO case, with lower relative improvement, and where there is minimal lexical overlap, embedding-based approaches might also be well-suited since they are optimized for semantic similarity.

### C Unseen mentions and short context scenario

**Benchmark.** The WikilinksNED Unseen-Mentions dataset was proposed in Onoe and Durrett (2020). It is derived from the WikilinksNED dataset (Eshel et al., 2017), ensuring that the mention-entity pairs are unseen at test time. The examples in the dataset consist of single sentences, resulting in highly limited context. The

Method	HO	MC	AS	LO
Baseline	87.64	77.27	75.89	71.46
BLINK	94.30	75.40	79.95	73.50
Uni-MPR	91.43	79.07	75.60	73.53
Bi-MPR	92.84	81.93	77.37	73.88
ReS	94.42	81.29	77.80	76.51
<b>LELA (ours)</b>				
w/ <code>Magistral-Small-2509</code>	<b>95.63</b>	<b>85.19</b>	<b>84.95</b>	<b>77.99</b>
w/ <code>Qwen3-30B-A3B</code>	88.39	79.16	76.88	68.31
w/ <code>Qwen3-4B</code>	79.03	70.83	70.68	60.92

Table 6: Category-specific macro-averaged normalized accuracy on the test set of ZESHEL, results for the other methods as reported in Xu et al. (2023).

target knowledge base is Wikipedia.

**LELA configuration.** We retrieve 64 candidates per mention through dense retrieval over Wikipedia entities, using the ELEVANT EL evaluation tool (Bast et al., 2022) Wikipedia mappings<sup>8</sup>, which achieves a recall of 83%. We also experiment with prior-based candidates using the candidate list from De Cao et al. (2021b), which yields a recall of 84%.

**Results.** Results reported in Table 7 show that BLINK leads on the benchmark. Regarding LELA, using the prior-based candidates leads to the best results. In particular, these candidates allow LELA to outperform the true zero-shot ET4EL, as well as the classical GENRE and ReFinED methods, which also use prior-based candidates. In fact, for all three models considered, using dense retrieval systematically underperforms on this benchmark. Therefore, we conducted an error analysis to analyze LELA’s behavior when using dense retrieval in this setting. Four out of the ten analyzed errors come from candidate retrieval, and another four come from imprecise ground truth annotations, where we believe that our approach’s predicted Wikipedia entity could also be considered as correct. This is caused by the large number of highly similar entities in Wikipedia. Only the two remaining cases correspond to the reranker losing the correct candidate and a wrong LLM final answer selection

### D The Recall vs Selection Accuracy Trade-off

The performance of LELA is a product of the candidate retrieval recall (*is the correct answer within the retrieved candidates?*) and the candidate selection accuracy (*does the LLM select the correct*

<sup>8</sup><https://github.com/ad-freiburg/elevant/wiki/Mapping-Files#wikipedia-mappings>

Method	Type	Acc.
Most Frequent Entity Baseline (Onoe and Durrett, 2020)	baseline	<b>54.1</b>
BLINK (Wu et al., 2020)		
Wikipedia	classical	75.2
Training set	classical	74.7
Training set + Wikipedia	classical	<b>76.8</b>
GENRE (De Cao et al., 2021b)	classical	63.5
ReFinED (Ayoola et al., 2022)	classical	66.5
ET4EL (Onoe and Durrett, 2020)	true zero-shot	62.2
<b>LELA (ours)</b>	true zero-shot	
<b>w/ Magistral-Small-2509</b>		
Dictionary		<b>68.0</b>
Dense Retrieval		62.8
w/ Qwen3-30B-A3B		
Dictionary		68.7
Dense Retrieval		61.9
w/ Qwen3-4B		
Dictionary		67.0
Dense Retrieval		59.0
w/o LLM selection		
Dictionary		61.9
Dense Retrieval		54.5

Table 7: Accuracy on the WikilinksNED Unseen Mentions benchmark, best per category in bold, globally best underlined.

*answer?*). Regarding the first retrieval stage, the number of retrieved candidates should be exclusively limited by computational cost constraints, given the finer subsequent stages and that the recall can only improve with the number of retrieved candidates. However, for the next two stages, the value of  $k$  results in a trade-off between recall and selection accuracy, as shown in Figure 4: lower values of  $k$  allow for a high selection accuracy, but may result in the correct entity not being among the candidates. Higher values of  $k$  reduce the selection accuracy but improve the candidate recall. The impact of  $k$  on these two components depends on the reranker and selection LLM in use. For instance, the selection accuracy experiences a steeper drop when  $k$  increases for Qwen3-30B-A3B than Magistral-Small-2509. Reranker-specific results are given in Section H. Figure 15 shows an example of selection output for  $k = 10$  and  $k = 64$ .

## E Self-consistency Study

As shown by our ablation studies, Section 4.5, self-consistency systematically allows for performance improvements over single selection LLM answer sampling. Figure 6a reports the selection accuracy over the number of sampled answers,  $k_{sc}$ , and shows that a plateau is reached across mod-

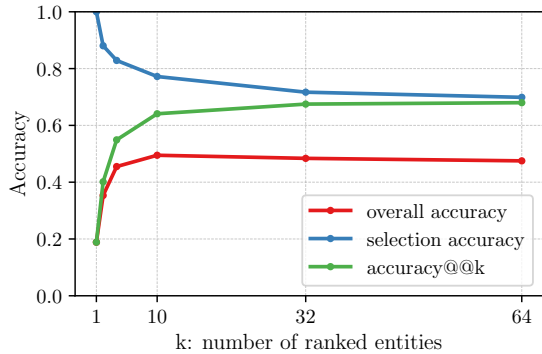
els for  $k_{sc} = 3$ . The best improvement comes from  $k_{sc} = 2$ . Furthermore, Figure 6b shows how the invalid output rate, which also significantly drops for  $k_{sc} = 2$ . In order to check that the impact of self-consistency is not strictly limited to higher robustness towards invalid model outputs, we show in Figure 5 that for our default  $k_{sc} = 10$  choice, the models select the same candidate across all samples in approximately 60% of the cases, two answers in approximately 25% of cases and three answers in less than 10% of the cases.

## F Positional Bias Study

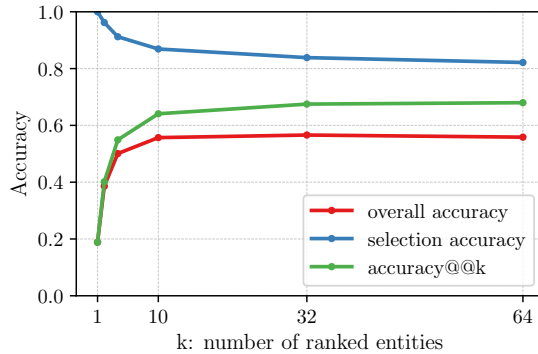
To study the impact of the positional bias (Liu et al., 2024a) on the LLM selection accuracy, we report the results on ZESHEL for five different top-k candidate sorting approaches:

- **Reranker:** We rank the top-k candidates according to their reranker-assigned score (the standard LELA approach)
- **Answer first:** The correct answer is moved to the first position.
- **Answer last:** The correct answer is moved to the last position.
- **Random:** The top-k candidates are shuffled.
- **BM25:** We rank the top-k candidates according to their BM25 score (the standard LELA approach)

The results in Table 8a show varying effects of candidate ordering across models. For Magistral-Small-2509, performance differences between the sorting approaches are minimal, indicating a low sensitivity to candidate position. Among these, the oracle *Answer first* sorting achieves the highest accuracy, closely followed by the *Reranker* sorting. For Qwen3-30B-A3B, the performance gap between sorting strategies becomes more pronounced. Again, the *Answer first* sorting yields the highest accuracy, with the *Reranker* sorting ranking second. In contrast, *Random* sorting performs worst, trailing the oracle setting by 2.51 percentage points. This effect further increases for Qwen3-4B, which exhibits the largest performance spread across sorting strategies. For this smaller model, the *Reranker* sorting achieves the highest accuracy, followed by the *Answer first* sorting, while *Random* sorting again results in the lowest performance, 4.84 percentage points behind the best-performing approach.



(a) Overall (micro-averaged) accuracy, selection accuracy and accuracy@k w.r.t. k for Qwen3-30B-A3B



(b) Overall (micro-averaged) accuracy, selection accuracy and accuracy@k w.r.t. k for Magistral-Small-2509

Figure 4: Comparison of (micro-averaged) accuracy, selection accuracy and accuracy@k on ZESHEL.

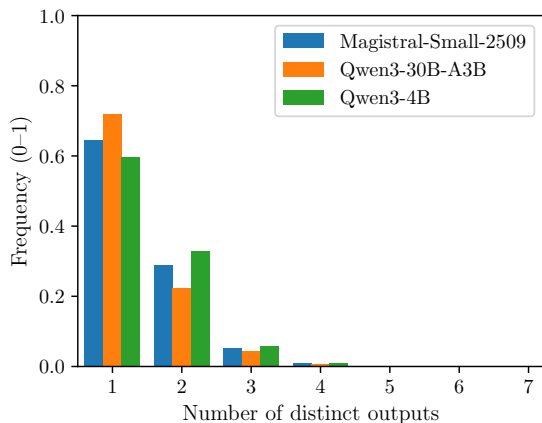


Figure 5: Output diversity histogram, using 10 samples for self consistency, on ZESHEL.

Overall, these results demonstrate that candidate position has a measurable impact on selection accuracy, with sensitivity increasing as model size decreases. We hypothesize that larger models are better able to compare candidates independently of their order, while smaller models rely more heavily on positional cues. The intermediate behavior of the 30B MoE Qwen3-30B-A3B model is consistent with this explanation. Importantly, the proposed *Reranker* sorting strategy consistently ranks first among the non-oracle methods and achieves the best performance for the smallest model, justifying its use in LELA. In contrast, *BM25* sorting underperforms relative to the reranker-based approaches, while *Random* sorting yields the lowest accuracy across all models.

Furthermore, we also report the results on ESCO to study the impact in the case of *None* prediction, considering three different positions for the UNK

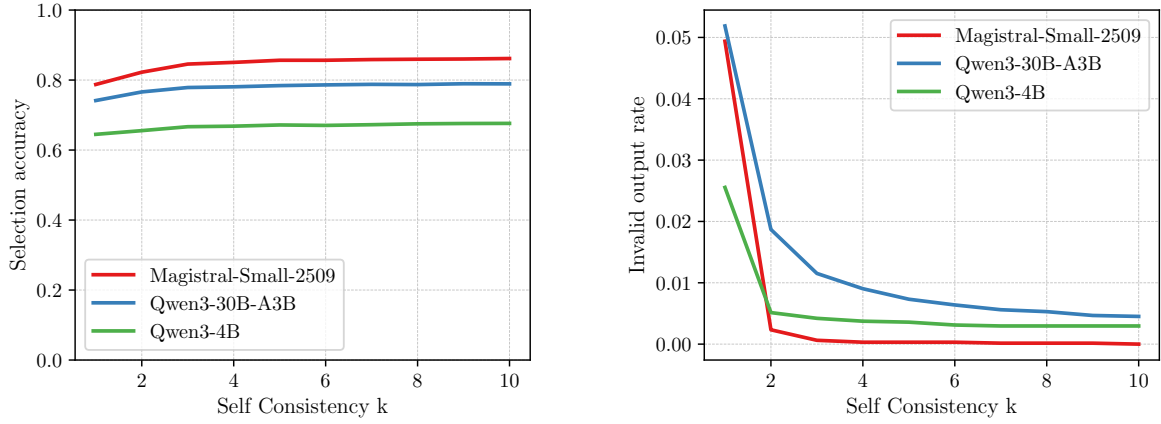
candidate, among the candidates according to the reranker-assigned scores:

- **None first:** The *None* option is in first position (the standard LELA approach)
- **None last:** The *None* option is in last position
- **Random:** The *None* option is in a random position

The results are reported in Table 8b. Across all models, changing the position of the *None* candidate produces only marginal differences in overall accuracy, indicating no statistically significant effect. We observe a consistent trend in which placing *None* first slightly improves UNK accuracy at the cost of marginally reduced non-UNK. Overall, model performance appears robust to the ordering of the *None* candidate.

## G Runtime and Scalability Results

First of all, Figure 8 reports the selection accuracy as a function of the number of input tokens on ZESHEL. It shows how the selection accuracy decreases for all three models as the input context length increases. The drop in performance is the least pronounced for the dense Magistral-Small-2509 and Qwen3-4B models. The MoE Qwen3-30B-A3B model experiences a significantly steeper drop in performance as the context length increases, eventually achieving worse performance than Qwen3-4B on the longer inputs, despite both models having the same native context window. Figure 7a reports the selection accuracy as a function of the number of output tokens on ZESHEL. The three curves show a range,



(a) Selection accuracy based on the number of self-consistency samples.

(b) Invalid LLM output rate based on the number of self-consistency samples.

Figure 6: Effect of the number of self-consistency samples on selection accuracy and invalid output rate on ZESHEL.

Model	Reranker	BM25	Answer first	Answer last	Random
Magistral-Small-2509	<u>81.88</u>	81.26	<b>81.91</b>	81.60	81.56
Qwen3-30B-A3B	<u>74.42</u>	74.34	<b>75.43</b>	73.72	72.92
Qwen3-4B	<b>65.84</b>	63.72	<u>65.20</u>	61.61	61.00

(a) Micro-averaged normalized accuracy on ZESHEL for different reranker output list sorting approaches.

Model	Sorting	Accuracy	UNK Accuracy	Non-UNK Accuracy
Magistral-Small-2509	None first	<b>26.75</b> $\pm$ 2.03	<b>5.41</b> $\pm$ 1.56	43.92 $\pm$ 3.06
	None last	26.43 $\pm$ 2.03	3.94 $\pm$ 1.34	<b>44.51</b> $\pm$ 3.07
	Random	26.48 $\pm$ 2.03	4.18 $\pm$ 1.38	44.41 $\pm$ 3.07
Qwen3-30B-A3b	None first	<b>36.62</b> $\pm$ 2.21	<b>28.04</b> $\pm$ 3.09	43.52 $\pm$ 3.06
	None last	36.51 $\pm$ 2.21	27.55 $\pm$ 3.08	<b>43.72</b> $\pm$ 3.06
	Random	35.20 $\pm$ 2.19	25.09 $\pm$ 2.99	43.32 $\pm$ 3.06
Qwen3-4B	None first	26.59 $\pm$ 2.03	8.36 $\pm$ 1.91	<b>41.25</b> $\pm$ 3.04
	None last	<b>26.70</b> $\pm$ 2.03	<b>9.84</b> $\pm$ 2.05	40.26 $\pm$ 3.03
	Random	26.32 $\pm$ 2.02	8.12 $\pm$ 1.88	41.00 $\pm$ 3.04
$\emptyset$ (Reranker-only)	-	-	-	<b>35.71</b> $\pm$ 2.96

(b) Accuracy on ESCO for different UNK candidate positions.

Table 8: Effect of candidate position on selection accuracy across two datasets.

around 500 tokens, where the accuracy is the highest. For Qwen3-4B and Qwen3-30B-A3B models, shorter outputs are less accurate, which is not the case for Magistral-Small-2509 which does not generate these short outputs. For all three models, the accuracy drops past this range. This is an instance of the typical reasoning LLMs *underthinking* and *overthinking* issue (Su et al., 2025). Figure 7b shows a similar ideal range below 500 tokens on another dataset, where all three models achieve the highest, and highly similar, performance, but no *underthinking* phenomena.

Table 9 reports the average runtime per mention across datasets using our main experiments configuration. It shows that Magistral-Small-2509 consistently results in the longest runtime, with the MoE Qwen3-30B-A3B representing a particularly interesting middle-ground, generally achieving performance on par with Magistral-Small-2509, which is still needed for achieving the best results on ZESHEL or specialized domains, while also running faster than Qwen3-4B. In fact, based on our results, the smaller dense Qwen3-4B should only be preferred to Qwen3-30B-A3B in cases of VRAM limitations. Similarly, in terms of tokens usage, which directly impact the runtime of the approach, Magistral-Small-2509 generated the longest thoughts on average, with similar values for the other models. In addition, these numbers of thought tokens also vary based on the setting and domain. Using the token length insights on ZESHEL, as well as those from Section E, we also report the results for a runtime-optimized configuration on ZESHEL by sampling only three answers for self-consistency and by limiting the context length to 8192 tokens. The results show that these two settings effectively result in significant runtime improvements across the three models, but at the expense of some accuracy percentage points.

## H Comparison of different pointwise rerankers

- BM25, using the outputs from Logeswaran et al. (2019).
- The Qwen3-Embedding-4B embedder, computing the similarity between mentions and our entity verbalizations from the candidate list.
- The Qwen3-4B LLM with the same prompt and score computation as in Zhang et al.

LELA config	Time (s)	# Tokens	Acc.
w/ Magistral-Small-2509	44.4	1042	83.11
w/ Magistral-Small-2509 <sub>opt</sub>	26.7	926	77.73
w/ Qwen3-30B-A3B	26.0	947	76.79
w/ Qwen3-30B-A3B <sub>opt</sub>	11.7	785	72.92
w/ Qwen3-4B	31.1	1059	68.00
w/ Qwen3-4B <sub>opt</sub>	19.1	798	64.45
w/o LLM selection	1.3	-	27.15

(a) ZESHEL dataset.

Method	Time (s)	# Tokens	Acc.
BLINK	0.9	-	6.4
GENRE	4.0	-	25.0
<b>LELA (ours)</b>			
w/ Magistral-Small-2509	57.2	1408	62.3
w/ Qwen3-30B-A3B	19.2	847	62.4
w/ Qwen3-4B	14.7	847	55.3
w/o LLM Selection	2.3	-	29.8

(b) GLADIS dataset.

LELA config	Time (s)	# Tokens	Acc.
w/ Magistral-Small-2509	35.8	826	68.0
w/ Qwen3-30B-A3B	10.0	581	68.7
w/ Qwen3-4B	19.6	652	67.0
w/o LLM selection	1.3	-	61.9

(c) Unseen Mentions dataset (prior-based candidates).

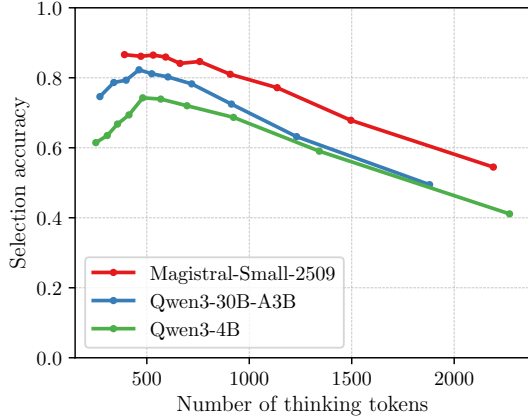
LELA config	Time (s)	# Tokens	Acc.
w/ Magistral-Small-2509	54.6	1368	26.21
w/ Qwen3-30B-A3B	21.7	882	39.36
w/ Qwen3-4B	27.2	854	29.44
w/o LLM selection	0.5	-	19.35

(d) ESCO dataset (BM25 candidates).

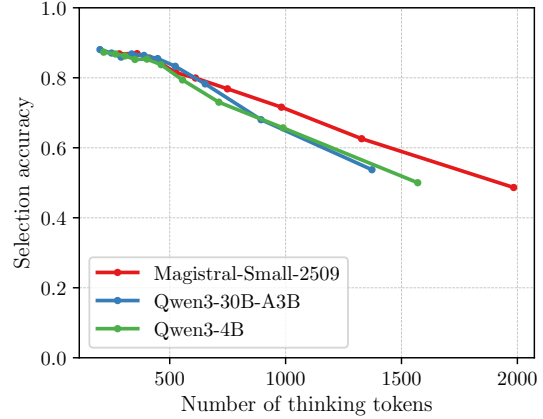
Table 9: Macro-average mention processing time, and thought token usage per self-consistency sample, across benchmarks.

(2025) but without any additional training.

The results show that BM25 underperforms compared to the three other approaches, while Qwen3-Reranker-4B achieves the highest recall@10 and Qwen3-Embedding-4B yield competitive results, despite the latter not being provided with the context. Furthermore, Figure 9 shows the ground truth rank distribution across the four approaches. Interestingly, Qwen3-Embedding-4B assigns the higher score to the ground truth in more than 60% of the cases, while Qwen3-4B-Reranker does so in less than 30% of cases, it catches up with the embedder’s recall at  $k = 6$ . It should be noted that when using models which are highly susceptible to positional bias, such as Qwen3-4B (see Appendix F), the accuracy@k metric is not sufficient to compare different rerankers, as it is



(a) ZESHEL.



(b) Unseen Mentions.

Figure 7: Selection accuracy as a function of the number of generated thinking tokens on two datasets.

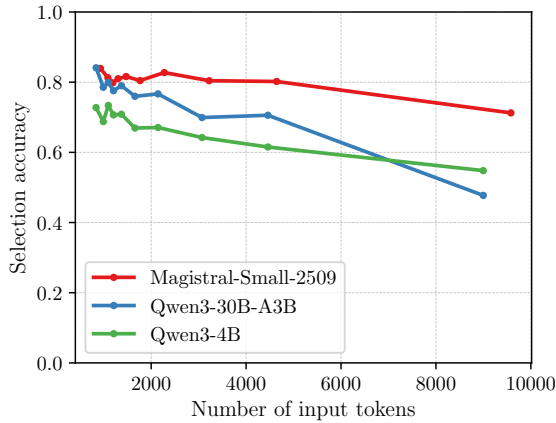


Figure 8: Selection accuracy based on the number of input tokens, on ZESHEL.

Reranker	Acc@10
BM25	64.99
Qwen3-Embedding-4B	89.94
Qwen3-4B	89.12
Qwen3-Reranker-4B	<b>93.57</b>

Table 10: Macro-averaged normalized accuracy@10 for different candidate reranking models on ZESHEL.

invariant to the output candidate position.

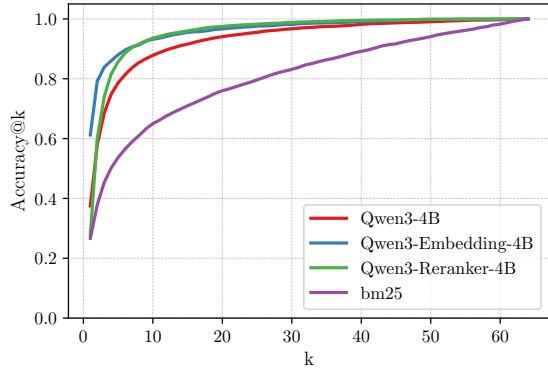
## I Candidate Retention Funnel

To better understand the contribution of each stage (Retrieval, Reranking, and Selection) to the final performance, we visualize the flow of the ground-truth entity through the LELA pipeline using Sankey diagrams (Figure 10). The diagrams clearly distinguish between retrieval-bound, reranking-bound, and reasoning-bound challenges

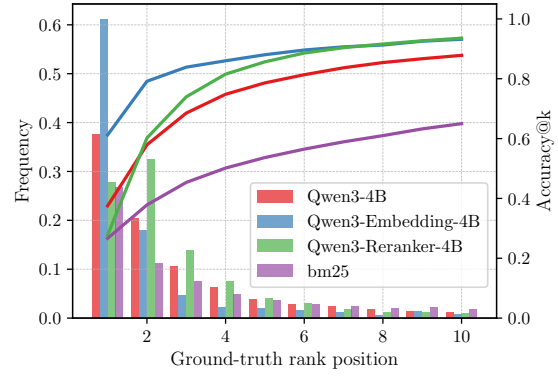
across different benchmarks:

- On ZESHEL, the dominant error source is consistently the initial BM25 retrieval. Once the correct entity is retrieved, LELA is highly effective at retaining and selecting it.
- On GLADIS, the bottleneck shifts to the reranking stage, particularly in specialized domains where the density of similar long forms is high.
- On WikilinksNED Unseen Mentions (specifically with dense retrieval), the challenge lies primarily in the final selection, suggesting that the reasoning model struggles to disambiguate between the highly similar entities typical of Wikipedia.

A key finding is the high retention rate of the reranking stage across most datasets. For example, on the ZESHEL/Forgotten Realms domain, the reranker preserves 98% of the recoverable recall (977 retained vs. 22 lost), while compressing the candidate list by a factor of 6.4 (from 64 to 10). These results demonstrate the robustness of our choice of  $k = 10$ , validating that aggressive filtering does not compromise the upper bound of the system. However, the results also highlight domain-dependent variations. While the reranker achieves a 90.5% retention rate on the GLADIS/General domain, this drops to 63.2% on the GLADIS/Scientific domain. This significant drop indicates that scientific acronyms pose a harder challenge for the reranker, likely due to the fine-grained technical distinctions required to separate valid candidates. Crucially, the Selection LLM



(a) Micro-averaged normalized accuracy@k.



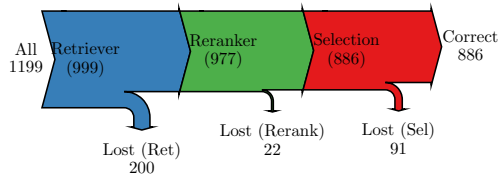
(b) Ground truth position in reranker output histogram.

Figure 9: Comparison of rerankers on ZESHEL using (a) accuracy@k and (b) ground-truth position histogram.

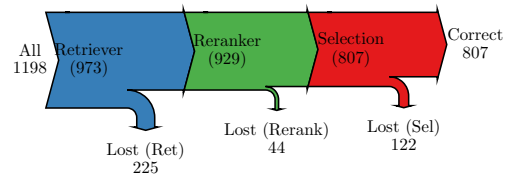
proves significantly more robust in this technical setting, successfully identifying the correct entity among the top-10 candidates in 74.1% of the cases passed to it (6593 out of 8894).

## J Examples

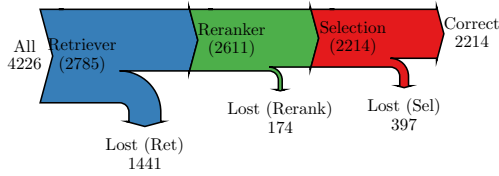
Figures 11, 12, 13, 14, 15, and 16 show examples of reasoning traces in the LLM candidate selection phase of LELA.



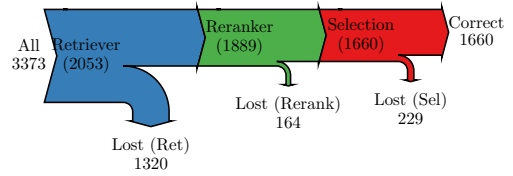
(a) ZESHEL/Forgotten Realms



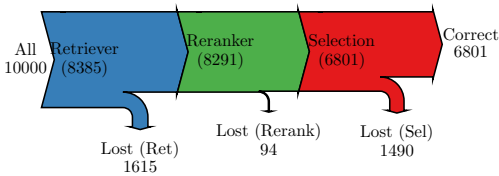
(b) ZESHEL/Lego



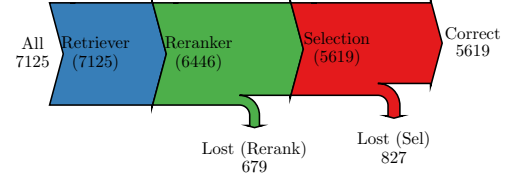
(c) ZESHEL/Star Trek



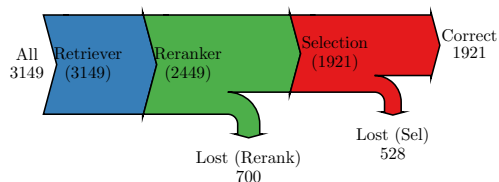
(d) ZESHEL/YuGiOh



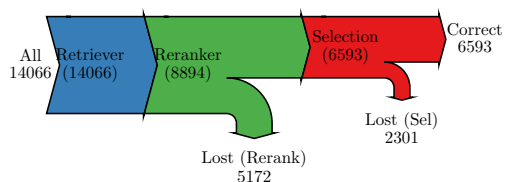
(e) WikilinksNED Unseen Mentions



(f) GLADIS/General



(g) GLADIS/Biomedical



(h) GLADIS/Scientific

Figure 10: Candidate retention funnel across datasets, using Magistral-Small-2509. The diagrams visualize the attrition of the ground-truth entity at each stage of the LELA pipeline: (1) from the total dataset to the Retriever, (2) to the Reranker, and (3) to the final Selection.

**Input sentence and mention:** [Develop tools] for our internal teams ( sales support operations finance ) to help run the business more efficiently .",

**LELA:** <think> Okay, let's tackle this problem step by step. The user has provided a sentence: "[Develop tools] for our internal teams ( sales support operations finance ) to help run the business more efficiently." The task is to determine which of the given candidate entities best fits the mention "[Develop tools]". First, I need to understand the context. The sentence is about developing tools for internal teams in sales, support, operations, and finance to improve business efficiency. The key here is "develop tools" in the context of internal business operations. Looking at the list of candidates: 0. None of the listed candidates 1. develop promotional tools - This seems related to marketing materials, not internal tools. 2. produce customised tools - This could be relevant, but the description mentions artisanal crafting or restoration, which might not fit internal business tools. 3. database development tools- Focuses on databases, which is a specific area, but the original sentence is broader. 4. KDevelop - A specific software development tool, but the mention is general. 5. business process modelling - Involves modeling business processes, which could relate to efficiency. 6. create business process models - Similar to 5, but more about creating models rather than tools. 7. develop film - Unrelated to internal business tools. 8. software frameworks - Tools for software development, which might be relevant if the tools are software-related. 9. Eclipse - Another specific software development tool, like KDevelop. 10. apply tools for content development - Focuses on content, not internal business processes. The original mention is "Develop tools" for internal teams to run the business more efficiently. The most general and fitting option seems to be 0 (None of the listed candidates) if none of the options perfectly match. However, looking again, option 5 and 6 are about business process modelling and creating models, which could be tools to improve efficiency. However, the mention is about "tools" in general, not specifically process models. Option 8 is software frameworks, which are tools for development. But the original sentence doesn't specify software. Option 10 is about content development tools, which doesn't fit. Alternatively, maybe "None of the listed candidates" is correct if none of the options accurately describe developing general internal tools. However, looking at the descriptions again, **option 5 and 6 are about business process modelling, which might be considered tools. But the original mention is more general. Since the options don't have a general "develop tools for internal efficiency" category, perhaps 0 is the answer. But I need to check if any of the options are close. Alternatively, maybe option 5 or 6. But the original text says "tools" not "models". So perhaps 0 is the correct choice here.</think> **"answer": 0.** (✓)**

Figure 11: LLM selection example from ESCO for the UNK entity. The model correctly chooses the None candidate as shown in blue instead of the other strong candidates.

**Input sentence and mention:** The Golden Rule is arguably the most essential basis for the modern concept of human rights , in which each individual has a [right] to just treatment, and a reciprocal responsibility to ensure justice for others. [ 4 ] A key element of the Golden Rule is

**LELA w/o LLM selection:** **Right to keep and bear arms:** The right to keep and bear arms (often referred to as the right to bear arms) is a legal right for people to possess weapons (arms) for the preservation of life, liberty, and property (X)

**LELA:** <think>Okay, let's try to figure out which entity is the correct answer here. The input text mentions the Golden Rule as the basis for the modern concept of human rights. The marked mention is [right], and we need to choose from the candidates. First, the context is about human rights and the Golden Rule. The sentence says each individual has a [right] to just treatment and a responsibility to ensure justice. The options include several types of rights. Let's look at the candidates. Option 4 is Human rights, which is already mentioned in the text. But the question is about the specific entity that fits the [right] in the sentence. The sentence is talking about an individual's right, so maybe a more general term. Option 6 is "Rights" in general. The description says rights are principles of freedom or entitlement. **The sentence uses "right" in a general sense, not a specific type like the right to bear arms or civil rights.** So maybe option 6. But wait, the text mentions "human rights" earlier, but the marked term is [right], which is more general. Let me check the options again. Option 6 is "Rights" which is the general term. The sentence says "a [right] to just treatment", which would fit under the general definition of rights. However, option 4 is Human rights, but the sentence is about individual rights within human rights. Hmm. But the answer might be 6 because it's the general term. Alternatively, maybe 7 is civil and political rights, but the text doesn't specify that. The Golden Rule is linked to human rights, but the specific mention here is "right" as a general concept. So I think the answer is 6.</think> **"answer": 6.** (✓)

Figure 12: Reranker and LLM selection example from WikilinksNED Unseen Mentions. When using the selection LLM, as shown in blue, the model is able to select the correct candidate, unlike the reranker which ranks a wrong candidate (shown in red) first.

**Input sentence and mention:** Fourvière , annual festival in June and August. The Biennale d'art contemporain . The Nuits Sonores . Lille ( French pronunciation: [lil] ; **[Dutch]** : Rijsel) is a city in northern France . It is the principal city of the Lille Métropole , the fourth-largest metropolitan area in the

**LELA w/o LLM selection:** **Dutch Language:** Dutch ( ) is a West Germanic language spoken by about 25 million people as a first language and 5 million as a second language. (✓)

**LELA:** <think>Okay, let's tackle this problem. The input text mentions "Lille" and describes it as a city in northern France, the principal city of the Lille Métropole, which is the fourth-largest metropolitan area. The task is to determine which of the given candidates best fits the context. First, I need to check the list of candidates. The options are from 1 to 10. **The mention in the text is "Lille,"** but the candidates don't include Lille directly. However, the text also mentions "French pronunciation: [lil]" and "[Dutch]: Rijsel." So, the city's Dutch name is Rijsel. But none of the candidates are Lille or Rijsel. Looking at the candidates, the options are about regions, languages, communities, etc. For example, option 2 is the Low Countries, which is a region in Northwestern Europe, including the Netherlands, Belgium, and Luxembourg. Option 5 is the Flemish Community, which is part of Belgium. Option 9 is the Netherlands. But the text is about Lille, which is in France, not the Netherlands or Belgium. However, since the Dutch name for Lille is Rijsel, maybe there's a connection. Wait, the question is to disambiguate the entity in the text. The text mentions Lille, which is a city in France. But the candidates don't have Lille. However, the Dutch name for Lille is Rijsel. But none of the candidates are Rijsel. The options are about regions, languages, etc. So perhaps the correct answer is related to the Low Countries (option 2), since Lille is in northern France, near the border with the Low Countries. Alternatively, maybe the Flemish Community (option 5) since Lille is near the French region of Flanders, which is part of the Flemish Community. But Lille is actually in the Nord-Pas-de-Calais region of France, which is part of the French Flanders, not the Flemish Community (which is in Belgium). However, the text mentions that Lille is in northern France, and the Dutch pronunciation is given. The mention of Dutch (option 1) might be relevant, but the text is about the city, not the language. Alternatively, maybe the correct answer is the Low Countries (option 2) because Lille is in northern France, close to the Low Countries region. But the candidates are entities, and the question is to pick the most appropriate one based on context. However, none of the candidates are directly Lille. The only possible connection is that the Dutch name for Lille is Rijsel, but there's no candidate for that. The options are about regions, languages, etc. Since the text is about Lille's location and the Dutch name, perhaps the correct answer is the Low Countries (option 2), as Lille is near that region. Alternatively, the Netherlands (option 9) but Lille is in France. Alternatively, the Flemish Community (option 5) but again, Lille is in France, not Belgium. Wait, maybe the answer is option 2, the Low Countries, since the text mentions the Dutch pronunciation and the city's location in northern France, which is near the Low Countries. But I'm not entirely sure. Alternatively, maybe there's a mistake here. However, given the options, the best fit might be the Low Countries (option 2) as a geographical region. But I need to check the descriptions again. Option 2: Low Countries - a region in Northwestern Europe, including the Benelux countries. Since Lille is in northern France, close to the Benelux area, maybe that's the connection. So the answer is 2.</think>**"answer": 2 (X)**

Figure 13: LLM selection failure example from WikilinksNED Unseen Mentions. The reranker ranks the correct candidate first, However, the selection LLM considers the wrong mention, as shown in red, overthinks, and selects a wrong candidate.

**Input sentence and mention:** Ann Lee Ann Lee is a minifigure introduced in the Monster Fighters theme . Description . Ann has sand blue legs printed with a clove of garlic , wooden stakes , and a belt . Her torso is white , and is printed with a sand blue desk , a red and white bodice , a white collar , and a silver necklace . She has white arms with yellow hands . Her hairpiece is reused from Cave Woman with a crossbow bolt instead of a bone . She has a facial scar from the Witch Nail . Background . Ann Lee is a Monster Fighter , proficient in the ancient art of broomstickajitsu . Her monster arch - enemy are the Ghosts , even though she is shown in videos with the [Mummy] . She met Dr . Rodney when Dr . Rodney was looking for a suitable pilot brave enough to fly into the Monster Realms , and was described as waiting in the air hangar for him .

**LELA w/o entity descriptions:** "<think> Okay, let's see. The input text is about Ann Lee, a minifigure from the Monster Fighters theme. The question is about disambiguating the mention of [Mummy] in the text. The context says that Ann Lee's monster arch-enemies are the Ghosts, but she's shown in videos with the [Mummy]. The task is to pick the correct entity from the list provided. First, I need to look at the candidates. The list includes various Mummy-related entries, like Mummy (Adventurers), The Mummy, Mummy (Scooby-Doo), Part numbers, and others. The key here is to figure out which Mummy is being referred to in the context of Ann Lee's story. The text mentions that Ann Lee is a Monster Fighter and her arch-enemy is the Ghosts, but she's seen with the [Mummy]. Since the options include different Mummy characters, I need to consider which one fits best. The Mummy in the context of LEGO minifigures might be a specific theme. Looking at the candidates, options like "Mummy (Minifigures)" or "9462 The Mummy" could be relevant. However, the mention of "Part 10546" or "Part 3626bpx112" might refer to specific parts, but the text doesn't mention parts, it's about a character. Also, "The Mummy" as a standalone entry could be a possible candidate. However, since the original text is about a minifigure, the most likely answer would be "Mummy (Minifigures)" which is option 6. But wait, there's also "9462 The Mummy" which might be a specific set. However, the context doesn't mention a set number. Alternatively, the text says "the [Mummy]" which is a general term, so maybe "Mummy (Minifigures)" is the correct choice. But I should also check if any other options are more specific. The options 1, 2, 3, 5, 6, 8, 9 are all Mummy-related. Since the original text is in the context of LEGO Monster Fighters, which is a theme, and the Mummy could be another minifigure in that theme. The correct answer is likely option 6: Mummy ( Minifigures ). **Alternatively, maybe "The Mummy" (option 2 or 8). But considering the list, option 6 is more specific to minifigures.** So I think the answer is 6. </think> **"answer": 6 (X)**

**LELA:** <think> Okay, let's try to figure out which candidate entity is the correct one here. The input text is about Ann Lee, a minifigure from the Monster Fighters theme. The mention in question is [Mummy], and we have a list of 10 candidates. First, I need to look at the context. The text says that Ann Lee is a Monster Fighter and her monster arch-enemy are the Ghosts, even though she's shown in videos with the [Mummy]. The key here is that the Mummy is mentioned as someone she's seen with, but her main enemies are the Ghosts. Now, looking at the candidates, most of them are different versions of the Mummy in various LEGO themes. The options range from different sets, like 9462 The Mummy, which is mentioned in the text. The text also mentions that the set 9462 includes Ann Lee and the Mummy. Wait, the text says "the Monster Fighter Ann Lee" is part of the set 9462 The Mummy. So that set includes both Ann Lee and the Mummy. But the question is about the [Mummy] in the text. The text says she's shown with the [Mummy], so the Mummy here is likely the one from the same theme, Monster Fighters. Looking at the candidates, option 2 is "The Mummy - The Mummy The Mummy King is a Monster Fighters minifigure..." and option 9 is "9462 The Mummy" which is a set. But the mention is [Mummy], which is a minifigure. Option 2 is a minifigure, and the set 9462 includes the Mummy. However, the text says "the Monster Fighter Ann Lee" is in the set, implying that the Mummy is another minifigure in that set. So the Mummy here is likely the one from the Monster Fighters theme. Looking at the candidates, option 2 is The Mummy from Monster Fighters, and option 9 is the set. Since the text refers to the Mummy as a character (since Ann Lee is a minifigure and is shown with the Mummy), it's more likely referring to the minifigure. So option 2. But wait, there's also option 8, which is "Mummy - Mummy The Mummy is a Studios minifigure." But the text mentions Monster Fighters, so the correct one would be the one from Monster Fighters. Additionally, option 9 is the set, not the minifigure. So the answer is probably 2. However, let me check the other options. Option 1 is Mummy (Adventurers), which is from a different game. Option 3 is Scooby-Doo, which is a different theme. **Option 5 and 6 are other minifigures but not from Monster Fighters.** Option 4 is a part, not a minifigure. Option 7 is a head part. Option 8 is a Studios minifigure, but the text mentions Monster Fighters. So yes, option 2 is the correct one. </think> **"answer": 2 (✓)**

Figure 14: LLM selection example from ZESHEL/Lego when using or omitting entity descriptions. Without descriptions, the LLM selects the entity with the most relevant title regarding the context. as shown in red. Using entity descriptions, it excludes this wrong candidate, as shown in blue.

**Input sentence and mention:** Dog People The Dog People , also called the " North Ones " , were primitive , savage , nomadic tribesmen who lived in the deserts of northern Maztica . Traits . Dog People tended to act emotionless and did not seem bothered by poverty or pain . They were dexterous and especially tough but often lacked the charisma or intellect of other races . Society . The Dog People lived a nomadic life in the harsh , northern lands, organized into villages . The separate tribes of Dog People never cooperated . A typical village had huts of mud or else the people lived in caves , and the populations ranged from 50 to 500 persons . They survived by growing crops in the brush and savannah regions , until a drought would occur and they were forced to find other fields . Others set traps for birds or hunted or fished for food . While they did not have an ordered civilization , they had a vibrant oral tradition . In Dog People society , men and women were equivalent ; for example , women were as likely to be warriors as men were . On rare occasions , a foreigner might impress a Dog People village and be welcomed into the tribe as a brother or sister . In such cases, it was expected that the new member offer his devotion to the tribe in exchange for their protection . Religion . The Dog People practiced a shamanistic religion , and revered all of the gods of the , especially [Maztica] and Tezca . They did not, however , practice human sacrifice . Combat . The Dog People were renowned for their skill at stealth . They were fearless fighters , but if pressed , Dog People would often retreat deeper into their harsh territory , not out of cowardice , but trusting in the environment to drive their enemies away . Most city - dwelling Maztican s thought of all Dog People as foes , and they were said to capture travelers and torture them . Magic . Some Dog People became plumaweaver s or hishnashaper s . Druid s of Maztica and priests of Tezca were also common . History . The Nexalan s were related to the Dog People and originated in the same region . The people of Huacli also had the blood of Dog People .

**LELA ( $k = 64$ ):** <think>Okay, let's tackle this problem. The user has given a text about the Dog People in Maztica and a list of candidate entities. The task is to determine which entity corresponds to the mention "[Maztica]" in the text. First, I'll read through the provided text again. The relevant part is in the "Religion" section: "The Dog People practiced a shamanistic religion, and revered all of the gods of the , especially [Maztica] and Tezca." So the mention here is "[Maztica]", and we need to find the correct entity from the list. Looking at the candidate entities, there are several entries related to Maztica. Let's check each one. Candidate 1 is "Maztica" itself, described as a continent west of Faerûn with jungles and mystery. The text mentions that the Dog People lived in the deserts of northern Maztica, so this seems like a strong candidate. **Other candidates like "Maztica trilogy" (2), "Maztican" (3), "Kukul" (4), etc., don't fit because they refer to different things—like a book series, a term for inhabitants, a god, etc.** Looking at the description of candidate 1, it says Maztica is a continent, which matches the context of the Dog People living in its deserts. Also, the mention of "Tezca" in the text might refer to a deity, but the question is about "Maztica" as a place. Other entries like "Maztican" (3) are collective terms for inhabitants, but the text refers to Maztica as a land. The answer should be the continent itself. Checking the other candidates, like "Maztica (goddess)" (13) is a deity, which doesn't fit. So the correct choice is candidate 1, "Maztica".</think> **"answer": 1 (X)**

**LELA ( $k = 10$ ):** think>Okay, let's tackle this problem. The user provided a text about the Dog People in Maztica and a list of candidate entities. The task is to determine which candidate corresponds to the mention of [Maztica] in the text. First, I need to look at the context where [Maztica] is mentioned. The sentence says: "They did not, however, practice human sacrifice." Wait, no, the actual sentence is: "religion. The Dog People practiced a shamanistic religion, and revered all of the gods of the , especially [Maztica] and Tezca." So the mention is in the context of gods. The Dog People revered the gods of the, especially Maztica and Tezca. Now, **looking at the candidates, there are several entries related to Maztica. Let's check each one. Candidate 1 is Maztica (the continent). Candidate 3 is Maztican (inhabitants). Candidate 7 is Maztica (goddess). Candidate 10 is Nexalan, which is a tribe in Maztica. But the mention here is [Maztica], and the text says they revered the gods of the, especially Maztica and Tezca. So Maztica here is likely a deity, not the continent or the people. Looking at the candidates, candidate 7 is Maztica (goddess), which is described as the ancient mother of the Maztican pantheon. The text mentions that the Dog People revered all the gods of the, especially Maztica and Tezca. So Tezca is another deity, and Maztica here is another deity. Therefore, the correct candidate should be 7.** But wait, let me check if there's another possibility. Candidate 1 is the continent, but the context is about gods. Candidate 3 is the people, but the sentence is about gods. Candidate 7 is the goddess, which fits. Also, in the text, the Dog People revered the gods of the (probably the Maztican gods), and specifically mentioned Maztica and Tezca. So Maztica here is a deity. Therefore, the answer is 7.</think> **"answer": 7. (✓)**

Figure 15: LLM selection example from ZESHEL/Forbidden Realms for  $k = 10$  and  $k = 64$  (all retrieved candidates). When considering all retrieved candidates, the model overlooks the seventh candidate, which is the correct one, and takes shortcuts, as highlighted in red. When only considering ten candidates, it successfully considers the seventh candidate, among other options, as highlighted in blue, and selects it.

**Input sentence and mention:** hydrogen is less than 25%. [ 47 ] [ 48 ] [ 49 ] [ edit ] Production For more details on this topic, see [Hydrogen production] . The molecular hydrogen needed as an on-board fuel for hydrogen vehicles can be obtained through many thermochemical methods utilizing natural gas , coal (by

**LELA (Qwen3-4B):** <think>nOkay, let's tackle this problem. The user provided a text snippet and a list of candidate entities. The task is to determine which entity from the list best fits the context given in the text. First, I need to understand the input text. The text mentions "hydrogen is less than 25% [47] [48] [49] [edit] Production For more details on this topic, see [Hydrogen production]." Then it talks about producing molecular hydrogen for hydrogen vehicles through thermochemical methods using natural gas and coal. The mention of "natural gas" and "coal" suggests methods of production. The candidates include various hydrogen-related terms like "Steam reforming," "Sulfur-iodine cycle," "Copper-chlorine cycle," etc. The key here is to find which candidate entity is related to the production methods mentioned. The text specifically says "thermochemical methods utilizing natural gas, coal." Looking at the candidates, "Steam reforming" (option 2) is a method that uses natural gas and water to produce hydrogen. The Sulfur-iodine cycle (4) and Copper-chlorine cycle (8) are also thermochemical cycles, but the text mentions natural gas and coal, which are more likely associated with steam reforming. **The other options like "Hydrogen production" (1) is a general term, but the text is discussing specific methods.** The mention of "natural gas" and "coal" points towards steam reforming as the primary method here. So the answer should be option 2.</think> **"answer": 2 (X)**

**LELA (Qwen3-30B-A3B):** <think>Okay, let's see. The user provided a text snippet that mentions hydrogen production and some numbers like "less than 25%". The marked mention is [47], [48], [49], but those might be references or footnote markers. The task is to disambiguate the entity based on the context and the list of candidates. Looking at the input text: "Hydrogen is less than 25%. [47] [48] [49] [edit] Production For more details on this topic, see [Hydrogen production]. The molecular hydrogen needed as an on-board fuel for hydrogen vehicles can be obtained through many thermochemical methods utilizing natural gas, coal by..." The key part here is the mention of "Hydrogen production" in the text. The user's instruction says that the answer should be the index number of the most appropriate entity. **The candidate list includes "Hydrogen production" as entry 1. The text explicitly says "see [Hydrogen production]" which is a direct reference. So even though there are other entries like steam reforming (6) or copper-chlorine cycle (1), the context here is pointing towards the general topic of hydrogen production.** The numbers in brackets might be references to sources or sections, but the main entity being discussed is "Hydrogen production". So the answer should be 1.</think> **"answer": 1. (✓)**

Figure 16: LLM selection example from WikilinksNED Unseen Mentionws for two selection LLMs. Qwen3-4B choses the wrong candidate, following the thoughts in red, while Qwen3-30B-A3B chooses the right answer, following the blue thoughts.