Data & Knowledge Engineering 161 (2026) 102504

Contents lists available at ScienceDirect

Data & Knowledge Engineering

journal homepage: www.elsevier.com/locate/datak

ELEVATE-ID: Extending Large Language Models for End-to-End Entity Linking Evaluation in Indonesian

Ria Hari Gusmita [a,b,*], Asep Fajar Firmansyah [a,b], Hamada M. Zahera [a], Axel-Cyrille Ngonga Ngomo [a]

[a] Data Science Group (DICE), Heinz Nixdorf Institute, Paderborn University, Warburger Street 100, Paderborn, 33098, North Rhine-Westphalia, Germany

[b] The State Islamic University Syarif Hidayatullah Jakarta, Ir. H. Juanda Street 95, Ciputat, South Tangerang, 15412, Banten, Indonesia

ARTICLE INFO

Keywords:
LLMs
Evaluation
End-to-end EL
Indonesian

ABSTRACT

Large Language Models (LLMs) have demonstrated remarkable performance across a wide range of natural language processing tasks. However, their effectiveness in low-resource languages remains underexplored, particularly in complex tasks such as end-to-end Entity Linking (EL), which requires both mention detection and disambiguation against a knowledge base (KB). In earlier work, we introduced IndEL — the first end-to-end EL benchmark dataset for the Indonesian language — covering both a general domain (news) and a specific domain (religious text from the Indonesian translation of the Quran), and evaluated four traditional end-to-end EL systems on this dataset. In this study, we propose ELEVATE-ID, a comprehensive evaluation framework for assessing LLM performance on end-to-end EL in Indonesian. The framework evaluates LLMs under both zero-shot and fine-tuned conditions, using multilingual and Indonesian monolingual models, with Wikidata as the target KB. Our experiments include performance benchmarking, generalization analysis across domains, and systematic error analysis. Results show that GPT-4 and GPT-3.5 achieve the highest accuracy in zero-shot and fine-tuned settings, respectively. However, even fine-tuned GPT-3.5 underperforms compared to DBpedia Spotlight — the weakest of the traditional model baselines — in the general domain. Interestingly, GPT-3.5 outperforms Babelfy in the specific domain. Generalization analysis indicates that fine-tuned GPT-3.5 adapts more effectively to cross-domain and mixed-domain scenarios. Error analysis uncovers persistent challenges that hinder LLM performance: difficulties with non-complete mentions, acronym disambiguation, and full-name recognition in formal contexts. These issues point to limitations in mention boundary detection and contextual grounding. Indonesian-pretrained LLMs, Komodo and Merak, reveal core weaknesses: template leakage and entity hallucination, respectively—underscoring architectural and training limitations in low-resource end-to-end EL.[1]

* Corresponding author at: Data Science Group (DICE), Heinz Nixdorf Institute, Paderborn University, Warburger Street 100, Paderborn, 33098, North Rhine-Westphalia, Germany.

E-mail addresses: ria.hari.gusmita@uni-paderborn.de (R.H. Gusmita), asep.fajar.firmansyah@uni-paderborn.de (A.F. Firmansyah), hamada.zahera@uni-paderborn.de (H.M. Zahera), axel.ngonga@uni-paderborn.de (A.-C. Ngonga Ngomo).

[1] Code and dataset are available at https://github.com/dice-group/ELEVATE-ID.

https://doi.org/10.1016/j.datak.2025.102504
Received 16 November 2024; Received in revised form 9 July 2025; Accepted 11 August 2025
Available online 19 August 2025

1. Introduction

Named Entity Recognition (NER) and Entity Linking (EL) are two closely related tasks in Natural Language Processing (NLP). NER identifies spans of text that refer to entities (e.g., persons, organizations, locations) [1], while EL grounds these mentions to unique entries in a Knowledge Base (KB) [2], such as Wikidata [3] or DBpedia [4]. These tasks are often integrated in a unified setting known as *end-to-end Entity Linking* (end-to-end EL), where a system simultaneously identifies entity mentions and links them to their corresponding KB entries without relying on a separate NER stage [5,6]. For instance, in the sentences *Affandi bergabung dengan kelompok Lima Bandung sekitar tahun 30-an* (Affandi joined the Lima Bandung group around the 1930s) and *Affandi berhasil memaksimalkan peran badan amil zakat kabupaten* (Affandi succeeded in enhancing the role of a district amil zakat foundation), end-to-end EL not only detects the mention *Affandi* but also disambiguates it as the renowned Indonesian painter (wd:Q2826050) in the first sentence and as a local regent (wd:Q20426359) in the second [7]. This joint approach extends the utility of NER by enabling deeper semantic understanding and supports downstream applications such as knowledge base population [8], question answering [9], and information extraction [10].

While spoken by over 280 million people,[2] Indonesian has limited NLP resources until recently, primarily due to the lack of robust benchmark datasets that accommodate both the general linguistic attributes of Indonesian and its specialized contexts. Since 2020, collaborative efforts have introduced over 140 Indonesian NLP datasets [11] and models like IndoBERT [12,13], with the addition of IndQNER [14] as the first NER dataset for a specific Indonesian domain. However, there remained a gap in EL benchmark datasets for both general and specific domains. To address this, we presented IndEL [7] in our previous work as the first EL benchmark dataset for Indonesian, covering both general and specific domains with carefully annotated entities linked to Wikidata.

Large Language Models (LLMs) have demonstrated impressive capabilities across various NLP tasks, including identifying entity mentions [15–20] and linking them to unique KB entries [21,22]. They have also shown strong performance in relation extraction [23], speech recognition [24], machine translation [25], and language understanding [26,27]. However, their evaluation has primarily focused on high-resource languages (HrLs), particularly English, limiting our understanding of their effectiveness in low-resource settings. Building on our prior work with IndEL — which introduced a benchmark suitable for end-to-end EL in Indonesian — we extend this research by systematically evaluating the ability of LLMs to perform end-to-end EL. Specifically, the LLMs are tasked with identifying mentions in text and linking them to corresponding entries in Wikidata. The evaluation encompasses both multilingual and monolingual Indonesian LLMs, assessed in zero-shot and fine-tuning settings across general and specific domains. The analyses are structured into three components: performance evaluation, generalization analysis [28], and error analysis. Generalization analysis includes cross-domain evaluations, where models fine-tuned on one domain (general or specific) are tested on the other to assess their ability to adapt to unseen contexts. Mixed-domain evaluations are also conducted, involving models fine-tuned on combined data from both domains and evaluated within their respective original training contexts to measure robustness and consistency. To further explore the challenges LLMs face in performing end-to-end EL tasks for Indonesian, human evaluation is conducted to qualitatively analyze the results from both zero-shot and fine-tuning experiments. Our contributions can be summarized as follows:

1. We benchmark the performance of multilingual LLMs, including GPT-3.5 [29], GPT-4 [30], and LLaMA-3 [31], as well as two monolingual Indonesian LLMs, Komodo [32] and Merak,[3] in end-to-end EL tasks on IndEL. In the zero-shot setting, all models perform poorly, with GPT-4 slightly outperforming the others. Despite being trained on Indonesian, Komodo and Merak also struggle in this setting. In contrast, fine-tuned GPT-3.5 achieves the highest F1-scores in both domains, while Komodo and Merak show competitive gains in the specific domain.
2. We identify key challenges faced by LLMs in end-to-end EL, including difficulties in handling incomplete mentions, acronym disambiguation, and failures to detect full-name entities in formal contexts. Additionally, some models struggle with salutation over-inclusion, hallucinating unrelated entities, or generating unreplaced prompt placeholders.

Our findings challenge the prevailing assumption that strong multilingual LLMs can generalize effectively to low-resource languages (LrLs) like Indonesian. Results from ELEVATE-ID show that even with fine-tuning, models such as GPT-3.5 fall short of traditional end-to-end EL systems in key performance metrics. This highlights important limitations in current LLM architectures when applied to LrL contexts. ELEVATE-ID thus fills a critical gap: it not only benchmarks LLM-based end-to-end EL for Indonesian, but also serves as a diagnostic tool to pinpoint and explain where these models struggle. By providing a reusable and extensible evaluation framework, we contribute a scalable methodology that can be applied to other LrLs, advancing the broader goal of inclusive and equitable NLP research.

2. Related works

2.1. Traditional methods for named entity recognition and entity linking

Traditional approaches to NER are commonly classified into three categories: *rule-based systems*, *unsupervised methods*, and *feature-based supervised models* [33]. Rule-based systems employ handcrafted linguistic rules, typically constructed using domain-specific gazetteers [34] and syntactic-lexical patterns [35]. While effective when supported by comprehensive lexicons, these

[2] https://www.bps.go.id/en/statistics-table/2/MTk3NSMy/jumlah-penduduk-pertengahan-tahun–ribu-jiwa-.html
[3] https://huggingface.co/Ichsan2895/Merak-7B-v4-GGUF

systems often exhibit limited recall and poor generalization, despite achieving high precision in narrow domains. Unsupervised approaches often use clustering [36] and distributional statistics [37] over unlabeled data to group and identify entities, reducing the need for annotated corpora. Feature-based supervised models such as Hidden Markov Models [38], Conditional Random Fields [39], Maximum Entropy Models [40], and Support Vector Machines [41] utilize carefully engineered features — like POS tags, orthographic cues, and gazetteer matches — to model named entities as sequence labeling tasks. While these methods have achieved strong performance in resource-rich settings, they rely heavily on annotated datasets and feature engineering, making them less suitable for LrLs.

EL is commonly modeled as a three-stage process: *candidate generation*, *candidate ranking*, and *result selection* [42]. Candidate generation aims to construct a set of possible entities E_m for each mention m, typically by comparing the surface form of the mention with names in a KB. According to [43], candidate generation approaches can be grouped into three main categories: (i) name dictionary-based techniques [44,45], which construct mappings using features from Wikipedia such as entity pages, redirect pages, disambiguation pages, bold phrases, and hyperlinks; (ii) surface form expansion methods [46,47], which expand acronyms or incomplete names using local document context through heuristic patterns or supervised learning methods; and (iii) search engine-based methods [48,49], which leverage web or Wikipedia search APIs to retrieve candidate entities based on the mention string and its surrounding context. Candidate ranking aims to select the most appropriate entity from the candidate set E_m for a given mention m. Since the size of E_m is often greater than one, this module plays a crucial role in disambiguating the correct entity. Ranking approaches can be broadly divided into two categories: (i) supervised ranking methods, which rely on annotated training data to learn ranking models—such as binary classifiers [50,51], learning-to-rank methods [52,53], probabilistic models [54,55], or graph-based approaches [56–58]; and (ii) unsupervised ranking methods, which operate without labeled data and use techniques such as Vector Space Models [59,60] or statistical information retrieval [44,61,62]. Alternatively, candidate ranking methods can also be classified based on how they handle interdependencies between mentions: (i) independent ranking methods treat mentions independently and typically compute similarity between local context and entity descriptions [63,64]; (ii) collective ranking methods assume mentions in a document refer to related entities and exploit topical coherence across mentions for joint disambiguation [65–67]; and (iii) collaborative ranking methods consider similar mentions and contexts across documents to enhance ranking decisions by sharing contextual cues [68,69]. Result selection is the final stage in the EL pipeline, tasked with determining whether the top-ranked entity e_{top} from the candidate set E_m should be assigned to the mention m, or if the system should return `NIL`, indicating no suitable match. Based on how this decision is made, result selection approaches can be grouped into three main categories [42,43]: (i) Threshold-based methods, which compare the score of the top-ranked candidate entity s_{top} against a fixed or learned threshold τ. If $s_{\text{top}} < \tau$, the system returns `NIL` for the mention m [70,71]. While simple and widely used, manual thresholding can lead to missed links when valid entities fall below the threshold. (ii) Classifier-based methods, which treat the result selection task as a binary classification problem [72,73]. These methods evaluate the pair $\langle m, e_{\text{top}} \rangle$ to determine whether e_{top} is a valid mapping for m. Features used often overlap with those in candidate ranking, such as contextual similarity and NER confidence scores [74,75]. (iii) Joint prediction methods, which integrate unlinkable mention prediction into the ranking process by augmenting the candidate set with a synthetic `NIL` entity [55,76]. If the ranker selects `NIL` as the top entity, the mention is considered unlinkable [49,77]. Probabilistic models extend this strategy by comparing the generative likelihood of the mention given the NIL entity versus that given by other candidate entities. A mention is predicted as unlinkable when the NIL entity yields higher likelihood than any real entity [55].

2.2. End-to-end entity linking models

End-to-end EL refers to systems that simultaneously detect entity mentions and link them to KB entries, eliminating the need for separate NER and EL pipelines [5,6]. Early heuristic-based systems like DBpedia Spotlight [78] and TagMe [79] applied mention detection and disambiguation in a single pass. TagMe's underlying algorithm was later refined into WAT, which improved precision and disambiguation speed [80]. A major milestone in the field was the work by Kolitsas et al. [5], which was among the first to introduce a neural end-to-end EL architecture. Their model considers all possible spans as candidate mentions and jointly optimizes mention detection and entity linking. The model learns contextual mention and entity embeddings, using a probabilistic mention–entity mapping, and achieves notable performance gains in the GERBIL benchmarking framework [81] when sufficient training data is available. Following this, several models have extended the end-to-end EL paradigm. ReFinED [82] proposes a type-aware architecture that jointly performs mention detection and disambiguation, leveraging fine-grained type information to prune the candidate space early and improve both efficiency and accuracy. Unlike pipeline-based approaches, ReFinED enables real-time *end-to-end* EL with large knowledge bases, making it suitable for industrial applications. Laskar et al. [83] extend the BLINK framework by integrating it with Elasticsearch to support scalable, real-time end-to-end EL in business conversation settings. These works demonstrate the effectiveness and scalability of fully neural, end-to-end EL systems in both open-domain and domain-specific settings.

2.3. End-to-end entity linking benchmark datasets

TweetNERD [84] is a large-scale benchmark dataset designed for evaluating end-to-end EL on social media texts, particularly tweets. It comprises over 340,000 English tweets annotated with both entity mention spans and their corresponding Wikidata entity links, enabling comprehensive evaluation of end-to-end EL systems in noisy, user-generated contexts. The dataset captures temporal diversity (spanning 2010–2021), making it suitable for assessing robustness to linguistic drift and emerging entities. Unlike prior

datasets that separate NER and linking stages, TweetNERD facilitates joint evaluation of mention detection and disambiguation, aligning well with modern, fully end-to-end EL models. Its release supports reproducible research in low-context, high-noise environments—conditions typical of real-world applications such as misinformation detection and social media analytics.

KORE^{DYWC} was introduced as an extension of the KORE 50 data set to include YAGO, Wikidata, and Crunchbase [85]. The goal is to provide an evaluation data set that addresses the limitations of existing data sets and can be easily used by other developers. The KORE 50 data set was chosen as a foundation because it is popular and covers a broad range of topics in English. Three sub-data sets are released for each KB: YAGO, Wikidata, and Crunchbase. YAGO and Wikidata cover general knowledge, while Crunchbase focuses on technology and business. To perform the annotation, the authors used WebAnno, a web-based annotation tool, to manually annotate the KORE 50 data set using entities from different KBs. Each document was manually annotated by searching for entities in the respective KB. The annotations were exported using the WebAnno TSV3 format. There are some peculiarities of the annotation. Some entities were available in YAGO and Wikidata, but not in Crunchbase. YAGO offers a larger number of resources for annotation compared to DBpedia. Wikidata provides information for a broader range of mentions than DBpedia. Crunchbase has a tech-focused domain, resulting in fewer entities compared to DBpedia.

2.4. Large language model-based for named entity recognition, entity linking, and end-to-end entity linking

The emergence of LLMs has redefined NER, EL, and end-to-end EL by enabling models to learn rich contextual representations. In NER, LLMs fine-tuned on token classification tasks have consistently outperformed traditional models across various domains and languages. Complementing this trend, hybrid architectures like LinkNER [15] combine small fine-tuned NER models with LLMs using uncertainty guidance: the local NER model predicts spans and estimates uncertainty, and ambiguous spans are then re-classified by an LLM (e.g., GPT-3.5 or LLaMA-2) using in-context prompts, resulting in robust detection particularly for unseen or noisy entities. In EL, LLM-based approaches typically rely on gold mention spans and focus solely on the linking stage. EntGPT [21] models EL as an instruction-following task. Given a pre-annotated mention and its surrounding context, EntGPT uses a GPT-style language model to generate or select the correct entity title from a list of candidates derived from a KB. This approach simplifies the disambiguation process by avoiding explicit retrieval during inference, but it does not address the mention detection step. LLMAEL [86] adopts a hybrid pipeline where LLMs augment mention contexts, which are then processed by traditional EL systems such as BLINK, GENRE, or ReFinED. This three-stage architecture — comprising LLM-based context enrichment, data fusion, and disambiguation — has shown strong performance for rare and long-tail entities. However, LLMAEL assumes gold mentions and does not address mention detection. Vollmers et al. [87] propose a jointly fine-tuned NER+EL model based on T5, enhanced with LLM-driven mention expansion. They prompt LLaMA-3 to expand ambiguous mentions — e.g., rewriting ‘‘Angelina’’ as ‘‘Angelina Jolie’’ — and then constrain outputs via a filtered Wikipedia-derived dictionary to reduce hallucination. The model is evaluated in both EL-only (with gold mentions) and joint NER+EL settings. Their tests show that augmentation improves performance significantly, especially on out-of-domain datasets.

Although recent methods have advanced the integration of LLMs into NER, EL, and end-to-end EL, none fully support true end-to-end EL — where mention detection and linking to KB entries are performed jointly — in LrL contexts. In contrast, ELEVATE-ID addresses this gap by evaluating LLMs as complete end-to-end EL systems for Indonesian: given raw text, models are required to detect entity spans and link them to corresponding KB entries. Using the IndEL benchmark, we assess both multilingual and Indonesian monolingual LLMs under zero-shot and fine-tuning settings, providing the first comprehensive evaluation of LLM-based end-to-end EL in an LrL environment.

3. IndEL dataset

The IndEL dataset is designed to support both EL and end-to-end EL across general and specific domains in Indonesian. To clarify this distinction, we define the two domains as follows. The *general domain* refers to news-based content that reflects everyday language use, current events, and public discourse, and is represented by the NER UI dataset.[4] In contrast, the *specific domain* encompasses religious and culturally grounded content derived from the Indonesian translation of the Quran, represented by the IndQNER dataset.[5] These domain definitions help capture both broad linguistic usage and specialized vocabulary, which are essential for evaluating domain-sensitive end-to-end EL systems.

NER UI includes 5055 entities across the classes *Person* (1870 entities), *Organization* (1949 entities), and *Location* (1236 entities). Among the first two Indonesian NER benchmark datasets introduced in 2020, NER UI has proven to be the most effective, as fine-tuning IndoBERT on it achieved the highest F1 score of 90.1% [13]. Despite its utility, the NER UI dataset also presents several annotation challenges that affect the quality and reliability of end-to-end EL evaluation. These include misspelled entities, incorrect entity spans, and missing entities. Misspellings, such as *Lea Iacocca* instead of *Lee Iacocca* and *Lentang* instead of *lenteng*, can make it difficult for end-to-end EL systems to find and accurately link these entities to the correct entries in knowledge bases (KBs). Additionally, the dataset sometimes incorrectly labels entity spans or fails to capture the complete span of an entity, such as treating *Fakultas Ekonomi* (Economics Faculty) and *Universitas Indonesia* (University of Indonesia) as separate entities instead of combining them into *Fakultas Ekonomi Universitas Indonesia* (Economics Faculty at the University of Indonesia). Similarly, entities like *Pemkot*

[4] https://github.com/indolem/indolem/tree/main/ner/data/nerui
[5] https://github.com/dice-group/IndQNER

Table 1
Distribution of total entities, unique entities, sentences with nested entities, and the average of entities in sentences in IndEL.

Property	General	Specific
Total entities	4765	2453
Unique entities	55	16
Sentences with nested entities	1488	141
Entities in sentence	2.4	1.6

(city/local government) and *Surabaya* should be recognized together as *Pemkot Surabaya*. Finally, some entities are entirely omitted, such as *Hye-kyo* (*Person*) and *Korea Times* (*Organization*), which can lead to incomplete end-to-end EL results.

On the other hand, IndQNER comprises 3117 sentences and 2475 entities across 18 entity types, as detailed in [14]. Evaluation of an Indonesian NER system using BiLSTM and CRF on IndQNER yielded an F1 score of 98% [14], indicating high annotation consistency. Since IndQNER is based on a well-defined source (the Indonesian translation of the Qur'an), it features more controlled vocabulary and fewer noisy or ambiguous mentions compared to the general domain. This makes it a valuable complement to NER UI in evaluating end-to-end EL performance across diverse linguistic and domain-specific settings.

3.1. Human annotation

The IndEL annotation involved six native Indonesian speakers. Annotators were grouped into pairs, forming two annotation groups for the general domain and one for the specific domain. Each group independently annotated the same source data within their assigned domain. An additional independent annotator performed quality control and adjudicated disagreements across both domains. The annotation process followed meticulously prepared guidelines.[6]

A trial round using 20 sentences from the NER UGM dataset[7] was first conducted to align annotators' understanding and ensure consistency. During the main annotation phase, entities were categorized as *Agreed* (both annotators linked to the same entity), *Disagreed* (conflicting links), or *OneNoLink* (only one annotator linked the entity). For *OneNoLink* cases, annotators revisited the entity to reach a decision. Disagreements were resolved by the third annotator, who reviewed Wikidata entries and, when needed, added new ones.

In the general domain, remaining *OneNoLink* entities were first validated against NER UI; those not found were marked as named entity candidates. Both valid entities and candidates were then manually verified by the third annotator. In the specific domain, remaining *OneNoLink* cases were manually reviewed, and no valid links were found. For *Disagreed* entities in both domains, the third annotator selected the correct link or proposed a new one. All *Agreed* entities were also re-verified. This process resulted in 4765 verified entities in the general domain and 2453 in the specific domain.

Although inter-annotator agreement scores, such as Cohen's kappa, were not reported in [7], the annotation process followed a structured double-annotation protocol with clearly defined disagreement categories and third-party adjudication. The grouping of annotators and multi-stage validation steps described here provide a strong qualitative basis for annotation consistency and reliability.

3.2. Dataset analysis, format, and usage

Table 1 summarizes key statistics of the IndEL dataset used in this work, including the distribution of total and unique entities, the number of sentences with nested entities, and the average number of entities per sentence across general and specific domains. The general domain, as expected, features a significantly broader array of entities, with about 1.2% being unique. It also includes more complex sentence structures, reflected in the higher proportion of nested entities and an average of 2.4 entities per sentence. In contrast, the specific domain, derived from religious texts, contains fewer unique entities and lower average entity density, indicating its more focused scope.

These statistics reflect the verified entities obtained through the human annotation process introduced in our prior work [7], and provide a foundation for evaluating end-to-end EL in both general and specific domain contexts. The dataset is formatted in the NLP Interchange Format (NIF) and integrated into the GERBIL platform [81], enabling reproducible evaluation of end-to-end EL systems. In this study, we independently evaluate LLM-based approaches using the same dataset.

4. LLMs evaluation framework for end-to-end entity linking

In this paper, we evaluated the performance of LLMs on the IndEL benchmark across both general and specific domains. We then compared their performances to those of traditional end-to-end EL models previously evaluated on IndEL [7]. As shown in Fig. 1, we conducted experiments with multilingual and monolingual LLMs on IndEL in both zero-shot and fine-tuning settings. Subsequently, we evaluated the results through performance analysis, generalization analysis, and error analysis. Performance analysis assessed how effectively the LLMs identified and linked entities to the correct entries on Wikidata. Generalization analysis examined the ability of LLMs to generalize across domains (cross-domain evaluation) and perform in mixed-domain settings, where they were

[6] Annotation Guidelines (IndEL)
[7] https://github.com/indolem/indolem/tree/main/ner/data/nerugm

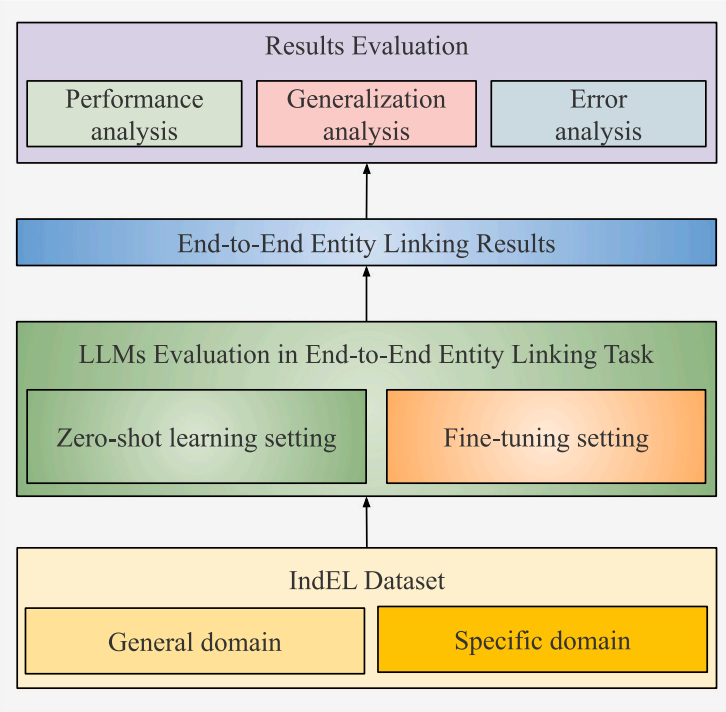


Fig. 1. The framework of LLMs evaluation on IndEL.

Table 2
Instruction example.

Task description	Find entities and their corresponding entry links in Wikidata within the following sentence. Use the context of the sentence to determine the correct entries in Wikidata.
Output format	The output should be formatted as: [entity1=link1, entity2=link2]. No explanations are needed.
Sample sentence	*Pria kelahiran Bogor, 16 Maret 60 tahun silam itu juga ditunjuk sebagai salah satu direktur Indofood dalam RUPS Juni 2008 silam.* (A man born in Bogor, 60 years ago on March 16, was also appointed as one of the directors of Indofood in the General Meeting of Shareholders in June 2008.)

fine-tuned on combined data from general and specific domains and tested on particular domain. Finally, error analysis identified common types of mistakes made by LLMs, such as misidentifying entities or failing to link them correctly, providing insights to develop better error mitigation strategies for LLM-based end-to-end EL systems. To ensure standardization and consistency, we defined relevant prompts for fine-tuning LLMs, similar to annotation guidelines used in human-based annotation tasks. Table 2 outlines the prompts, which comprised two parts: the task description and the desired outputs.

5. Experiments

We evaluated four LLMs, GPT-4, Komodo (Komodo-7b-base[8]), LLaMA-3 (LLaMA-3-8B-Instruct[9]), and Merak (Merak-7B-v4[10]) in end-to-end EL tasks using the IndEL dataset. Our experiments focused on two scenarios: zero-shot and fine-tuning settings. For the fine-tuning setting, GPT-3.5 was included in place of GPT-4 to assess its adaptability.[11] Through the experiments, we were interested in addressing the following questions:

RQ1: How do GPT-4, Komodo, LLaMA-3, and Merak perform in the zero-shot setting on the IndEL dataset?

[8] https://huggingface.co/Yellow-AI-NLP/komodo-7b-base
[9] https://huggingface.co/meta-llama/Meta-LLaMA-3-8B-Instruct
[10] https://huggingface.co/Ichsan2895/Merak-7B-v4
[11] At the time of this research, fine-tuning GPT-4 was not accessible to us.

RQ2: How does fine-tuning affect the performance of GPT-3.5, Komodo, LLaMA-3, and Merak, and how do they compare to each other after fine-tuning?
RQ3: How well do the models generalize to unseen entities or contexts in the IndEL dataset?
RQ4: What types of errors are most common for each model in both the zero-shot and fine-tuning settings?

To address RQ1, we evaluated GPT-4, Komodo, LLaMA-3, and Merak using the IndEL dataset, covering both general and specific domains. For RQ2, we fine-tuned GPT-3.5, Komodo, LLaMA-3, and Merak with training and validation sets from IndEL, then evaluated the fine-tuned models with the IndEL test set. We compared their fine-tuned performance to their zero-shot performance. To address RQ3, we performed cross-domain and mixed-domain evaluations involving GPT-3.5, Komodo, LLaMA-3, and Merak. To address RQ4, we conducted a detailed analysis of the results from both zero-shot and fine-tuning settings both in general and specific domains.

5.1. Experiment setup

5.1.1. Zero-shot learning setting

In the zero-shot setting, we prompted the LLMs using the instruction format shown in Table 2, where the prompt includes only the task description and output format. Zero-shot learning in this context means that the model is performing the end-to-end EL task without being explicitly trained for it. Instead of being trained specifically to identify entities and link them to Wikidata, the model uses its general understanding of language and knowledge encoded during pre-training to infer the mentions and their correct links. We evaluated the LLMs using the test set obtained from IndEL, covering both general and specific domains.

5.1.2. Fine-tuning setting

Fine-tuning LLMs involves adapting a pre-trained model to a specific task, such as end-to-end EL, by using smaller, task-specific datasets. In this process, the LLMs were provided with detailed prompts and example sentences from the dataset, as outlined in Table 2. The fine-tuning leveraged both general and specific domain datasets from IndEL, ensuring the model could handle a wide range of contexts. Training and validation sets from both domains were used to refine the model's parameters, enhancing its ability to accurately identify mentions and link them within sentences. Key hyperparameters during the fine-tuning process included a batch size of 8 with gradient accumulation steps of 4, a learning rate of 2e−4, and training over 3 epochs. Additionally, a warmup ratio of 0.03 and 100 warmup steps were used to stabilize the initial learning rate, and gradient clipping with a max norm of 0.3 was applied to maintain training stability. The model utilized 4-bit quantization via BitsAndBytes and Low-Rank Adaptation (LoRA) with specific settings (alpha of 16, dropout of 0.1, and rank of 64) to optimize memory and computation efficiency. Once the fine-tuning was complete, the model was evaluated using test sets from both general and specific domains to measure its performance.

5.2. Dataset setup

In our experiments, we split IndEL into training, validation, and test sets with an 8:1:1 ratio [88]. In the general domain, this resulted in 1673, 229, and 212 sentences for training, validation, and testing, respectively. For the specific domain, the dataset was split into 2075, 283, and 263 sentences for training, validation, and testing, respectively.

5.3. Evaluation metric

We employed two evaluation metrics: automatic evaluation for quantitative performance assessment and human evaluation for qualitative analysis. The details of each metric are outlined below.

5.3.1. Automatic evaluation

The automatic evaluation relies on standard metrics, including precision, recall, and F1-score, to objectively quantify the performance of LLMs in end-to-end EL tasks:

- **Precision** (P): Represents the proportion of correctly linked entities among all entities predicted as linked. This metric highlights the model's accuracy in minimizing incorrect links.

$$P = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

  where TP being true positives (correct links) and FP being false positives (incorrect links).
- **Recall** (R): Denotes the proportion of correctly linked entities relative to all ground truth entities. Recall measures the model's ability to comprehensively capture relevant entities.

$$R = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

  where TP being true positives (correctly linked mentions) and FN being false negatives (mentions that should have been linked but were missed).

- **F1-score** ($F1$): Calculated as the harmonic mean of precision and recall, this metric provides a balanced evaluation of the model's performance, especially in cases where precision and recall are equally critical.

$$F1 = 2 \cdot \frac{P \cdot R}{P + R} \tag{3}$$

where P is precision and R is recall.

5.3.2. Human evaluation

The human evaluation was conducted to qualitatively examine the challenges encountered by LLMs during the end-to-end EL process. In doing so, the experiment results were categorized into three distinct groups:

A: Sentences where all identified entities and their corresponding links exactly match the IndEL references.
B: Sentences where all entities are correctly identified, but some or all links are inaccurate.
C: Sentences where no correct entities are identified.

For each category, a comprehensive manual analysis was conducted to identify and evaluate the underlying factors affecting the model's performance. This qualitative investigation aims to uncover specific challenges and limitations of LLMs in performing end-to-end EL tasks.

5.3.3. Evaluation consistency across systems

While the LLMs were evaluated using our custom Python scripts and the traditional end-to-end EL systems were evaluated through the GERBIL platform, the same core metric — exact match of predicted entity links to gold-standard Wikidata QIDs — was applied across all systems. For both settings, we computed micro-averaged precision, recall, and F1-score, based solely on exact QID matches. No partial credit or fuzzy matching was used. This alignment ensures a fair comparison between LLM-based and traditional end-to-end EL models despite differences in implementation and execution environments.

5.4. Baselines

To benchmark the performance of LLMs in end-to-end EL tasks, we compared them against four traditional end-to-end EL models previously evaluated on the IndEL dataset [7]. These models were assessed using the GERBIL framework [81], which provides a standardized platform for evaluating end-to-end EL systems. The baseline models include Babelfy [89], DBpedia Spotlight [78], OpenTapioca [90], and WAT [80]. Each system's performance was tested on the IndEL benchmark, covering both general and specific domains. Details of the evaluation results are available in the repository.[12]

6. Results and analysis

6.1. Performance analysis (RQ1 and RQ2)

Tables 3 and 4 compare the performance of various LLMs, including GPT-4, LLaMA-3, Komodo, and Merak in the zero-shot setting and GPT-3.5, LLaMA-3, Komodo, and Merak in the fine-tuning setting, with the IndEL dataset in both general and specific domains, respectively. In the zero-shot setting, both GPT-4 and LLaMA-3 exhibit very low performance across both domains, but GPT-4 achieves slightly better results with an F1-score of 0.083 in the general domain and 0.012 in the specific domain, compared to LLaMA-3's F1-scores of 0.003 and 0.000, respectively. For Indonesian-trained LLMs, Komodo and Merak, unexpectedly show very poor performance with F1-scores of 0.000 in both domains. In the fine-tuning setting, all LLMs generally perform better in the specific domain compared to the general domain. GPT-3.5 significantly outperforms the other LLMs in the general domain, achieving an F1-score of 0.373, while the other LLMs do not reach an F1-score of 0.1. In the specific domain, GPT-3.5 shows the best performance among all LLMs, achieving an F1-score of 0.611. Merak shows notable performance by achieving an F1-score of 0.407, which is slightly lower than LLaMA-3's F1-score. Despite Komodo has the worst performance among all models based on F1-scores, it outperforms Merak and LLaMA-3 models according to recall measurement. These results demonstrate that the fine-tuning process significantly benefits the LLMs, as their performances increase substantially compared to their performances in the zero-shot setting. Table 3 also shows that the performance of most LLMs in end-to-end EL tasks in Indonesian still lags behind that of the four end-to-end EL systems evaluated on IndEL [7]. In the general domain, the best performer in the fine-tuning setting, GPT-3.5, still underperforms compared to DBpedia Spotlight, which is the weakest among the four traditional end-to-end EL systems. This comparison is intended to highlight that even the least effective traditional system still outperforms state-of-the-art LLMs, underscoring the substantial performance gap that remains. This indicates that even with fine-tuning, significant challenges persist in the ability of LLMs to effectively handle end-to-end EL in LrLs like Indonesian.

[12] https://github.com/dice-group/IndEL/blob/main/README.md

Table 3
Precision, recall, and F1-score from zero-shot (*) and fine-tuning (**) experiments with GPT-3.5, GPT-4, LLaMA-3, Komodo, and Merak, compared to traditional end-to-end EL systems also evaluated using the IndEL dataset in the general domain.

Model (#parameters)	Precision	Recall	F1-score
General domain			
Babelfy	0.727	0.372	0.492
DBpedia Spotlight	0.675	0.358	0.468
OpenTapioca	**0.798**	0.410	0.542
WAT	0.612	**0.555**	**0.582**
GPT−4*	**0.083**	**0.089**	**0.083**
LLaMA-3 (8B)*	0.003	0.003	0.003
Komodo (7B)*	0.000	0.000	0.000
Merak (7B)*	0.000	0.000	0.000
GPT-3.5 (175B)**	**0.385**	**0.373**	**0.373**
LLaMA-3 (8B)**	0.084	0.117	0.093
Komodo (7B)**	0.018	0.026	0.021
Merak (7B)**	0.045	0.039	0.041

Table 4
Precision, recall, and F1-score from zero-shot (*) and fine-tuning (**) experiments with GPT-3.5, GPT-4, LLaMA-3, Komodo, and Merak, compared to traditional end-to-end EL systems also evaluated using the IndEL dataset in the specific domain.

Model (#parameters)	Precision	Recall	F1-score
Specific domain			
Babelfy	0.805	0.473	0.595
DBpedia Spotlight	**0.847**	0.673	0.750
OpenTapioca	0.618	0.031	0.059
WAT	0.772	**0.750**	**0.761**
GPT−4*	**0.010**	**0.016**	**0.012**
LLaMA-3 (8B)*	0.000	0.000	0.000
Komodo (7B)*	0.000	0.000	0.000
Merak (7B)*	0.000	0.000	0.000
GPT-3.5 (175B)**	**0.616**	**0.610**	**0.611**
LLaMA-3 (8B)**	0.415	0.444	0.409
Komodo (7B)**	0.221	0.471	0.285
Merak (7B)**	0.446	0.393	0.407

6.2. Generalization analysis (RQ3)

Table 5 exhibits the cross-domain and mixed-domain evaluation results in the fine-tuning scenario, involving GPT-3.5, LLaMA-3, Komodo, and Merak. We refer to the *General domain to specific domain* scenario as the model being fine-tuned with the general domain set and evaluated with the specific domain set, and vice-versa for the *Specific domain to general domain* scenario. GPT-3.5 consistently outperforms LLaMA-3 in both scenarios. Specifically, in the *General domain to specific-domain* evaluation, GPT-3.5 achieves a precision of 0.003 and both recall and F1-scores of 0.004, while LLaMA-3 scores zero across all metrics. In the *Specific domain to general-domain* evaluation, GPT-3.5 attained a precision of 0.044, recall of 0.042, and F1-score of 0.043, while LLaMA-3 scored 0.002 for precision, 0.004 for recall, and 0.003 for F1-score. Komodo and Merak are unexpectedly not able to perform in any scenarios. These findings indicate that GPT-3.5 has better adaptability and performance in cross-domain tasks compared to LLaMA-3, with the highest performance observed when fine-tuned with the specific domain set and evaluated with the counterpart. Fine-tuning with a combination of both domains significantly enhances the models' performance compared to one-domain fine-tuning, resulting in 4 out of 8 F1-scores exceeding 0.4. Specifically, the LLMs show better performance when the fine-tuned models are evaluated with the specific domain set. GPT-3.5 outperforms all other LLMs with a precision of 0.571, and both recall and F1-score at 0.566.

6.3. Error analysis (RQ4)

Table 6 outlines the percentage of sentences in each category (A, B, and C), as defined in Section 5.3.2, based on human evaluation of GPT-4 (zero-shot), GPT-3.5 (fine-tuned), and other LLMs (LLaMA-3, Komodo, and Merak) evaluated under both zero-shot and fine-tuned settings, across both general and specific domains of the IndEL dataset.

In category A — *where all entities and their corresponding Wikidata links are correctly identified* — fine-tuned GPT-3.5 achieves the highest performance, correctly resolving 58.6% of sentences in the specific domain and 17.0% in the general domain. Among the remaining models, Merak — an Indonesian monolingual LLM — outperforms the multilingual LLaMA-3 in the specific domain, achieving 36.5% compared to LLaMA-3's 32.3%. This suggests that language-specific pretraining can be advantageous in domain-relevant tasks, particularly when entity names and structures follow culturally and linguistically specific patterns. Komodo, another

Table 5
The precision, recall, and F1-score from cross-domain and mixed-domain evaluations of fine-tuning GPT-3.5, LLaMA-3, Komodo, and Merak.

Model	Precision	Recall	F1-score
General domain to specific-domain			
GPT-3.5	**0.003**	**0.004**	**0.004**
LLaMA-3	0.000	0.000	0.000
Komodo	0.000	0.000	0.000
Merak	0.000	0.000	0.000
Specific domain to general-domain			
GPT-3.5	**0.044**	**0.042**	**0.043**
LLaMA-3	0.002	0.004	0.003
Komodo	0.000	0.000	0.000
Merak	0.000	0.000	0.000
Mix training data to general-domain			
GPT-3.5	**0.405**	**0.412**	**0.404**
LLaMA-3	0.094	0.123	0.103
Komodo	0.067	0.106	0.077
Merak	0.050	0.044	0.046
Mix training data to specific-domain			
GPT-3.5	**0.571**	**0.566**	**0.566**
LLaMA-3	0.425	0.436	0.415
Komodo	0.105	0.165	0.122
Merak	0.486	0.421	0.441

Indonesian model, performs notably lower at 2.3%. In the zero-shot setting, GPT-4 achieves only 0.8% and 0.5% in the specific and general domains, respectively, highlighting the substantial gains enabled through fine-tuning.

In category B — *sentences where all entities are correctly identified, but some or all links are inaccurate* — fine-tuned LLaMA-3 achieves the highest performance in the general domain (84.4%), followed by Merak (69.3%), GPT-3.5 (50.0%), and Komodo (42.5%). In the specific domain, LLaMA-3 again leads (16.4%), with Merak (8.4%), Komodo (8.0%), and GPT-3.5 (3.8%) following. These findings suggest that while mention detection is largely successful after fine-tuning, accurate entity disambiguation — especially linking to the correct Wikidata entries — remains a significant challenge. This difficulty persists even for models with strong language alignment (such as Komodo and Merak, which are trained specifically on Indonesian) and models that benefit from large-scale fine-tuning (such as GPT-3.5 and LLaMA-3). The results indicate that mastering the linguistic form of the input (e.g., Indonesian) or even optimizing the model through task-specific supervision does not automatically translate to robust knowledge resolution. This points to limitations in current models' reasoning over entity candidates and suggests a need for more structured, knowledge-aware training or hybrid approaches that can combine surface-level language understanding with deeper semantic linking capabilities [91].

In category C — *sentences where no correct entities are identified* — fine-tuning substantially reduces the number of such cases across most models. In the specific domain, GPT-3.5 achieves the lowest error rate (0.4%), followed by LLaMA-3 (0.8%), Komodo (4.2%), and Merak (12.6%). In the general domain, Merak shows the most notable improvement, with its error rate dropping from 20.3% (zero-shot) to 1.4% (fine-tuned). Similarly, LLaMA-3 and GPT-3.5 exhibit strong gains, with reductions from 7.6% and 5.2% to 0.0% and 0.9%, respectively. In contrast, Komodo's error rate increases in the general domain after fine-tuning — from 5.2% to 10.4% — suggesting limitations in its ability to generalize across domains. Notably, Komodo fails to generate any valid output for this category in the specific domain, indicating a critical breakdown in its entity recognition and linking pipeline under fine-tuned conditions.

In summary, fine-tuning substantially boosts performance in both entity detection and linking—especially in the specific domain. GPT-3.5 stands out in end-to-end accuracy, while LLaMA-3 shows strength in partial linking. Among the Indonesian-trained models, Merak demonstrates stronger generalization than Komodo, particularly in the general domain. While both models perform well in mention detection, their performance lags in entity disambiguation and domain transfer compared to the multilingual LLMs. These results highlight the promise of localized models, while pointing to key areas — such as entity linking reasoning and robustness across domains — where further development is needed.

While the quantitative evaluation across categories A, B, and C highlights performance differences among models and the effects of fine-tuning, it does not fully reveal the underlying causes of model failure—particularly in zero-shot settings. To gain deeper insights into the challenges faced by LLMs in performing end-to-end EL in Indonesian, we conducted a detailed analysis of two groups of problematic cases in the zero-shot setting : (i) sentences that were not processed by the models (i.e., no output was returned), and (ii) sentences that fell into category C, where no correct entities were identified. In the general domain, we found that LLMs failed to process or correctly resolve entities in these cases due to the following factors:

1. The entities exist in non-complete form such as, first names, nicknames, or aliases in sentences. Mega in the sentence "*Apa sikap Mega itu bisa disebut egois karena kadernya tidak ada yang jadi menteri?* (Can Mega's attitude be considered selfish because none of her party members were appointed as ministers?)" is considered as a non-complete entity as it is the nickname of *Megawati Soekarno Putri*.

Table 6
Statistics of the detailed human evaluation conducted on the results of GPT-3.5, GPT-4, LLaMA-3, Komodo, and Merak, assessed against IndEL in both general and specific domains.

Category	General Domain (212 sentences)		Specific Domain (263 sentences)	
	Zero-shot	Fine-tuning	Zero-shot	Fine-tuning
GPT-4 & GPT-3.5				
A	**0.5%**	**17.0%**	**0.8%**	**58.6%**
B	59.0%	50.0%	46.8%	3.8%
C	**5.2%**	0.9%	**5.3%**	**0.4%**
LLaMA-3				
A	0.0%	0.9%	0.0%	32.3%
B	**60.0%**	**84.4%**	**50.0%**	**16.4%**
C	7.6%	**0.0%**	6.5%	0.8%
Komodo				
A	0.0%	0.0%	0.0%	2.3%
B	0.5%	42.5%	0.0%	8.0%
C	**5.2%**	10.4%	0.0%	4.2%
Merak				
A	0.0%	0.9%	0.0%	36.5%
B	33.5%	69.3%	31.6%	8.4%
C	20.3%	1.4%	20.5%	12.6%

2. The entities exist as their acronyms in sentences. TNGL in "*Menurut laman resmi TNGL, DiCaprio datang bersama dua aktor lain, yakni Adrien Brody dan Fisher Stevens, bersama sejumlah kru.* (According to TNGL's official website, DiCaprio came with two other actors, namely Adrien Brody and Fisher Stevens, along with several crew members.)" is the acronym of *Taman Nasional Gunung Leuser* (Gunung Leuser National Park).
3. The entities are in their full-name form, and the sentences are written in a formal style (sourced from the general domain), yet the LLMs still fail to identify them. This issue can be seen in the following examples for each entity class covered in the general domain:

   (a) Person: *Dengan jaminan dua menteri yang memiliki integritas, Menhub Ignasius Jonan dan Menkominfo Rudiantara, proses akan mudah.* (With the assurance of two ministers who have integrity, Minister of Transportation Ignasius Jonan and Minister of Communication and Information Technology Rudiantara, the process will be easy.)
   (b) Location: *Sebagai Bandara, lalu lintas ke dan dari Bandara HLP sudah sangat padat.* (As an airport, the traffic to and from HLP Airport is already very congested.)
   (c) Organization: *Tantangan kian besar karena Sociedad akan berusaha mengeksploitasi fisik Barca setelah melakoni laga tengah pekan di Liga Champions.* (The challenge is even greater because Sociedad will try to exploit Barca's physical condition after playing a midweek match in the Champions League.)

According to the analysis of sentences in category C — where no correct entities were identified — in the general domain under the zero-shot setting, we identified the most frequent error type for each model based on the actual number of annotated error cases:

1. GPT-4: The dominant error involves partial entity extraction, which accounts for 6 out of 11 annotated cases (54.5%). In these cases, the model identifies only a portion of the correct entity or embeds the entity within a broader, imprecise phrase. For instance, instead of recognizing *BUMN* as a standalone organizational entity, GPT-4 returns *menteri pemberdayaan BUMN* (Minister of State-Owned Enterprises), incorporating a role title. Similarly, in another case, the entities *PSG* and *Ibra* are collapsed into *PSG TV*, which refers to a different concept entirely. These boundary issues reduce linking precision by distorting the intended semantic reference and impairing disambiguation against KBs like Wikidata.
2. LLaMA-3: The most common error involves the inclusion of salutations or official titles within entity spans, found in 5 out of 16 sentences (31.3%). The model frequently prepends role designations to named entities, resulting in overly broad or incorrect spans. For example, instead of correctly identifying *Yasonna Laoly*, it returns *menteri hukum dan HAM Yasonna Laoly* (Minister of Law and Human Rights Yasonna Laoly). Similarly, *Ignasius Jonan* is detected as *menhub Ignasius Jonan* (Minister of Transportation), and *Megawati Soekarnoputri* appears as *mantan presiden Megawati Soekarnoputri* (former president Megawati Soekarnoputri). This over-inclusion of titles reduces entity linking accuracy, especially when role names are not part of the canonical entity label in Wikidata.
3. Komodo: All 11 of its errors (100.0%) are due to unreplaced template placeholders such as *'entity1'*, *'entity2'*, *'link1'*, or *'link2'*, which are directly copied from the prompt format into the output. For example, instead of generating valid named entities from the sentence — such as *Telkom* and its corresponding Wikidata link — the model simply outputs generic tokens like *entity1* and *link1*, showing no evidence of sentence understanding or mention detection. These failures suggest an over-reliance on prompt structure during fine-tuning and highlight a critical weakness in semantic grounding and output generation.

4. Merak: Hallucination is the most frequent error, occurring in 18 out of 43 annotated sentences (41.9%). In these cases, the model generates entities that are semantically unrelated to the sentence content or simply fabricated. For example, in the sentence "*Tapi kita juga punya kemampuan", tandas mantan Menteri Pemberdayaan BUMN ini.*" (But we also have capabilities", asserted the former Minister of State-Owned Enterprises), Merak hallucinates *mantan menteri pemberdayaan bumno*, a distorted phrase not grounded in the input and unrelated to the gold entity *BUMN*. In another case, for the sentence mentioning government aid in *Halmahera Utara*, the model predicts a list of invented or irrelevant programs such as *PKH*, *ASLU*, and *Beras Rastra*—none of which appear in the original sentence. These errors reveal a pattern of over-generation and poor contextual grounding, where the model fails to anchor its predictions in the actual input.

These results demonstrate that each model exhibits a distinct primary weakness when applied in zero-shot settings on general-domain text. GPT-4 most often returns incomplete mention spans, LLaMA-3 frequently includes extraneous titles within entity boundaries, Komodo struggles with placeholder substitution due to template leakage, and Merak is prone to hallucinating entities that are semantically unrelated to the input. Understanding these model-specific failure profiles is essential for improving end-to-end EL performance, particularly in LrLs like Indonesian.

6.4. Comparative analysis: LLMs vs. Traditional end-to-end EL systems

Despite the growing capabilities of LLMs, our experiments show that traditional end-to-end EL systems such as Babelfy, DBpedia Spotlight, OpenTapioca, and WAT still outperform LLMs across both general and specific domains. For instance, in the general domain (Table 3), WAT achieves the highest F1-score of 0.582, followed by OpenTapioca (0.542), Babelfy (0.492), and DBpedia Spotlight (0.468). In contrast, the best-performing fine-tuned LLM — GPT-3.5 — achieves an F1-score of only 0.373. In the specific domain (Table 4), WAT and DBpedia Spotlight again dominate with F1-scores of 0.761 and 0.750, respectively, while GPT-3.5, the strongest fine-tuned LLM, achieves 0.611. Other LLMs, including Indonesian-trained Komodo and Merak, and multilingual LLaMA-3, consistently fall below the 0.5 mark in both domains.

This persistent gap can be attributed to several key factors:

- Lack of Explicit Knowledge Grounding: Traditional systems explicitly rely on structured KB access through deterministic components—mention detection, candidate generation, and coherence-based entity disambiguation. This enables them to directly retrieve and rank candidate entities based on symbolic signals. In contrast, LLMs rely on latent knowledge encoded during pretraining and lack transparent KB access, which limits their ability to resolve ambiguous mentions unless explicitly guided via prompt engineering or external retrieval modules [31].
- Mention Boundary Detection Consistency: Traditional end-to-end EL systems typically rely on rule-based or statistical NER modules that are fine-tuned to accurately segment entity spans. As a result, they rarely encounter span boundary issues. In contrast, LLMs often struggle to delineate entity boundaries in Indonesian sentences, as observed in category C and non-processed sentence errors. For instance, LLaMA-3 frequently includes salutations (e.g., *menteri* or *mantan presiden*) as part of the entity span—components that traditional systems typically exclude through gazetteer-based heuristics or supervised training on curated corpora.
- Template Leakage and Format Robustness: Errors such as template leakage — observed in Komodo's outputs with tokens like *entity1* and *link1* — are not present in traditional systems, which follow strict formatting rules throughout their pipelines. These systems enforce structured output schemas by design, whereas LLMs may revert to training-time templates when the prompt is not fully grounded in semantic context, especially in fine-tuned small-scale models [92].
- Domain and Language Adaptation: While traditional systems generally perform consistently across domains due to broad KB coverage and modular pipeline tuning, LLMs require fine-tuning to achieve reasonable performance, and still fall short in domain transfer. For example, although Komodo and Merak are pretrained on Indonesian, they struggle in the general domain due to insufficient exposure to diverse linguistic forms, whereas DBpedia Spotlight performs competitively in both domains without task-specific adaptation.
- Disambiguation Accuracy: Even when LLMs successfully detect all entity mentions — as in category B — they often fail to assign the correct Wikidata entity IDs, particularly in domain-specific texts. This highlights a weakness in the disambiguation component of the end-to-end EL pipeline, where traditional end-to-end EL systems outperform LLMs due to their use of structured signals such as entity descriptions, popularity scores, and coherence-based reranking [5,89,93].

ELEVATE-ID reveals that although fine-tuned LLMs demonstrate promising improvements over their zero-shot counterparts, they still lag significantly behind traditional end-to-end EL systems in terms of F1 performance, mention-link coherence, and cross-domain robustness. These findings highlight the importance of developing hybrid approaches that combine the contextual adaptability of LLMs with the structured precision, reliability, and interpretability of traditional end-to-end EL architectures [21,86].

7. Practical implications

The findings from ELEVATE-ID demonstrate practical relevance for real-world applications of end-to-end EL in Indonesian and potentially other LrLs. For example, improved end-to-end EL performance in the specific domain has strong implications for religious information systems, where disambiguating scriptural and historical entities is essential. In the general domain, ELEVATE-ID can support digital journalism, public discourse monitoring, and personalized content delivery by enabling accurate linkage of ambiguous

mentions (e.g., "Hatta", "Anies") to KB entries. The framework also facilitates the development of local NLP tools — such as Indonesian-language virtual assistants, search engines, and question answering systems — by identifying end-to-end EL bottlenecks that LLMs still face, particularly in handling acronyms, incomplete mentions, and culturally grounded expressions.

8. Limitations

While ELEVATE-ID provides a comprehensive evaluation of LLM-based end-to-end EL in Indonesian, several limitations should be acknowledged. First, the study focuses exclusively on a single LrL (Indonesian), which may limit the generalizability of the findings to other LrLs with different linguistic characteristics. Second, while IndEL includes both general and specific domain data, its relatively small scale compared to high-resource benchmarks may restrict the diversity of contexts available during training and evaluation. Third, the evaluation is limited to Wikidata as the target KB, which offers rich textual descriptions, structured relations, and standardized identifiers [3]. These characteristics make Wikidata particularly suitable for LLM-based linking, especially since earlier models in the LLaMA family were trained on multilingual Wikipedia dumps [31], which embed Wikidata-aligned content. Although the training data for LLaMA-3 has not been disclosed, it may have retained similar exposure to such entity distributions. In contrast, other KBs such as DBpedia or YAGO provide sparser or less standardized entity representations [94], and adapting ELEVATE-ID to these resources may require alternative alignment or linking strategies. Fourth, our experiments evaluate zero-shot and fine-tuned settings but do not include advanced prompting strategies (e.g., few-shot or chain-of-thought), nor hybrid models that combine symbolic and neural methods. Future work could explore these directions to further enhance performance and robustness.

9. Conclusion and future works

This paper presents ELEVATE-ID, a framework for evaluating end-to-end EL in Indonesian — a low-resource language — using multilingual and Indonesian-pretrained LLMs. Leveraging the IndEL benchmark, we assess model performance under zero-shot and fine-tuned settings across general and specific domains. While GPT-4 and GPT-3.5 outperform others, our analysis reveals that persistent disambiguation and linking errors remain across all models. Indonesian LLMs show reasonable mention detection but suffer from hallucinations and format-related issues. Compared to traditional end-to-end EL systems, LLMs still underperform in F1-score and linking precision. The limited size of IndEL further underscores the need to develop more comprehensive datasets for robust evaluation. Future work may explore LLM-based data augmentation [95,96], mention paraphrasing [97], or back translation [98] to improve generalization and mitigate data scarcity in low-resource end-to-end EL tasks.

CRediT authorship contribution statement

Ria Hari Gusmita: Writing – review & editing, Writing – original draft, Validation, Resources, Methodology, Formal analysis, Data curation, Conceptualization. **Asep Fajar Firmansyah:** Writing – review & editing, Writing – original draft, Methodology, Data curation. **Hamada M. Zahera:** Writing – review & editing, Supervision. **Axel-Cyrille Ngonga Ngomo:** Supervision.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT in order to ensure the readability of the paper. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Ria Hari Gusmita reports administrative support, article publishing charges,equipment, drugs, or supplies, statistical analysis, and writing assistance were provided by Paderborn University. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We acknowledge the support of the German Federal Ministry of Research, Technology and Space (BMFTR) within the project KI-OWL under grant no. 01IS24057B, as well as the Mora Scholarship from the Ministry of Religious Affairs, Republic of Indonesia.

Data availability

I have shared the data on the paper.

References

[1] S. Tedeschi, S. Conia, F. Cecconi, R. Navigli, Named entity recognition for entity linking: What works and what's next, in: M.-F. Moens, X. Huang, L. Specia, S.W.-t. Yih (Eds.), Findings of the Association for Computational Linguistics, EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 2584–2596, http://dx.doi.org/10.18653/v1/2021.findings-emnlp.220, URL https://aclanthology.org/2021.findings-emnlp.220/.

[2] W. Shen, Y. Li, Y. Liu, J. Han, J. Wang, X. Yuan, Entity linking meets deep learning: Techniques and solutions, IEEE Trans. Knowl. Data Eng. 35 (3) (2023) 2556–2578, http://dx.doi.org/10.1109/TKDE.2021.3117715.

[3] D. Vrandečić, M. Krötzsch, Wikidata: a free collaborative knowledgebase, Commun. ACM 57 (10) (2014) 78–85, http://dx.doi.org/10.1145/2629489.

[4] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ives, DBpedia: a nucleus for a web of open data, in: Proceedings of the 6th International the Semantic Web and 2nd Asian Conference on Asian Semantic Web Conference, in: ISWC'07/ASWC'07, Springer-Verlag, Berlin, Heidelberg, 2007, pp. 722–735.

[5] N. Kolitsas, O.-E. Ganea, T. Hofmann, End-to-end neural entity linking, in: A. Korhonen, I. Titov (Eds.), Proceedings of the 22nd Conference on Computational Natural Language Learning, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 519–529, http://dx.doi.org/10.18653/v1/K18-1050, URL https://aclanthology.org/K18-1050/.

[6] S. Broscheit, Investigating entity knowledge in BERT with simple neural end-to-end entity linking, in: M. Bansal, A. Villavicencio (Eds.), Proceedings of the 23rd Conference on Computational Natural Language Learning, (CoNLL), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 677–685, http://dx.doi.org/10.18653/v1/K19-1063, URL https://aclanthology.org/K19-1063/.

[7] R.H. Gusmita, M.F.A. Abshar, D. Moussallem, A.-C.N. Ngomo, IndEL: Indonesian entity linking benchmark dataset for general and specific domains, in: A. Rapp, L. Di Caro, F. Meziane, V. Sugumaran (Eds.), Natural Language Processing and Information Systems, Springer Nature Switzerland, Cham, 2024, pp. 500–513.

[8] C. Peng, F. Xia, M. Naseriparsa, F. Osborne, Knowledge graphs: Opportunities and challenges, Artif. Intell. Rev. 56 (11) (2023) 13071–13102, http://dx.doi.org/10.1007/s10462-023-10465-9.

[9] Z. Hu, Y. Xu, W. Yu, S. Wang, Z. Yang, C. Zhu, K.-W. Chang, Y. Sun, Empowering language models with knowledge graph reasoning for open-domain question answering, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 9562–9581, http://dx.doi.org/10.18653/v1/2022.emnlp-main.650, URL https://aclanthology.org/2022.emnlp-main.650/.

[10] F. Fatahi Bayat, N. Bhutani, H. Jagadish, CompactIE: Compact facts in open information extraction, in: M. Carpuat, M.-C. de Marneffe, I.V. Meza Ruiz (Eds.), Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States, 2022, pp. 900–910, http://dx.doi.org/10.18653/v1/2022.naacl-main.65, URL https://aclanthology.org/2022.naacl-main.65/.

[11] S. Cahyawijaya, H. Lovenia, A.F. Aji, G. Winata, B. Wilie, F. Koto, R. Mahendra, C. Wibisono, A. Romadhony, K. Vincentio, J. Santoso, D. Moeljadi, C. Wirawan, F. Hudi, M.S. Wicaksono, I. Parmonangan, I. Alfina, I.F. Putra, S. Rahmadani, Y. Oenang, A. Septiandri, J. Jaya, K. Dhole, A. Suryani, R.A. Putri, D. Su, K. Stevens, M.N. Nityasya, M. Adilazuarda, R. Hadiwijaya, R. Diandaru, T. Yu, V. Ghifari, W. Dai, Y. Xu, D. Damapuspita, H. Wibowo, C. Tho, I. Karo Karo, T. Fatyanosa, Z. Ji, G. Neubig, T. Baldwin, S. Ruder, P. Fung, H. Sujaini, S. Sakti, A. Purwarianti, NusaCrowd: Open source initiative for Indonesian NLP resources, in: Findings of the Association for Computational Linguistics, ACL 2023, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 13745–13818, http://dx.doi.org/10.18653/v1/2023.findings-acl.868.

[12] B. Wilie, K. Vincentio, G.I. Winata, S. Cahyawijaya, X. Li, Z.Y. Lim, S. Soleman, R. Mahendra, P. Fung, S. Bahar, A. Purwarianti, IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding, in: K.-F. Wong, K. Knight, H. Wu (Eds.), Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, Association for Computational Linguistics, Suzhou, China, 2020, pp. 843–857, URL https://aclanthology.org/2020.aacl-main.85.

[13] F. Koto, A. Rahimi, J.H. Lau, T. Baldwin, IndoLEM and indoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP, in: D. Scott, N. Bel, C. Zong (Eds.), Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 757–770, http://dx.doi.org/10.18653/v1/2020.coling-main.66.

[14] R.H. Gusmita, A.F. Firmansyah, D. Moussallem, A.-C. Ngonga Ngomo, IndQNER: Named entity recognition benchmark dataset from the Indonesian translation of the quran, in: E. Métais, F. Meziane, V. Sugumaran, W. Manning, S. Reiff-Marganiec (Eds.), Natural Language Processing and Information Systems, Springer Nature Switzerland, Cham, 2023, pp. 170–185.

[15] Z. Zhang, Y. Zhao, H. Gao, M. Hu, LinkNER: Linking local named entity recognition models to large language models using uncertainty, in: Proceedings of the ACM Web Conference 2024, WWW '24, Association for Computing Machinery, New York, NY, USA, 2024, pp. 4047–4058, http://dx.doi.org/10.1145/3589334.3645414.

[16] F. Yan, P. Yu, X. Chen, LTNER: Large language model tagging for named entity recognition with contextualized entity marking, 2024, arXiv:2404.05624, URL https://arxiv.org/abs/2404.05624.

[17] N. Kholodna, S. Julka, M. Khodadadi, M.N. Gumus, M. Granitzer, LLMs in the loop: Leveraging large language model annotations for active learning in low-resource languages, in: Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track: European Conference, ECML PKDD 2024, Vilnius, Lithuania, September 9–13, 2024, Proceedings, Part X, Springer-Verlag, Berlin, Heidelberg, 2024, pp. 397–412, http://dx.doi.org/10.1007/978-3-031-70381-2_25.

[18] Y. Heng, C. Deng, Y. Li, Y. Yu, Y. Li, R. Zhang, C. Zhang, ProgGen: Generating named entity recognition datasets step-by-step with self-reflexive large language models, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Findings of the Association for Computational Linguistics, ACL 2024, Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 15992–16030, http://dx.doi.org/10.18653/v1/2024.findings-acl.947.

[19] H. Kang, H. Seo, J. Jung, S. Jung, D.-S. Chang, R. Chung, Guidance-based prompt data augmentation in specialized domains for named entity recognition, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 665–672, http://dx.doi.org/10.18653/v1/2024.acl-short.61.

[20] P. Elchafei, A. Fashwan, Arabic NER evaluation: Pre-trained models via contrastive learning vs. LLM few-shot prompting, Procedia Comput. Sci. 244 (2024) 229–237, http://dx.doi.org/10.1016/j.procs.2024.10.196, URL https://www.sciencedirect.com/science/article/pii/S1877050924029971, 6th International Conference on AI in Computational Linguistics.

[21] Y. Ding, A. Poudel, Q. Zeng, T. Weninger, B. Veeramani, S. Bhattacharya, EntGPT: Linking generative large language models with knowledge bases, 2024, arXiv:2402.06738.

[22] R. Vīksna, I. Skadiņa, D. Deksne, R. Rozis, Large language models for multilingual slavic named entity linking, in: Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023, (SlavicNLP 2023), 2023, pp. 172–178.

[23] L. Xue, D. Zhang, Y. Dong, J. Tang, AutoRE: Document-level relation extraction with large language models, in: Y. Cao, Y. Feng, D. Xiong (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 211–220, http://dx.doi.org/10.18653/v1/2024.acl-demos.20.

[24] Y. Hu, C. Chen, C. Qin, Q. Zhu, E. Chng, R. Li, Listen again and choose the right answer: A new paradigm for automatic speech recognition with large language models, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Findings of the Association for Computational Linguistics, ACL 2024, Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 666–679, http://dx.doi.org/10.18653/v1/2024.findings-acl.37.

[25] X. Huang, Z. Zhang, X. Geng, Y. Du, J. Chen, S. Huang, Lost in the source language: How large language models evaluate the quality of machine translation, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Findings of the Association for Computational Linguistics, ACL 2024, Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 3546–3562, http://dx.doi.org/10.18653/v1/2024.findings-acl.211.

[26] F. Koto, N. Aisyah, H. Li, T. Baldwin, Large language models only pass primary school exams in Indonesia: A comprehensive test on indoMMLU, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 12359–12374, http://dx.doi.org/10.18653/v1/2023.emnlp-main.760.

[27] H. Nomoto, Issues surrounding the use of chatGPT in similar languages: The case of Malay and Indonesian, in: J.C. Park, Y. Arase, B. Hu, W. Lu, D. Wijaya, A. Purwarianti, A.A. Krisnadhi (Eds.), Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Nusa Dua, Bali, 2023, pp. 76–82, http://dx.doi.org/10.18653/v1/2023.ijcnlp-short.9.

[28] S. Cahyawijaya, H. Lovenia, F. Koto, R.A. Putri, E. Dave, J. Lee, N. Shadieq, W. Cenggoro, S.M. Akbar, M.I. Mahendra, D.A. Putri, B. Wilie, G.I. Winata, A.F. Aji, A. Purwarianti, P. Fung, Cendol: Open instruction-tuned generative large language models for Indonesian languages, 2024, arXiv:2404.06138.

[29] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P.F. Christiano, J. Leike, R. Lowe, Training language models to follow instructions with human feedback, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), in: Advances in Neural Information Processing Systems, vol. 35, Curran Associates, Inc., 2022, pp. 27730–27744, URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

[30] OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F.L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, R. Avila, I. Babuschkin, S. Balaji, V. Balcom, P. Baltescu, H. Bao, M. Bavarian, J. Belgum, I. Bello, J. Berdine, G. Bernadett-Shapiro, C. Berner, L. Bogdonoff, O. Boiko, M. Boyd, A.-L. Brakman, G. Brockman, T. Brooks, M. Brundage, K. Button, T. Cai, R. Campbell, A. Cann, B. Carey, C. Carlson, R. Carmichael, B. Chan, C. Chang, F. Chantzis, D. Chen, S. Chen, R. Chen, J. Chen, M. Chen, B. Chess, C. Cho, C. Chu, H.W. Chung, D. Cummings, J. Currier, Y. Dai, C. Decareaux, T. Degry, N. Deutsch, D. Deville, A. Dhar, D. Dohan, S. Dowling, S. Dunning, A. Ecoffet, A. Eleti, T. Eloundou, D. Farhi, L. Fedus, N. Felix, S.P. Fishman, J. Forte, I. Fulford, L. Gao, E. Georges, C. Gibson, V. Goel, T. Gogineni, G. Goh, R. Gontijo-Lopes, J. Gordon, M. Grafstein, S. Gray, R. Greene, J. Gross, S.S. Gu, Y. Guo, C. Hallacy, J. Han, J. Harris, Y. He, M. Heaton, J. Heidecke, C. Hesse, A. Hickey, W. Hickey, P. Hoeschele, B. Houghton, K. Hsu, S. Hu, X. Hu, J. Huizinga, S. Jain, S. Jain, J. Jang, A. Jiang, R. Jiang, H. Jin, D. Jin, S. Jomoto, B. Jonn, H. Jun, T. Kaftan, L. ukasz Kaiser, A. Kamali, I. Kanitscheider, N.S. Keskar, T. Khan, L. Kilpatrick, J.W. Kim, C. Kim, Y. Kim, J.H. Kirchner, J. Kiros, M. Knight, D. Kokotajlo, L. ukasz Kondraciuk, A. Kondrich, A. Konstantinidis, K. Kosic, G. Krueger, V. Kuo, M. Lampe, I. Lan, T. Lee, J. Leike, J. Leung, D. Levy, C.M. Li, R. Lim, M. Lin, S. Lin, M. Litwin, T. Lopez, R. Lowe, P. Lue, A. Makanju, K. Malfacini, S. Manning, T. Markov, Y. Markovski, B. Martin, K. Mayer, A. Mayne, B. McGrew, S.M. McKinney, C. McLeavey, P. McMillan, J. McNeil, D. Medina, A. Mehta, J. Menick, L. Metz, A. Mishchenko, P. Mishkin, V. Monaco, E. Morikawa, D. Mossing, T. Mu, M. Murati, O. Murk, D. Mély, A. Nair, R. Nakano, R. Nayak, A. Neelakantan, R. Ngo, H. Noh, L. Ouyang, C. O'Keefe, J. Pachocki, A. Paino, J. Palermo, A. Pantuliano, G. Parascandolo, J. Parish, E. Parparita, A. Passos, M. Pavlov, A. Peng, A. Perelman, F.d.B. Peres, M. Petrov, H.P.d. Pinto, Michael, Pokorny, M. Pokrass, V.H. Pong, T. Powell, A. Power, B. Power, E. Proehl, R. Puri, A. Radford, J. Rae, A. Ramesh, C. Raymond, F. Real, K. Rimbach, C. Ross, B. Rotsted, H. Roussez, N. Ryder, M. Saltarelli, T. Sanders, S. Santurkar, G. Sastry, H. Schmidt, D. Schnurr, J. Schulman, D. Selsam, K. Sheppard, T. Sherbakov, J. Shieh, S. Shoker, P. Shyam, S. Sidor, E. Sigler, M. Simens, J. Sitkin, K. Slama, I. Sohl, B. Sokolowsky, Y. Song, N. Staudacher, F.P. Such, N. Summers, I. Sutskever, J. Tang, N. Tezak, M.B. Thompson, P. Tillet, A. Tootoonchian, E. Tseng, P. Tuggle, N. Turley, J. Tworek, J.F.C. Uribe, A. Vallone, A. Vijayvergiya, C. Voss, C. Wainwright, J.J. Wang, A. Wang, B. Wang, J. Ward, J. Wei, C. Weinmann, A. Welihinda, P. Welinder, J. Weng, L. Weng, M. Wiethoff, D. Willner, C. Winter, S. Wolrich, H. Wong, L. Workman, S. Wu, J. Wu, M. Wu, K. Xiao, T. Xu, S. Yoo, K. Yu, Q. Yuan, W. Zaremba, R. Zellers, C. Zhang, M. Zhang, S. Zhao, T. Zheng, J. Zhuang, W. Zhuk, B. Zoph, GPT-4 technical report, 2024, arXiv:2303.08774.

[31] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, LLaMA: Open and efficient foundation language models., 2023, CoRR, abs/2302.13971, URL http://dblp.uni-trier.de/db/journals/corr/corr2302.html#abs-2302-13971.

[32] L. Owen, V. Tripathi, A. Kumar, B. Ahmed, Komodo: A linguistic expedition into Indonesia's regional languages, 2024, arXiv:2403.09362, URL https://arxiv.org/abs/2403.09362.

[33] J. Li, A. Sun, J. Han, C. Li, A survey on deep learning for named entity recognition, IEEE Trans. Knowl. Data Eng. 34 (1) (2022) 50–70, http://dx.doi.org/10.1109/TKDE.2020.2981314.

[34] D. Hanisch, K. Fundel, H.-T. Mevissen, R. Zimmer, J. Fluck, ProMiner: rule-based protein and gene entity recognition, BMC Bioinformatics (2005) URL http://nbn-resolving.de/urn/resolver.pl?urn=nbn:de:bvb:19-epub-23765-3.

[35] J.-H. Kim, P.C. Woodland, A rule-based named entity recognition system for speech input, (ICSLP 2000), in: 6th International Conference on Spoken Language Processing, vol. 1, 2000, pp. 528–531, http://dx.doi.org/10.21437/ICSLP.2000-131.

[36] J. Li, D. Zhou, Y. Duan, X. Li, H. Yao, A clustering-oriented method for open-domain named entity recognition, in: CCGrid, 2024, pp. 189–195, http://dx.doi.org/10.1109/CCGrid59990.2024.00030.

[37] V. G., V. Kanjirangat, D. Gupta, AGRONER: An unsupervised agriculture named entity recognition using weighted distributional semantic model, Expert Syst. Appl. 229 (2023) 120440, http://dx.doi.org/10.1016/j.eswa.2023.120440, URL https://www.sciencedirect.com/science/article/pii/S0957417423009429.

[38] Y. Li, L. Song, C. Zhang, Sparse conditional hidden Markov model for weakly supervised named entity recognition, in: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '22, Association for Computing Machinery, New York, NY, USA, 2022, pp. 978–988, http://dx.doi.org/10.1145/3534678.3539247.

[39] S. Singh, U.S. Tiwary, ACRF: Aggregated conditional random field for out of vocab (OOV) token representation for hindi NER, IEEE Access 12 (2024) 22707–22717, http://dx.doi.org/10.1109/ACCESS.2024.3362645.

[40] O. Bender, F.J. Och, H. Ney, Maximum entropy models for named entity recognition, in: Proceedings of the Seventh Conference on Natural Language Learning At HLT-NAACL 2003, 2003, pp. 148–151, URL https://aclanthology.org/W03-0420/.

[41] R. Sasano, S. Kurohashi, Japanese named entity recognition using structural natural language processing, in: Proceedings of the Third International Joint Conference on Natural Language Processing, Vol. II, 2008, URL https://aclanthology.org/I08-2080/.

[42] W. Shen, J. Wang, J. Han, Entity linking with a knowledge base: Issues, techniques, and solutions, IEEE Trans. Knowl. Data Eng. 27 (2) (2015) 443–460, http://dx.doi.org/10.1109/TKDE.2014.2327028.

[43] G. Wu, Y. He, X. Hu, Entity linking: An issue to extract corresponding entity with knowledge base, IEEE Access 6 (2018) 6220–6231, http://dx.doi.org/10.1109/ACCESS.2017.2787787.

[44] T. Lai, H. Ji, C. Zhai, Improving candidate retrieval with entity profile generation for wikidata entity linking, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Findings of the Association for Computational Linguistics, ACL 2022, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 3696–3711, http://dx.doi.org/10.18653/v1/2022.findings-acl.292, URL https://aclanthology.org/2022.findings-acl.292/.

[45] R. Harige, P. Buitelaar, Generating a large-scale entity linking dictionary from wikipedia link structure and article text, in: N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Tenth International Conference on Language Resources and Evaluation, LREC'16, European Language Resources Association (ELRA), Portorož, Slovenia, 2016, pp. 2431–2434, URL https://aclanthology.org/L16-1385/.

[46] L. Hebert, R. Makki, S. Mishra, H. Saghir, A. Kamath, Y. Merhav, Robust candidate generation for entity linking on short social media texts, in: Proceedings of the Eighth Workshop on Noisy User-Generated Text, (W-NUT 2022), Association for Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 83–89, URL https://aclanthology.org/2022.wnut-1.8/.

[47] P. Le, I. Titov, Distant learning for entity linking with automatic noise detection, in: A. Korhonen, D. Traum, L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 4081–4090, http://dx.doi.org/10.18653/v1/P19-1400, URL https://aclanthology.org/P19-1400/.

[48] X. Fu, W. Shi, X. Yu, Z. Zhao, D. Roth, Design challenges in low-resource cross-lingual entity linking, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP, Association for Computational Linguistics, Online, 2020, pp. 6418–6432, http://dx.doi.org/10.18653/v1/2020.emnlp-main.521, URL https://aclanthology.org/2020.emnlp-main.521/.

[49] M. Dredze, P. McNamee, D. Rao, A. Gerber, T. Finin, Entity disambiguation for knowledge base population, in: C.-R. Huang, D. Jurafsky (Eds.), Proceedings of the 23rd International Conference on Computational Linguistics, (Coling 2010), Coling 2010 Organizing Committee, Beijing, China, 2010, pp. 277–285, URL https://aclanthology.org/C10-1032/.

[50] S. Li, Y. Zhang, Improving entity linking by combining semantic entity embeddings and cross-attention encoder, J. Intell. Fuzzy Syst. 46 (1) (2024) 2899–2910, http://dx.doi.org/10.3233/JIFS-233124.

[51] A. Pilz, G. Paaß, From names to entities using thematic context distance, in: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, CIKM '11, Association for Computing Machinery, New York, NY, USA, 2011, pp. 857–866, http://dx.doi.org/10.1145/2063576.2063700.

[52] J.-Y. Jiang, W.-C. Chang, J. Zhang, C.-J. Hsieh, H.-F. Yu, Entity disambiguation with extreme multi-label ranking, in: Proceedings of the ACM Web Conference 2024, WWW '24, Association for Computing Machinery, New York, NY, USA, 2024, pp. 4172–4180, http://dx.doi.org/10.1145/3589334.3645498.

[53] W. Shen, J. Wang, P. Luo, M. Wang, LIEGE: link entities in web lists with knowledge base, in: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12, Association for Computing Machinery, New York, NY, USA, 2012, pp. 1424–1432, http://dx.doi.org/10.1145/2339530.2339753.

[54] O.-E. Ganea, M. Ganea, A. Lucchi, C. Eickhoff, T. Hofmann, Probabilistic bag-of-hyperlinks model for entity linking, in: Proceedings of the 25th International Conference on World Wide Web, WWW '16, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 2016, pp. 927–938, http://dx.doi.org/10.1145/2872427.2882988.

[55] X. Han, L. Sun, A generative entity-mention model for linking entities with knowledge base, in: D. Lin, Y. Matsumoto, R. Mihalcea (Eds.), Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Portland, Oregon, USA, 2011, pp. 945–954, URL https://aclanthology.org/P11-1095/.

[56] P. Oza, S. Chatterjee, L. Dietz, Neural entity context models, in: Proceedings of the the 12th International Joint Conference on Knowledge Graphs, 2023, 12th International Joint Conference on Knowledge Graphs, IJCKG 2023 ; Conference date: 08-12-2023 Through 09-12-2023, URL https://ijckg2023.knowledge-graph.jp/.

[57] W. Ding, V.K. Chaudhri, N. Chittar, K. Konakanchi, JEL: Applying end-to-end neural entity linking in jpmorgan chase, 2024, ArXiv.

[58] X. Han, L. Sun, J. Zhao, Collective entity linking in web text: a graph-based method, in: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '11, Association for Computing Machinery, New York, NY, USA, 2011, pp. 765–774, http://dx.doi.org/10.1145/2009916.2010019.

[59] M. Asgari-Bidhendi, B. Janfada, A. Havangi, S.A. Hossayni, B. Minaei-Bidgoli, An unsupervised language-independent entity disambiguation method and its evaluation on the english and Persian languages, 2021, arXiv:2102.00395, URL https://arxiv.org/abs/2102.00395.

[60] Z.-B. Zhang, Z.-M. Zhong, P.-P. Yuan, H. Jin, Improving entity linking in Chinese domain by sense embedding based on graph clustering, J. Comput. Sci. Tech. 38 (1) (2023) 196–210, http://dx.doi.org/10.1007/s11390-023-2835-4.

[61] Y. Chen, Z. Xu, B. Hu, M. Zhang, Revisiting sparse retrieval for few-shot entity linking, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 12801–12806, http://dx.doi.org/10.18653/v1/2023.emnlp-main.789, URL https://aclanthology.org/2023.emnlp-main.789/.

[62] S. Gottipati, J. Jiang, Linking entities to a knowledge base with query expansion, in: R. Barzilay, M. Johnson (Eds.), Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Edinburgh, Scotland, UK., 2011, pp. 804–813, URL https://aclanthology.org/D11-1074/.

[63] B. Jia, C. Wang, H. Zhao, L. Shi, An entity linking algorithm derived from graph convolutional network and contextualized semantic relevance, Symmetry 14 (10) (2022) http://dx.doi.org/10.3390/sym14102060, URL https://www.mdpi.com/2073-8994/14/10/2060.

[64] T. Lin, Mausam, O. Etzioni, Entity linking at web scale, in: J. Fan, R. Hoffman, A. Kalyanpur, S. Riedel, F. Suchanek, P.P. Talukdar (Eds.), Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-Scale Knowledge Extraction, AKBC-WEKEX), Association for Computational Linguistics, Montréal, Canada, 2012, pp. 84–88, URL https://aclanthology.org/W12-3016/.

[65] Z. Chen, Y. Wu, Y. Feng, D. Zhao, Integrating manifold knowledge for global entity linking with heterogeneous graphs, Data Intell. 4 (1) (2022) 20–40, http://dx.doi.org/10.1162/dint_a_00116, arXiv:https://direct.mit.edu/dint/article-pdf/4/1/20/1985039/dint_a_00116.pdf.

[66] W. Shen, J. Wang, P. Luo, M. Wang, Linking named entities in tweets with knowledge base via user interest modeling, in: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '13, Association for Computing Machinery, New York, NY, USA, 2013, pp. 68–76, http://dx.doi.org/10.1145/2487575.2487686.

[67] K. Zaporojets, J. Deleu, Y. Jiang, T. Demeester, C. Develder, Towards consistent document-level entity linking: Joint models for entity linking and coreference resolution, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 778–784, http://dx.doi.org/10.18653/v1/2022.acl-short.88, URL https://aclanthology.org/2022.acl-short.88/.

[68] Z. Dong, M. Wang, S. deng, L. Dai, J. Li, X. Liu, R. Nong, Cross-document contextual coreference resolution in knowledge graphs, 2025, arXiv:2504.05767, URL https://arxiv.org/abs/2504.05767.

[69] Y. Guo, B. Qin, T. Liu, S. Li, Microblog entity linking by leveraging extra posts, in: D. Yarowsky, T. Baldwin, A. Korhonen, K. Livescu, S. Bethard (Eds.), Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Seattle, Washington, USA, 2013, pp. 863–868, URL https://aclanthology.org/D13-1085/.

[70] F. Zhu, J. Yu, H. Jin, L. Hou, J. Li, Z. Sui, Learn to not link: Exploring NIL prediction in entity linking, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Findings of the Association for Computational Linguistics, ACL 2023, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 10846–10860, http://dx.doi.org/10.18653/v1/2023.findings-acl.690, URL https://aclanthology.org/2023.findings-acl.690/.

[71] L. Chen, G. Varoquaux, F. Suchanek, A lightweight neural model for biomedical entity linking, in: AAAI-21 Technical Tracks 14, in: AAAI-21 Technical Tracks 14, vol. 35, (14) Association for the Advancement of Artificial Intelligence, Palo Alto (virtual), United States, 2021, pp. 12657–12665, URL https://hal.science/hal-03086044.

[72] R. Pozzi, R. Rubini, C. Bernasconi, M. Palmonari, Named entity recognition and linking for entity extraction from Italian civil judgements, in: R. Basili, D. Lembo, C. Limongelli, A. Orlandini (Eds.), AIxIA 2023 – Advances in Artificial Intelligence, Springer Nature Switzerland, Cham, 2023, pp. 187–201.

[73] A. Olieman, H. Azarbonyad, M. Dehghani, J. Kamps, M. Marx, Entity linking by focusing dbpedia candidate entities, in: Proceedings of the First International Workshop on Entity Recognition & Disambiguation, ERD '14, Association for Computing Machinery, New York, NY, USA, 2014, pp. 13–24, http://dx.doi.org/10.1145/2633211.2634353.

[74] Z. Zheng, F. Li, M. Huang, X. Zhu, Learning to link entities with knowledge base, in: R. Kaplan, J. Burstein, M. Harper, G. Penn (Eds.), Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Los Angeles, California, 2010, pp. 483–491, URL https://aclanthology.org/N10-1072/.

[75] L. Ratinov, D. Roth, D. Downey, M. Anderson, Local and global algorithms for disambiguation to wikipedia, in: D. Lin, Y. Matsumoto, R. Mihalcea (Eds.), Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Portland, Oregon, USA, 2011, pp. 1375–1384, URL https://aclanthology.org/P11-1138/.

[76] N. Heist, H. Paulheim, NASTyLinker: NIL-aware scalable transformer-based entity linker, in: C. Pesquita, E. Jimenez-Ruiz, J. McCusker, D. Faria, M. Dragoni, A. Dimou, R. Troncy, S. Hertling (Eds.), The Semantic Web, Springer Nature Switzerland, Cham, 2023, pp. 174–191.

[77] P. McNamee, HLTCOE efforts in entity linking at TAC KBP 2010, in: Proceedings of the Third Text Analysis Conference, TAC 2010, Gaithersburg, Maryland, USA, November 15-16, 2010, NIST, 2010, URL https://tac.nist.gov/publications/2010/participant.papers/hltcoe.proceedings.pdf.

[78] P.N. Mendes, M. Jakob, A. García-Silva, C. Bizer, DBpedia Spotlight: Shedding light on the web of documents, in: Proceedings of the 7th International Conference on Semantic Systems, in: I-Semantics '11, Association for Computing Machinery, New York, NY, USA, 2011, pp. 1–8, http://dx.doi.org/10.1145/2063518.2063519.

[79] P. Ferragina, U. Scaiella, TAGME: on-the-fly annotation of short text fragments (by wikipedia entities), in: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, CIKM '10, Association for Computing Machinery, New York, NY, USA, 2010, pp. 1625–1628, http://dx.doi.org/10.1145/1871437.1871689.

[80] F. Piccinno, P. Ferragina, From TagME to WAT: A new entity annotator, in: Proceedings of the First International Workshop on Entity Recognition & Disambiguation, ERD '14, Association for Computing Machinery, New York, NY, USA, 2014, pp. 55–62, http://dx.doi.org/10.1145/2633211.2634350.

[81] R. Verborgh, M. Röder, R. Usbeck, A.-C. Ngonga Ngomo, GERBIL – benchmarking named entity recognition and linking consistently, Semant. Web 9 (5) (2018) 605–625, http://dx.doi.org/10.3233/SW-170286.

[82] T. Ayoola, S. Tyagi, J. Fisher, C. Christodoulopoulos, A. Pierleoni, ReFinED: An efficient zero-shot-capable approach to end-to-end entity linking, in: A. Loukina, R. Gangadharaiah, B. Min (Eds.), Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track, Association for Computational Linguistics, Hybrid: Seattle, Washington + Online, 2022, pp. 209–220, http://dx.doi.org/10.18653/v1/2022.naacl-industry.24, URL https://aclanthology.org/2022.naacl-industry.24/.

[83] M.T.R. Laskar, C. Chen, A. Martsinovich, J. Johnston, X.-Y. Fu, S.B. Tn, S. Corston-Oliver, BLINK with elasticsearch for efficient entity linking in business conversations, in: A. Loukina, R. Gangadharaiah, B. Min (Eds.), Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track, Association for Computational Linguistics, Hybrid: Seattle, Washington + Online, 2022, pp. 344–352, http://dx.doi.org/10.18653/v1/2022.naacl-industry.38, URL https://aclanthology.org/2022.naacl-industry.38/.

[84] S. Mishra, A. Saini, R. Makki, S. Mehta, A. Haghighi, A. Mollahosseini, TweetNERD - end to end entity linking benchmark for tweets, in: Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22, Curran Associates Inc., Red Hook, NY, USA, 2022.

[85] K. Noullet, R. Mix, M. Färber, KORE 50ˆDYWC: An evaluation data set for entity linking based on DBpedia, YAGO, wikidata, and crunchbase, in: Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 2389–2395, URL https://aclanthology.org/2020.lrec-1.291.

[86] A. Xin, Y. Qi, Z. Yao, F. Zhu, K. Zeng, X. Bin, L. Hou, J. Li, LLMAEL: Large language models are good context augmenters for entity linking, 2024, arXiv:2407.04020, URL https://arxiv.org/abs/2407.04020.

[87] D. Vollmers, H. Zahera, D. Moussallem, A.-C. Ngonga Ngomo, Contextual augmentation for entity linking using large language models, in: O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B.D. Eugenio, S. Schockaert (Eds.), Proceedings of the 31st International Conference on Computational Linguistics, Association for Computational Linguistics, Abu Dhabi, UAE, 2025, pp. 8535–8545, URL https://aclanthology.org/2025.coling-main.570/.

[88] A. Hotho, J.L. Martinez-Rodriguez, A. Hogan, I. Lopez-Arevalo, Information extraction meets the semantic web: A survey, Semant. Web 11 (2) (2020) 255–335, http://dx.doi.org/10.3233/SW-180333.

[89] A. Moro, A. Raganato, R. Navigli, Entity linking meets word sense disambiguation: a unified approach, in: D. Lin, M. Collins, L. Lee (Eds.), Trans. Assoc. Comput. Linguist. 2 (2014) 231–244, http://dx.doi.org/10.1162/tacl_a_00179.

[90] A. Delpeuch, OpenTapioca: Lightweight entity linking for Wikidata, 2019, ArXiv, abs/1904.09131, URL https://api.semanticscholar.org/CorpusID:125953443.

[91] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, X. Wu, Unifying large language models and knowledge graphs: A roadmap., IEEE Trans. Knowl. Data Eng. 36 (7) (2024) 3580–3599, URL http://dblp.uni-trier.de/db/journals/tkde/tkde36.html#PanLWCWW24.

[92] Z. Li, B. Peng, P. He, X. Yan, Evaluating the instruction-following robustness of large language models to prompt injection, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 557–568, http://dx.doi.org/10.18653/v1/2024.emnlp-main.33, URL https://aclanthology.org/2024.emnlp-main.33/.

[93] J.A. Botha, Z. Shan, D. Gillick, Entity linking in 100 languages, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP, Association for Computational Linguistics, Online, 2020, pp. 7833–7845, http://dx.doi.org/10.18653/v1/2020.emnlp-main.630, URL https://aclanthology.org/2020.emnlp-main.630/.

[94] T. Rebele, F. Suchanek, J. Hoffart, J. Biega, E. Kuzey, G. Weikum, YAGO: A multilingual knowledge base from Wikipedia, Wordnet, and Geonames, in: The Semantic Web – ISWC 2016: 15th International Semantic Web Conference, Kobe, Japan, October 17–21, 2016, Proceedings, Part II, Springer-Verlag, Berlin, Heidelberg, 2016, pp. 177–185, http://dx.doi.org/10.1007/978-3-319-46547-0_19.

[95] B. Ding, C. Qin, R. Zhao, T. Luo, X. Li, G. Chen, W. Xia, J. Hu, A.T. Luu, S. Joty, Data augmentation using LLMs: Data perspectives, learning paradigms and challenges, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Findings of the Association for Computational Linguistics, ACL 2024, Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 1679–1705, http://dx.doi.org/10.18653/v1/2024.findings-acl.97, URL https://aclanthology.org/2024.findings-acl.97/.

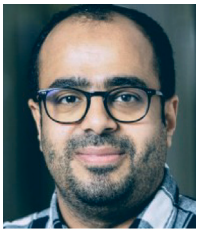
[96] C. Whitehouse, M. Choudhury, A.F. Aji, LLM-powered data augmentation for enhanced cross-lingual performance, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 671–686, http://dx.doi.org/10.18653/v1/2023.emnlp-main.44, URL https://aclanthology.org/2023.emnlp-main.44/.

[97] S. Sharma, A. Joshi, Y. Zhao, N. Mukhija, H. Bhathena, P. Singh, S. Santhanam, When and how to paraphrase for named entity recognition? in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 7052–7087, http://dx.doi.org/10.18653/v1/2023.acl-long.390, URL https://aclanthology.org/2023.acl-long.390/.

[98] Y. Li, X. Li, Y. Yang, R. Dong, A diverse data augmentation strategy for low-resource neural machine translation, Information 11 (5) (2020) http://dx.doi.org/10.3390/info11050255, URL https://www.mdpi.com/2078-2489/11/5/255.

Ria Hari Gusmita is a doctoral researcher at the Data Science Group (DICE), Heinz Nixdorf Institute, Paderborn University, Germany, and a lecturer at the State Islamic University Syarif Hidayatullah Jakarta, Indonesia. Her main research fields are natural language processing, knowledge graphs, and large language models, with a particular focus on low-resource languages such as Indonesian. She has developed frameworks and benchmark datasets that support question answering, named entity recognition, and entity linking for Indonesian. Her current research focuses on improving entity disambiguation and building knowledge graphs, with an emphasis on multilingual and domain-specific applications.

Asep Fajar Firmansyah is a doctoral researcher and research associate at the Data Science Group (DICE), Heinz Nixdorf Institute, Paderborn University, Germany, and a lecturer at the State Islamic University Syarif Hidayatullah Jakarta, Indonesia. His research centers on knowledge graph summarization, especially entity summarization. He has developed both extractive and abstractive summarization models and toolchains. His current work extends entity summarization from single-entity to multi-entity settings, aiming for coherent, relation-aware summaries across sets of entities. Beyond summarization, his research also addresses entity disambiguation and knowledge graph construction, with a focus on multilingual and domain-specific applications.



Dr. Hamada M. Zahera is a senior researcher and head of the NLP unit in the Data Science Group (DICE) at the Heinz Nixdorf Institute, Paderborn University, Germany. His research focuses on natural language processing (NLP), knowledge graphs, and large language models, with a strong emphasis on multilinguality and semantic representation. He has contributed to the development of several NLP frameworks and models. His current work explores multilingual representation learning, entity summarization, and question answering over knowledge graphs.

Prof. Dr. Axel-Cyrille Ngonga Ngomo is a full professor and head of the Data Science Group (DICE) at the Heinz Nixdorf Institute, Paderborn University, Germany. His research spans knowledge graphs, semantic web technologies, machine learning, and natural language processing, with a focus on the extraction, integration, and efficient management of large-scale graph data. He is best known for pioneering frameworks such as LIMES for scalable link discovery, AGDISTIS for graph-based entity disambiguation, and GERBIL for benchmarking entity annotation systems. His contributions have earned him numerous international accolades, including the Next Einstein Fellowship and multiple best paper awards. His current research emphasizes explainable artificial intelligence, responsible machine learning, and efficient question answering over knowledge graphs through advanced data-driven methods. Additionally, since mid-2024, he serves on the Board of Directors of the Heinz Nixdorf Institute.