Language Models are not a Panacea: Combining them with Domain Knowledge and Efficient Indexes for Entity Linking

Daniel Lucas Albuquerque, Vitória S. Santos, Pedro Nack, Renato Fileto, Carina F. Dorneles

[1]Dep. de Informática e Estatística (INE), Univ. Federal de Santa Catarina (UFSC)
Campus Trindade, Caixa Postal 476 – 88.040-370 – Florianópolis – SC – Brazil

corresponding e-mail: daniel.lucas.f@posgrad.ufsc.br

Abstract. *Language models enable cutting-edge solutions for many problems. However, they may not always be the best choice—at least not on their own—for certain tasks in specific contexts. In this paper, we propose a hybrid approach to entity linking (EL) that employs domain knowledge and efficient indexes for named entity recognition (NER), delegating only the disambiguation step (NED) to language models. We evaluated this hybrid approach on textual descriptions of invoice items from public medication purchases. The experiments showed that domain knowledge and indexes enabled efficient recognition of medications (NER), with accuracy superior to most state-of-the-art language models investigated and comparable to the GPT-4o reasoning language model. In addition, candidate medications recognized by our computationally efficient approach were disambiguated (NED) by GPT-4o with 90.55% precision.*

1. Introduction

Short textual descriptions of products in documents, such as invoice item descriptions, pose challenges for information extraction and data linkage. Accurately identifying a product and linking it to its corresponding description is hampered by the variability and incompleteness of these descriptions. Identifying and linking the textual descriptions manually to the respective product definitions is usually unfeasible due to the large volume of data. Some enterprises cope with millions of invoices daily, involving a wide range of products. Therefore, properly solving these problems automatically is crucial to enable downstream applications such as the detection of overpricing and other fraudulent practices in public procurement, based on detailed and precise information about the product characteristics, and sometimes even the usual price of the product.

Entity Linking (EL) is a fundamental task in Natural Language Processing (NLP), whose objective is mapping textual mentions to the respective entity representations in an information or knowledge base [Jurafsky and Martin 2024]. Usually, textual mentions of named entities are previously recognized through Named Entity Recognition (NER), a step responsible for detecting and categorizing these mentions according to their respective entity types (e.g., person, organization, product) [Oliveira et al. 2021]. In this context, a typical EL system is structured in three main steps: candidate entity generation, candidate entity ranking, and non-linkable mention (NIL) prediction [Shen et al. 2023].

Large Language Models (LLMs), such as GPT-4 [Achiam et al. 2023], LLaMA [Touvron et al. 2023], Flan-T5-11B [Chung et al. 2024], have received enormous at-

tention in recent years. They have established the state of the art for several NLP tasks. Pre-trained Language models can be fine-tuned to solve a myriad of problems [Wang et al. 2023, Santos and Dorneles 2024, Cao et al. 2021, Barba et al. 2022, Xiao et al. 2023, Nascimento and Casanova 2024]. However, LLMs remain underexplored and present unique challenges for tasks such as entity linking (EL). They are more naturally suited to text-to-text tasks such as translation, summarization, and question answering [Xiao et al. 2023]. The main challenge lies in the fact that LLMs tend to produce mentions that link to entities that do not exist in the KB [Xin et al. 2024, Xiao et al. 2023, Liu et al. 2024], a phenomenon commonly referred to as hallucination.

In this context, [Liu et al. 2024] proposes an approach for EL using LLMs in *few-shot* scenarios, without the need for fine-tuning and with the application of a consistency algorithm to reduce hallucinations. [Xin et al. 2024] presents a solution using LLMs with Retrieval Augmented Generation (RAG) to enrich the context of mentions. [Liu and Fang 2023] leverages zero-shot, in-context learning (ICL) and chain-of-thought (CoT) strategies to disambiguate homonymous named entities. These strategies are centered on the use of LLMs, which can be expensive, as they require intensive use of computation power, even when they exploit domain knowledge with approaches like RAG. The current Reasoning Language Models (RLMs) [Besta et al. 2025] can apply some of these techniques, along with other ones like reinforcement learning, making them even more expensive than LLMs.

This work proposes an EL solution that combines efficient indexes, domain knowledge and language models. The proposed approach relies on efficient indexes of full text for doing NER, by efficiently finding in the text surface names from a domain-specific knowledge base. Then, language models are used to disambiguate candidate links as they can handle the variability and incompleteness commonly found in textual descriptions. This method differs from others that make extensive use of LLM calls, as it recognizes the named entity mentions and selects candidates for entity linking solely via indexes, relying on language models just for disambiguation. It alleviates the number of expensive calls to an RLM or an LLM using strategies like few-shot [Brown et al. 2020], RAG [Lewis et al. 2021], IOCT [Trivedi et al. 2023] or CoT [Wei et al. 2022].

This approach is effective in situations where NER can be tackled by exploiting lexical or semantic similarity matching with a dictionary of surface names. We evaluated it in experiments using real textual descriptions of invoice items referring to medications purchased by public authorities of the state of Santa Catarina, Brazil. The domain knowledge about the medications approved for use in the country and their price was adapted from a dataset of medications approved by the Brazilian National Health Surveillance Agency (Anvisa) and a dataset with medication prices over time from the Chamber for the Regulation of the Pharmaceutical Market (CMED). The NER results obtained by using our proposal were compared to those of the *baselines*: DeepSeek R1, Claude 3.7, Qwen 2.5, LLaMA 3.3 and GPT-4o. GPT-4o yielded the best NER accuracy (91.03%) to find proper candidates for entity linking, followed by our proposal (88.46%). The NER candidates obtained without using language models were then disambiguated using the language models, reaching accuracies close to 90%, with much lower use of language models, confirming the viability of the proposal in the case study.

The remainder of this work is organized as follows. Section 2 presents the foun-

dations that support the proposal and discusses related work. Section 3 presents the proposed solution. Section 4 reports the experimental evaluation. Section 5 discusses the experimental results and the limitations. Finally, section 6 presents the conclusion and suggestions for future work.

2. Related Work

Entity Linking (EL) is a key component in various Information Extraction (IE) workflows, since it addresses ambiguities in entity mentions and assigns them the correct referent depending on the context [Sevgili et al. 2022]. A traditional EL system typically consists of three stages: (i) candidate entity generation, (ii) candidate ranking, and (iii) unlinkable mention prediction [Shen et al. 2023]. In the candidate generation stage, a set of potential entities is associated with each identified mention. During ranking, these candidates are ordered based on various types of contextual evidence. Finally, the unlinkable mention prediction step identifies mentions that cannot be linked to any known entity and classifies them as NIL.

Traditional EL methods based on Machine Learning (ML) can be categorized into supervised and unsupervised approaches, particularly regarding the ranking stage [Shen et al. 2023]. Supervised approaches include binary classification, learning to rank, probabilistic models, and graph-based methods [Shen et al. 2023]. In contrast, unsupervised approaches comprise vector space models and information retrieval techniques. Among the main limitations of ML-based methods are the need for extensive manual feature engineering and the limited generalization capability of the models, which often stems from their reliance on domain-specific knowledge.

Although LLMs have demonstrated strong performance across various NLP tasks, their application to EL remains challenging. The main difficulty lies in the inaccuracy of LLMs when generating unique entity identifiers [Xin et al. 2024, Xiao et al. 2023, Liu et al. 2024]. In fact, studies have shown that these models often produce fictitious entity identifiers, a phenomenon referred to as hallucination [Xin et al. 2024]. Furthermore, LLMs face substantial limitations in the candidate entity generation phase, as EL typically involves returning a large set of candidate entities along with their corresponding descriptions [Liu et al. 2024].

The work by [Liu et al. 2024] introduces OneNet, an approach to EL in few-shot scenarios that eliminates the need for fine-tuning. The proposed process consists of three main steps: (i) filtering out irrelevant entities, (ii) leveraging contextual information combined with prior knowledge to perform linking, and (iii) applying a consistency algorithm to reduce hallucinations. Meanwhile, [Xin et al. 2024] presents LLMAEL, a plug-and-play solution for EL that employs LLMs as an additional source for data augmentation. Instead of directly asking LLMs to perform entity linking, the authors propose using them to enrich the context provided to traditional EL models by adding complementary information about specific mentions.

The study by [Liu and Fang 2023] introduces an approach aimed at disambiguating homonymous named entities extracted from academic knowledge graphs. Their work evaluates the performance of LLMs using various strategies, including zero-shot, in-context learning, and chain-of-thought. [Ding et al. 2024] proposes ChatEL, a three-step framework designed to guide LLMs toward producing more accurate results during

the entity linking process. The work of [Vollmers et al. 2025] presents a fine-tuning-based approach that integrates the recognition and disambiguation stages into a unified framework, leveraging LLMs to enrich the context of entity mentions. E-BELA [Pereira and Ferreira 2024] is a lightweight entity linking approach that uses literal-based embeddings to project mentions and entities into a shared vector space. By measuring similarity or distance, the method links mentions to their most likely entity, considering semantic context.

[Romero et al. 2025] evaluate LLMs for extracting medications and clinical attributes such as dosage and adverse effects, proposing ensemble strategies that outperform fine-tuned baselines like BERT and BioBERT. The study also includes an entity linking component that maps extracted terms to standard terminologies such as SNOMED-CT and BNF. [Santos and Dorneles 2024] also evaluate LLMs for segmenting invoice item descriptions into structured attributes using Zero-Shot learning. Their experiments focus on attribute extraction, highlighting the potential of LLMs for pre-processing in entity resolution tasks.

Finally, the work by [Miranda et al. 2024] proposes OntoDrug, an ontology designed to enhance drug management in the Brazilian healthcare context. OntoDrug integrates pharmaceutical terminologies and national regulatory frameworks to support medication identification and enable interoperability with other health information systems.

Unlike ML-based approaches, which require model training or prompt engineering steps and other approaches for effective and intensive use of LLMs, this work proposes the combination of efficient indexes, domain knowledge, and language to solve entity linking in particular domains. The former enables efficient and effective NER by exact or similarity matching full text with entity surface names from a knowledge base (KB). This method allows selecting candidate entities from the KB at lower costs than using LLMs or RLMs. In our proposal, an LLM or RLM is used only for disambiguating candidate entities.

3. Intelligent Information Extraction from Medication Purchases

3.1. Overview

We propose Intelligent Information Extraction (I^2E), a framework designed to support the integration of domain knowledge and machine learning techniques for efficient and effective information extraction. In this study, we apply the I^2E framework to the task of identifying medication mentions in short textual descriptions of invoice items and linking them to their corresponding entries in a Knowledge Graph (KG).

To support the linking process, a domain-specific ontology was developed to formally represent pharmaceutical concepts such as drug names, active ingredients, manufacturers, and product presentations. This ontology served as the basis for instantiating the KG using structured data extracted from official drug registration sources. Each medication was represented as an individual with semantically rich attributes, including commercial name, active ingredients, manufacturer, and presentation details.

The textual descriptions of medications were indexed to enable efficient candidate retrieval. By leveraging lexical similarity and domain-specific matching strategies, a set of candidate entities was generated for each description of medication. These candidates

were then enriched with contextual attributes—such as manufacturer and presentation—to support downstream disambiguation. This pipeline was designed to be robust to variations in how drugs are described in real-world invoice data. Figure 1 illustrates the conceptual workflow of this I^2E-based application.

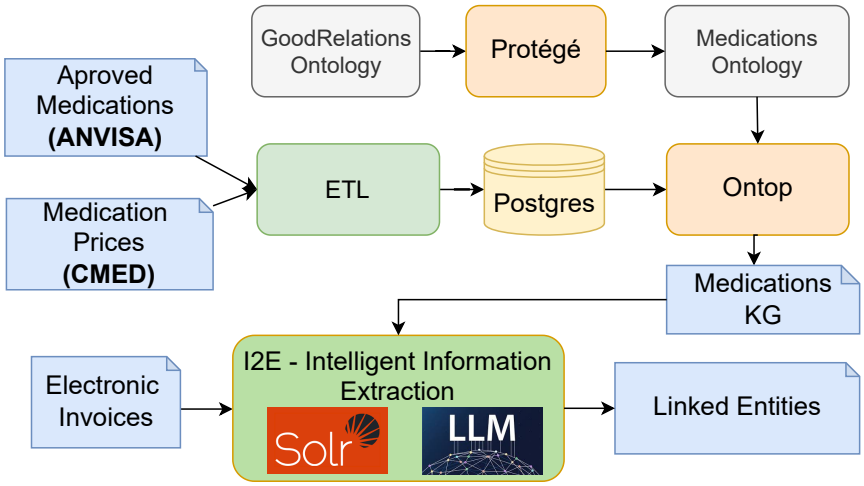

Figure 1. The I^2E workflow applied to medications

The preliminary version of the I^2E framework used in this work performs entity linking in two phases: generation of candidate entities for each invoice item description and entity disambiguation. These phases are detailed in the following subsections.

3.2. Generation of Candidate Entities using Indexes

The Candidate Entity Generation stage of the proposed method employs an efficient, domain-oriented indexing strategy to simulate the traditional NER process. Instead of relying on supervised NER models or LLMs, which typically require large volumes of annotated data, struggle to generalize across specific domains, and incur significant computational costs, this approach is based on token matching over the set of drug descriptions.

The drug descriptions from the electronic invoices were indexed for lexical segmentation. This process enabled the efficient recovery of candidate entities based on textual similarity, serving as a basis for the entity generation stage in the solution flow. For each drug description in invoice items, an initial lexical match is performed between the tokens in the description and the trade names of the drugs. If no match is found in this first attempt, a second search is conducted using the tokens of the active ingredients. This hierarchical strategy, prioritizing trade names followed by active ingredients, is designed to maximize mention coverage, considering the variability in how medications are described in invoices.

Finally, a data augmentation strategy was applied to enrich the list of candidate entities by incorporating complementary information related to the corresponding manufacturers and drug presentations. This semantic enrichment aims to provide a more robust context for the subsequent Entity Disambiguation stage, helping LLMs to perform more

accurate inferences by considering not only the drug name or active ingredient, but also the contextual attributes associated with each candidate entity.

3.3. Entity Disambiguation using Language Models

Following the generation of candidate entities, the next step in the proposed method involves Entity Disambiguation using LLMs. Unlike traditional EL approaches, which rely on the direct retrieval of unique identifiers, this work leverages the reasoning capabilities of LLMs over textual evidence to infer the correct entity. Such an approach allows for greater flexibility and adaptability, particularly in scenarios marked by high textual variability and semantic ambiguity, such as those found in item descriptions from electronic invoices. To guide the reasoning process of the language model, a structured zero-shot prompt was designed to prioritize key attributes during the disambiguation step. The prompt is shown below:

> You will receive a drug description and a list of possible matches.
> For each description, identify the number of the most appropriate option by following the priority order below:
> 1. Manufacturer
> 2. Concentration
> 3. Presentation
>
> Return only the number of the most suitable option, without any additional explanation.
>
> **Description:** "{`description`}"
> **Options:**
> {`candidates_list`}

4. Experiments

This section details the datasets, data sampling methods, language models and their configurations, and performance evaluation criteria used in our experiments.

4.1. Datasets

This study relies on three datasets: (i) drug descriptions found in electronic invoices (NFe) of public purchases issued in the state of Santa Catarina; (ii) official drug records retrieved through APIs provided by the Brazilian Health Regulatory Agency (Anvisa); and (iii) drug prices from the Drug Market Regulation Chamber (CMED). Table 1 provides an overview of these datasets. Drug description texts from invoice items in NFe were extracted from

Table 1. Summary of the datasets used in the experiments

Dataset	Records	Description
NFe	2,535	textual descriptions of invoice items
Anvisa	34,202	drug names, active ingredients, dose, packaging, etc.
CMED	24,642	other drug features and temporal series of their prices

public purchase invoices, filtering a database of general descriptions using the Mercosur

Common Nomenclature (NCM) 3004 number. The work used a statistical extract of 2,535 descriptions, applying normalization to their length.

Anvisa's data detail drugs in general terms of notified, registered, active and inactive registrations. The data were obtained from the Open Data portal[1] directly in CSV format, containing 34,202 drug registrations. In terms of main name, the dataset contains 18,813 distinct drugs and 2,346 distinct active ingredients. The information available for each registration includes *surface name*, *active ingredient(s)*, *9-digit registration number*, *regulatory category*, *manufacturing laboratory*, among others.

The CMED data officially define the Factory Price (PF) and the Maximum Consumer Price (PMC) for all drugs registered by Anvisa. In addition, the data are detailed in terms of their respective presentations, that is, their concentrations, pharmaceutical forms (tablet, injectable, cream, etc.) and volumes. The work obtained the CMED data directly in CSV format [2]. The dataset contains 24,642 drug records, consisting of 5,826 distinct drugs and 2,030 different active ingredients, both in terms of main name. Information for each registration includes *surface name*, *active ingredient(s)*, *13-digit registration number*, *manufacturing laboratory*, *presentation*, *therapeutic class*, in addition to PFs and PMCs, according to each state's tax rate.

4.2. Sampling Method

To support manual evaluation and enable a rigorous assessment of both the candidate entity generation phase and the subsequent entity disambiguation performed by LLMs, a representative sample of 156 medication descriptions was drawn from the filtered dataset of 2,535 descriptions obtained from NFe. The sampling procedure accounted for the heterogeneity and frequency of the descriptions, reflecting the non-probabilistic nature of the underlying population. The sample size was determined following the methodological framework of [Rea and Parker 2012], applying the formula designed for small populations:

$$sample = \frac{Z^2 \cdot (0.25) \cdot N}{Z^2 \cdot (0.25) + (N-1) \cdot M_E^2}.$$

In this equation, M_E represents the proportional margin of error, Z is the Z-score corresponding to the selected confidence level, and N is the population size. For this study, we adopted a confidence level of 99% ($Z = 2.575$) and a margin of error of 10% ($M_E = 0.1$), which yielded the final sample. This approach ensured that the subset was sufficiently diverse and statistically robust, while remaining tractable for manual annotation and comprehensive benchmarking of both the candidate entity generation process and the entity linking performance.

4.3. Language Models

Five language models were selected for evaluating the Entity Disambiguation stage, including those with advanced reasoning and contextual understanding capabilities. Table 2 presents the models used and their respective context window sizes.

[1]https://dados.gov.br/dados/conjuntos-dados/medicamentos-registrados-no-brasil (accessed on April 14, 2025)

[2]https://dados.anvisa.gov.br/dados/TA_PRECO_MEDICAMENTO.csv (accessed on April 14, 2025)

Model	*API model name*	*Context window (Cw)*
DeepSeek	deepseek-r1:70b	128k tokens
Claude	claude-3-7-sonnet-20250219	up to 128k tokens
GPT-4	gpt-4o-2024-11-20	128k tokens
Qwen	qwen2.5:72b	128k tokens
LLaMA	llama3.3:70b	8k tokens

Table 2. Models and context window sizes employed

4.4. Evaluation Criteria

The evaluation of the language models was conducted in two distinct stages: Named Entity Recognition (NER) and Entity Linking (EL). In the NER stage, the performance of the LLMs was also compared to the results produced by the I^2E framework. For both stages, true positives (T_P) were defined according to the specific objectives of each task:

- **NER:** a true positive occurred when the LLM correctly extracted the medication name from the description or when I^2E successfully generated a list of candidate entities.
- **EL:** a true positive corresponded to cases where the LLM correctly identified the corresponding medication among the candidate entities.

The performance assessment was carried out through a line-by-line comparison between the predicted outputs and the reference data. Accuracy was adopted as the primary evaluation metric, calculated as the proportion of true positives relative to the total number of evaluated cases:

$$accuracy = \frac{T_P}{Total}.$$

The variable $Total$ represents the total number of cases assessed. The decision to focus exclusively on true positives reflects the study's emphasis on measuring the models' ability to produce correct and complete outputs. This approach was selected for its clear interpretability and its suitability for providing an overall assessment of predictive performance. The evaluation strategy employed in this study follows established methodological recommendations described in [Géron 2019], supporting transparency in the performance analysis.

5. Results and Discussion

5.1. Results

This section presents the results of the experiments conducted to evaluate the performance in the tasks of NER and EL. Table 3 presents the NER performance of the language models and the I^2E framework. Accuracy was computed separately for two types of entity mentions: the commercial name of the medication and the active ingredient.

GPT-4o achieved the highest accuracy for commercial medication names (85.26%), closely followed by the I^2E framework (84.61%). Claude 3.7 and LLaMA 3.3 also performed well, surpassing 75% in this category. In contrast, Qwen 2.5 and DeepSeek R1 obtained lower scores, with 62.18% and 56.41%, respectively.

Table 3. NER Accuracy

Model	Medication Name (%)	Active Ingredient (%)
DeepSeek	56.41	41.67
Claude	77.56	53.85
GPT-4	**85.26**	**55.77**
Qwen	62.18	34.62
LLaMA	75.64	41.67
I^2E	84.61	50.97

In the identification of active ingredients, overall performance was lower across all models. GPT-4o again led the results (55.77%), followed by Claude 3.7 (53.85%) and I^2E (50.97%). Notably, DeepSeek R1 and LLaMA 3.3 showed the weakest results for active ingredient recognition (41.67%), suggesting greater difficulty in recognizing this type of entity.

Table 4 presents the results of a complementary experiment focused on residual extraction, aimed at better understanding the models' ability to recover information in challenging cases. Specifically, it shows how often the active ingredient was correctly identified when the commercial name of the medication was not, and how this residual recovery contributes to overall recognition coverage.

Table 4. Residual NER Coverage

Model	Active Ingredient on Residue (%)	Combined (%)
DeepSeek	33.82	71.15
Claude	**40.00**	86.54
GPT-4	39.13	**91.03**
Qwen	22.03	70.51
LLaMA	21.05	80.77
I^2E	25.00	88.46

The first column of Table 4 shows the residual recovery rate—i.e., the percentage of cases where the active ingredient was correctly extracted despite failure in medication name recognition. GPT-4o and Claude 3.7 demonstrated the highest recovery in this aspect (39.13% and 40.00%, respectively), with DeepSeek R1 also performing moderately well (33.82%). Qwen 2.5 and LLaMA 3.3 achieved lower residual coverage, indicating reduced fallback capacity in such cases.

The second column presents the combined coverage, considering successful recognition of either the medication name or the active ingredient. GPT-4o achieved the highest combined coverage (91.03%), followed closely by I^2E (88.46%) and Claude 3.7 (86.54%). These results highlight the importance of leveraging both entity types for robust NER performance in real-world texts.

Table 5 presents the accuracy of each model in the EL task, based on a controlled NER input generated by the I^2E framework. By isolating the EL stage, the experiment ensures that all models receive the same candidate lists and descriptions, enabling a fair

comparison of their reasoning and selection capabilities.

Table 5. EL General Accuracy Comparison

Model	Accuracy (%)
GPT-4	**90.55**
Claude	84.25
Deepseek	59.06
Qwen	53.54
LLaMA	51.97

Finally, an additional experiment was conducted under varying levels of ambiguity to further assess the robustness of the entity linking process. As shown in Figure 2, the evaluated language models were tested in three scenarios: (i) when only one correct entity is present among the candidates, (ii) when up to two entities are valid, and (iii) when up to five correct entities are included.

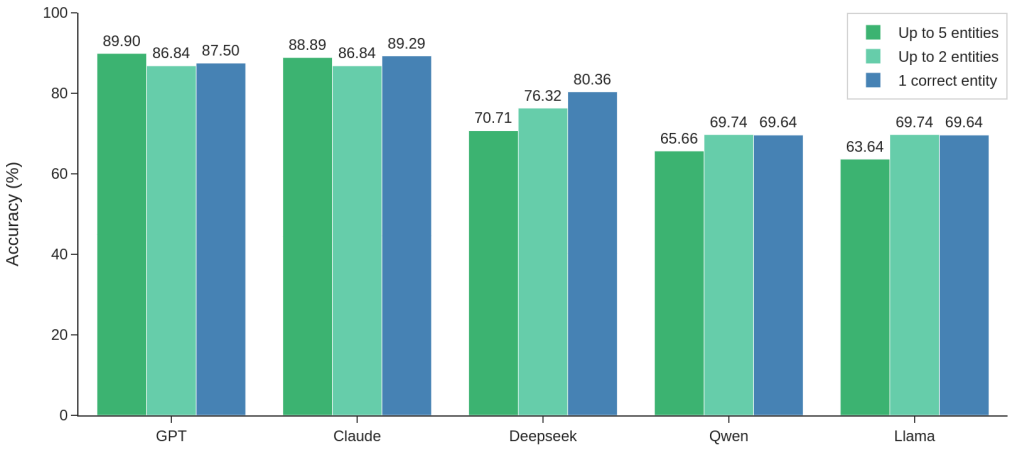

Figure 2. EL accuracy results for each language model across different levels of candidate ambiguity.

GPT-4o and Claude 3.7 consistently outperformed the other models across all levels of ambiguity, maintaining high accuracy even when the number of valid candidates increased. GPT-4o achieved a peak accuracy of 89.90% in the most ambiguous scenario (up to five correct entities), while Claude 3.7 reached 88.89%. DeepSeek R1 showed moderate degradation in performance as ambiguity increased, while LLaMA 3.3 and Qwen 2.5 exhibited lower resilience, with noticeable drops in accuracy in multi-match cases.

Despite the overall dominance of GPT-4o and Claude 3.7, the performance gap between models becomes particularly pronounced under higher ambiguity conditions. While GPT-4o and Claude 3.7 maintained accuracies above 85% even in the most complex scenario, the remaining models demonstrated greater sensitivity to the increasing number of valid entities. DeepSeek R1, although outperforming Qwen 2.5 and LLaMA 3.3, experienced a drop of nearly 10 percentage points between the exact-match and five-candidate settings. In contrast, Qwen 2.5 and LLaMA 3.3 not only started with lower

baseline performance but also exhibited less stability, suggesting potential limitations in contextual disambiguation or retrieval alignment under uncertain conditions.

5.2. Computational Cost

We measured I^2E NER costs on a Maxsun MS-Challenger B760M-N D5 server with a 13th Gen Intel Core i7-13700 (24 cores), 64 GB RAM, 1.0 TB disk, running Ubuntu 24.04.1 LTS. Matching medication names in 2k invoice descriptions using Apache Solr version 9.7 took 337.57 seconds. Since this cost is mainly due to checking 34,202 medicine names on the set of descriptions, it grows very slowly with increasing numbers of descriptions. Meanwhile, the average time to recognize medicine names on the same 2k descriptions was 385 seconds on average when using Claude 3.7, GPT-4o, Qwen 2.5, DeepSeek R1, or LLaMA 3.3. Howevr, this time grows linearly with the number of descriptions, much faster than using Solr.

In addition, the I^2E NER costs are limited to part time use of our server, while the cost for NER with LLMs varies with the model. The estimated API cost for processing 156 descriptions ranges from US$ 0.35 with Claude 3.7 to US$ 0.47 with GPT-4o. Thus, the current costs for doing NER using Claude 3.7 or GPT-4o, which give around the same accuracy as I^2E, on around 10 million invoice item descriptions that we have to process each year is expected to be between US$ 22,435.90 and US$ 30,128.20. Open-source models like DeepSeek R1 and LLaMA 3.3 run cheaper, but do not perform so well.

The EL stage involves larger prompts due to the candidate lists, increasing computational demand. Average response times range from 4 to 12 seconds per string, leading to a total estimated runtime of 20 to 30 minutes for the 156 descriptions. The cost for disambiguating all these descritptions was approximately US$1.30.

5.3. Discussion

The I^2E competitive performance on NER shows that cheaper solutions can be a good alternative to the extensive use of language models. The relatively low performance of all models in the recognition of active ingredients is probably due to the absence of mentions to them in a considerable portion of the textual description of invoice items. They usually appear in purchases of generic medications, while commercial medication names are more common in the NFe dataset.

On the other hand, the EL results confirm the language models' ability to handle the variability and semantic ambiguity of textual descriptions of complementary information semantic ambiguity, such as dosage and packaging of medications, during entity disambiguation. This challenge occurs specially in real-world scenarios as the one considered in our case study, where multiple correct interpretations may coexist.

5.4. Limitations of the proposal

One of the main limitations of the proposed approach is its reliance on domain-specific datasets. Although this dependency requires structured and up-to-date sources, such as drug registration data, it offers a practical advantage: maintaining and updating such datasets is typically more feasible than retraining large language models or implementing complex fine-tuning procedures. However, the effectiveness of the approach is strongly tied to the availability of formalized domain knowledge. In domains where

well-structured ontologies or consolidated data sources are not available, additional effort is required to model and curate the necessary semantic structures, which may limit the scalability of the solution across different application contexts.

6. Conclusion

This work presented I^2E, a hybrid solution for entity linking that combines efficient full text indexes and domain knowledge for candidate generation, leaving to language models like LLMs and RLM only the task of entity disambiguation. This strategy can reduce costs associated with language model use, while maintaining competitive performance.

We evaluated this proposal on textual descriptions of public medication purchases. I^2E achieved NER accuracy comparable to the best-performing language model (GPT-4o), even without the use of fine-tuning or pre-training strategies. Moreover, some language models were able to disambiguate the candidate links efficiently obtained by I^2E with accuracy close to 90%.

Although the proposed approach depends on the availability of domain knowledge, it offers a sustainable and cost-effective alternative for real-world applications. These results reinforce the value of combining symbolic and neural methods: structured knowledge enables controlled, transparent candidate generation, while LLMs contribute with flexible and context-aware inference for final disambiguation.

Future work may explore alternative indexing strategies to improve the generation and ranking of candidate entities in distinct application domains. This research focus includes experimenting with different tokenization methods to better capture lexical variations and domain-specific patterns in the texts applied to EL. We also plan to investigate other strategies to combine domain knowledge with language models, such as applying deduction techniques in knowledge graphs and ontologies, which may include rules, to check and improve language model results. It may help to detect and correct hallucinations and lack of response of language models. Additionally, improvements in the disambiguation stage could involve exploring more sophisticated prompting techniques. While the current approach relies solely on zero-shot learning, future experiments could incorporate few-shot strategies or in-context learning to enhance model guidance, especially in more ambiguous or noisy scenarios.

Acknowledgements

This work has been supported by a 2022 CNPq Universal grant, FAPESC grant 2021TR1510, and by the Céos project, financed by the Public Ministry of Santa Catarina State (MPSC), which has been crucial to improve our working conditions.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Barba, E., Procopio, L., and Navigli, R. (2022). Extend: Extractive entity disambiguation. In *Proc.of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2478–2488.

Besta, M., Barth, J., Schreiber, E., Kubicek, A., Catarino, A., Gerstenberger, R., Nyczyk, P., Iff, P., Li, Y., Houliston, S., Sternal, T., Copik, M., Kwaśniewski, G., Müller, J., Łukasz Flis, Eberhard, H., Niewiadomski, H., and Hoefler, T. (2025). Reasoning language models: A blueprint.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Cao, N. D., Izacard, G., Riedel, S., and Petroni, F. (2021). Autoregressive entity retrieval.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al. (2024). Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Ding, Y., Zeng, Q., and Weninger, T. (2024). Chatel: Entity linking with chatbots. *arXiv preprint arXiv:2402.14858*.

Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly.

Jurafsky, D. and Martin, J. H. (2024). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Third edition draft edition.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., tau Yih, W., Rocktäschel, T., Riedel, S., and Kiela, D. (2021). Retrieval-augmented generation for knowledge-intensive nlp tasks.

Liu, S. and Fang, Y. (2023). Use large language models for named entity disambiguation in academic knowledge graphs. In *2023 3rd Intl. conf. on Education, Information Management and Service Science (EIMSS 2023)*, pages 681–691. Atlantis Press.

Liu, X., Liu, Y., Zhang, K., Wang, K., Liu, Q., and Chen, E. (2024). Onenet: A fine-tuning free framework for few-shot entity linking via large language model prompting. *arXiv preprint arXiv:2410.07549*.

Miranda, N., Machado, M. M., and Moreira, D. A. (2024). Ontodrug: Enhancing brazilian health system interoperability with a national medication ontology. In *Brazilian Symposium on Multimedia and the Web (WebMedia)*, pages 240–248. SBC.

Nascimento, E. and Casanova, M. A. (2024). Querying databases with natural language: The use of large language models for text-to-sql tasks. In *Anais Estendidos do XXXIX Simp. Brasileiro de Bancos de Dados*, pages 196–201, Porto Alegre, RS, Brasil. SBC.

Oliveira, I. L., Fileto, R., Speck, R., Garcia, L. P., Moussallem, D., and Lehmann, J. (2021). Towards holistic entity linking: Survey and directions. *Information Systems*, 95:101624.

Pereira, Í. M. and Ferreira, A. A. (2024). E-bela: Enhanced embedding-based entity linking approach. In *Brazilian Symposium on Multimedia and the Web (WebMedia)*, pages 115–123. SBC.

Rea, L. and Parker, R. (2012). *Designing and Conducting Survey Research: A Comprehensive Guide*. Wiley.

Romero, P., Han, L., and Nenadic, G. (2025). Medication extraction and entity linking using stacked and voted ensembles on LLMs. In Ananiadou, S., Demner-Fushman, D., Gupta, D., and Thompson, P., editors, *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health)*, pages 303–315, Albuquerque, New Mexico. Association for Computational Linguistics.

Santos, V. and Dorneles, C. (2024). Unveiling the segmentation power of llms: Zero-shot invoice item description analysis. In *Anais do XXXIX Simpósio Brasileiro de Bancos de Dados*, pages 549–561, Porto Alegre, RS, Brasil. SBC.

Sevgili, Ö., Shelmanov, A., Arkhipov, M., Panchenko, A., and Biemann, C. (2022). Neural entity linking: A survey of models based on deep learning. *Semantic Web*, 13(3):527–570.

Shen, W., Li, Y., Liu, Y., Han, J., Wang, J., and Yuan, X. (2023). Entity linking meets deep learning: Techniques and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 35(3):2556–2578.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Trivedi, H., Balasubramanian, N., Khot, T., and Sabharwal, A. (2023). Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions.

Vollmers, D., Zahera, H., Moussallem, D., and Ngomo, A.-C. N. (2025). Contextual augmentation for entity linking using large language models. In *Proc.of the 31st International Conference on Computational Linguistics*, pages 8535–8545.

Wang, S., Li, A. H., Zhu, H., Zhang, S., Hang, C.-W., Perera, P., Ma, J., Wang, W., Wang, Z., Castelli, V., et al. (2023). Benchmarking diverse-modal entity linking with generative models. *arXiv preprint arXiv:2305.17337*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., and Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Xiao, Z., Gong, M., Wu, J., Zhang, X., Shou, L., Pei, J., and Jiang, D. (2023). Instructed language models with retrievers are powerful entity linkers. *arXiv preprint arXiv:2311.03250*.

Xin, A., Qi, Y., Yao, Z., Zhu, F., Zeng, K., Bin, X., Hou, L., and Li, J. (2024). Llmael: Large language models are good context augmenters for entity linking. *arXiv preprint arXiv:2407.04020*.