CELESTA: Hybrid LLM Mention Expansion with Similarity-Based Selection for Fine-Tuning-Free Entity Disambiguation in Low-Resource Languages

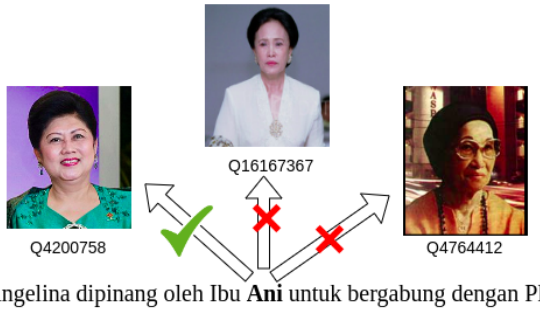
Anonymous ACL submission

Abstract

Entity Disambiguation (ED) in low-resource languages (LRLs) is severely degraded by incomplete mentions and data scarcity. We introduce CELESTA, a fine-tuning-free framework that enriches mentions using a hybrid pipeline: parallel multilingual and monolingual Large Language Models (LLMs) generate candidate expansions, a similarity-based filter selects the most contextually-faithful one, and an off-the-shelf ED model performs the final linking. Across three Indonesian datasets, CELESTA delivers consistent and measurable improvements over strong multilingual baselines like ReFinED and mGENRE. Ablation studies reveal two key drivers of this success: (1) The hybrid LLM design is critical, as monolingual models provide irreplaceable cultural context that multilingual models lack. (2) The similarity-based selection mechanism is highly effective, filtering noisy candidates to boost F1 scores by up to 0.060. The framework is language-agnostic and adaptable, offering a robust, generalizable solution for fine-tuning-free ED in LRLs. The code and dataset are available at https://anonymous.4open.science/r/CELESTA-4064.

1 Introduction

Entity Disambiguation (ED)—the task of linking entity mentions in text to their corresponding entries in a knowledge base (KB)—is a core component in downstream applications such as question answering (Tasawong et al., 2024; Xiong et al., 2024; Bouarroudj et al., 2022), information extraction (Fatahi Bayat et al., 2022), and knowledge graph population (Peng et al., 2023). Figure 1 illustrates an example where the mention *Ani* could refer to multiple individuals but is correctly mapped to Wikidata entity Q4200758 based on contextual clues. Recent advances in Large Language Models (LLMs) have significantly improved ED in high-resource languages (HRLs) with abundant annotated datasets and lexical resources (Ding et al., 2024b,a; Salman et al., 2025). However, *low-resource languages* (LRLs) continue to face persistent challenges (Huang et al., 2023), including limited annotated data, incomplete or ambiguous mention contexts, and the underperformance of ED models trained predominantly on HRL corpora (Datta and Pramanik, 2024). These include severely limited annotated data, scarce language-specific resources, and the inherent linguistic complexities that lead to incomplete or highly ambiguous mention contexts. Consequently, ED models trained predominantly on HRL corpora often underperform drastically when applied to LRLs (Datta and Pramanik, 2024).



Angelina dipinang oleh Ibu **Ani** untuk bergabung dengan PD, dan di sanalah ia bertemu dengan suaminya, Adjie.

Figure 1: Entity disambiguation process for mention *Ani* ("Angelina was invited by Mrs. Ani to join PD, where she met her husband, Adjie"), correctly linked to entity Q4200758 in Wikidata.

A particularly challenging scenario in LRL ED arises with underspecified mentions—incomplete names (e.g., *Barack* for *Barack Obama*) or abbreviations (e.g., *NYC* for *New York City*). Such mentions lack sufficient lexical cues for reliable disambiguation. In HRLs, these cases are frequently resolved using extensive alias dictionaries, acronym mappings, and popularity-based priors. However, in LRLs, these crucial resources are often nonexistent, making automatic resolution significantly

more challenging. Prior work distinguishes between *local context* (words immediately surrounding a mention) and *global context* (document-level coherence) (Ganea and Hofmann, 2017). Both signals diminish in effectiveness when the mention's surface form is incomplete or abbreviated, underscoring the need for mention expansion as a crucial preprocessing step to enrich mentions before applying broader contextual clues.

Recent LLM-based mention expansion methods have shown promise (Xin et al., 2024; Vollmers et al., 2025; Borchert et al., 2024), but often rely on a single LLM. This is insufficient for LRLs, as multilingual models may lack cultural grounding while monolingual models lack broad coverage and robustness (Qin et al., 2024; Tao et al., 2024). This leaves a significant gap for resource-scarce settings.

To address these limitations, we propose CELESTA, a hybrid ED framework for LRLs, and demonstrate it on Indonesian—a representative case where the lack of publicly accessible benchmarks constrains multi-language LRL research. Despite Indonesian's scale, reliance on a single public dataset makes it a realistic testbed for robust, fine-tuning-free methods; this gap motivates our design and underscores our contribution of a new evaluation dataset. In brief, CELESTA enriches mentions with parallel multilingual–monolingual LLMs, selects the most context-faithful expansion via semantic similarity, and links with an off-the-shelf multilingual ED system. The principles are generalizable to other LRLs where suitable monolingual LLMs exist.

Our main contributions are:

- We propose CELESTA, a hybrid multilingual-monolingual ED framework whose effectiveness for LRLs is demonstrated through a comprehensive case study on Indonesian.
- We introduce a dual-LLM mention expansion strategy with similarity-based selection to enrich underspecified mentions while reducing hallucinations.
- We systematically evaluate prompting strategies (zero-shot and few-shot) for mention expansion in an LRL setting.
- To help address the scarcity of public benchmarks that currently limits large-scale LRL evaluation, we release a new Indonesian ED dataset and show that CELESTA achieves competitive or superior performance compared to strong multilingual ED baselines (ReFinED, mGENRE) on three datasets without fine-tuning.

2 Related Works

To address the challenges of ED in LRLs—such as limited annotated datasets, scarce language resources, and incomplete mappings between mentions and KB entries—prior research has largely focused on improving the candidate generation stage. This stage is particularly critical in LRL scenarios, where resources such as alias dictionaries, surface-form look-up tables, and cross-lingual supervision signals are often sparse or unavailable.

Garcia-Durán et al. (Garcia-Duran et al., 2022) propose Pivot Token Indexes (PTI), a lightweight and high-recall candidate generation framework for LRLs. PTI constructs prior and posterior token–entity probability indexes using training data from both pivot and target languages, enabling efficient retrieval even when lexical overlap is minimal. Their results across multiple language pairs show that PTI outperforms state-of-the-art neural models like CHARAGRAM in both recall and computational efficiency, making it suitable for resource-constrained settings.

Barrena et al. (Barrena et al., 2021) investigate zero-shot ED methods that do not require target-language supervision. They introduce a word-expert-style classifier in which each mention string is treated as a separate classification task trained solely on English Wikipedia. While scalable, this approach depends heavily on surface-form overlap across languages and may struggle with morphologically rich or lexically distant LRLs.

Fu et al. (Fu et al., 2020) highlight the limitations of cross-lingual ED systems that rely heavily on Wikipedia interlanguage links, which are often sparse or incomplete for LRLs. They propose QuEL, a query log–based candidate generation system that retrieves candidate entities using multilingual web search combined with mapping heuristics. QuEL achieves substantial improvements in recall and linking accuracy, showing that external user-generated resources (e.g., query logs) can effectively augment traditional KB-based methods.

Eggertsson et al. (Eggertsson et al., 2023) present an ED system for Icelandic that uses the Icelandic Wikipedia as a KB, with candidate generation via alias tables and Wikipedia API search.

Their ranking model, IceLUKE, is an entity-aware Icelandic language model trained on fine-grained entity typing and Wikipedia data. IceLUKE outperforms several Icelandic transformer-based baselines, achieving 95.2% accuracy. Their results also show that combining fine-grained typing with strong priors (e.g., entity in-degree) can serve as robust rule-based baselines in low-resource environments.

Recent advances in LLM-based ED have been predominantly English-centric. ChatEL (Ding et al., 2024b) introduces a zero-shot prompting framework that ranks candidate entities using a ChatGPT-style model. EntGPT (Ding et al., 2024a) treats ED as a sequence generation task with entity typing constraints, generating entity names autoregressively. Both achieve strong results but assume rich English-centric training data and high-quality entity descriptions. In a complementary direction, Vollmers et al. (Vollmers et al., 2025) propose a unified ED and mention detection model, fine-tuned from T5, and use LLaMA-3 for contextual augmentation of mentions (e.g., expanding "Angelina" to "Angelina Jolie"). They report that LLM-based mention expansion is particularly effective for ambiguous or incomplete mentions.

Despite these advances, most methods assume well-resourced KBs and HRLs. In contrast, our approach, CELESTA, directly targets mention incompleteness in LRLs by combining multilingual and monolingual LLMs for complementary mention expansion. This focus on enhancing surface forms before linking sets it apart from prior work that concentrates primarily on candidate retrieval.

3 The Architecture

Figure 2 illustrates the architecture of our approach, dubbed CELESTA. The system takes as input a sentence with an identified entity mention and returns the most appropriate linked entity from a KB.

3.1 Hybrid-LLM and Similarity-Based Mention Expansion

This component performs the core task of expanding an identified mention into a more explicit and contextually enriched form. To leverage their complementary strengths, CELESTA processes the mention and its sentence context in parallel through two distinct LLMs:

- A multilingual LLM, which offers broad generalization across languages and can capture entities known globally, and
- A monolingual LLM, which provides richer cultural and linguistic grounding specifically for the target LRL.

Each LLM independently generates a candidate expansion (e_1 and e_2) for the mention. To best leverage their respective strengths, the multilingual model is prompted in English, capitalizing on its robust cross-lingual instruction-following, while the monolingual model is prompted in the target LRL to activate its deep, language-specific knowledge. Our architecture ensures that both these perspectives are generated concurrently, making them readily available for subsequent selection.

Similarity-Based Selection The two candidate expansions are evaluated by a similarity-based selection module, which determines the final expanded mention. This selection is made by measuring the semantic similarity between an expanded sentence (formed using each candidate) and the original sentence context. Specifically, for each candidate expansion e_i, we construct an expanded sentence s'_i by replacing the original mention in the input sentence s with e_i. We then encode both the expanded sentence s'_i and the original sentence s using a multilingual sentence embedding model, obtaining their respective embeddings $\phi(s'_i)$ and $\phi(s)$. The candidate expansion whose corresponding expanded sentence s'_i yields the highest cosine similarity with the original sentence s is selected as the final expansion:

$$\text{sim}(s'_i, s) = \frac{\phi(s'_i) \cdot \phi(s)}{\|\phi(s'_i)\| \cdot \|\phi(s)\|} \quad \text{for } i = 1, 2 \quad (1)$$

$$\hat{e} = \arg \max_{e_i \in \{e_1, e_2\}} \text{sim}(s'_i, s) \quad (2)$$

Here, $\phi(\cdot)$ denotes the sentence embedding function. Cosine similarity is chosen for its robustness and model-agnostic nature, providing a direct measure of contextual alignment between the expanded mention within s'_i and the original sentence s.

3.2 Entity Disambiguation

The selected expansion $\hat{e}$ and its enriched sentence context (i.e., the selected expanded sentence s') are passed to an off-the-shelf multilingual ED module. We employ this module given the limited availability of LRL-specific ED models and the prior work

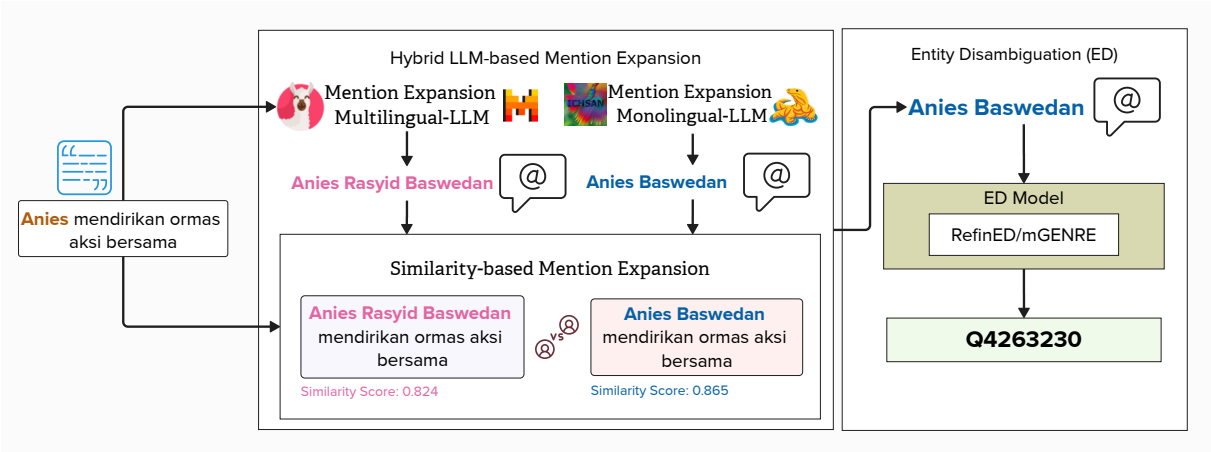


Figure 2: Overview of the CELESTA architecture for entity disambiguation in low-resource languages, using Indonesian as a case study. Given a sentence with an identified entity mention (e.g., "*Anies mendirikan ormas aksi bersama*" — *Anies founded the joint action organization*), the mention is sent in parallel to a multilingual LLM and a monolingual LLM, each producing a candidate expansion (e.g., "**Anies Rasyid Baswedan**" and "**Anies Baswedan**"). A similarity-based selection module compares each expanded mention in its sentence context to the original sentence using cosine similarity, selecting the most contextually appropriate expansion (i.e. "**Anies Baswedan**"). The selected expansion is then passed to an off-the-shelf ED model (e.g., ReFinED or mGENRE) to generate and link to the correct Wikidata entity (e.g., Q4263230).

****Example Outputs:****
- Input: mention = "PBB", sentence = "Indonesia adalah anggota aktif PBB sejak awal."
  Response= Perserikatan Bangsa-Bangsa
- Input: mention = "UI", sentence = "Dia lulus dari UI dengan predikat cumlaude."
  Response= Universitas Indonesia
- Input: mention = "Joko", sentence = "Presiden Joko menghadiri KTT G20 di Bali."
  Response= Joko Widodo
- Input: mention = "Bung Karno", sentence = "Bung Karno dikenal sebagai proklamator kemerdekaan Indonesia."
  Response= Soekarno
- Input: mention = "Presiden Prabowo Subianto", sentence = "Presiden Prabowo Subianto akan mengunjungi Malaysia bulan ini."
  Response= Complete"""

task_instruction = f""" Prompt: Expand the following entity mention "**{mention}**" based on the context: "**{sentence}**".

****Instructions:****
- Provide a meaningful expansion of the given mention in **Indonesian**, based on the provided context.
- If the mention is an **abbreviation or acronym**, return its full form.
- If the mention is an **incomplete name** (e.g., only a first name or partial name), return the full name.
- If the mention is already complete (a full name or a full term), return: Response= Complete
- If no suitable expansion is found, return: Response= None.
- Only return the response in the format:

Response=

Figure 3: CELESTA prompting template for mention expansion in Indonesian using few-shot prompting.

suggesting that multilingual ED can be effective in low-resource settings (Zeng et al., 2020). In this stage, only the disambiguation component of the ED module is invoked—mention detection and typing are not used—since mention boundaries are assumed to be given in our setup. The module ranks entity candidates and outputs a knowledge-base identifier (e.g., a Wikidata Q-ID) without additional fine-tuning. See Section 4.4 for the concrete instantiation used in our experiments.

4 Experiments

Our experiments aims to answer the following Research Questions (RQ): evaluate CELESTA's performance against strong baselines (RQ1), assess the contribution of its hybrid design over individual LLMs (RQ2), measure the effectiveness of similarity-based selection (RQ3), and compare prompting strategies (RQ4).

4.1 Experimental Settings

4.1.1 Datasets

A primary obstacle for advancing ED research in LRLs is the severe scarcity of high-quality, publicly accessible evaluation benchmarks. This challenge directly informed our choice of Indonesian as a case study. To date, INDEL (Gusmita et al., 2024) is the sole publicly available dataset for this task, with other known resources like those from the LORELEI program[1] remaining non-public. This reliance on a single benchmark not only makes Indonesian a challenging and realistic test case but also underscores the difficulty of conducting multi-LRL evaluations. Our work confronts this issue by leveraging the existing benchmark and contributing a new one, INDEL-WIKI, to foster more robust evaluation for the community.

[1] https://catalog.ldc.upenn.edu/LDC2023T07

Table 1: Entity counts, sentence stats, and average entities per sentence in IndEL.

Property	IndGEL	IndQEL	IndEL-Wiki
Total Entities	4,765	2,453	24,678
Unique Entities	55	16	24,678
Sentences	2,114	2,621	24,678
Entities in Sentence	2.4	1.6	1.0

IndEL To evaluate our approach, we use the benchmark dataset from IndEL (Gusmita et al., 2024), which comprises two domains: a general domain and a specific domain, referred to as INDGEL and INDQEL, respectively, as shown in Table 1. The INDGEL dataset consists of 1,673 training, 229 validation, and 212 test sentences. In contrast, the INDQEL dataset contains 2,075 training, 283 validation, and 263 test sentences.[2]

IndEL-Wiki To complement INDEL, the only publicly available benchmark for ED in Indonesian, we introduce INDEL-WIKI for improved coverage. We constructed contextually grounded sentences by sampling from the summary sections of Indonesian Wikipedia pages for entities sourced from a multilingual ED resource.[3] After replacing pronouns and ensuring each sentence contained a single mention, we applied transformation strategies inspired by Indonesian discourse to create more realistic, underspecified mentions. For example, two-word names were commonly shortened to the first name, while names with titles (e.g., *Dr.*) were preserved to avoid ambiguity. The dataset underwent a final human annotation phase to ensure quality and comprises 17,172 training, 4,958 validation, and 2,548 test sentences (Table 1).

4.2 Evaluation Protocol

We evaluate all models using standard Precision, Recall, and F1-score. For all experiments, we use only the core disambiguation component of the ED backends, as mention boundaries are provided.

4.3 Baselines

We compared CELESTA's performance against two ED models, i.e., REFINED (Ayoola et al., 2022) and MGENRE (De Cao et al., 2022). Further details of the baseline models are as follows:

1. REFINED is a fast, end-to-end entity linking model that jointly performs mention detection, typing, and disambiguation. In our experiments, we use only its ED component, bypassing mention detection and typing, since mentions are provided. It combines type prediction and semantic similarity to enable accurate linking, including in zero-shot settings, and supports large-scale KBs like Wikidata with high efficiency.

2. MGENRE formulates ED as a sequence generation task. Given a mention and its context, it directly generates the target entity ID (e.g., a Wikidata QID) using constrained decoding. It supports multilingual and zero-shot settings by bypassing candidate generation and leveraging autoregressive language modeling for high recall.

4.4 Implementation Detail

Our selection of LLMs was guided by a commitment to using state-of-the-art, publicly accessible models that offer diversity in both scale and language specialization. This allows for a robust and reproducible evaluation of our CELESTA framework.

Multilingual LLMs We selected LLAMA-3 (70B)[4] and MISTRAL (7B)[5]. LLAMA-3 represents a powerful, state-of-the-art large model known for its strong multilingual reasoning and instruction-following capabilities (Dubey et al., 2024). In contrast, MISTRAL is a highly capable and widely-used smaller model, allowing us to assess the effectiveness and scalability of our approach with more resource-efficient models (Jiang et al., 2023).

Monolingual LLMs For the Indonesian monolingual component, we chose two specialized open-source models: KOMODO (7B)[6] (Owen et al., 2024), a Llama-2 model adapted for the Indonesian linguistic ecosystem, and MERAK (7B)[7] (Ichsan, 2023), a Mistral-7B model fine-tuned on Indonesian Wikipedia. Their inclusion is crucial for

[2] https://github.com/dice-group/IndEL

[3] This foundational multilingual ED dataset, comprising 18 languages including Indonesian, was initially accessed from Zenodo as a data.zip archive. Its original Zenodo publication could not be precisely re-located despite extensive efforts.

[4] https://huggingface.co/meta-llama/Meta-Llama-3-70B-Instruct

[5] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

[6] https://huggingface.co/suayptalha/Komodo-7B-Instruct

[7] https://huggingface.co/Ichsan2895/Merak-7B-v4-GGUF

testing our hypothesis that specialized models provide deep cultural grounding absent in multilingual LLMs. MERAK's training on encyclopedic text is particularly well-suited for recognizing canonical entity names. As both are instruction-tuned, they are highly suitable for our prompting-based expansion task.

Our experiments are structured into a three-stage pipeline that implements the architecture described in Section 3:

- *Parallel Dual-LLM Mention Expansion.* The input sentence and its identified mention are processed concurrently by a multilingual LLM and a monolingual LLM. Each model generates an independent candidate expansion using either zero-shot or few-shot prompting (the few-shot template is shown in Figure 3). To ensure deterministic and high-fidelity outputs, the generation temperature is set to a low value of 0.1. This stage implements our core hypothesis: combining the broad cross-lingual generalization of a multilingual model with the deep, culturally-grounded understanding of a monolingual model yields a more robust set of candidate expansions.

- *Similarity-Based Selection.* To select the most contextually appropriate expansion, we employ a semantic filtering step. For each generated candidate, we create an expanded sentence by replacing the original mention. We then compute sentence-level embeddings for both the original and expanded sentences using a multilingual Sentence Transformer. Cosine similarity is used to rank the candidates. To handle cases where an LLM fails to generate a novel expansion, candidates identical to the original mention are filtered out before scoring. Furthermore, a similarity threshold ($\tau = 0.8$) acts as a safeguard: if the highest-scoring candidate's similarity does not exceed this threshold, the original mention is retained to prevent the injection of low-quality or hallucinated expansions. For our experiments, we use LaBSE (Feng et al., 2022) as our sentence encoder due to its strong performance on multilingual semantic representation tasks.

- *Entity Disambiguation.* In the final stage, the selected, contextually-enriched mention is passed to a downstream ED model. We evaluate our pipeline using two distinct, state-of-the-art ED models as backends: the zero-shot ReFinED model (Ayoola et al., 2022) and the mGENRE model (De Cao et al., 2022). For both, we bypass their mention detection and typing modules, as mention boundaries are provided in our setup. Disambiguation is performed against Wikidata, linking the expanded mention to a globally recognized entity identifier. This final step directly benefits from the high-quality, semantically-grounded inputs produced by the preceding expansion and selection stages.

Table 2: Performance comparison of CELESTA versus baseline models

Dataset	Model	Precision	Recall	F1
IndGEL	ReFinED	**0.749**	0.547	0.633
	mGENRE	0.742	0.718	0.730
	CELESTA (ours)	0.748	**0.722**	**0.735**
IndQEL	ReFinED	0.208	0.160	0.181
	mGENRE	0.298	0.298	0.298
	CELESTA (ours)	**0.298**	**0.298**	**0.298**
IndEL-Wiki	ReFinED	**0.627**	0.327	0.430
	mGENRE	0.601	0.489	0.539
	CELESTA (ours)	0.595	**0.495**	**0.540**

5 Results and Discussion

Table 2 shows that CELESTA consistently outperforms or matches the strong mGENRE and ReFinED baselines across all three datasets. On the general domain (INDGEL) and our larger, more heterogeneous INDEL-WIKI dataset, CELESTA establishes a new state-of-the-art F1 score. On the highly ambiguous INDQEL dataset, it matches the performance of the best baseline, mGENRE, demonstrating its robustness even in challenging, low-context scenarios. This consistent improvement is driven by CELESTA's balanced precision-recall profile, confirming its ability to deliver accurate predictions without sacrificing coverage.

5.1 Performance of CELESTA Compared to Individual LLMs (RQ2)

Our results confirm the superiority of the hybrid CELESTA framework over its individual component LLMs (RQ2), as shown in Tables 3 and 4. The performance gains are particularly notable under few-shot prompting and are most consistent when paired with the high-recall MGENRE backend. Analysis of the LLM contributions (Table 5) validates our core hypothesis: the multilin-

Table 3: Performance comparison of CELESTA versus individual LLMs in zero-shot and few-shot prompts using multilingual ED model REFINED

Dataset	Model	Zero-shot			Few-shot		
		P	R	F1	P	R	F1
INDGEL	LLAMA-3	0.727	**0.499**	**0.592**	0.777	0.531	0.631
	MISTRAL	0.699	0.411	0.517	**0.806**	0.310	0.448
	KOMODO	0.709	0.447	0.548	0.704	0.527	0.603
	MERAK	0.654	0.441	0.526	0.749	0.547	0.633
	CELESTA with REFINED						
	LLAMA-3 & KOMODO	**0.731**	0.437	0.547	0.757	0.513	0.612
	LLAMA-3 & MERAK	0.688	0.431	0.530	0.802	**0.586**	**0.677**
	MISTRAL & KOMODO	0.719	0.390	0.506	0.781	0.344	0.478
	MISTRAL & MERAK	0.678	0.402	0.505	0.779	0.503	0.611
INDQEL	LLAMA-3	0.154	0.051	0.077	**0.327**	0.058	0.099
	MISTRAL	0.179	0.131	0.151	0.072	0.029	0.042
	KOMODO	0.158	0.116	0.134	0.208	**0.160**	**0.181**
	MERAK	**0.203**	**0.149**	**0.172**	0.142	0.106	0.121
	CELESTA with REFINED						
	LLAMA-3 & KOMODO	0.138	0.047	0.071	0.282	0.073	0.116
	LLAMA-3 & MERAK	0.160	0.113	0.132	0.130	0.098	0.112
	MISTRAL & KOMODO	0.138	0.095	0.112	0.107	0.047	0.066
	MISTRAL & MERAK	0.196	0.146	0.167	0.128	0.095	0.109
INDEL-WIKI	LLAMA-3	0.581	0.243	0.342	0.627	0.327	0.430
	MISTRAL	0.565	0.232	0.329	0.552	0.201	0.294
	KOMODO	0.592	0.256	0.357	0.591	0.270	0.370
	MERAK	0.591	**0.285**	**0.385**	0.548	0.293	0.382
	CELESTA with REFINED						
	LLAMA-3 & KOMODO	0.577	0.234	0.332	0.639	0.322	0.428
	LLAMA-3 & MERAK	**0.596**	0.273	0.374	**0.641**	**0.355**	**0.457**
	MISTRAL & KOMODO	0.576	0.231	0.330	0.575	0.219	0.317
	MISTRAL & MERAK	0.564	0.248	0.345	0.581	0.270	0.369

Table 4: Performance comparison of CELESTA versus individual LLMs in zero-shot and few-shot prompts using multilingual ED model MGENRE

Dataset	Model	Zero-shot			Few-shot		
		P	R	F1	P	R	F1
INDGEL	LLAMA-3	**0.720**	**0.694**	**0.707**	0.742	0.718	0.730
	MISTRAL	0.667	0.640	0.653	0.607	0.584	0.595
	KOMODO	0.702	0.668	0.685	0.740	0.698	0.718
	MERAK	0.611	0.576	0.594	0.696	0.672	0.684
	CELESTA with MGENRE						
	LLAMA-3 & KOMODO	0.695	0.660	0.677	0.741	0.708	0.724
	LLAMA-3 & MERAK	0.631	0.596	0.613	**0.748**	**0.722**	**0.735**
	MISTRAL & KOMODO	0.657	0.632	0.644	0.623	0.602	0.612
	MISTRAL & MERAK	0.620	0.588	0.603	0.702	0.676	0.686
INDQEL	LLAMA-3	0.298	0.298	0.298	**0.274**	**0.273**	**0.273**
	MISTRAL	0.258	0.258	0.258	0.185	0.182	0.183
	KOMODO	0.252	0.251	0.251	0.269	0.269	0.269
	MERAK	0.233	0.233	0.233	0.255	0.255	0.255
	CELESTA with MGENRE						
	LLAMA-3 & KOMODO	**0.298**	**0.298**	**0.298**	0.266	0.266	0.266
	LLAMA-3 & MERAK	0.276	0.276	0.276	0.256	0.255	0.255
	MISTRAL & KOMODO	0.262	0.262	0.262	0.185	0.182	0.183
	MISTRAL & MERAK	0.236	0.236	0.236	0.202	0.200	0.201
INDEL-WIKI	LLAMA-3	0.516	**0.415**	0.460	0.601	0.489	0.539
	MISTRAL	0.457	0.360	0.403	0.447	0.363	0.401
	KOMODO	0.542	0.401	0.461	0.547	0.422	0.476
	MERAK	0.474	0.371	0.417	0.428	0.353	0.387
	CELESTA with MGENRE						
	LLAMA-3 & KOMODO	**0.548**	0.411	**0.470**	**0.618**	0.481	0.537
	LLAMA-3 & MERAK	0.521	0.412	0.460	0.595	**0.495**	**0.540**
	MISTRAL & KOMODO	0.500	0.368	0.424	0.484	0.382	0.427
	MISTRAL & MERAK	0.447	0.349	0.392	0.507	0.413	0.455

gual and monolingual models offer complementary strengths. While the multilingual model often provides the majority of correct expansions, the monolingual model (Merak) frequently resolves cases the other misses, particularly in culturally-grounded contexts, demonstrating the clear benefit of the dual-LLM architecture.

5.2 Ablation Study (Component-Wise Analysis) (RQ3)

To better understand the contribution of each component in CELESTA, we perform an ablation study by incrementally removing or modifying individual modules.

5.2.1 Impact of Hybrid LLMs (Multilingual and Monolingual)

To quantify our hybrid architecture's contribution, we compare CELESTA against single and dual-model baselines, all using the mGENRE backend (Table 6). The results confirm our hybrid design is critical: CELESTA achieves the highest F1-score (0.735), outperforming all other configurations. The analysis also validates that these gains are not a mere artifact of using a dual-model system, as both *Two Monolinguals* and *Two Multilinguals* setups underperform. Notably, for the dual-multilingual configuration, prompting in Indonesian causes a substantial performance degradation (0.551 F1) compared to English (0.684 F1). This confirms that multilingual models, best leveraged in English, cannot substitute for a genuine monolingual's deep, native-language understanding. CELESTA's necessity lies in its specific design: it effectively combines the multilingual model's high-precision global coverage with the monolingual's irreplaceable high-recall, culturally-grounded knowledge to achieve a balanced and superior performance.

Table 5: Contribution of each LLM and the original mention to correct predictions in the dual multilingual–monolingual mention expansion setup.

LLM Pair	LLM1 (%)	LLM2 (%)	Original (%)
CELESTA with REFINED			
LLaMA-3 & Komodo	64.49	12.32	23.19
LLaMA-3 & Merak	41.46	56.71	1.83
Mistral & Komodo	79.28	16.22	4.5
Mistral & Merak	43.24	56.76	0.0
CELESTA with MGENRE			
LLaMA-3 & Komodo	59.83	14.53	25.64
LLaMA-3 & Merak	41.62	57.54	1.12
Mistral & Komodo	78.86	14.43	6.38
Mistral & Merak	46.87	53.43	0.0

5.2.2 Effectiveness of Similarity-Based Selection

Our ablation study confirms the crucial role of the similarity-based selection mechanism (RQ3). As

Table 6: Ablation study on the contribution of different LLM configurations to CELESTA on the INDGEL dataset under few-shot prompting. All experiments use the mGENRE backend.

LLM Configuration	Precision	Recall	F1
Multilingual only (Llama-3)	0.742	0.718	0.730
Monolingual only (Merak)	0.740	0.698	0.718
Two Multilinguals (Llama-3 & Mistral, EN Prompt)	0.696	0.672	0.684
Two Multilinguals (Llama-3 & Mistral, ID Prompt)	0.591	0.515	0.551
Two Monolinguals (Komodo & Merak, ID Prompt)	0.683	0.658	0.670
CELESTA (Hybrid: Multi-EN + Mono-ID)	**0.748**	**0.722**	**0.735**

shown in Table 7, our semantic filtering approach consistently outperforms a random-selection baseline across all LLM combinations, improving the F1 score by up to +0.060. These results demonstrate the value of ranking expansions by contextual fit. The largest improvements are seen in lower-recall scenarios, indicating that the selection module is especially effective at filtering noise from diverse candidate sets.

Table 7: An ablation study comparing similarity-based and random selection of LLM-generated mention expansions on the INDGEL dataset under few-shot prompting. In similarity-based selection, the expansion with the highest contextual cosine similarity to the original sentence is chosen, while random selection picks an expansion uniformly at random.

Model	Precision	Recall	F1
CELESTA with SIMILARITY-BASED MENTION EXPANSION			
LLAMA-3 & KOMODO	0.757	0.513	**0.612**
LLAMA-3 & MERAK	0.802	0.586	**0.677**
MISTRAL & KOMODO	0.781	0.344	**0.478**
MISTRAL & MERAK	0.779	0.503	**0.611**
CELESTA with RANDOM-BASED MENTION EXPANSION			
LLAMA-3 & KOMODO	0.761	0.507	0.609
LLAMA-3 & MERAK	0.794	0.560	0.656
MISTRAL & KOMODO	0.758	0.328	0.458
MISTRAL & MERAK	0.753	0.435	0.551

5.3 Prompting Strategy Evaluation (RQ4)

As shown in Table 8, the optimal prompting strategy is context-dependent (RQ4). Few-shot prompting delivers significant performance gains on datasets with richer context and broader entity coverage (INDGEL and INDEL-WIKI). However, for the highly ambiguous, low-recall INDQEL benchmark, zero-shot prompting proves to be more robust and stable. This finding suggests that while in-context examples are generally beneficial, they may introduce noise or unhelpful biases when the disambiguation context is extremely sparse.

Table 8: Prompting strategy evaluation of CELESTA with zero-shot and few-shot prompts across all datasets using the MGENRE ED model.

Dataset	Prompt Type	Precision	Recall	F1
IndGEL	Zero-shot Prompt	0.695	0.660	0.677
IndGEL	Few-shot Prompt	**0.741**	**0.708**	**0.724**
IndQEL	Zero-shot Prompt	**0.298**	**0.298**	**0.298**
IndQEL	Few-shot Prompt	0.266	0.266	0.266
IndEL-Wiki	Zero-shot Prompt	0.548	0.411	0.470
IndEL-Wiki	Few-shot Prompt	**0.618**	**0.481**	**0.537**

Generalizability to Other LRLs. CELESTA's architecture is fundamentally language-agnostic. It relies on multilingual LLMs, sentence encoders, and ED backends. The primary prerequisite for applying CELESTA to another LRL is the availability of a suitable monolingual LLM, a condition increasingly being met.

6 Conclusion and Future Work

We introduced CELESTA, a hybrid, fine-tuning-free framework that enhances mention expansion for ED in LRLs. By combining multilingual and monolingual LLMs, CELESTA achieves state-of-the-art or competitive performance on three Indonesian datasets. Our ablation studies confirm the hybrid design is critical for capturing cultural nuances and that similarity-based selection effectively filters noise, boosting F1 scores by up to 0.060.

For future work, we plan to explore retrieval-augmented generation (RAG) for highly ambiguous cases, an adaptive strategy to select prompting methods, and fine-tuning the sentence encoder to create a more robust semantic filter.

Limitations

Our evaluation of CELESTA is focused on a single low-resource language, Indonesian, a constraint reflecting the widespread lack of standardized ED datasets for most LRLs. A crucial avenue for future work is to extend this evaluation to other languages as new benchmarks become available—an effort we hope our release of the INDEL-WIKI dataset will help catalyze. While CELESTA's principles are designed for generalizability, its empirical validation on a broader range of LRLs remains a crucial next step.

References

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

FinED: 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.

Ander Barrena, Aitor Soroa, and Eneko Agirre. 2021. Towards zero-shot cross-lingual named entity disambiguation. *Expert Syst. Appl.*, 184(C).

Florian Borchert, Ignacio Llorca, and Matthieu-P Schapranow. 2024. Improving biomedical entity linking for complex entity mentions with llm-based text simplification. *Database*, 2024:baae067.

Wissem Bouarroudj, Zizette Boufaida, and Ladjel Bellatreche. 2022. Named entity disambiguation in short texts over knowledge graphs. *Knowl. Inf. Syst.*, 64(2):325–351.

Debarghya Datta and Soumajit Pramanik. 2024. Unsupervised named entity disambiguation for low resource domains. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14922–14928, Miami, Florida, USA. Association for Computational Linguistics.

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

Yifan Ding, Amrit Poudel, Qingkai Zeng, Tim Weninger, Balaji Veeramani, and Sanmitra Bhattacharya. 2024a. Entgpt: Linking generative large language models with knowledge bases. *CoRR*, abs/2402.06738.

Yifan Ding, Qingkai Zeng, and Tim Weninger. 2024b. ChatEL: Entity linking with chatbots. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3086–3097, Torino, Italia. ELRA and ICCL.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 82 others. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Valdimar Ágúst Eggertsson, Benedikt Geir Jóhannesson, Hafsteinn Einarsson, and Hrafn Loftsson. 2023. Effective entity disambiguation in low-resource languages: A study of icelandic. In *2023 IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, pages 318–324.

Farima Fatahi Bayat, Nikita Bhutani, and H. Jagadish. 2022. CompactIE: Compact facts in open information extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 900–910, Seattle, United States. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Xingyu Fu, Weijia Shi, Xiaodong Yu, Zian Zhao, and Dan Roth. 2020. Design challenges in low-resource cross-lingual entity linking. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6418–6432, Online. Association for Computational Linguistics.

Octavian-Eugen Ganea and Thomas Hofmann. 2017. Deep joint entity disambiguation with local neural attention. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2619–2629, Copenhagen, Denmark. Association for Computational Linguistics.

Alberto Garcia-Duran, Akhil Arora, and Robert West. 2022. Efficient entity candidate generation for low-resource languages. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6429–6438, Marseille, France. European Language Resources Association.

Ria Hari Gusmita, Muhammad Faruq Amiral Abshar, Diego Moussallem, and Axel-Cyrille Ngonga Ngomo. 2024. Indel: Indonesian entity linking benchmark dataset for general and specific domains. In *Natural Language Processing and Information Systems*, pages 500–513, Cham. Springer Nature Switzerland.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12365–12394, Singapore. Association for Computational Linguistics.

Muhammad Ichsan. 2023. Merak-7b: The llm for bahasa indonesia. *Hugging Face Repository*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Louis Owen, Vishesh Tripathi, Abhay Kumar, and Biddwan Ahmed. 2024. Komodo: A linguistic expedition into indonesia's regional languages. *Preprint*, arXiv:2403.09362.

Ciyuan Peng, Feng Xia, Mehdi Naseriparsa, and Francesco Osborne. 2023. Knowledge graphs: Opportunities and challenges. *Artif. Intell. Rev.*, 56(11):13071–13102.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S. Yu. 2024. Multilingual large language model: A survey of resources, taxonomy and frontiers. *CoRR*, abs/2404.04925.

Muhammad Salman, Haoting Chen, Sergio José Rodríguez Méndez, and Armin Haller. 2025. An llm-sparql hybrid framework for named entity linking and disambiguation to wikidata. In *China Conference on Knowledge Graph and Semantic Computing and International Joint Conference on Knowledge Graphs*, pages 210–226, Singapore. Springer Nature Singapore.

Mingxu Tao, Chen Zhang, Quzhe Huang, Tianyao Ma, Songfang Huang, Dongyan Zhao, and Yansong Feng. 2024. Unlocking the potential of model merging for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8705–8720, Miami, Florida, USA. Association for Computational Linguistics.

Panuthep Tasawong, Peerat Limkonchotiwat, Potsawee Manakul, Can Udomcharoenchaikit, Ekapol Chuangsuwanich, and Sarana Nutanong. 2024. Efficient overshadowed entity disambiguation by mitigating shortcut learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15313–15321, Miami, Florida, USA. Association for Computational Linguistics.

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

Amy Xin, Yunjia Qi, Zijun Yao, Fangwei Zhu, Kaisheng Zeng, Xu Bin, Lei Hou, and Juanzi Li. 2024. Llmael: Large language models are good context augmenters for entity linking. *Preprint*, arXiv:2407.04020.

Guanming Xiong, Junwei Bao, and Wen Zhao. 2024. Interactive-KBQA: Multi-turn interactions for knowledge base question answering with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10561–10582, Bangkok, Thailand. Association for Computational Linguistics.

Weixin Zeng, Xiang Zhao, Jiuyang Tang, Zhen Tan, and Xuqian Huang. 2020. CLEEK: A Chinese long-text corpus for entity linking. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2026–2035, Marseille, France. European Language Resources Association.