CELESTA: Dual-LLM Mention Expansion for Entity Disambiguation in Low-Resource Languages

Anonymous ACL submission

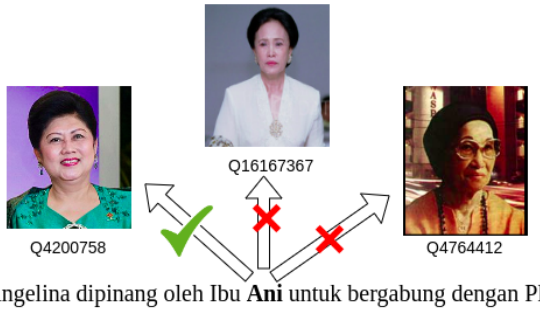
Abstract

Entity Disambiguation (ED) in low-resource languages (LRLs) is challenged by underspecified mentions and insufficient cultural grounding in standard multilingual models. We introduce CELESTA, a fine-tuning-free framework that separates mention expansion from knowledge base grounding. The framework uses a parallel dual-LLM architecture that combines a broad multilingual model with a culturally grounded monolingual model to generate candidate expansions. To mitigate the propagation of hallucinations, a similarity-based safeguard enforces a strict fallback threshold, ensuring that only high-fidelity candidates are routed to an off-the-shelf linker. Additionally, three novel evaluation suites for Southeast Asian LRLs are introduced: IndEL-Wikidata, MalayEL, and TagalogEL. Extensive evaluations demonstrate that CELESTA achieves state-of-the-art performance, with improvements up to $+0.21$ absolute F1 over direct baselines and $+0.11$ over the strongest individual LLMs. Furthermore, its language-agnostic selection mechanism enables robust fallback performance with dual multilingual models when specialized regional LLMs are unavailable. Code and data: https://anonymous.4open.science/r/CELESTA-4064.

1 Introduction

Entity Disambiguation (ED), which involves linking mentions to knowledge base (KB) entities, is foundational for downstream tasks 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 demonstrates the resolution of an ambiguous mention *Ani* to the Wikidata entity Q4200758 using contextual information.

Although Large Language Models (LLMs) have substantially improved ED for high-resource languages (HRLs) by leveraging extensive resources



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.

(Ding et al., 2024b,a; Salman et al., 2025), low-resource languages (LRLs) continue to pose significant challenges (Huang et al., 2023). The absence of a comprehensive lookup table (alias dictionaries) in LRLs renders underspecified mentions, such as partial names and initialisms, highly ambiguous (Ganea and Hofmann, 2017). Consequently, ED models trained primarily on HRL corpora and zero-shot LLMs often fail due to insufficient cultural grounding, while fine-tuning remains both costly and domain-restricted (Datta and Pramanik, 2024; Gusmita et al., 2026). Reliance on an LLM's parametric memory for end-to-end ED is impractical in data-scarce environments. Consequently, mention expansion, which enriches surface forms prior to KB retrieval, emerges as a crucial processing step.

Recent LLM-based mention expansion methods (Xin et al., 2024; Vollmers et al., 2025; Borchert et al., 2024) predominantly employ single-model paradigms, which are inadequate for LRLs. Large multilingual models often lack regional cultural specificity, while specialized monolingual LLMs face challenges in aligning local entities with globally-structured KBs (Qin et al., 2024; Tao et al., 2024). Exclusive reliance on either archi-

tecture increases the risk of generating hallucinated expansions. To address these challenges, we introduce CELESTA, a fine-tuning-free, hybrid ED framework for LRLs evaluated on three Southeast Asian languages: Indonesian, Malay, and Tagalog. The framework generates expansions using parallel multilingual and monolingual LLMs, filters hallucinations through a hybrid similarity threshold, and routes the highest-fidelity candidate to an off-the-shelf multilingual entity linker. Our main contributions are:

- **Fine-Tuning-Free Hybrid Framework:** CELESTA bridges local cultural context and global KBs for LRLs, completely circumventing computationally expensive fine-tuning.
- **Parallel Expansion & Safeguards:** Independent multilingual and monolingual LLM expansions are coupled with a hybrid similarity-based judge that explicitly penalizes hallucinations, ensuring high-fidelity downstream routing.
- **New Multilingual LRL Benchmarks:** We release an ED evaluation suite for Southeast Asian: INDEL-WIKIDATA (a human-validated gold benchmark for Indonesian), alongside strictly balanced, cross-domain silver datasets for Malay and Tagalog synthesized via a novel language-aware substitution pipeline.

2 Related Work

Prior ED research addressing the scarcity of language resources and alias dictionaries in LRLs has primarily focused on candidate generation. For instance, Garcia-Durán et al. (Garcia-Duran et al., 2022) introduce Pivot Token Indexes for efficient cross-lingual retrieval, while Eggertsson et al. (Eggertsson et al., 2023) demonstrate that fine-grained typing and statistical priors improve ranking robustness in Icelandic. However, these traditional architectures still encounter significant challenges when processing highly underspecified surface forms.

Recent LLM paradigms, including ChatEL (Ding et al., 2024b) and EntGPT (Ding et al., 2024a), exhibit strong zero-shot and autoregressive ED performance but are predominantly trained on English-centric data. These end-to-end approaches yield suboptimal results when applied to LRLs. The ELEVATE-ID benchmark for Indonesian (Gusmita et al., 2026) indicates that zero-shot prompting is particularly ineffective for acronyms and incomplete names. Furthermore, even with resource-intensive fine-tuning, monolingual models frequently produce entity hallucinations when mapping localized mentions to a global KB.

To mitigate the unreliability of end-to-end generation, recent studies have investigated the use of LLMs exclusively for preprocessing tasks. Vollmers et al. (Vollmers et al., 2025) show that employing models such as LLaMA-3 for contextual mention expansion effectively resolves ambiguities in high-resource settings. CELESTA extends this approach to LRLs by focusing on mention incompleteness. It integrates multilingual and monolingual LLMs to achieve complementary, hallucination-filtered mention expansion, thereby linking resource-scarce surface forms to standard linkers without requiring extensive dictionaries or model fine-tuning.

3 Preliminaries

Let document $D \in \mathcal{D}$ contain n NER-extracted mentions, $\mathcal{M}_D = \{m_1, \ldots, m_n\}$. Let $\mathcal{G}$ be a knowledge graph with entity vocabulary $\mathcal{E}$. For each mention $m_i \in \mathcal{M}_D$, an initial step retrieves candidate matches $\mathcal{C} \subseteq \mathcal{E}$. Unsupervised ED aims to link each m_i to a canonical entity $c_j \in \mathcal{E}$ without task-specific fine-tuning. Given the input tuple $\Psi = (D, \mathcal{M}_D, \mathcal{G})$, the task resolves pairs via:

$$\Psi = \{(m_i, c_j) \mid m_i \in \mathcal{M}_D, c_j \in \mathcal{E}, m_i \leftrightarrow c_j\} \tag{1}$$

where $m_i \leftrightarrow c_j$ denotes semantic equivalence (i.e., both refer to the same real-world entity).

4 The Approach

Figure 2 presents the CELESTA architecture, which processes a sentence context containing an identified mention and produces the corresponding linked KB entity.

4.1 Parallel Dual-LLM Mention Expansion

This stage expands an underspecified mention into a contextually enriched form. CELESTA processes the mention and sentence context in parallel using a general-purpose multilingual LLM, prompted in English to facilitate cross-lingual instruction, and a specialized regional LLM, prompted in the target LRL to enhance localized cultural grounding. The models independently generate candidate expansions e_1 and e_2 for subsequent filtering.

Dynamic Semantic Context Selection Rather than relying on static or arbitrary token truncation

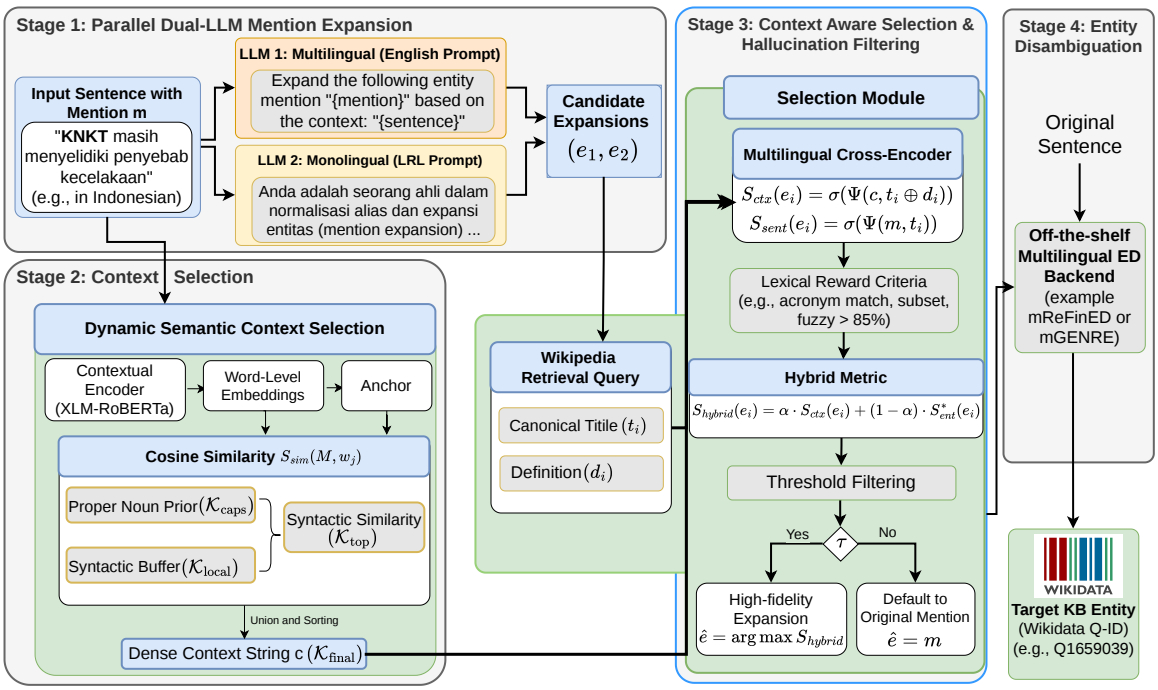


Figure 2: Overview of the CELESTA architecture for ED in LRLs. Given a mention (e.g., the acronym "KNKT"), parallel multilingual and monolingual LLMs generate candidate expansions. A selection module filters hallucinations using a hybrid similarity score that combines contextual embedding similarity with surface-level heuristics over retrieved Wikipedia titles and definitions. The highest-fidelity expansion ("Komisi Nasional Keselamatan Transportasi") is passed to an off-the-shelf linker (e.g., mReFinED) to resolve the target Wikidata entity (Q1659039).

Few-Shot Exemplars *(Context-based)*

Input 1: `mention` = "UI"
Sentence: "Dia lulus dari UI dengan predikat cum-laude."
Response: `Universitas Indonesia`

Input 2: `mention` = "Bung Karno"
...

Task Instruction Template:
`Prompt: Expand the following entity mention "{mention}" based on the context: "{sentence}".`

Formatting Instructions:

- Provide a meaningful expansion in **Indonesian**.
- If the mention is an **abbreviation/acronym**, return full form.
- ...
- Only return the response in the designated format below:

Response=

Figure 3: Representative CELESTA few-shot prompting template for mention expansion. Verbatim templates for all languages are in Appendix X.

windows, which risk omitting non-contiguous but highly informative tokens, CELESTA implements a dynamic, hybrid semantic-structural context engine to construct the dense local context representation c from an input sentence $S = \{w_1, w_2, \ldots, w_N\}$.

First, the sentence is processed through a pre-trained contextual encoder (e.g., XLM-RoBERTa (Conneau et al., 2020)). Because words may be split into multiple subword tokens, we align subword representations with individual words via mean pooling. Let P_j denote the set of subword token indices corresponding to word w_j, and $\mathbf{H}_k \in \mathbb{R}^d$ represent the hidden state of subword token k. The word-level embedding $\mathbf{h}_j$ is computed as:

$$\mathbf{h}_j = \frac{1}{|P_j|} \sum_{k \in P_j} \mathbf{H}_k \tag{2}$$

Next, let $M \subset \{1, \ldots, N\}$ denote the set of word indices comprising the target ambiguous mention m. An anchor representation for the mention is derived by aggregating its constituent word embeddings: $\mathbf{h}_M = \frac{1}{|M|} \sum_{j \in M} \mathbf{h}_j$. To isolate the top-$k$ most semantically relevant words across the sentence, we calculate the cosine similarity between the target anchor and every word vector:

$$S_{\text{sim}}(M, w_j) = \frac{\mathbf{h}_M \cdot \mathbf{h}_j}{\|\mathbf{h}_M\| \|\mathbf{h}_j\|} \tag{3}$$

The indices of the top-k highest-scoring words are extracted to form the semantic core subset, denoted as $\mathcal{K}_{\text{top}}$. To ensure syntactic continuity and capitalize on named entity clues common in low-resource environments, the engine augments $\mathcal{K}_{\text{top}}$ using two deterministic structural priors:

1. **Proper Noun Prior ($\mathcal{K}_{\text{caps}}$):** Any unstopped word featuring a leading capital letter is automatically retained, capturing hidden secondary entities or localized spatial landmarks elsewhere in the sentence.
2. **Syntactic Buffer ($\mathcal{K}_{\text{local}}$):** A local symmetric sliding window of ± 3 words immediately surrounding any word in M is explicitly preserved to protect critical local dependencies, such as adjacent verbs or prepositions.

The final dense contextual string c is compiled by taking the union of these indexes, $\mathcal{K}_{\text{final}} = \mathcal{K}_{\text{top}} \cup \mathcal{K}_{\text{caps}} \cup \mathcal{K}_{\text{local}}$, and reconstructing the sequence of words sorted chronologically according to their original relative position in S.

Context-Aware Selection and Hallucination Filtering The primary objective of the selection module is to evaluate parallel LLM outputs and dynamically identify the most contextually accurate expansion, while also serving as a strict filter against model-induced hallucinations. For each generated candidate expansion e_i, Wikipedia is queried to retrieve its corresponding canonical title t_i and the first 100 words of its introductory abstract d_i, following established practices in dense retrieval (Logeswaran et al., 2019; Wu et al., 2020). Although bi-encoder architectures encode sequences independently for efficiency, they lack the deep interaction required for nuanced disambiguation. Given that the candidate pool per mention is tightly constrained to the outputs of the parallel LLMs, this efficiency trade-off can be bypassed. Instead, a multilingual Cross-Encoder (Ψ) (Reimers and Gurevych, 2019) is employed to compute a hybrid similarity score. By applying full joint self-attention to both inputs, the Cross-Encoder captures fine-grained semantic overlaps that independent embeddings do not capture, systematically balancing deep contextual alignment with surface-level morphological fidelity.

To evaluate semantic appropriateness, the bounded contextual similarity score $S_{ctx}(e_i)$ assesses the dynamically extracted local sentence context c against the expanded entity definition. This score is defined as $S_{ctx}(e_i) = \sigma(\Psi(c, t_i \oplus d_i))$, where σ denotes a Sigmoid activation function and $\oplus$ denotes string concatenation.

To prevent the selection of semantically plausible but lexically disconnected hallucinations, the surface similarity score $S_{ent}(e_i)$ measures direct alignment between the original ambiguous mention m and the retrieved title t_i: $S_{ent}(e_i) = \sigma(\Psi(m, t_i))$. Since LRL mentions often appear as highly truncated aliases or initialisms, $S_{ent}(e_i)$ is augmented by a deterministic lexical reward. If the mention satisfies specific morphological criteria against the title or definition, such as exact acronym alignment, word-level subsetting, or a high fuzzy string similarity ratio (greater than 85%), the surface score is strictly set to a minimum of 0.95.

$$S^*_{ent}(e_i) = \begin{cases} \max(S_{ent}(e_i), 0.95), & \text{if criteria met} \\ S_{ent}(e_i), & \text{otherwise} \end{cases} \tag{4}$$

The final hybrid metric is computed as a linear combination governed by the hyperparameter α.

$$S_{hybrid}(e_i) = \alpha \cdot S_{ctx}(e_i) + (1-\alpha) \cdot S^*_{ent}(e_i) \tag{5}$$

To enforce the module's secondary function of suppressing low-confidence noise and preventing error propagation to the downstream linker, a rigid filtering threshold τ is applied. If the highest-scoring candidate does not meet this threshold, the pipeline rejects the LLM expansions and defaults to the unexpanded original mention m.

$$\hat{e} = \begin{cases} \arg\max_{e_i} S_{hybrid}(e_i), & \text{if } \max S_{hybrid} \geq \tau \\ m, & \text{otherwise} \end{cases} \tag{6}$$

4.2 Entity Disambiguation

In the final stage, the selected high-fidelity mention expansion $\hat{e}$ and its corresponding original sentence are routed to an off-the-shelf multilingual ED backend, such as mReFinED (Limkonchotiwat et al., 2023) or mGENRE (De Cao et al., 2022). A pre-trained multilingual linker is employed due to the scarcity of LRL-specific ED models, building on prior work demonstrating their utility in cross-lingual settings (Zeng et al., 2020). Since gold mention boundaries (Kolitsas et al., 2018) are provided in this setup, the linker's mention detection and typing components are bypassed, isolating only the candidate ranking and disambiguation functions. Using the enriched expansion $\hat{e}$ rather than the ambiguous original mention m grounds the search

space, enabling precise zero-shot mapping to standardized KB identifiers (e.g., a Wikidata Q-ID) without task-specific fine-tuning.

5 Experiments

The experiments are designed to address three research questions (RQs): **(RQ1)** How does CELESTA perform compared to strong, direct zero-shot ED baselines? **(RQ2)** What is the specific contribution of the parallel dual-LLM architecture compared to standalone LLM expansions across different model scales (Large vs. Small)? **(RQ3)** How effectively does the similarity-based selection module filter hallucinations and preserve precision?

5.1 Experimental Settings and Datasets

Given that CELESTA is a fine-tuning-free framework, evaluation is conducted exclusively on the test splits of three Southeast Asian Language benchmarks. To test the robustness of the framework against linguistic domain shifts and prompt sensitivity, we implement a multi-dimensional experimental design:

- **Datasets & Domain Robustness:** Evaluation is performed on IndGEL and IndEL-Wikidata for Indonesian, and our newly introduced MalayEL and TagalogEL challenge sets. To assess domain robustness, each dataset is curated into three distinct styles: *Wikipedia* (encyclopedic), *News* (formal), and *Social Media* (noisy, colloquial slang), providing insights into how CELESTA performs across varying registers of linguistic formality.
- **Linguistic Prompting Strategies:** We evaluate two prompting paradigms to determine the impact of language alignment between the LLM and the prompt instruction:
  1. **Uniform English Prompting:** All LLMs (multilingual and monolingual) receive English instructions, testing the framework's reliance on the multilingual reasoning capabilities of the base models.
  2. **Mixed-Language Prompting:** Global multilingual LLMs receive English instructions, while regional monolingual LLMs receive instructions in their native language (Indonesian, Malay, or Tagalog), testing if native-language prompting enhances the extraction of culturally-grounded entity expansions.

5.2 Baselines & Evaluation Protocol

Core disambiguation performance is isolated using standard Precision (P), Recall (R), and F1-score metrics based on gold mention boundaries. CELESTA is evaluated against two categories of baselines:

1. **Direct Zero-Shot ED Baselines:** Off-the-shelf multilingual systems evaluated in a zero-shot setting, where the unexpanded ambiguous mention is fed directly to the linker without any dataset-specific fine-tuning. We evaluate mReFinED (Limkonchotiwat et al., 2023), a discriminative bi-encoder representing dense-retrieval paradigms, and mGENRE (De Cao et al., 2022), an autoregressive sequence-to-sequence generative model.
2. **Standalone LLM Baselines:** To isolate the synergistic gains of our parallel architecture, we evaluate every LLM individually, routing its standalone expanded mention through the ED backends without the dual-model fallback mechanism.

5.3 Implementation Details

Model Selection & Two-Phase Configurations: To comprehensively evaluate the dual-LLM architecture, our selection strategy is divided into two phases. In Phase 1 (Upper Bound)/PI, we pair state-of-the-art large multilingual models—LLAMA-3 (70B) (Dubey et al., 2024) and MISTRAL (7B)[1] (Jiang et al., 2023)—with specialized regional models to establish the framework's maximum capability. In Phase 2 (Efficiency)/PII, we replace the large models with highly efficient small multilingual LLMs (3B–4B parameters), specifically LLAMA-3.2-3B, QWEN2.5-3B, and SAILOR-4B, to test architectural robustness in resource-constrained environments. Across both phases, these global models are paired with the following regional monolingual engines: *Komodo-7B*[2] (Owen et al., 2024) and *Merak-7B* (Ichsan, 2023) for Indonesian; *Malaysian-Llama-3*[3] and *Malaysian-Mistral*[4] (Zolkepli et al., 2024) for Malay; and *SeaLLM-2.5*[5]

[1] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3
[2] https://huggingface.co/suayptalha/Komodo-7B-Instruct
[3] https://huggingface.co/mesolitica/malaysian-llama-3-8b-instruct-16k
[4] https://huggingface.co/mesolitica/malaysian-mistral-7b-32k-instructions-v2
[5] https://huggingface.co/SeaLLMs/SeaLLMs-v2.5-7B-Chat

(Nguyen et al., 2024) and *Sea-Lion-v3*[6] for Tagalog.

Pipeline Execution: Stage 1 (Mention Expansion) employs 4-bit NF4 quantization (Dettmers et al., 2023) and greedy decoding to ensure high-fidelity deterministic outputs. Models are prompted using few-shot In-Context Learning (Figure 3) with $k = 5$ language-specific exemplars sampled from disjoint development sets to prevent data leakage. In Stage 3 (Similarity Selection), hybrid scores are computed using the `mmarco-mMiniLMv2-L12-H384-v1` cross-encoder evaluated across a grid search of fallback thresholds $\tau \in [0.1, 0.5]$. Finally, in Stage 4 (Entity Disambiguation), the ED backends explicitly bypass internal mention detection mechanisms to strictly evaluate the mapping of the τ-filtered expansion to a Wikidata identifier.

6 Results and Discussion

6.1 Main Results (RQ1)

To address RQ1, Table 1 presents an evaluation of CELESTA against robust direct zero-shot ED baselines, demonstrating strong cross-lingual generalization. In all eight evaluation settings, both the CELESTA PI and the highly efficient CELESTA PII configurations consistently outperform zero-shot baselines, which struggle with unexpanded and underspecified mentions. The framework yields the most substantial improvements on challenging datasets. For example, CELESTA PI achieves an absolute F1 gain of $+0.21$ over the direct ED baseline on IndEL-Wikidata (mReFinED: 0.64 vs. 0.43) and a $+0.11$ increase on IndGEL (mGENRE: 0.84 vs. 0.73). Notably, even with compact parameter pools, CELESTA PII remains highly competitive. It frequently matches the performance of its larger Phase I counterparts and significantly surpasses the zero-shot pipeline. This is evidenced by a $+0.08$ F1 boost on MalayEL (mReFinED: 0.64 vs. 0.56) and a $+0.05$ boost on TagalogEL (mReFinED: 0.68 vs. 0.63). These results demonstrate that CELESTA effectively resolves the ambiguity of underspecified mentions, establishes a new state of the art across the evaluated LRL datasets, and provides a scalable solution for resource-constrained environments.

[6] `https://huggingface.co/aisingapore/sea-lion-7b-instruct-v3`

Table 1: Consolidated F1 performance of CELESTA. Bold and underlining indicate the highest and the second-highest scores. "Phase I (PI)" baselines represent the maximum score achieved by either the large multilingual or the monolingual model independently. "Phase II (PII)" baselines represent the maximum independent score from the small multilingual or monolingual pool. Optimal configurations are in Appendix A.

Model	mReFinED (F1)	mGENRE (F1)
IndGEL		
Baseline: Direct ED	0.66	0.73
PI Standalone (LLAMA-3 / KOMODO)	<u>0.78</u>	<u>0.80</u>
PII Standalone (MERAK / KOMODO)	0.70	<u>0.80</u>
CELESTA PI (LLAMA-3 & KOMODO)	**0.79**	**0.84**
CELESTA PII (SAILOR-4B & KOMODO)	0.73	0.79
IndEL-Wikidata		
Baseline: Direct ED	0.43	0.54
PI Standalone (LLAMA-3)	0.56	0.64
PII Standalone (MERAK)	0.54	0.62
CELESTA PI (LLAMA-3 & MERAK)	**0.64**	**0.70**
CELESTA PII (SAILOR-4B & MERAK)	<u>0.60</u>	<u>0.66</u>
Malay EL		
Baseline: Direct ED	0.56	0.45
PI Standalone (MAL-LLAMA-3)	0.54	**0.51**
PII Standalone (MAL-LLAMA-3)	0.54	**0.51**
CELESTA PI (LLAMA-3 & MAL-MISTRAL)	**0.65**	<u>0.50</u>
CELESTA PII (LLAMA-3.2 / QWEN2.5 & MAL-MISTRAL)	<u>0.64</u>	0.49
TagalogEL		
Baseline: Direct ED	0.63	0.52
PI Standalone (SEA-LION / MISTRAL)	0.64	0.53
PII Standalone (SEALLM-2.5 / SEA-LION-V3	0.64	0.53
CELESTA PI (MISTRAL & SEA-LLM)	**0.70**	**0.61**
CELESTA PII (Qwen2.5-3B, Sea-Lion-v3) / (LLaMA-3.2-3B, SeaLLM-2.5)	<u>0.68</u>	<u>0.59</u>

6.2 Contribution of the Dual-LLM (RQ2)

The dual-LLM architecture was evaluated against standalone expansions at both Phase I (Large) and Phase II (Small) scales. Across all datasets, the dual-LLM pipeline consistently outperformed standalone models, indicating that integrating a global perspective with local expertise enhances reasoning capabilities. For example, on IndGEL (Table 1), LLAMA-3 and KOMODO achieved the best standalone Phase I performance with mReFinED and mGENRE, respectively. However, CELESTA PI (LLAMA-3 and KOMODO) surpassed both individual results. Although improvements over the best standalone models are sometimes modest, such as a peak of 0.79 F1, CELESTA consistently leverages synergy to resolve edge-case ambiguities. This approach reduces blind spots: the local model interprets cultural nuances, while the global model provides cross-lingual structure. Additionally, this parallel architecture scales efficiently, with Phase II

Table 2: Ablation study of CELESTA using mReFinED. Datasets are abbreviated as IndGEL (**IG**), IndEL-Wiki (**IW**), Malay (**MY**), and Tagalog (**TG**). Δ indicates the average F1 degradation compared to the full pipeline.

Config.	IG	IW	MY	TG	Avg.	Δ
CELESTA	**0.79**	**0.64**	**0.65**	**0.70**	**0.70**	–
Selection Ablations						
Lexical (TF-IDF)	0.76	0.64	0.65	0.69	0.69	-0.01
Internal (Logprobs)	0.77	0.58	0.53	0.62	0.62	-0.07
Pipeline Ablations						
No Fallback ($\tau = 0$)	0.73	0.58	0.62	0.65	0.65	-0.05
Static (Best Indiv.)	0.78	0.56	0.54	0.64	0.63	-0.07

Table 3: Attribution of correct predictions for PI dual-LLM settings (mReFinED backend). Percentages denote whether the expansion was generated exclusively by the Global model, Local model, Both models, or if the pipeline defaulted to the unexpanded input (Orig.).

Dataset	LLM Pair	Global (%)	Local (%)	Both (%)	Orig. (%)
IndGEL	LLaMA-3 & Komodo	26.75	11.15	53.60	8.51
IndEL-Wikidata	LLaMA-3 & Merak	14.41	21.08	60.37	4.15
MalayEL	LLaMA-3 & Mal-Mistral	8.78	20.03	61.35	9.84
TagalogEL	Mistral & SeaLLM-2.5	8.07	16.29	57.31	18.32

configurations remaining highly competitive under strict parameter constraints.

6.3 Ablation Study: Effectiveness of the Selection Module (RQ3)

The similarity-based selection module serves as a critical safeguard, filtering hallucinations by applying a fallback threshold (τ) to the hybrid metric (S_{hybrid}). This mechanism enables CELESTA to revert to the unexpanded baseline when candidate expansions lack sufficient contextual or lexical grounding. Ablation experiments using mReFinED (Table 2) demonstrate the necessity of this approach. Eliminating the fallback ($\tau = 0$) results in unconditional acceptance of all expansions, which introduces substantial downstream noise and reduces the average F1 score by -0.05 (from 0.70 to 0.65), with particularly negative effects on IndGEL and IndEL-Wiki. Additionally, the hybrid metric outperforms alternative filters. Substituting S_{hybrid} with either LLM internal confidence (Logprobs) or purely lexical similarity (TF-IDF) reduces the average F1 score by -0.07 and -0.01, respectively. These results underscore the importance of jointly computing contextual and surface similarities. Moreover, static routing, which relies exclusively on the best individual LLM rather than dynamically selecting from the dual-LLM pool, results in a 0.07 F1 reduction. In summary, CELESTA's dynamic selection effectively integrates the complementary strengths of parallel LLMs while mitigating the risk of hallucinations.

6.4 Discussions

Dual-LLM Synergy and Hallucination Fallback CELESTA achieves state-of-the-art performance across all LRL datasets, with results such as 0.79 F1 on IndGEL and 0.70 on TagalogEL. Single LLMs consistently underperform: multilingual models lack sufficient cultural grounding, while localized models struggle to map entities to global KBs. Integrating both approaches addresses these limitations; for example, linking SAILOR-4B with MERAK increases IndEL-Wikidata Phase II F1 from 0.54 to 0.60. Source attribution (Table 3) shows that these improvements arise from mutually exclusive knowledge sets. On IndEL-Wikidata, MERAK uniquely resolves 21.08% of correct predictions that LLAMA-3 misses. Importantly, this recall gain does not compromise precision: the similarity module (τ) reverts to the original mention in 8.51% of correct disambiguations, effectively isolating localized hallucinations. In zero-resource scenarios where regional LLMs are unavailable, parallel multilingual models (LLAMA-3 and MISTRAL) remain highly competitive (Table 4), demonstrating the architecture's generalizability.

Context Dynamics and Morphological Fidelity Static context truncation frequently introduces semantic noise. CELESTA mitigates this issue by dynamically deriving a contextual core ($\mathcal{K}_{\text{top}}$) using XLM-RoBERTa, and supplementing it with structural priors ($\mathcal{K}_{\text{caps}}$, $\mathcal{K}_{\text{local}}$). This approach significantly outperforms fixed-window baselines, achieving an F1 score of 0.79 compared to 0.66 on IndGEL dataset with mReFinED backend (see Table 1). Additionally, LRLs often exhibit heavily truncated initialisms. To prevent the Cross-Encoder (S_{ctx}) from favoring semantically plausible but morphologically incorrect hallucinations, the selection module applies a deterministic lexical reward (S^*_{ent}), which increases scores for exact acronym or subset matches. This dual-constraint strategy enables robust mention normalization. As a result, while standard linkers experience substantial degradation on noisy colloquial text, CELESTA's lo-

Table 4: Empirical evaluation of CELESTA's generalizability to other LRLs using the language-agnostic fallback configuration (dual-multilingual LLAMA-3 & MISTRAL). We report the highest F1 score achieved for each context-window strategy using the **mReFinED** backend, with the required threshold (τ) noted in parentheses.

Context Strategy	IndGEL	IndEL-W	Malay	Tagalog
Full Sent.	**0.79** (0.2)	**0.60** (0.5)	0.63 (0.2)	**0.69** (0.5)
Semantic $k = 5$	**0.79** (0.2)	**0.60** (0.5)	**0.64** (0.2)	**0.69** (0.5)

calized expansion buffers this noise and maintains high-fidelity inputs across encyclopedic, formal (News), and informal (Social Media) domains (see Appendix C.B.

Scalability and Prompting Asymmetry CELESTA demonstrates strong robustness across different backends (mReFinED, mGENRE) and removes the requirement for large-scale fine-tuning. The hybrid routing approach enables compact 3B models to outperform standalone 70B baselines. Evaluation reveals distinct linguistic asymmetries in prompt sensitivity (Table 5): mixed-language prompting (English for global, native language for regional) optimizes performance in Indonesian, whereas uniform English prompting yields the best results for Malay and Tagalog. This finding highlights the importance of the global LLM as a universally responsive and stabilizing component. Ultimately, CELESTA offers a scalable framework for integrating deeply localized entities into global KBs without incurring the prohibitive costs of manual alias-dictionary curation.

7 Conclusion and Future Work

We introduced CELESTA, a fine-tuning-free hybrid ED framework for LRLs. By separating mention expansion from knowledge base grounding, CELESTA utilizes both multilingual and monolingual LLMs in parallel to combine broad cross-lingual generalization with deep regional grounding. To reduce the risk of propagating LLM hallucinations, a similarity-based selection module applies a strict fallback threshold, dynamically filtering ungrounded expansions. CELESTA achieves a new state-of-the-art across eight configurations in Indonesian, Malay, and Tagalog, with up to $+0.21$ absolute F1 improvements over direct baselines. Ablation studies demonstrate that the dual-LLM approach and similarity safeguard are essential for addressing structural blind spots and ensuring high-fidelity disambiguation. In addition to the architecture, we release IndEL-Wikidata, MalayEL, and TagalogEL to help address the significant scarcity of LRL evaluation data. We plan to integrate retrieval-augmented generation (RAG) to improve grounding for highly ambiguous tail entities, develop adaptive prompting strategies, and extend this robust, fine-tuning-free architecture to a wider range of severely underrepresented languages.

8 Limitations

Reliance on Wikipedia Coverage The selection module determines contextual similarity by dynamically querying the introductory abstract of each candidate entity. When a highly localized entity lacks a corresponding Wikipedia entry, the Cross-Encoder cannot compute a semantic score and must revert to the unexpanded mention. As a result, CELESTA is fundamentally constrained by the extent of Wikipedia coverage in the target language.

Sensitivity to Instruction Tuning The experiments demonstrate asymmetric sensitivities to native-language and English prompting across regional LLMs. Since the system relies on few-shot in-context learning, its effectiveness is constrained by the quality and semantic coverage of the selected monolingual model's instruction-tuning corpus.

Rigid Lexical Heuristics To evaluate surface similarity, the framework employs hardcoded morphological heuristics, such as exact acronym matching and fuzzy string similarity exceeding 85%. Although these deterministic rewards and rigid thresholds, including the fallback threshold τ, are effective for Indonesian, Malay, and Tagalog, they may not generalize effectively to languages with substantially different morphological typologies, such as agglutinative or logographic languages, without significant retuning.

References

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.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco

Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

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.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

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.

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, Asep Fajar Firmansyah, Hamada M. Zahera, and Axel-Cyrille Ngonga Ngomo. 2026. Elevate-id: Extending large language models for end-to-end entity linking evaluation in indonesian. *Data & Knowledge Engineering*, 161:102504.

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.

Nikolaos Kolitsas, Octavian-Eugen Ganea, and Thomas Hofmann. 2018. End-to-end neural entity linking. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 519–529, Brussels, Belgium. Association for Computational Linguistics.

Peerat Limkonchotiwat, Weiwei Cheng, Christos Christodoulopoulos, Amir Saffari, and Jens Lehmann. 2023. mReFinED: An efficient end-to-end multilingual entity linking system. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15080–15089, Singapore. Association for Computational Linguistics.

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.

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Zhiqiang Hu, Chenhui Shen, Yew Ken

Chia, Xingxuan Li, Jianyu Wang, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, Hang Zhang, and Lidong Bing. 2024. SeaLLMs - large language models for Southeast Asia. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 294–304, Bangkok, Thailand. Association for Computational Linguistics.

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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

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.

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

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.

Husein Zolkepli, Aisyah Razak, Kamarul Adha, and Ariff Nazhan. 2024. Large malaysian language model based on mistral for enhanced local language understanding. *Preprint*, arXiv:2401.13565.

A Operational Configurations

Table 5 details the optimal operational configurations (fallback threshold τ, prompt language, and context window k) identified for CELESTA across datasets and backends. These settings were determined through an empirical search over the pipeline's decision logic, ensuring peak performance without requiring gradient-based optimization or weight updates.

B Complete Master Macro Results Reference

This section houses the complete, un-distilled macro evaluation results across both backends.

- Table 6: Performance comparison of CELESTA and large individual LLMs cross-evaluated across targeted model pairs on mReFinED and mGENRE backends (Large Multilingual configurations).
- Table 7): Complete performance comparison of CELESTA and compact individual LLMs across mReFinED and mGENRE backends under Phase II (Small Multilingual configurations).
- Table 8: Performance comparison of CELESTA and standalone LLMs across the mReFinED and mGENRE backends on the IndGEL dataset (Phase II: Small Multilingual Configurations).
- Table 9: Performance comparison of CELESTA and standalone LLMs across the mReFinED and mGENRE backends on the IndEL-Wikidata dataset (Phase II: Small Multilingual Configurations).
- Table 10: Performance comparison of CELESTA and standalone LLMs across the mReFinED and mGENRE backends on the MalayEL dataset (Phase II: Small Multilingual Configurations).
- Table 11: Performance comparison of CELESTA and standalone LLMs across the mReFinED and mGENRE backends on the TagalogEL dataset (Phase II: Small Multilingual Configurations).

C Exhaustive Ablation Grids

The following subsections document the complete grid sweeps across selection thresholds ($\tau \in [0.1, 0.5]$), prompting language frameworks, and context constraint shifts.

Appendix C.A Semantic Token Window Ablations (`semantic_k` = 5)

C.A.1 Language-Aligned Mixed-Prompt Tiers

- Table 12: Performance comparison of CELESTA and individual LLMs across mReFinED and mGENRE backends on IndGEL.
- Table 13: Performance comparison of CELESTA and individual LLMs across mReFinED and mGENRE backends on IndEL_Wikidata.
- Table 14: Performance comparison of CELESTA and individual LLMs across mReFinED and mGENRE backends on MalayEL.
- Table 15: Performance comparison of CELESTA and individual LLMs across mReFinED and mGENRE backends on TagalogEL.

C.A.2 Uniform All-English Prompt Control Baselines

- Table 16: Performance comparison of CELESTA and individual LLMs across mReFinED and mGENRE backends on IndGEL using an English prompt for both multilingual and monolingual LLMs.
- Table 17: Performance comparison of CELESTA and individual LLMs across mReFinED and mGENRE backends on IndEL-Wikidata using an English prompt for both multilingual and monolingual LLMs.
- Table 18: Performance comparison of CELESTA and individual LLMs across mReFinED and mGENRE backends on MalayEL using an English prompt for both multilingual and monolingual LLMs.
- Table 19: Performance comparison of CELESTA and individual LLMs across mReFinED and mGENRE backends on TagalogEL using an English prompt for both multilingual and monolingual LLMs.

Table 5: Optimal operational configurations (fallback threshold τ, prompt language, and context window k) identified for CELESTA across datasets and backends for both Phase I (Standard Multilingual) and Phase II (Small Multilingual) models.

Dataset	Backend	τ	Prompt Language	Context Window
Phase I: Standard Multilingual Configurations				
IndGEL	mReFinED	0.1	Mixed	$k = 5$
IndGEL	mGENRE	0.2	Mixed	$k = 5$
IndEL-Wikidata	mReFinED	0.5	Mixed	$k = 5$
IndEL-Wikidata	mGENRE	0.5	Mixed	$k = 5$
MalayEL	mReFinED	0.5	English	Full Sentence
MalayEL	mGENRE	0.2	English	Full Sentence
TagalogEL	mReFinED	0.5	English	Full Sentence
TagalogEL	mGENRE	0.5	English	Full Sentence
Phase II: Small Multilingual Configurations				
IndGEL	mReFinED	0.2	English	$k = 5$
IndGEL	mGENRE	0.2	English	$k = 5$
IndEL-Wikidata	mReFinED	0.5	English	$k = 5$
IndEL-Wikidata	mGENRE	0.5	English	$k = 5$
MalayEL	mReFinED	0.5	English	$k = 5$
MalayEL	mGENRE	0.5	English	$k = 5$
TagalogEL	mReFinED	0.5	English	$k = 5$
TagalogEL	mGENRE	0.5	English	$k = 5$

Appendix C.B Full Sentence Context Window Ablations

- Table 20: Performance comparison of CELESTA and standalone LLMs across the mReFinED and mGENRE backends on the IndGEL dataset.
- Table 21: Performance comparison of CELESTA and standalone LLMs across the mReFinED and mGENRE backends on the IndEL-Wikidata dataset.
- Table 22: Performance comparison of CELESTA and individual LLMs across mReFinED and mGENRE backends on MalayEL using an English prompt for both multilingual and monolingual LLMs.
- Table 23: Performance comparison of CELESTA and individual LLMs across mReFinED and mGENRE backends on TagalogEL using an English prompt for both multilingual and monolingual LLMs.
- Table 24: Detailed domain-level performance comparison for the **Malay** datasets (MALAYEL-WIKI, MALAYEL-NEWS, and MALAYEL-SOCIAL) across the REFINED and MGENRE backends.

Table 6: Performance comparison of CELESTA and large individual LLMs cross-evaluated across targeted model pairs on mReFinED and mGENRE backends (Large Multilingual configurations). Indonesian results are per dataset; Malay and Tagalog are domain macro-averages. Bold indicates the top score for each metric within each backend group, with backend-specific configurations listed above each column.

	mReFinED Backend			mGENRE Backend		
Model	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
IndGEL	*Config:* [$\tau = 0.1$ | Mixed | $k = 5$]			*Config:* [$\tau = 0.2$ | Mixed | $k = 5$]		
Baseline: Direct ED Model	0.77	0.57	0.66	0.75	0.72	0.73
Individual LLM Expansion						
LLaMA-3	**0.87**	0.70	0.78	0.82	0.79	0.80
Mistral	0.73	0.56	0.64	0.76	0.73	0.75
Komodo	0.79	0.57	0.67	0.81	0.78	0.80
Merak	0.78	0.63	0.70	0.79	0.76	0.78
CELESTA *(Dual-LLM Mention Expansion)*						
LLaMA-3 & Komodo	0.87	**0.73**	**0.79**	**0.85**	**0.83**	**0.84**
LLaMA-3 & Merak	0.83	0.69	0.76	0.81	0.80	0.80
Mistral & Komodo	0.82	0.68	0.75	0.82	0.80	0.81
Mistral & Merak	0.81	0.67	0.73	0.81	0.79	0.80
IndEL-Wikidata	*Config:* [$\tau = 0.5$ | Mixed | $k = 5$]			*Config:* [$\tau = 0.5$ | Mixed | $k = 5$]		
Baseline: Direct ED Model	0.63	0.33	0.43	0.60	0.49	0.54
Individual LLM Expansion						
LLaMA-3	0.73	0.45	0.56	0.71	0.58	0.64
Mistral	0.58	0.28	0.37	0.53	0.42	0.47
Komodo	0.60	0.29	0.40	0.56	0.41	0.47
Merak	0.71	0.44	0.54	0.68	0.57	0.62
CELESTA *(Dual-LLM Mention Expansion)*						
LLaMA-3 & Komodo	0.72	0.53	0.61	0.70	0.64	0.67
LLaMA-3 & Merak	**0.74**	**0.56**	**0.64**	**0.73**	**0.68**	**0.70**
Mistral & Komodo	0.64	0.45	0.53	0.62	0.56	0.59
Mistral & Merak	0.69	0.51	0.59	0.67	0.62	0.65
MalayEL-Average	*Config:* [$\tau = 0.5$ | English | Full Sent.]			*Config:* [$\tau = 0.2$ | English | Full Sent.]		
Baseline: Direct ED Model	0.78	0.44	0.56	0.55	0.37	0.45
Individual LLM Expansion						
LLaMA-3	0.78	0.37	0.50	0.61	0.38	0.47
Mistral	0.76	0.40	0.52	0.61	0.40	0.49
Malaysian-Llama-3	**0.79**	0.41	0.54	**0.62**	0.43	**0.51**
Malaysian-Mistral	0.77	0.39	0.52	0.61	0.40	0.49
CELESTA *(Dual-LLM Mention Expansion)*						
LLaMA-3 & Malaysian-LLaMA-3	0.76	0.55	0.64	0.57	0.43	0.49
LLaMA-3 & Malaysian-Mistral	0.79	0.55	**0.65**	0.59	0.43	0.50
Mistral & Malaysian-LLaMA-3	0.75	0.55	0.64	0.57	0.42	0.48
Mistral & Malaysian-Mistral	0.77	**0.56**	0.65	0.58	0.43	0.49
TagalogEL-Average	*Config:* [$\tau = 0.5$ | English | Full Sent.]			*Config:* [$\tau = 0.5$ | English | Full Sent.]		
Baseline: Direct ED Model	0.82	0.52	0.63	0.66	0.43	0.52
Individual LLM Expansion						
LLaMA-3	0.71	0.43	0.54	0.68	0.41	0.51
Mistral	0.80	0.45	0.58	**0.72**	0.42	0.53
SEA-LLM	0.82	0.48	0.60	0.71	0.42	0.53
SEA-Lion	0.81	0.53	0.64	0.67	0.42	0.52
CELESTA *(Dual-LLM Mention Expansion)*						
LLaMA-3 & SEA-LLM	0.80	0.57	0.67	0.69	0.50	0.58
LLaMA-3 & SEA-Lion	0.81	0.59	0.68	0.70	0.51	0.59
Mistral & SEA-LLM	**0.84**	**0.60**	**0.70**	0.71	**0.53**	**0.61**
Mistral & SEA-Lion	0.82	0.60	0.69	0.70	0.52	0.60

Table 7: Complete performance comparison of CELESTA and compact individual LLMs across mReFinED and mGENRE backends under Phase II (Small Multilingual configurations). Standalone expansions are limited to efficient edge-scale architectures ($\leq$4B parameters) and are cross-evaluated across targeted model pairs.

	mReFinED Backend			mGENRE Backend		
Model	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
IndGEL	*Config:* [$\tau = 0.2$ | English | $k = 5$]			*Config:* [$\tau = 0.2$ | English | $k = 5$]		
Baseline: Direct ED Model	0.77	0.57	0.66	0.75	0.72	0.73
Individual LLM Expansion						
LLaMA-3.2-3B	0.78	0.55	0.64	0.72	0.68	0.70
Qwen2.5-3B	0.74	0.53	0.62	0.70	0.67	0.69
Sailor-4B	0.74	0.52	0.61	0.69	0.67	0.68
Komodo	0.79	0.57	0.67	**0.81**	**0.78**	**0.80**
Merak	0.78	0.63	0.70	0.79	0.76	0.78
CELESTA *(Dual-LLM Mention Expansion)*						
LLaMA-3.2-3B & Komodo	0.80	0.65	0.72	0.79	0.77	0.78
LLaMA-3.2-3B & Merak	0.80	0.65	0.72	0.79	**0.78**	0.78
Qwen2.5-3B & Komodo	**0.82**	0.65	0.72	0.79	0.76	0.77
Qwen2.5-3B & Merak	0.81	**0.66**	**0.73**	0.80	**0.78**	0.79
Sailor-4B & Komodo	**0.82**	**0.66**	**0.73**	0.80	**0.78**	0.79
Sailor-4B & Merak	0.81	**0.66**	0.72	0.79	**0.78**	0.78
IndEL-Wikidata	*Config:* [$\tau = 0.5$ | English | $k = 5$]			*Config:* [$\tau = 0.5$ | English | $k = 5$]		
Baseline: Direct ED Model	0.63	0.33	0.43	0.60	0.49	0.54
Individual LLM Expansion						
LLaMA-3.2-3B	0.61	0.28	0.39	0.54	0.42	0.47
Qwen2.5-3B	0.57	0.24	0.34	0.52	0.39	0.44
Sailor-4B	0.60	0.32	0.42	0.57	0.45	0.50
Komodo	0.60	0.29	0.40	0.56	0.41	0.47
Merak	0.71	0.44	0.54	0.68	0.57	0.62
CELESTA *(Dual-LLM Mention Expansion)*						
LLaMA-3.2-3B & Komodo	0.63	0.42	0.50	0.60	0.53	0.57
LLaMA-3.2-3B & Merak	0.70	0.51	0.59	0.68	0.63	0.65
Qwen2.5-3B & Komodo	0.63	0.39	0.48	0.59	0.51	0.54
Qwen2.5-3B & Merak	0.71	0.50	0.59	0.68	0.62	0.65
Sailor-4B & Komodo	0.63	0.44	0.52	0.62	0.55	0.58
Sailor-4B & Merak	**0.71**	**0.52**	**0.60**	**0.69**	**0.64**	**0.66**
MalayEL-Average	*Config:* [$\tau = 0.5$ | English | $k = 5$]			*Config:* [$\tau = 0.5$ | English | $k = 5$]		
Baseline: Direct ED Model	0.78	0.44	0.56	0.55	0.37	0.45
Individual LLM Expansion						
LLaMA-3.2-3B	0.78	0.36	0.49	0.54	0.35	0.42
Qwen2.5-3B	0.77	0.39	0.52	0.56	0.37	0.44
Sailor-4B	0.75	0.37	0.49	0.53	0.34	0.41
Malaysian-LLaMA-3	**0.79**	0.41	0.54	**0.62**	**0.43**	**0.51**
Malaysian-Mistral	0.77	0.39	0.52	0.61	0.40	0.49
CELESTA *(Dual-LLM Mention Expansion)*						
LLaMA-3.2-3B & Malaysian-Llama-3	0.76	0.54	0.63	0.56	0.42	0.48
LLaMA-3.2-3B & Malaysian-Mistral	0.78	**0.55**	**0.64**	0.57	0.42	0.48
Qwen2.5-3B & Malaysian-Llama-3	0.75	0.54	0.63	0.56	0.42	0.48
Qwen2.5-3B & Malaysian-Mistral	0.78	0.54	**0.64**	0.58	**0.43**	0.49
Sailor-4B & Malaysian-Llama-3	0.76	**0.55**	**0.64**	0.58	**0.43**	0.49
Sailor-4B & Malaysian-Mistral	0.78	0.54	**0.64**	0.59	0.42	0.49
TagalogEL-Average	*Config:* [$\tau = 0.5$ | English | $k = 5$]			*Config:* [$\tau = 0.5$ | English | $k = 5$]		
Baseline: Direct ED Model	0.82	0.52	0.63	0.66	0.43	0.52
Individual LLM Expansion						
LLaMA-3.2-3B	0.78	0.45	0.57	0.65	0.41	0.50
Qwen2.5-3B	0.80	0.44	0.56	0.66	0.38	0.48
Sailor-4B	0.80	0.43	0.56	0.67	0.39	0.49
SeaLLM-2.5	0.81	0.53	0.64	0.67	0.42	0.52
Sea-Lion-v3	0.82	0.48	0.60	**0.71**	0.42	0.53
CELESTA *(Dual-LLM Mention Expansion)*						
LLaMA-3.2-3B & SeaLLM-2.5	0.82	**0.58**	**0.68**	0.70	**0.51**	**0.59**
LLaMA-3.2-3B & Sea-Lion-v3	0.83	0.57	**0.68**	0.69	0.49	0.58
Qwen2.5-3B & SeaLLM-2.5	0.82	**0.58**	**0.68**	0.70	0.50	0.58
Qwen2.5-3B & Sea-Lion-v3	**0.84**	**0.58**	**0.68**	0.70	0.50	0.58
Sailor-4B & SeaLLM-2.5	0.82	**0.58**	**0.68**	0.70	0.50	0.58
Sailor-4B & Sea-Lion-v3	**0.84**	**0.58**	**0.68**	0.69	0.50	0.58

Table 8: Performance comparison of CELESTA and standalone LLMs across the mReFinED and mGENRE backends on the IndGEL dataset (Phase II: Small Multilingual Configurations). Both multilingual and monolingual LLMs are prompted in English under a semantic $k = 5$ context window. We evaluate the impact of varying the fallback threshold (τ) within the similarity-based selection module.

	mReFinED Backend			**mGENRE Backend**		
Model	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Direct Zero-Shot ED Baselines						
Direct ED Model (No Expansion)	0.77	0.57	0.66	0.75	0.72	0.73
Standalone LLM Baselines						
LLAMA-3.2-3B	0.78	0.55	0.64	0.72	0.68	0.70
QWEN2.5-3B	0.74	0.53	0.62	0.70	0.67	0.69
SAILOR-4B	0.74	0.52	0.61	0.69	0.67	0.68
KOMODO	0.79	0.57	0.67	**0.81**	**0.78**	**0.80**
MERAK	0.78	0.63	0.70	0.79	0.76	0.78
CELESTA *(Dual-LLM Mention Expansion)*						
Threshold $\tau = 0.5$						
LLAMA-3.2-3B & KOMODO	0.80	**0.65**	0.72	0.80	0.77	0.78
LLAMA-3.2-3B & MERAK	0.80	**0.65**	0.72	0.79	0.77	0.78
QWEN2.5-3B & KOMODO	0.81	0.64	0.72	0.79	0.76	0.77
QWEN2.5-3B & MERAK	0.80	**0.65**	0.72	0.80	**0.78**	0.79
SAILOR-4B & KOMODO	**0.82**	**0.65**	**0.73**	0.80	**0.78**	0.79
SAILOR-4B & MERAK	0.80	**0.65**	0.72	0.79	**0.78**	0.78
Threshold $\tau = 0.4$						
LLAMA-3.2-3B & KOMODO	0.80	**0.65**	0.72	0.80	0.77	0.78
LLAMA-3.2-3B & MERAK	0.80	**0.65**	0.72	0.79	0.77	0.78
QWEN2.5-3B & KOMODO	0.81	0.64	0.72	0.79	0.76	0.77
QWEN2.5-3B & MERAK	0.80	**0.65**	0.72	0.80	**0.78**	0.79
SAILOR-4B & KOMODO	**0.82**	**0.65**	**0.73**	0.80	**0.78**	0.79
SAILOR-4B & MERAK	0.80	**0.65**	0.72	0.79	**0.78**	0.78
Threshold $\tau = 0.3$						
LLAMA-3.2-3B & KOMODO	0.80	**0.65**	0.71	0.79	0.76	0.78
LLAMA-3.2-3B & MERAK	0.80	**0.65**	0.72	0.79	0.77	0.78
QWEN2.5-3B & KOMODO	0.81	0.64	0.72	0.79	0.76	0.77
QWEN2.5-3B & MERAK	0.80	**0.65**	0.72	0.80	**0.78**	0.79
SAILOR-4B & KOMODO	**0.82**	**0.65**	**0.73**	0.80	**0.78**	0.79
SAILOR-4B & MERAK	0.80	**0.65**	0.72	0.79	**0.78**	0.78
Threshold $\tau = 0.2$						
LLAMA-3.2-3B & KOMODO	0.80	0.65	0.72	0.79	0.77	0.78
LLAMA-3.2-3B & MERAK	0.80	0.65	0.72	0.79	**0.78**	0.78
QWEN2.5-3B & KOMODO	**0.82**	0.65	0.72	0.79	0.76	0.77
QWEN2.5-3B & MERAK	0.81	**0.66**	**0.73**	0.80	**0.78**	0.79
SAILOR-4B & KOMODO	**0.82**	**0.66**	**0.73**	0.80	**0.78**	0.79
SAILOR-4B & MERAK	0.81	**0.66**	0.72	0.79	**0.78**	0.78
Threshold $\tau = 0.1$						
LLAMA-3.2-3B & KOMODO	0.80	**0.66**	0.72	0.79	0.76	0.78
LLAMA-3.2-3B & MERAK	0.80	**0.66**	0.72	0.79	0.77	0.78
QWEN2.5-3B & KOMODO	0.81	0.65	0.72	0.78	0.75	0.77
QWEN2.5-3B & MERAK	0.81	**0.66**	**0.73**	0.79	**0.78**	0.78
SAILOR-4B & KOMODO	**0.82**	**0.66**	**0.73**	0.80	0.77	0.79
SAILOR-4B & MERAK	0.80	**0.66**	0.72	0.79	**0.78**	0.78

Table 9: Performance comparison of CELESTA and standalone LLMs across the mReFinED and mGENRE backends on the IndEL-Wikidata dataset (Phase II: Small Multilingual Configurations). Both multilingual and monolingual LLMs are prompted in English under a semantic $k = 5$ context window. We evaluate the impact of varying the fallback threshold (τ) within the similarity-based selection module.

	mReFinED Backend			**mGENRE Backend**		
Model	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Direct Zero-Shot ED Baselines						
Direct ED Model (No Expansion)	0.63	0.33	0.43	0.60	0.49	0.54
Standalone LLM Baselines						
LLAMA-3.2-3B	0.61	0.28	0.39	0.54	0.42	0.47
QWEN2.5-3B	0.57	0.24	0.34	0.52	0.39	0.44
SAILOR-4B	0.60	0.32	0.42	0.57	0.45	0.50
KOMODO	0.60	0.29	0.40	0.56	0.41	0.47
MERAK	**0.71**	0.44	0.54	0.68	0.57	0.62
CELESTA *(Dual-LLM Mention Expansion)*						
Threshold $\tau = 0.5$						
LLAMA-3.2-3B & KOMODO	0.63	0.42	0.50	0.60	0.53	0.57
LLAMA-3.2-3B & MERAK	0.70	0.51	0.59	0.68	0.63	0.65
QWEN2.5-3B & KOMODO	0.63	0.39	0.48	0.59	0.51	0.54
QWEN2.5-3B & MERAK	0.71	0.50	0.59	0.68	0.62	0.65
SAILOR-4B & KOMODO	0.63	0.44	0.52	0.62	0.55	0.58
SAILOR-4B & MERAK	**0.71**	**0.52**	**0.60**	**0.69**	**0.64**	**0.66**
Threshold $\tau = 0.4$						
LLAMA-3.2-3B & KOMODO	0.62	0.41	0.50	0.60	0.53	0.56
LLAMA-3.2-3B & MERAK	0.69	0.51	0.59	0.68	**0.63**	0.65
QWEN2.5-3B & KOMODO	0.63	0.39	0.48	0.58	0.51	0.54
QWEN2.5-3B & MERAK	**0.71**	0.50	0.59	0.68	0.62	0.65
SAILOR-4B & KOMODO	0.63	0.44	0.52	0.62	0.55	0.58
SAILOR-4B & MERAK	**0.71**	**0.52**	**0.60**	**0.69**	**0.63**	**0.66**
Threshold $\tau = 0.3$						
LLAMA-3.2-3B & KOMODO	0.62	0.41	0.50	0.60	0.53	0.56
LLAMA-3.2-3B & MERAK	0.69	0.51	0.59	0.68	**0.63**	0.65
QWEN2.5-3B & KOMODO	0.63	0.39	0.48	0.58	0.51	0.54
QWEN2.5-3B & MERAK	**0.71**	0.50	0.59	0.68	0.62	0.65
SAILOR-4B & KOMODO	0.63	0.44	0.52	0.62	0.55	0.58
SAILOR-4B & MERAK	**0.71**	**0.52**	**0.60**	**0.69**	**0.63**	**0.66**
Threshold $\tau = 0.2$						
LLAMA-3.2-3B & KOMODO	0.62	0.41	0.50	0.60	0.53	0.56
LLAMA-3.2-3B & MERAK	0.69	0.51	0.59	0.68	0.62	0.65
QWEN2.5-3B & KOMODO	0.62	0.39	0.48	0.58	0.50	0.54
QWEN2.5-3B & MERAK	0.70	0.50	0.59	0.68	0.62	0.65
SAILOR-4B & KOMODO	0.63	0.44	0.52	0.61	0.55	0.58
SAILOR-4B & MERAK	**0.71**	**0.52**	**0.60**	**0.69**	**0.63**	**0.66**
Threshold $\tau = 0.1$						
LLAMA-3.2-3B & KOMODO	0.62	0.41	0.50	0.59	0.53	0.56
LLAMA-3.2-3B & MERAK	0.69	0.51	0.58	**0.68**	0.62	0.65
QWEN2.5-3B & KOMODO	0.62	0.39	0.48	0.58	0.50	0.54
QWEN2.5-3B & MERAK	0.70	0.50	0.59	**0.68**	0.62	0.65
SAILOR-4B & KOMODO	0.63	0.44	0.51	0.61	0.55	0.58
SAILOR-4B & MERAK	**0.70**	**0.52**	**0.60**	**0.68**	**0.63**	**0.66**

Table 10: Performance comparison of CELESTA and standalone LLMs across the mReFinED and mGENRE backends on the MalayEL dataset (Phase II: Small Multilingual Configurations). Both multilingual and monolingual LLMs are prompted in English under a semantic $k = 5$ context window. We evaluate the impact of varying the fallback threshold (τ) within the similarity-based selection module.

	mReFinED Backend			**mGENRE Backend**		
Model	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Direct Zero-Shot ED Baselines						
Direct ED Model (No Expansion)	0.78	0.44	0.56	0.55	0.37	0.45
Standalone LLM Baselines						
LLAMA-3.2-3B	0.78	0.36	0.49	0.54	0.35	0.42
QWEN2.5-3B	0.77	0.39	0.52	0.56	0.37	0.44
SAILOR-4B	0.75	0.37	0.49	0.53	0.34	0.41
MALAYSIAN-LLAMA-3	**0.79**	0.41	0.54	**0.62**	**0.43**	**0.51**
MALAYSIAN-MISTRAL	0.77	0.39	0.52	0.61	0.40	0.49
CELESTA *(Dual-LLM Mention Expansion)*						
Threshold $\tau = 0.5$						
LLAMA-3.2-3B & MALAYSIAN-LLAMA-3	0.76	0.54	0.63	0.56	0.42	0.48
LLAMA-3.2-3B & MALAYSIAN-MISTRAL	0.78	**0.55**	**0.64**	0.57	0.42	0.48
QWEN2.5-3B & MALAYSIAN-LLAMA-3	0.75	0.54	0.63	0.56	0.42	0.48
QWEN2.5-3B & MALAYSIAN-MISTRAL	0.78	0.54	**0.64**	0.58	**0.43**	0.49
SAILOR-4B & MALAYSIAN-LLAMA-3	0.76	**0.55**	**0.64**	0.58	**0.43**	0.49
SAILOR-4B & MALAYSIAN-MISTRAL	0.78	0.54	**0.64**	0.59	0.42	0.49
Threshold $\tau = 0.4$						
LLAMA-3.2-3B & MALAYSIAN-LLAMA-3	0.76	0.54	0.63	0.56	0.42	0.48
LLAMA-3.2-3B & MALAYSIAN-MISTRAL	0.78	**0.55**	**0.64**	0.57	0.42	0.48
QWEN2.5-3B & MALAYSIAN-LLAMA-3	0.75	0.54	0.63	0.56	0.42	0.48
QWEN2.5-3B & MALAYSIAN-MISTRAL	0.78	0.54	**0.64**	0.58	**0.43**	0.49
SAILOR-4B & MALAYSIAN-LLAMA-3	0.76	**0.55**	0.63	0.58	**0.43**	0.49
SAILOR-4B & MALAYSIAN-MISTRAL	0.78	0.54	**0.64**	0.58	0.42	0.49
Threshold $\tau = 0.3$						
LLAMA-3.2-3B & MALAYSIAN-LLAMA-3	0.75	0.54	0.63	0.56	0.42	0.48
LLAMA-3.2-3B & MALAYSIAN-MISTRAL	0.78	**0.55**	**0.64**	0.57	0.42	0.48
QWEN2.5-3B & MALAYSIAN-LLAMA-3	0.74	0.54	0.63	0.56	0.42	0.48
QWEN2.5-3B & MALAYSIAN-MISTRAL	0.78	**0.55**	**0.64**	0.58	**0.43**	0.49
SAILOR-4B & MALAYSIAN-LLAMA-3	0.76	**0.55**	**0.64**	0.58	**0.43**	0.49
SAILOR-4B & MALAYSIAN-MISTRAL	0.78	0.54	**0.64**	0.59	0.42	0.49
Threshold $\tau = 0.2$						
LLAMA-3.2-3B & MALAYSIAN-LLAMA-3	0.75	0.54	0.63	0.56	0.42	0.48
LLAMA-3.2-3B & MALAYSIAN-MISTRAL	0.78	**0.55**	**0.64**	0.57	0.42	0.48
QWEN2.5-3B & MALAYSIAN-LLAMA-3	0.74	0.54	0.63	0.56	0.42	0.48
QWEN2.5-3B & MALAYSIAN-MISTRAL	0.78	0.54	**0.64**	0.58	**0.43**	0.49
SAILOR-4B & MALAYSIAN-LLAMA-3	0.75	**0.55**	0.63	0.58	0.42	0.49
SAILOR-4B & MALAYSIAN-MISTRAL	0.77	0.54	**0.64**	0.58	0.42	0.49
Threshold $\tau = 0.1$						
LLAMA-3.2-3B & MALAYSIAN-LLAMA-3	0.75	0.54	0.63	0.56	0.42	0.48
LLAMA-3.2-3B & MALAYSIAN-MISTRAL	0.77	**0.55**	**0.64**	0.57	0.42	0.48
QWEN2.5-3B & MALAYSIAN-LLAMA-3	0.74	0.54	0.63	0.56	0.42	0.48
QWEN2.5-3B & MALAYSIAN-MISTRAL	0.78	0.54	**0.64**	0.58	0.42	0.49
SAILOR-4B & MALAYSIAN-LLAMA-3	0.75	**0.55**	0.63	0.58	0.42	0.49
SAILOR-4B & MALAYSIAN-MISTRAL	0.77	0.54	**0.64**	0.58	0.42	0.48

Table 11: Performance comparison of CELESTA and standalone LLMs across the mReFinED and mGENRE backends on the TagalogEL dataset (Phase II: Small Multilingual Configurations). Both multilingual and monolingual LLMs are prompted in English under a semantic $k = 5$ context window. We evaluate the impact of varying the fallback threshold (τ) within the similarity-based selection module.

	mReFinED Backend			**mGENRE Backend**		
Model	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Direct Zero-Shot ED Baselines						
Direct ED Model (No Expansion)	0.82	0.52	0.63	0.66	0.43	0.52
Standalone LLM Baselines						
LLAMA-3.2-3B	0.78	0.45	0.57	0.65	0.41	0.50
QWEN2.5-3B	0.80	0.44	0.56	0.66	0.38	0.48
SAILOR-4B	0.80	0.43	0.56	0.67	0.39	0.49
SEALLM-2.5	0.81	0.53	0.64	0.67	0.42	0.52
SEA-LION-V3	0.82	0.48	0.60	**0.71**	0.42	0.53
CELESTA *(Dual-LLM Mention Expansion)*						
Threshold $\tau = 0.5$						
LLAMA-3.2-3B & SEALLM-2.5	0.82	**0.58**	**0.68**	0.70	**0.51**	**0.59**
LLAMA-3.2-3B & SEA-LION-V3	0.83	0.57	**0.68**	0.69	0.49	0.58
QWEN2.5-3B & SEALLM-2.5	0.82	**0.58**	**0.68**	0.70	0.50	0.58
QWEN2.5-3B & SEA-LION-V3	**0.84**	**0.58**	**0.68**	0.70	0.50	0.58
SAILOR-4B & SEALLM-2.5	0.82	**0.58**	**0.68**	0.70	0.50	0.58
SAILOR-4B & SEA-LION-V3	**0.84**	**0.58**	**0.68**	0.69	0.50	0.58
Threshold $\tau = 0.4$						
LLAMA-3.2-3B & SEALLM-2.5	**0.83**	0.57	**0.68**	0.70	**0.51**	**0.59**
LLAMA-3.2-3B & SEA-LION-V3	0.82	**0.58**	**0.68**	0.69	0.50	0.58
QWEN2.5-3B & SEALLM-2.5	**0.83**	**0.58**	**0.68**	0.70	0.50	0.58
QWEN2.5-3B & SEA-LION-V3	0.82	**0.58**	**0.68**	0.69	0.50	0.58
SAILOR-4B & SEALLM-2.5	**0.83**	0.57	**0.68**	0.70	0.50	0.58
SAILOR-4B & SEA-LION-V3	0.82	**0.58**	**0.68**	0.69	0.50	0.58
Threshold $\tau = 0.3$						
LLAMA-3.2-3B & SEALLM-2.5	0.82	0.57	0.67	0.70	**0.51**	**0.59**
LLAMA-3.2-3B & SEA-LION-V3	0.82	**0.58**	**0.68**	0.69	0.50	0.58
QWEN2.5-3B & SEALLM-2.5	**0.83**	**0.58**	**0.68**	0.70	0.50	0.58
QWEN2.5-3B & SEA-LION-V3	0.82	**0.58**	**0.68**	0.69	0.50	0.58
SAILOR-4B & SEALLM-2.5	**0.83**	0.57	**0.68**	0.69	0.50	0.58
SAILOR-4B & SEA-LION-V3	0.82	**0.58**	**0.68**	0.69	0.50	0.58
Threshold $\tau = 0.2$						
LLAMA-3.2-3B & SEALLM-2.5	0.82	0.57	0.67	0.70	**0.51**	**0.59**
LLAMA-3.2-3B & SEA-LION-V3	0.82	**0.58**	**0.68**	0.69	0.50	0.58
QWEN2.5-3B & SEALLM-2.5	**0.83**	0.57	**0.68**	0.70	0.50	0.58
QWEN2.5-3B & SEA-LION-V3	0.82	0.57	0.67	0.69	0.50	0.58
SAILOR-4B & SEALLM-2.5	**0.83**	0.57	0.67	0.69	0.50	0.58
SAILOR-4B & SEA-LION-V3	0.82	0.57	0.67	0.69	0.50	0.58
Threshold $\tau = 0.1$						
LLAMA-3.2-3B & SEALLM-2.5	0.82	0.56	**0.67**	0.69	**0.51**	0.58
LLAMA-3.2-3B & SEA-LION-V3	0.81	**0.57**	**0.67**	0.69	0.50	0.58
QWEN2.5-3B & SEALLM-2.5	0.82	**0.57**	**0.67**	0.70	**0.51**	**0.59**
QWEN2.5-3B & SEA-LION-V3	0.81	**0.57**	**0.67**	0.69	0.50	0.58
SAILOR-4B & SEALLM-2.5	**0.83**	0.56	**0.67**	0.69	**0.51**	0.58
SAILOR-4B & SEA-LION-V3	0.81	**0.57**	**0.67**	0.69	0.50	0.58

Table 12: Performance comparison of CELESTA and individual LLMs across mReFinED and mGENRE backends on IndGEL (`semantic_k = 5`). We evaluate: (1) language-aligned prompting (English for multilingual and Indonesian for monolingual LLMs) against an all-English baseline, and (2) the impact of varying the fallback threshold (τ) within the language-aligned configuration.

	mReFinED Backend			**mGENRE Backend**		
Model	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Configuration 1: All LLMs with English Prompt (Threshold $\tau = 0.5$)						
Baseline: Direct ED Model	0.77	0.57	0.66	0.75	0.72	0.73
Individual LLM Expansion						
LLaMA-3	**0.87**	0.70	**0.78**	0.82	0.79	0.80
Mistral	0.73	0.56	0.64	0.76	0.73	0.75
Komodo	0.79	0.57	0.67	0.74	0.71	0.72
Merak	0.78	0.63	0.70	0.78	0.75	0.77
CELESTA *(Dual-LLM Mention Expansion)*						
LLaMA-3 & Komodo	0.85	**0.71**	0.77	**0.84**	**0.81**	**0.83**
LLaMA-3 & Merak	0.84	0.69	0.76	0.83	0.81	0.82
Mistral & Komodo	0.83	0.67	0.74	0.81	0.78	0.80
Mistral & Merak	0.82	0.66	0.73	0.80	0.78	0.79
Configuration 2: Multi LLMs (English Prompt), Mono LLMs (Indonesian Prompt)						
Individual LLM Expansion						
LLaMA-3	**0.87**	0.70	**0.78**	0.82	0.79	0.80
Mistral	0.73	0.56	0.64	0.76	0.73	0.75
Komodo	0.79	0.57	0.67	0.81	0.78	0.80
Merak	0.78	0.63	0.70	0.79	0.76	0.78
CELESTA *(Dual-LLM Mention Expansion)*						
Threshold $\tau = 0.5$						
LLaMA-3 & Komodo	0.86	**0.71**	**0.78**	**0.85**	**0.82**	**0.83**
LLaMA-3 & Merak	0.83	0.68	0.75	0.81	0.79	0.80
Mistral & Komodo	0.83	0.67	0.74	0.82	0.80	0.81
Mistral & Merak	0.82	0.66	0.73	0.80	0.79	0.79
Threshold $\tau = 0.4$						
LLaMA-3 & Komodo	0.86	**0.71**	**0.78**	**0.85**	**0.83**	**0.84**
LLaMA-3 & Merak	0.83	0.68	0.75	0.81	0.79	0.80
Mistral & Komodo	0.83	0.67	0.74	0.82	0.80	0.81
Mistral & Merak	0.82	0.66	0.73	0.80	0.79	0.79
Threshold $\tau = 0.3$						
LLaMA-3 & Komodo	0.86	**0.72**	**0.78**	**0.85**	**0.83**	**0.84**
LLaMA-3 & Merak	0.83	0.68	0.75	0.81	0.79	0.80
Mistral & Komodo	0.83	0.68	0.74	0.82	0.80	0.81
Mistral & Merak	0.82	0.67	0.73	0.80	0.79	0.79
Threshold $\tau = 0.2$						
LLaMA-3 & Komodo	0.87	**0.72**	**0.79**	**0.85**	**0.83**	**0.84**
LLaMA-3 & Merak	0.83	0.69	0.75	0.81	0.80	0.80
Mistral & Komodo	0.83	0.68	0.75	0.82	0.80	0.81
Mistral & Merak	0.82	0.67	0.74	0.81	0.79	0.80
Threshold $\tau = 0.1$						
LLaMA-3 & Komodo	0.87	**0.73**	**0.79**	**0.85**	**0.83**	**0.84**
LLaMA-3 & Merak	0.83	0.69	0.76	0.81	0.80	0.80
Mistral & Komodo	0.82	0.68	0.75	0.82	0.80	0.81
Mistral & Merak	0.81	0.67	0.73	0.80	0.79	0.79

Table 13: Performance comparison of CELESTA and individual LLMs across mReFinED and mGENRE backends on IndEL_Wikidata (`semantic_k` = 5). We evaluate: (1) language-aligned prompting (English for multilingual and Indonesian for monolingual LLMs) against an all-English baseline, and (2) the impact of varying the fallback threshold (τ) within the language-aligned configuration.

	mReFinED Backend			**mGENRE Backend**		
Model	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Configuration 1: All LLMs with English Prompt (Threshold $\tau = 0.5$)						
Baseline: Direct ED Model	0.63	0.33	0.43	0.60	0.49	0.54
Individual LLM Expansion						
LLaMA-3	0.73	0.45	0.56	0.71	0.58	0.64
Mistral	0.58	0.28	0.37	0.53	0.42	0.47
Komodo	0.60	0.29	0.40	0.56	0.41	0.47
Merak	0.71	0.44	0.54	0.68	0.57	0.62
CELESTA *(Dual-LLM Mention Expansion)*						
LLaMA-3 & Komodo	0.72	0.52	0.60	0.69	0.62	0.65
LLaMA-3 & Merak	**0.74**	**0.56**	**0.64**	**0.73**	**0.68**	**0.70**
Mistral & Komodo	0.62	0.42	0.50	0.60	0.54	0.57
Mistral & Merak	0.70	0.51	0.59	0.68	0.63	0.65
Configuration 2: Multi LLMs (English Prompt), Mono LLMs (Indonesian Prompt)						
Individual LLM Expansion						
LLaMA-3	0.73	0.45	0.56	0.71	0.58	0.64
Mistral	0.58	0.28	0.37	0.53	0.42	0.47
Komodo	0.64	0.39	0.48	0.61	0.50	0.55
Merak	0.70	0.44	0.54	0.69	0.57	0.62
CELESTA *(Dual-LLM Mention Expansion)*						
Threshold $\tau = 0.5$						
LLaMA-3 & Komodo	0.72	0.53	0.61	0.70	0.64	0.67
LLaMA-3 & Merak	**0.74**	**0.56**	**0.64**	**0.73**	**0.68**	**0.70**
Mistral & Komodo	0.64	0.45	0.53	0.62	0.56	0.59
Mistral & Merak	0.69	0.51	0.59	0.67	0.62	0.65
Threshold $\tau = 0.4$						
LLaMA-3 & Komodo	0.72	0.53	0.61	0.70	0.64	0.67
LLaMA-3 & Merak	**0.74**	**0.56**	**0.64**	**0.73**	**0.67**	**0.70**
Mistral & Komodo	0.64	0.45	0.53	0.62	0.56	0.59
Mistral & Merak	0.69	0.51	0.59	0.67	0.62	0.65
Threshold $\tau = 0.3$						
LLaMA-3 & Komodo	0.72	0.53	0.61	0.70	0.64	0.66
LLaMA-3 & Merak	**0.74**	**0.56**	**0.64**	**0.73**	**0.67**	**0.70**
Mistral & Komodo	0.64	0.45	0.53	0.62	0.56	0.59
Mistral & Merak	0.69	0.51	0.59	0.67	0.62	0.65
Threshold $\tau = 0.2$						
LLaMA-3 & Komodo	0.72	0.53	0.61	0.69	0.63	0.66
LLaMA-3 & Merak	**0.74**	**0.56**	**0.64**	**0.73**	**0.67**	**0.70**
Mistral & Komodo	0.64	0.45	0.53	0.61	0.56	0.58
Mistral & Merak	0.69	0.51	0.59	0.67	0.62	0.64
Threshold $\tau = 0.1$						
LLaMA-3 & Komodo	0.72	0.53	0.61	0.69	0.63	0.66
LLaMA-3 & Merak	**0.74**	**0.56**	**0.64**	**0.73**	**0.67**	**0.70**
Mistral & Komodo	0.63	0.45	0.52	0.61	0.55	0.58
Mistral & Merak	0.69	0.51	0.59	0.67	0.62	0.64

Table 14: Performance comparison of CELESTA and individual LLMs across mReFinED and mGENRE backends on MalayEL (`semantic_k` = 5). We evaluate: (1) language-aligned prompting (English for multilingual and Malay for monolingual LLMs) against an all-English baseline, and (2) the impact of varying the fallback threshold (τ) within the language-aligned configuration.

	mReFinED Backend			**mGENRE Backend**		
Model	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Configuration 1: All LLMs with English Prompt (Threshold $\tau = 0.5$)						
Baseline: Direct ED Model	0.78	0.44	0.56	0.55	0.37	0.45
Individual LLM Expansion						
LLaMA-3	0.78	0.37	0.50	0.61	0.38	0.47
Mistral	0.76	0.40	0.52	0.61	0.40	0.49
Malaysian-LLaMA-3	**0.79**	0.41	0.54	**0.62**	**0.43**	**0.51**
Malaysian-Mistral	0.77	0.39	0.52	0.61	0.40	0.49
CELESTA *(Dual-LLM Mention Expansion)*						
LLaMA-3 & Malaysian-LLaMA-3	0.76	0.55	0.64	0.57	0.43	0.49
LLaMA-3 & Malaysian-Mistral	0.79	0.55	**0.65**	0.58	0.43	0.50
Mistral & Malaysian-LLaMA-3	0.75	0.55	0.63	0.56	0.42	0.48
Mistral & Malaysian-Mistral	0.77	**0.56**	0.65	0.58	0.43	0.49
Configuration 2: Multi LLMs (English Prompt), Mono LLMs (Malay Prompt)						
Individual LLM Expansion						
LLaMA-3	0.78	0.37	0.50	0.61	0.38	0.47
Mistral	0.76	0.40	0.52	0.61	0.40	0.49
Malaysian-LLaMA-3	**0.79**	0.41	0.54	**0.62**	**0.43**	**0.51**
Malaysian-Mistral	0.77	0.39	0.52	0.61	0.40	0.49
CELESTA *(Dual-LLM Mention Expansion)*						
Threshold $\tau = 0.5$						
LLaMA-3 & Malaysian-LLaMA-3	0.01	0.01	0.01	0.01	0.01	0.01
LLaMA-3 & Malaysian-Mistral	0.01	0.01	0.01	0.01	0.01	0.01
Mistral & Malaysian-LLaMA-3	0.01	0.01	0.01	0.01	0.01	0.01
Mistral & Malaysian-Mistral	0.01	0.01	0.01	0.01	0.01	0.01
Threshold $\tau = 0.4$						
LLaMA-3 & Malaysian-LLaMA-3	0.01	0.01	0.01	0.01	0.01	0.01
LLaMA-3 & Malaysian-Mistral	0.01	0.01	0.01	0.01	0.01	0.01
Mistral & Malaysian-LLaMA-3	0.01	0.01	0.01	0.01	0.01	0.01
Mistral & Malaysian-Mistral	0.01	0.01	0.01	0.01	0.01	0.01
Threshold $\tau = 0.3$						
LLaMA-3 & Malaysian-LLaMA-3	0.01	0.01	0.01	0.01	0.01	0.01
LLaMA-3 & Malaysian-Mistral	0.01	0.01	0.01	0.01	0.01	0.01
Mistral & Malaysian-LLaMA-3	0.01	0.01	0.01	0.01	0.01	0.01
Mistral & Malaysian-Mistral	0.01	0.01	0.01	0.01	0.01	0.01
Threshold $\tau = 0.2$						
LLaMA-3 & Malaysian-LLaMA-3	0.01	0.01	0.01	0.01	0.01	0.01
LLaMA-3 & Malaysian-Mistral	0.01	0.01	0.01	0.01	0.01	0.01
Mistral & Malaysian-LLaMA-3	0.01	0.01	0.01	0.01	0.01	0.01
Mistral & Malaysian-Mistral	0.01	0.01	0.01	0.01	0.01	0.01
Threshold $\tau = 0.1$						
LLaMA-3 & Malaysian-LLaMA-3	0.01	0.01	0.01	0.01	0.01	0.01
LLaMA-3 & Malaysian-Mistral	0.01	0.01	0.01	0.01	0.01	0.01
Mistral & Malaysian-LLaMA-3	0.01	0.01	0.01	0.01	0.01	0.01
Mistral & Malaysian-Mistral	0.01	0.01	0.01	0.01	0.01	0.01

Table 15: Performance comparison of CELESTA and individual LLMs across mReFinED and mGENRE backends on TagalogEL (`semantic_k` = 5). We evaluate: (1) language-aligned prompting (English for multilingual and Tagalog for monolingual LLMs) against an all-English baseline, and (2) the impact of varying the fallback threshold (τ) within the language-aligned configuration.

	mReFinED Backend			**mGENRE Backend**		
Model	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Configuration 1: All LLMs with English Prompt (Threshold $\tau = 0.5$)						
Baseline: Direct ED Model	0.82	0.52	0.63	0.66	0.43	0.52
Individual LLM Expansion						
LLaMA-3	0.71	0.43	0.54	0.68	0.41	0.51
Mistral	0.80	0.45	0.58	**0.72**	0.42	0.53
Sea-Lion-v3	0.82	0.48	0.60	0.71	0.42	0.53
SeaLLM-2.5	0.81	0.53	0.64	0.67	0.42	0.52
CELESTA *(Dual-LLM Mention Expansion)*						
LLaMA-3 & Sea-Lion-v3	0.80	0.57	0.67	0.69	0.50	0.58
LLaMA-3 & SeaLLM-2.5	0.81	0.58	0.68	0.70	0.51	0.59
Mistral & Sea-Lion-v3	**0.84**	**0.60**	**0.70**	0.71	**0.53**	**0.61**
Mistral & SeaLLM-2.5	0.82	0.60	0.69	0.70	0.52	0.60
Configuration 2: Multi LLMs (English Prompt), Mono LLMs (Tagalog Prompt)						
Individual LLM Expansion						
LLaMA-3	0.71	0.43	0.54	0.68	0.41	0.51
Mistral	0.80	0.45	0.58	**0.72**	0.42	**0.53**
Sea-Lion-v3	0.82	0.48	0.60	0.71	0.42	0.53
SeaLLM-2.5	0.81	**0.53**	**0.64**	0.67	0.42	0.52
CELESTA *(Dual-LLM Mention Expansion)*						
Threshold $\tau = 0.5$						
LLaMA-3 & Sea-Lion-v3	0.01	0.00	0.01	0.01	0.01	0.01
LLaMA-3 & SeaLLM-2.5	0.01	0.00	0.01	0.01	0.01	0.01
Mistral & Sea-Lion-v3	0.01	0.00	0.01	0.01	0.01	0.01
Mistral & SeaLLM-2.5	0.01	0.00	0.01	0.01	0.01	0.01
Threshold $\tau = 0.4$						
LLaMA-3 & Sea-Lion-v3	0.01	0.00	0.01	0.01	0.01	0.01
LLaMA-3 & SeaLLM-2.5	0.01	0.00	0.01	0.01	0.01	0.01
Mistral & Sea-Lion-v3	0.01	0.00	0.01	0.01	0.01	0.01
Mistral & SeaLLM-2.5	0.01	0.00	0.01	0.01	0.01	0.01
Threshold $\tau = 0.3$						
LLaMA-3 & Sea-Lion-v3	0.01	0.00	0.01	0.01	0.01	0.01
LLaMA-3 & SeaLLM-2.5	0.01	0.00	0.01	0.01	0.01	0.01
Mistral & Sea-Lion-v3	0.01	0.00	0.01	0.01	0.01	0.01
Mistral & SeaLLM-2.5	0.01	0.00	0.01	0.01	0.01	0.01
Threshold $\tau = 0.2$						
LLaMA-3 & Sea-Lion-v3	0.01	0.00	0.01	0.01	0.01	0.01
LLaMA-3 & SeaLLM-2.5	0.01	0.00	0.01	0.01	0.01	0.01
Mistral & Sea-Lion-v3	0.01	0.00	0.01	0.01	0.01	0.01
Mistral & SeaLLM-2.5	0.01	0.00	0.01	0.01	0.01	0.01
Threshold $\tau = 0.1$						
LLaMA-3 & Sea-Lion-v3	0.01	0.00	0.01	0.01	0.01	0.01
LLaMA-3 & SeaLLM-2.5	0.01	0.00	0.01	0.01	0.01	0.01
Mistral & Sea-Lion-v3	0.01	0.00	0.01	0.01	0.01	0.01
Mistral & SeaLLM-2.5	0.01	0.00	0.01	0.01	0.01	0.01

Table 16: Performance comparison of CELESTA and individual LLMs across mReFinED and mGENRE backends on IndGEL using an English prompt for both multilingual and monolingual LLMs under a fixed semantic context window (`semantic_k` = 5). For the CELESTA pipeline, we evaluate the impact of varying the hallucination fallback threshold (τ) within the similarity-based selection module.

	mReFinED Backend			**mGENRE Backend**		
Model	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Baseline: Direct ED Model	0.77	0.57	0.66	0.75	0.72	0.73
Individual LLM Expansion						
LLAMA-3	**0.87**	0.70	**0.78**	0.82	0.79	0.80
MISTRAL	0.73	0.56	0.64	0.76	0.73	0.75
KOMODO	0.79	0.57	0.67	0.74	0.71	0.72
MERAK	0.78	0.63	0.70	0.78	0.75	0.77
CELESTA *(Dual-LLM Mention Expansion)*						
Threshold $\tau = 0.5$						
LLAMA-3 & KOMODO	0.85	**0.71**	0.77	**0.84**	**0.81**	**0.83**
LLAMA-3 & MERAK	0.84	0.69	0.76	0.83	0.81	0.82
MISTRAL & KOMODO	0.83	0.67	0.74	0.81	0.78	0.80
MISTRAL & MERAK	0.82	0.66	0.73	0.80	0.78	0.79
Threshold $\tau = 0.4$						
LLAMA-3 & KOMODO	0.85	0.71	0.77	**0.84**	**0.82**	**0.83**
LLAMA-3 & MERAK	0.84	0.69	0.76	0.83	0.81	0.82
MISTRAL & KOMODO	0.83	0.67	0.74	0.81	0.79	0.80
MISTRAL & MERAK	0.82	0.66	0.73	0.80	0.78	0.79
Threshold $\tau = 0.3$						
LLAMA-3 & KOMODO	0.85	**0.71**	**0.78**	**0.84**	**0.81**	**0.83**
LLAMA-3 & MERAK	0.84	0.70	0.76	0.83	0.81	0.82
MISTRAL & KOMODO	0.82	0.67	0.74	0.81	0.78	0.80
MISTRAL & MERAK	0.82	0.66	0.73	0.80	0.78	0.79
Threshold $\tau = 0.2$						
LLAMA-3 & KOMODO	0.85	**0.72**	**0.78**	**0.84**	**0.82**	**0.83**
LLAMA-3 & MERAK	0.84	0.70	0.76	0.83	0.81	0.82
MISTRAL & KOMODO	0.83	0.67	0.74	0.82	0.79	0.80
MISTRAL & MERAK	0.82	0.67	0.74	0.80	0.78	0.79
Threshold $\tau = 0.1$						
LLAMA-3 & KOMODO	0.86	**0.72**	**0.78**	**0.84**	**0.82**	**0.83**
LLAMA-3 & MERAK	0.84	0.71	0.77	0.83	0.81	0.82
MISTRAL & KOMODO	0.82	0.67	0.74	0.81	0.78	0.79
MISTRAL & MERAK	0.82	0.66	0.73	0.80	0.78	0.79

Table 17: Performance comparison of CELESTA and individual LLMs across mReFinED and mGENRE backends on IndEL-Wikidata using an English prompt for both multilingual and monolingual LLMs under a fixed semantic context window (`semantic_k` = 5). For the CELESTA pipeline, we evaluate the impact of varying the hallucination fallback threshold (τ) within the similarity-based selection module.

	mReFinED Backend			**mGENRE Backend**		
Model	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Baseline: Direct ED Model	0.63	0.33	0.43	0.60	0.49	0.54
Individual LLM Expansion						
LLaMA-3	0.73	0.45	0.56	0.71	0.58	0.64
Mistral	0.58	0.28	0.37	0.53	0.42	0.47
Komodo	0.60	0.29	0.40	0.56	0.41	0.47
Merak	0.71	0.44	0.54	0.68	0.57	0.62
CELESTA *(Dual-LLM Mention Expansion)*						
Threshold $\tau = 0.5$						
LLaMA-3 & Komodo	0.72	0.52	0.60	0.69	0.62	0.65
LLaMA-3 & Merak	**0.74**	**0.56**	**0.64**	**0.73**	**0.68**	**0.70**
Mistral & Komodo	0.62	0.42	0.50	0.60	0.54	0.57
Mistral & Merak	0.70	0.51	0.59	0.68	0.63	0.65
Threshold $\tau = 0.4$						
LLaMA-3 & Komodo	0.72	0.52	0.60	0.69	0.62	0.65
LLaMA-3 & Merak	**0.74**	**0.56**	**0.64**	**0.73**	**0.68**	**0.70**
Mistral & Komodo	0.62	0.42	0.50	0.59	0.53	0.56
Mistral & Merak	0.70	0.51	0.59	0.68	0.63	0.65
Threshold $\tau = 0.3$						
LLaMA-3 & Komodo	0.72	0.51	0.60	0.69	0.62	0.65
LLaMA-3 & Merak	**0.74**	**0.56**	**0.64**	**0.73**	**0.68**	**0.70**
Mistral & Komodo	0.62	0.42	0.50	0.59	0.53	0.56
Mistral & Merak	0.70	0.51	0.59	0.68	0.62	0.65
Threshold $\tau = 0.2$						
LLaMA-3 & Komodo	0.72	0.52	0.60	0.69	0.62	0.65
LLaMA-3 & Merak	**0.74**	**0.56**	**0.64**	**0.73**	**0.67**	**0.70**
Mistral & Komodo	0.62	0.42	0.50	0.59	0.53	0.56
Mistral & Merak	0.70	0.51	0.59	0.67	0.62	0.65
Threshold $\tau = 0.1$						
LLaMA-3 & Komodo	0.72	0.51	0.60	0.68	0.62	0.65
LLaMA-3 & Merak	**0.74**	**0.56**	**0.64**	**0.73**	**0.67**	**0.70**
Mistral & Komodo	0.61	0.42	0.50	0.59	0.53	0.56
Mistral & Merak	0.70	0.51	0.59	0.67	0.62	0.65

Table 18: Performance comparison of CELESTA and individual LLMs across mReFinED and mGENRE backends on MalayEL using an English prompt for both multilingual and monolingual LLMs under a fixed semantic context window (`semantic_k` = 5). For the CELESTA pipeline, we evaluate the impact of varying the hallucination fallback threshold (τ) within the similarity-based selection module.

	mReFinED Backend			**mGENRE Backend**		
Model	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Baseline: Direct ED Model	0.78	0.44	0.56	0.55	0.37	0.45
Individual LLM Expansion						
LLaMA-3	0.78	0.37	0.50	0.61	0.38	0.47
Mistral	0.76	0.40	0.52	0.61	0.40	0.49
Malaysian-LLaMA-3	**0.79**	0.41	0.54	**0.62**	**0.43**	**0.51**
Malaysian-Mistral	0.77	0.39	0.52	0.61	0.40	0.49
CELESTA *(Dual-LLM Mention Expansion)*						
Threshold $\tau = 0.5$						
LLaMA-3 & Malaysian-LLaMA-3	0.76	0.55	0.64	0.57	0.43	0.49
LLaMA-3 & Malaysian-Mistral	0.79	0.55	**0.65**	0.58	0.43	0.50
Mistral & Malaysian-LLaMA-3	0.75	0.55	0.63	0.56	0.42	0.48
Mistral & Malaysian-Mistral	0.77	**0.56**	0.65	0.58	0.43	0.49
Threshold $\tau = 0.4$						
LLaMA-3 & Malaysian-LLaMA-3	0.75	0.55	0.63	0.57	0.43	0.49
LLaMA-3 & Malaysian-Mistral	0.78	0.55	**0.65**	0.58	0.43	0.50
Mistral & Malaysian-LLaMA-3	0.75	0.55	0.63	0.56	0.42	0.48
Mistral & Malaysian-Mistral	0.76	**0.56**	0.65	0.58	0.43	0.49
Threshold $\tau = 0.3$						
LLaMA-3 & Malaysian-LLaMA-3	0.75	**0.58**	0.63	0.57	0.43	0.49
LLaMA-3 & Malaysian-Mistral	0.78	0.55	0.65	0.58	0.43	0.50
Mistral & Malaysian-LLaMA-3	0.75	0.56	0.64	0.57	0.42	0.48
Mistral & Malaysian-Mistral	0.76	0.56	**0.65**	0.58	0.43	0.49
Threshold $\tau = 0.2$						
LLaMA-3 & Malaysian-LLaMA-3	0.75	0.55	0.64	0.57	0.43	0.49
LLaMA-3 & Malaysian-Mistral	0.78	0.56	**0.65**	0.59	0.43	0.50
Mistral & Malaysian-LLaMA-3	0.74	0.56	0.64	0.57	0.42	0.48
Mistral & Malaysian-Mistral	0.76	**0.56**	0.65	0.58	0.43	0.49
Threshold $\tau = 0.1$						
LLaMA-3 & Malaysian-LLaMA-3	0.75	0.55	0.63	0.57	0.43	0.49
LLaMA-3 & Malaysian-Mistral	0.77	0.55	**0.65**	0.58	0.43	0.49
Mistral & Malaysian-LLaMA-3	0.74	0.55	0.63	0.57	0.42	0.48
Mistral & Malaysian-Mistral	0.76	**0.56**	0.64	0.58	0.43	0.49

Table 19: Performance comparison of CELESTA and individual LLMs across mReFinED and mGENRE backends on TagalogEL using an English prompt for both multilingual and monolingual LLMs under a fixed semantic context window (`semantic_k` = 5). For the CELESTA pipeline, we evaluate the impact of varying the hallucination fallback threshold (τ) within the similarity-based selection module.

	mReFinED Backend			**mGENRE Backend**		
Model	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Baseline: Direct ED Model	0.82	0.52	0.63	0.66	0.43	0.52
Individual LLM Expansion						
LLAMA-3	0.71	0.43	0.54	0.68	0.41	0.51
MISTRAL	0.80	0.45	0.58	**0.72**	0.42	0.53
SEA-LION-V3	0.82	0.48	0.60	0.71	0.42	0.53
SEALLM-2.5	0.81	0.53	0.64	0.67	0.42	0.52
CELESTA *(Dual-LLM Mention Expansion)*						
Threshold $\tau = 0.5$						
LLAMA-3 & SEA-LION-V3	0.80	0.57	0.67	0.69	0.50	0.58
LLAMA-3 & SEALLM-2.5	0.81	0.58	0.68	0.70	0.51	0.59
MISTRAL & SEA-LION-V3	**0.84**	**0.60**	**0.70**	0.71	**0.53**	**0.61**
MISTRAL & SEALLM-2.5	0.82	0.60	0.69	0.70	0.52	0.60
Threshold $\tau = 0.4$						
LLAMA-3 & SEA-LION-V3	0.80	0.57	0.67	0.69	0.50	0.58
LLAMA-3 & SEALLM-2.5	0.81	0.58	0.68	0.70	0.51	0.59
MISTRAL & SEA-LION-V3	**0.83**	**0.60**	**0.70**	0.71	**0.53**	**0.61**
MISTRAL & SEALLM-2.5	0.82	0.60	0.69	0.70	0.52	0.60
Threshold $\tau = 0.3$						
LLAMA-3 & SEA-LION-V3	0.80	0.57	0.67	0.69	0.50	0.58
LLAMA-3 & SEALLM-2.5	0.81	0.58	0.68	0.70	0.51	0.59
MISTRAL & SEA-LION-V3	**0.83**	**0.60**	**0.70**	0.71	**0.53**	**0.61**
MISTRAL & SEALLM-2.5	0.82	0.60	0.69	0.70	0.52	0.60
Threshold $\tau = 0.2$						
LLAMA-3 & SEA-LION-V3	0.79	0.57	0.66	0.68	0.50	0.58
LLAMA-3 & SEALLM-2.5	0.80	0.58	0.67	0.69	0.51	0.59
MISTRAL & SEA-LION-V3	**0.83**	**0.59**	**0.69**	0.71	**0.53**	**0.61**
MISTRAL & SEALLM-2.5	0.82	0.59	0.69	0.70	0.52	0.60
Threshold $\tau = 0.1$						
LLAMA-3 & SEA-LION-V3	0.79	0.56	0.66	0.68	0.50	0.58
LLAMA-3 & SEALLM-2.5	0.79	0.57	0.66	0.69	0.51	0.59
MISTRAL & SEA-LION-V3	**0.83**	**0.59**	**0.69**	0.71	**0.54**	**0.61**
MISTRAL & SEALLM-2.5	0.81	0.59	0.68	0.70	0.52	0.60

Table 20: Performance comparison of CELESTA and standalone LLMs across the mReFinED and mGENRE backends on the IndGEL dataset. In this configuration, both multilingual and monolingual LLMs are prompted in English under a full sentence context window. For the dual-LLM CELESTA pipeline, we evaluate the impact of varying the fallback threshold (τ) within the similarity-based selection module.

	mReFinED Backend			**mGENRE Backend**		
Model	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Direct Zero-Shot ED Baselines						
Direct ED Model (No Expansion)	0.77	0.57	0.66	0.75	0.72	0.73
Standalone LLM Baselines						
LLAMA-3	**0.87**	0.70	**0.78**	0.82	0.79	0.80
MISTRAL	0.73	0.56	0.64	0.76	0.73	0.75
KOMODO	0.79	0.57	0.67	0.74	0.71	0.72
MERAK	0.78	0.63	0.70	0.78	0.75	0.77
CELESTA *(Dual-LLM Mention Expansion)*						
Threshold $\tau = 0.5$						
LLAMA-3 & KOMODO	0.85	**0.71**	0.78	**0.84**	**0.82**	**0.83**
LLAMA-3 & MERAK	0.84	0.70	0.76	0.83	0.81	0.82
MISTRAL & KOMODO	0.83	0.67	0.74	0.81	0.79	0.80
MISTRAL & MERAK	0.82	0.66	0.73	0.80	0.78	0.79
Threshold $\tau = 0.4$						
LLAMA-3 & KOMODO	0.85	**0.71**	0.78	**0.84**	**0.82**	**0.83**
LLAMA-3 & MERAK	0.84	0.70	0.76	0.83	0.81	0.82
MISTRAL & KOMODO	0.83	0.67	0.74	0.82	0.79	0.80
MISTRAL & MERAK	0.82	0.66	0.73	0.80	0.78	0.79
Threshold $\tau = 0.3$						
LLAMA-3 & KOMODO	0.85	**0.71**	0.78	**0.84**	**0.82**	**0.83**
LLAMA-3 & MERAK	0.84	0.70	0.76	0.83	0.81	0.82
MISTRAL & KOMODO	0.82	0.67	0.74	0.81	0.79	0.80
MISTRAL & MERAK	0.82	0.66	0.73	0.80	0.78	0.79
Threshold $\tau = 0.2$						
LLAMA-3 & KOMODO	0.85	**0.72**	**0.78**	**0.84**	**0.82**	**0.83**
LLAMA-3 & MERAK	0.84	0.70	0.77	0.83	0.82	0.82
MISTRAL & KOMODO	0.83	0.67	0.74	0.82	0.79	0.80
MISTRAL & MERAK	0.82	0.67	0.74	0.80	0.78	0.79
Threshold $\tau = 0.1$						
LLAMA-3 & KOMODO	0.86	**0.72**	**0.79**	**0.84**	**0.82**	**0.83**
LLAMA-3 & MERAK	0.84	0.71	0.77	0.83	0.82	0.82
MISTRAL & KOMODO	0.82	0.67	0.74	0.81	0.78	0.79
MISTRAL & MERAK	0.82	0.67	0.73	0.80	0.78	0.79

Table 21: Performance comparison of CELESTA and standalone LLMs across the mReFinED and mGENRE backends on the IndEL-Wikidata dataset. In this configuration, both multilingual and monolingual LLMs are prompted in English under a full sentence context window. For the dual-LLM CELESTA pipeline, we evaluate the impact of varying the fallback threshold (τ) within the similarity-based selection module.

	mReFinED Backend			**mGENRE Backend**		
Model	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Direct Zero-Shot ED Baselines						
Direct ED Model (No Expansion)	0.63	0.33	0.43	0.60	0.49	0.54
Standalone LLM Baselines						
LLAMA-3	0.73	0.45	0.56	0.71	0.58	0.64
MISTRAL	0.58	0.28	0.37	0.53	0.42	0.47
KOMODO	0.60	0.29	0.40	0.56	0.41	0.47
MERAK	0.71	0.44	0.54	0.68	0.57	0.62
CELESTA *(Dual-LLM Mention Expansion)*						
Threshold $\tau = 0.5$						
LLAMA-3 & KOMODO	0.72	0.52	0.60	0.69	0.62	0.66
LLAMA-3 & MERAK	**0.74**	**0.56**	**0.64**	**0.73**	**0.67**	**0.70**
MISTRAL & KOMODO	0.62	0.42	0.50	0.60	0.54	0.57
MISTRAL & MERAK	0.70	0.51	0.59	0.68	0.62	0.65
Threshold $\tau = 0.4$						
LLAMA-3 & KOMODO	0.72	0.52	0.60	0.69	0.62	0.65
LLAMA-3 & MERAK	**0.74**	**0.56**	**0.64**	**0.73**	**0.67**	**0.70**
MISTRAL & KOMODO	0.62	0.42	0.50	0.59	0.53	0.56
MISTRAL & MERAK	0.70	0.51	0.59	0.67	0.62	0.65
Threshold $\tau = 0.3$						
LLAMA-3 & KOMODO	0.72	0.51	0.60	0.69	0.62	0.65
LLAMA-3 & MERAK	**0.74**	**0.56**	**0.64**	**0.73**	**0.67**	**0.70**
MISTRAL & KOMODO	0.62	0.42	0.50	0.59	0.53	0.56
MISTRAL & MERAK	0.70	0.51	0.59	0.67	0.62	0.65
Threshold $\tau = 0.2$						
LLAMA-3 & KOMODO	0.72	0.51	0.60	0.69	0.62	0.65
LLAMA-3 & MERAK	**0.74**	**0.56**	**0.64**	**0.73**	**0.67**	**0.70**
MISTRAL & KOMODO	0.62	0.42	0.50	0.59	0.53	0.56
MISTRAL & MERAK	0.70	0.51	0.59	0.67	0.62	0.64
Threshold $\tau = 0.1$						
LLAMA-3 & KOMODO	0.71	0.51	0.60	0.68	0.62	0.65
LLAMA-3 & MERAK	**0.74**	**0.56**	**0.64**	**0.73**	**0.67**	**0.70**
MISTRAL & KOMODO	0.61	0.42	0.50	0.59	0.53	0.56
MISTRAL & MERAK	0.70	0.51	0.59	0.67	0.62	0.64

Table 22: Performance comparison of CELESTA and individual LLMs across mReFinED and mGENRE backends on MalayEL using an English prompt for both multilingual and monolingual LLMs. For the CELESTA pipeline, we evaluate the impact of varying the hallucination fallback threshold (τ) with a full-sentence context window in the similarity-based selection module.

	mReFinED Backend			**mGENRE Backend**		
Model	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Baseline: Direct ED Model	0.78	0.44	0.56	0.55	0.37	0.45
Individual LLM Expansion						
LLaMA-3	0.78	0.37	0.50	0.61	0.38	0.47
Mistral	0.76	0.40	0.52	0.61	0.40	0.49
Malaysian-LLaMA-3	**0.79**	0.41	0.54	**0.62**	**0.43**	**0.51**
Malaysian-Mistral	0.77	0.39	0.52	0.61	0.40	0.49
CELESTA *(Dual-LLM Mention Expansion)*						
Threshold $\tau = 0.5$						
LLaMA-3 & Malaysian-LLaMA-3	0.76	0.55	0.64	0.57	0.43	0.49
LLaMA-3 & Malaysian-Mistral	0.79	0.55	**0.65**	0.58	0.43	0.50
Mistral & Malaysian-LLaMA-3	0.75	0.55	0.64	0.57	0.42	0.48
Mistral & Malaysian-Mistral	0.77	**0.56**	0.65	0.58	0.43	0.49
Threshold $\tau = 0.4$						
LLaMA-3 & Malaysian-LLaMA-3	0.75	0.55	0.63	0.57	0.43	0.49
LLaMA-3 & Malaysian-Mistral	0.78	0.55	**0.65**	0.58	0.43	0.50
Mistral & Malaysian-LLaMA-3	0.75	0.55	0.64	0.56	0.42	0.48
Mistral & Malaysian-Mistral	0.76	**0.56**	0.65	0.58	0.43	0.49
Threshold $\tau = 0.3$						
LLaMA-3 & Malaysian-LLaMA-3	0.75	**0.58**	0.63	0.57	0.43	0.49
LLaMA-3 & Malaysian-Mistral	0.78	0.55	0.65	0.58	0.43	0.50
Mistral & Malaysian-LLaMA-3	0.75	0.56	0.64	0.57	0.42	0.48
Mistral & Malaysian-Mistral	0.77	0.56	**0.65**	0.58	0.43	0.49
Threshold $\tau = 0.2$						
LLaMA-3 & Malaysian-LLaMA-3	0.75	0.55	0.64	0.57	0.43	0.49
LLaMA-3 & Malaysian-Mistral	0.78	0.56	**0.65**	0.59	0.43	0.50
Mistral & Malaysian-LLaMA-3	0.74	0.56	0.64	0.57	0.42	0.48
Mistral & Malaysian-Mistral	0.77	**0.56**	0.65	0.58	0.43	0.49
Threshold $\tau = 0.1$						
LLaMA-3 & Malaysian-LLaMA-3	0.75	0.55	0.63	0.57	0.43	0.49
LLaMA-3 & Malaysian-Mistral	0.77	0.55	**0.65**	0.58	0.43	0.49
Mistral & Malaysian-LLaMA-3	0.74	0.56	0.64	0.57	0.42	0.48
Mistral & Malaysian-Mistral	0.76	**0.56**	0.64	0.58	0.43	0.49

Table 23: Performance comparison of CELESTA and individual LLMs across mReFinED and mGENRE backends on TagalogEL using an English prompt for both multilingual and monolingual LLMs. For the CELESTA pipeline, we evaluate the impact of varying the hallucination fallback threshold (τ) with a full-sentence context window in the similarity-based selection module.

	mReFinED Backend			**mGENRE Backend**		
Model	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Baseline: Direct ED Model	0.82	0.52	0.63	0.66	0.43	0.52
Individual LLM Expansion						
LLaMA-3	0.71	0.43	0.54	0.68	0.41	0.51
Mistral	0.80	0.45	0.58	**0.72**	0.42	0.53
Sea-Lion-v3	0.82	0.48	0.60	0.71	0.42	0.53
SeaLLM-2.5	0.81	0.53	0.64	0.67	0.42	0.52
CELESTA *(Dual-LLM Mention Expansion)*						
Threshold $\tau = 0.5$						
LLaMA-3 & Sea-Lion-v3	0.80	0.57	0.67	0.69	0.50	0.58
LLaMA-3 & SeaLLM-2.5	0.81	0.59	0.68	0.70	0.51	0.59
Mistral & Sea-Lion-v3	**0.84**	**0.60**	**0.70**	0.71	**0.53**	**0.61**
Mistral & SeaLLM-2.5	0.82	0.60	0.69	0.70	0.52	0.60
Threshold $\tau = 0.4$						
LLaMA-3 & Sea-Lion-v3	0.80	0.57	0.67	0.68	0.50	0.58
LLaMA-3 & SeaLLM-2.5	0.81	0.58	0.68	0.70	0.51	0.59
Mistral & Sea-Lion-v3	**0.83**	**0.60**	**0.70**	0.71	**0.53**	**0.61**
Mistral & SeaLLM-2.5	0.82	0.60	0.69	0.70	0.52	0.60
Threshold $\tau = 0.3$						
LLaMA-3 & Sea-Lion-v3	0.80	0.57	0.67	0.69	0.50	0.58
LLaMA-3 & SeaLLM-2.5	0.81	0.58	0.68	0.70	0.51	0.59
Mistral & Sea-Lion-v3	**0.83**	**0.60**	**0.70**	0.71	**0.53**	**0.61**
Mistral & SeaLLM-2.5	0.82	0.59	0.69	0.70	0.52	0.60
Threshold $\tau = 0.2$						
LLaMA-3 & Sea-Lion-v3	0.79	0.57	0.66	0.68	0.50	0.58
LLaMA-3 & SeaLLM-2.5	0.80	0.58	0.67	0.69	0.51	0.59
Mistral & Sea-Lion-v3	**0.83**	**0.59**	**0.69**	0.71	**0.53**	**0.61**
Mistral & SeaLLM-2.5	0.82	0.59	0.69	0.70	0.52	0.60
Threshold $\tau = 0.1$						
LLaMA-3 & Sea-Lion-v3	0.79	0.56	0.66	0.68	0.50	0.58
LLaMA-3 & SeaLLM-2.5	0.79	0.57	0.66	0.69	0.51	0.59
Mistral & Sea-Lion-v3	**0.83**	**0.59**	**0.69**	0.71	**0.54**	**0.61**
Mistral & SeaLLM-2.5	0.81	0.59	0.68	0.70	0.52	0.60

Table 24: Detailed domain-level performance comparison for the **Malay** datasets (MALAYEL-WIKI, MALAYEL-NEWS, and MALAYEL-SOCIAL) across the REFINED and MGENRE backends. Bold indicates the highest score per metric for each dataset within that specific backend.

		REFINED Backend			MGENRE Backend		
Dataset	**Model**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
MALAYEL-WIKI	**Baseline:** Direct ED Model	0.776	0.441	0.563	0.549	0.373	0.443
	Individual LLM Expansion						
	LLAMA-3	0.776	0.441	0.563	0.587	0.378	0.459
	MISTRAL	0.745	0.373	0.497	0.558	0.371	0.445
	MALAYSIAN-LLAMA-3	**0.804**	0.402	0.536	0.561	0.373	0.447
	MALAYSIAN-MISTRAL	0.750	0.382	0.507	0.562	0.373	0.448
	CELESTA *(Dual-LLM Mention Expansion)*						
	LLAMA-3 & MAL-LLAMA-3	0.761	0.555	0.642	0.581	0.431	0.495
	LLAMA-3 & MAL-MISTRAL	0.801	0.561	**0.660**	**0.588**	**0.435**	**0.500**
	MISTRAL & MAL-LLAMA-3	0.747	0.557	0.638	0.559	0.422	0.480
	MISTRAL & MAL-MISTRAL	0.777	**0.569**	0.657	0.568	0.426	0.486
MALAYEL-NEWS	**Baseline:** Direct ED Model	0.804	0.435	0.565	0.558	0.392	0.460
	Individual LLM Expansion						
	LLAMA-3	0.769	0.392	0.520	0.592	0.400	0.476
	MISTRAL	**0.808**	0.412	0.546	0.565	0.380	0.454
	MALAYSIAN-LLAMA-3	**0.808**	0.412	0.546	0.550	0.377	0.446
	MALAYSIAN-MISTRAL	0.800	0.392	0.526	0.548	0.375	0.444
	CELESTA *(Dual-LLM Mention Expansion)*						
	LLAMA-3 & MAL-LLAMA-3	0.766	0.559	0.646	0.592	0.455	0.514
	LLAMA-3 & MAL-MISTRAL	0.792	0.567	**0.661**	**0.597**	**0.457**	**0.517**
	MISTRAL & MAL-LLAMA-3	0.751	0.563	0.643	0.581	0.443	0.502
	MISTRAL & MAL-MISTRAL	0.780	**0.569**	0.658	0.580	0.441	0.500
MALAYEL-SOCIAL	**Baseline:** Direct ED Model	0.765	0.428	0.548	0.555	0.353	0.431
	Individual LLM Expansion						
	LLAMA-3	0.714	0.294	0.417	0.571	0.355	0.437
	MISTRAL	0.750	0.353	0.480	0.580	0.355	0.440
	MALAYSIAN-LLAMA-3	**0.824**	0.412	0.549	0.572	0.361	0.442
	MALAYSIAN-MISTRAL	0.735	0.353	0.477	0.572	0.355	0.437
	CELESTA *(Dual-LLM Mention Expansion)*						
	LLAMA-3 & MAL-LLAMA-3	0.747	0.531	0.621	**0.595**	0.418	**0.491**
	LLAMA-3 & MAL-MISTRAL	0.765	0.535	**0.630**	0.586	**0.420**	0.489
	MISTRAL & MAL-LLAMA-3	0.747	0.537	0.625	0.563	0.396	0.465
	MISTRAL & MAL-MISTRAL	0.750	**0.541**	0.629	0.566	0.394	0.464