Journal of Information and Data Management, 2026, 17:1, doi: 10.5753/jidm.2026.7440
This work is licensed under a Creative Commons Attribution 4.0 International License.

Combining Domain Knowledge, Full-Text Search, Heuristics and LLM for Efficient Accurate Entity Linking in Medicine Invoices

Daniel Lucas Albuquerque [**Universidade Federal de Santa Catarina** | ***daniel.lucas.f@posgrad.ufsc.br***]
Vitória S. Santos [**Universidade Federal de Santa Catarina** | ***vit2santoss@gmail.com***]
Pedro Nack [**Universidade Federal de Santa Catarina** | ***pedro.nackm@gmail.com***]
Carina F. Dorneles [**Universidade Federal de Santa Catarina** | ***carina.dorneles@ufsc.br***]
Renato Fileto [**Universidade Federal de Santa Catarina** | ***r.fileto@ufsc.br***]

Programa de Pós-graduação em Ciência da Computação (PPGCC), Departamento de Informática e Estatística (INE), Univ. Federal de Santa Catarina (UFSC), Campus Trindade, Caixa Postal 476, Florianópolis, SC, 88.040-370, Brazil.

Received: 21 January 2026 • **Published:** 27 November 2026

Abstract Large Language Models (LLMs) have enabled cutting-edge solutions for many problems. However, they are not always the best choice for certain tasks in specific contexts, at least not on their own. This paper presents a new hybrid approach to entity linking (EL) that employs efficient full-text search and lightweight regex-based heuristics to find and filter EL candidates in a domain knowledge base (KB), delegating to the LLM only the disambiguation of pre-filtered candidates. Our proposal is implemented as a framework that can be adapted to different domains by changing the KB and adjusting domain-specific heuristics and prompts. We evaluate it on invoice item descriptions from public purchases of medicines. Our experiments show that the full-text search allowed efficient recognition of medicines and active ingredients in the pharmaceutical domain KB with accuracy superior to most state-of-the-art models investigated, and comparable to the best performing one in our experiments: GPT-4o. Candidate medicines, recognized by using only our computationally efficient index-based approach, were disambiguated by GPT-4o (best results), Claude, DeepSeek, Qwen and Llama with accuracies between 90.55% and 51.97%. Further experiments with regex-based heuristics revealed that they allow considerable reduction of EL candidate sets, improving performance and reducing LLM costs. Therefore, our hybrid pipeline contributes to both computational efficiency and effectiveness. The index-based search serves as an efficient and much cheaper alternative to LLMs for accurately performing NER and EL candidate list generation. The candidate filtering using heuristics is also very cost-effective, eliminating a considerable portion of candidates and thereby reducing LLM costs for disambiguation with high accuracy.

Keywords: Entity Linking, Large Language Model (LLM), Domain Knowledge, Full-text Indexing, Heuristic Filtering.

1 Introduction

Short textual descriptions of products, such as invoice item descriptions, pose challenges for information extraction and data linkage. Accurately identifying a product and linking it to the corresponding entity in a knowledge base (KB) is hampered by the variability and incompleteness of these descriptions. Doing this manually 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, automatically linking each description to detailed and precise information of the specific product it refers to is crucial to enable downstream applications such as the detection of overpricing and other frauds in public procurement.

Entity Linking (EL) is a fundamental task in Natural Language Processing (NLP), which maps textual mentions of a document to the respective entity representations in an KB [Jurafsky and Martin, 2024]. EL is traditionally performed in three main steps: candidate entity generation, candidate entity ranking, and non-linkable mention (NIL) prediction [Shen *et al*., 2023]. The candidate generation can be realized by a Named Entity Recognition (NER) method that matches words in a text to surface names of entities of certain types (e.g., person, product) in a KB [Oliveira *et al*., 2021].

Pre-trained Large Language Models (LLMs), such as GPT-4 [Achiam *et al*., 2023] and LLaMA [Touvron *et al*., 2023], have received enormous attention and established the state of the art for several NLP tasks. They 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], including NER. However, EL in specific domains is challenging even for these models, which are more naturally suited to text-to-text tasks such as translation, summarization, and question answering [Xiao *et al*., 2023]. LLMs tend to produce links to wrong entities or 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.

Liu *et al*. [2024] proposes an approach for EL using LLMs without fine-tuning. They apply *few-shot* [Brown *et al*., 2020] and a consistency algorithm to reduce hallucinations. Xin *et al*. [2024] presents a solution that uses LLMs with Retrieval Augmented Generation (RAG) Lewis *et al*. [2021] to enrich the context of mentions. Liu and Fang [2023] leverages zero-shot, in-context learning (ICL) and chain-of-thought (CoT) [Wei *et al*., 2022] strategies to disambiguate homonymous named entities. These LLM-centered proposals can be expensive due to the high computational power re-

quirements of these models. In addition, LLLs tend to yield non-factual EL results in specific domains, even when exploiting domain knowledge in strategies like RAG or IRCoT (Interleaved Retrieval guided by Chain-of-Thought) [Trivedi *et al*., 2023]. The current Reasoning Language Models (RLMs) [Besta *et al*., 2025] can apply a myriad of strategies to improve responses, but making them more expensive and much slower to respond with limited improvements.

This work proposes an EL solution that combines domain knowledge, efficient indexes, heuristics and language models. Our method differs from others that make extensive use of LLMs because it recognizes named entity mentions and selects candidates for entity linking solely via indexes and heuristics, relying on a language model only for disambiguation. While the index-based search serves as a highly efficient alternative for EL candidate list generation, our core contribution lies in the overall EL framework. The hybrid pipeline leverages LLMs effectively for EL: by using full-text indexing and heuristic-based methods for fast and effective EL candidate generation and filtering, we offload the LLM and achieve state-of-the-art effectiveness, while drastically reducing the computational costs associated with pure LLM-based solutions.

The underlying architecture is designed with modularity and reuse in mind. It was implemented as a framework that can be adapted to different domains by changing the KB, the domain-specific heuristics and the LLM prompts. Its instantiation and the experimental results presented in this paper are restricted to the pharmaceutical domain due to space limitations. Driven by the challenges of large-scale public auditing, our experiments were designed to address the following overarching research question: How can we perform accurate, computationally efficient, and economically viable information extraction within a real-world scenario comprising millions of electronic invoices?

Our approach is effective in situations where NER and candidate entity generation can be tackled by matching words of the textual documents with entity surface names in a KB. It is easily applicable to the precise nomenclature of pharmaceutical products, and also to other domains where there are structured data sources with standardized entity names (e.g., vehicle manufacturers and models). Our experiments used real textual descriptions of invoice items referring to medicines purchased by public authorities of the state of Santa Catarina, Brazil. The KB was built from two existing datasets: (i) medicines approved by the Brazilian National Health Surveillance Agency (Anvisa) and (ii) medicine presentations and prices over time from the Chamber for the Regulation of the Pharmaceutical Market (CMED).

The performance of our proposal for doing NER was compared to pure LLM baselines (DeepSeek R1, Claude 3.7, Qwen 2.5, LLaMA 3.3, and GPT-4o). While GPT-4o yielded the highest accuracy (91.03%), our index-based approach followed closely (88.46%), but at a fraction of the computational cost. The filtered candidates were then disambiguated by these LLMs, reaching EL accuracies close to 90%. This comparative analysis of LLMs and our index-based solution for NER highlights our framework's robustness and efficiency, particularly on its NER stage, rather than merely benchmarking the models.

This article extends [Albuquerque *et al*., 2025] by providing a more detailed description of our proposal, Intelligent Information Extraction (I^2E). In addition, in this paper we introduce regex-based heuristics to reduce the number of EL candidates prior to their submission to LLMs for disambiguation. New LLMs were evaluated on a larger dataset and an extended KB to evaluate the effectiveness of the heuristics in a more challenging setting. New experimental results show that these heuristics reduce the number of EL candidates and LLM costs, allowing higher performance.

In the following, section 2 discusses the foundations of our proposal and related work. Section 3 describes the proposal. Section 4 reports the experimental evaluation. Section 5 discusses the experimental results and the limitations. Finally, section 6 presents the conclusion and future work.

2 Related Work

Entity Linking (EL) resolves ambiguities in Information Extraction (IE) by assigning mentions to their correct contextual referents [Sevgili *et al*., 2022]. Supervised machine learning (ML) models for EL include classification, ranking, probabilistic, and graph-based techniques [Shen *et al*., 2023], whereas unsupervised methods rely on vector spaces and information retrieval. However, these models suffer from limited generalization and require extensive feature engineering. LLMs, on the other hand, struggle to generate massive sets of candidate entities [Liu *et al*., 2024] and frequently hallucinate by producing fictitious entity identifiers [Xin *et al*., 2024; Xiao *et al*., 2023; Liu *et al*., 2024]. Liu *et al*. [2024] proposes OneNet, a training-free, few-shot EL approach that filters candidate links using contextual knowledge, and applies a consistency algorithm to mitigate hallucinations. LLMAEL [Xin *et al*., 2024] uses LLMs strictly for data augmentation, enriching mention contexts for traditional EL models. Other notable approaches include [Liu and Fang, 2023], which evaluates zero-shot, in-context, and CoT prompting to disambiguate homonymous entities in academic Knowledge Graphs (KGs), and ChatEL [Ding *et al*., 2024], a three-step framework for LLM-guided EL. Furthermore, [Vollmers *et al*., 2025] unifies recognition and disambiguation via fine-tuning and contextual enrichment. Offering a lightweight alternative, E-BELA [Pereira and Ferreira, 2024] links entities by mapping mentions and KGs into a shared literal-based embedding space to compute semantic similarity.

Romero *et al*. [2025] uses an ensemble of LLM strategies to extract clinical attributes and link terms to medical terminologies, outperforming fine-tuned baselines. In a related context, [Santos and Dorneles, 2024] highlights the pre-processing potential of zero-shot LLMs by segmenting invoice descriptions into structured attributes for downstream entity resolution. Miranda *et al*. [2024] proposes OntoDrug, a Brazilian healthcare ontology integrating pharmaceutical terminologies and regulatory frameworks to enhance medicine identification and system interoperability.

Richter-Pechanski *et al*. [2025] achieves state-of-the-art performance on end-to-end information extraction from clinical notes using parameter-efficient fine-tuning of local models. Alternatively, Islam *et al*. [2025] performs NER on electronic health records using GPT-4o and DeepSeek-R1

by relying on costly prompt ensembles and majority voting, while Yu *et al*. [2026] benchmark multimodal LLMs for information extraction from noisy, faxed medical PDFs. Wang *et al*. [2024] proposes a modular entity extraction pipeline using quantized, low-parameter LLMs (such as Qwen-14B) on clinical admission records. However, their method involves a computationally heavy sequence of multiple LLM prompts.

Finally, though the experiments reported in this work focus on pharmaceutical product identification in short textual descriptions of medicine purchases, our approach may eventually be applied to extract information from other kinds of documents, such as clinical textual data. We plan experiments with public corpora such as SemClinBR Oliveira *et al*. [2022], to compare the performance and the costs of our approach with related work on this domain. Ruas *et al*. [2026] provides an overview of the recent scenario of EL and knowledge management in the biomedical and clinical domains, highlighting key limitations of existing approaches and identifying promising directions for future research in this area. [de Almeida *et al*., 2026] shows that the multilingual BERT-based model mmBERT outperforms LLMs such as GPT-5 and Gemini-2.5 for NER on the Portuguese language SemClinBr corpus and a private breast cancer dataset. These lighter models can run on-premises with limited computational resources to maintain data privacy at lower costs. Balanced data-splitting, among other strategies, further enhances the BERT-based model's performance.

To clarify the positioning of I^2E relative to prior work, we highlight key similarities, differences, advantages, and limitations. Similarly to recent LLM-guided frameworks [Ding *et al*., 2024; Xin *et al*., 2024], we leverage the advanced reasoning capabilities of LLMs to resolve complex semantic ambiguities during EL candidate disambiguation. However, the core difference lies in the architecture of our solution: unlike end-to-end LLM extraction models [Richter-Pechanski *et al*., 2025; Islam *et al*., 2025] or computationally heavy prompting sequences [Wang *et al*., 2024], we strictly isolate the LLM from the EL candidate generation stage. We perform NER and candidate filtering using deterministic full-text indexing and heuristics. The primary advantage of this division of labor is that it overcomes the known LLM bottleneck of processing massive candidate sets [Liu *et al*., 2024], drastically reducing API costs and latency while maintaining state-of-the-art accuracy. As a limitation, unlike open-domain zero-shot LLMs that require no prior data, our approach's effectiveness is directly bounded by the availability and quality of a (semi-)structured, domain-specific KB and predefined surface name dictionaries.

3 Intelligent Information Extraction

We propose Intelligent Information Extraction (I^2E), a framework designed to support the integration of domain knowledge and language models for efficient and effective information extraction. In this study, we apply the I^2E framework to the task of identifying medicine mentions in short textual descriptions of invoice items and linking them to corresponding entries in a KG. Figure 1 illustrates the application of the I^2E framework for the EL of medicine mentions in NF-e items to their corresponding entities, specifically official descriptions from ANVISA and CMED.

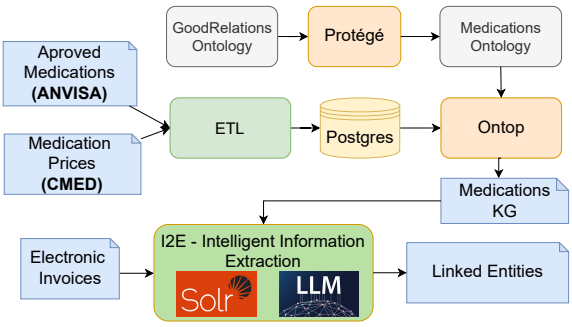

Figure 1. The I^2E workflow applied to medicines.

To support the linking process, a domain-specific ontology was developed to formally represent pharmaceutical concepts such as drug names, active ingredients, and manufacturers. This ontology was built as an extension of the GoodRelations (GR) [Hepp, 2008] vocabulary, using the Protégé [Musen, 2015] tool. Each medicine is modeled as a subclass of *gr:ProductOrService* to ensure a unified semantic view across different domains. Semi-structured data from ANVISA and CMED underwent an ETL (Extract, Transform, Load) process for integration and were stored in a PostgreSQL relational database. Ontop [Xiao *et al*., 2020] was used to map the relational data to the proposed ontology and provide a Virtual Knowledge Graph (VKG), the Medicines KG. It enables semantic data retrieval via a SPARQL endpoint, allowing our EL framework to access medicine attributes like commercial names, active ingredients, prices, and presentation formats (dosages, packaging, etc.).

The textual descriptions of medicines were indexed to enable efficient matching of medicine names, active ingredients and manufacturers to the respective entity surface names in the KB. A set of matching candidate entities was generated for each medical description. Each candidate was then enriched with contextual attributes, such as manufacturer and presentation, to help downstream disambiguation.

Though this pipeline was tailored to handle variations in real-world pharmaceutical data, the underlying framework is domain-agnostic. Its versatile architecture can be seamlessly adapted to other domains by simply replacing the source datasets, the target ontology, the reference KG, domain-specific heuristics, and the LLM prompt templates used in the disambiguation phase.

Figure 2 details the I^2E process of Figure 1. After selecting the NF-e items that refer to medicines (Mercosur Common Nomenclature - NCM code 3004) from a relational database, the plain text descriptions of these items are indexed using Solr. Next, depending on the method chosen, the indexed descriptions are either matched against the medicine records from the medicines KG or used as input for LLM-based EL. These two alternatives represent distinct mechanisms for generating candidate entities. The subsequent phase, *Filtering the Candidate List*, applies regex-based heuristics to reduce the number of candidates and decrease the likelihood of errors in the *Entity Disambiguation using Language Models* stage.

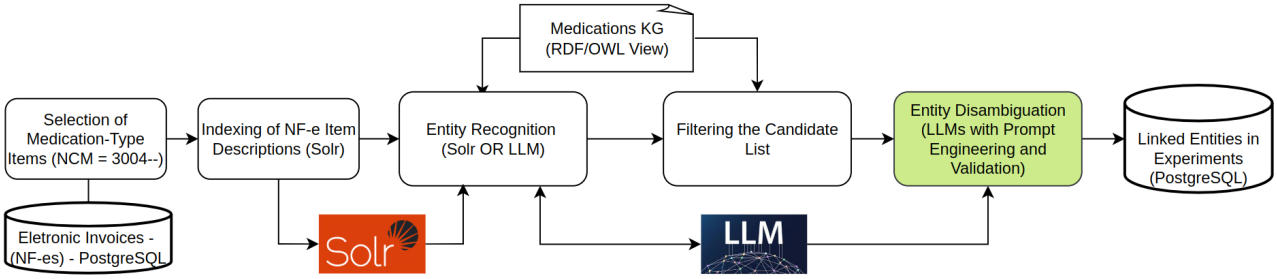

Figure 2. The I^2E workflow detailed.

3.1 Identification of EL Candidate Entities via Efficient Full-Text Search

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 medicines 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.

3.2 Filtering the EL Candidate List

To reduce the candidates forwarded to the LLM disambiguation phase, I^2E introduces an intermediate regex-based filtering step. This step extracts key attributes from invoice descriptions, particularly highly informative numeric values like active ingredient concentration and quantity, and discards incompatible candidates. To achieve this, we evaluate two numeric matching heuristics: (i) retaining candidates whose presentation text shares a numeric value with the invoice description, and (ii) strictly requiring matches for both the numbers and their associated measurement units (e.g., ml, mg) when present. While heuristic (i) provides broad numeric filtering, heuristic (ii) targets pharmaceutical formatting more precisely. Concurrently, we are improving surface-name dictionaries to resolve synonyms and acronyms. By drastically narrowing the candidate space, this filtering improves efficiency and prevents noise in the LLM decision process. The regex patterns for these heuristics are shown in table 1:

Table 1. Regex expressions for the heuristics

Heuristic	Regex Expression
1	`\s+(C/)?[0-9]+([\.,][0-9]+)*\s?(?!/)`
2	`([0-9]+([\.,][0-9]+)*\s*(MGG|G|MG|UI|%|MCG)\s*(/\s*[0-9]*?\s*(G|ML|MG|UI|%)+)?)`

These heuristics significantly narrow the candidate space, ensuring that only the compatible EL candidates are forwarded to the LLM-based disambiguation step. This filtering procedure not only improves efficiency but also increases robustness by removing candidates that would otherwise introduce noise into the decision process.

3.3 Entity Disambiguation

Following the filtering 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: manufacturer,
active ingredient(s) concentration, presentation.
Description: ''{description}''
Options: {candidates_list}
```

4 Experiments

4.1 Datasets

This study relies on three datasets: (i) drug descriptions found in electronic invoices (NF-e) 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 2 provides an overview of these datasets.

Table 2. Summary of the datasets used in the experiments

Dataset	Records	Description
NF-e	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

Drug description texts from invoice items in NF-e were extracted from public purchase invoices, filtering a database of general descriptions using the Mercosur Common Nomenclature (NCM) 3004 code. 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 EL Gold Standard Sampling Method

The candidate entity generation was executed on the entire dataset of 2,535 filtered medicine descriptions of NF-e items. However, to rigorously assess the subsequent entity disambiguation performed by LLMs, it was necessary to construct a manual gold standard. Because a single description could yield over 100 candidate records, all of which required careful manual evaluation to establish the gold standard, we drew a representative sample of 156 descriptions for this specific phase. 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% ($M_E = 0.1$). This approach ensured that the final sample was sufficiently diverse and statistically robust, while tractable for manual annotation and comprehensive benchmarking of both the candidate entity generation process and the overall entity linking performance.

4.3 Selected Language Models

Five language models were selected for evaluation, balancing local deployment capability with advanced reasoning performance. As our framework processes sensitive public procurement, data confidentiality is critical. Thus, we prioritized prominent, freely available models (DeepSeek-R1, Qwen 2.5, and LLaMA 3.3) that are computationally efficient enough to be executed in local infrastructure. We included leading commercial API-driven models (GPT-4o and Claude 3.7) for comparison purposes. Table 3 lists the models used and their architectural specifications.

Table 3. Models and context windows employed

Model	Version	Type	Arch.	Context
DeepSeek	deepseek-r1:70b	RLM	MoE	128k
Claude	claude-3-7-sonnet-20250219	Hybrid	MoE*	128k
GPT-4o	gpt-4o-2024-11-20	LLM	MoE	128k
Qwen	qwen2.5:72b	LLM	Dense	128k
LLaMA	llama3.3:70b	LLM	Dense	8k

* Inferred from public analyses; not officially disclosed.

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 occurs when the LLM or the I^2E full-text search correctly extracts the medicine name and generates the candidate entities list.
- **NED:** a true positive corresponds to cases where the LLM correctly identifies the corresponding medicine among the candidate entities.

Accuracy was adopted as the primary evaluation metric, calculated as the proportion of true positives relative to the total number of evaluated cases ($T_p/Total$), 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.

[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)

4.5 Candidate Filtering using Heuristics

To evaluate the filtering heuristics (Section 3.2), we utilized 1,000 manually annotated NF-e item descriptions and an extended KG comprising 223,410 Anvisa-approved historical medicine registries. This expanded dataset enables a robust impact analysis across numerous descriptions and larger EL candidate sets.

We assessed two metrics: candidate reduction per description and its impact on downstream LLM disambiguation accuracy. Comparing accuracy before and after filtering reveals whether the heuristics harmfully discard correct candidates or successfully boost performance by reducing noise.

To isolate filtering impacts and prevent variability from external APIs, these experiments strictly employed locally executed small LLMs, as detailed in Table 4.

Table 4. Locally executed models used in the filtering experiments

Model	API model name	Context window
Mistral	mistral:7b	8k tokens
Gemma	gemma3:27b	128k tokens
Qwen	qwen2.5:72b	128k tokens
LLaMA	llama3.3:70b	128k tokens

5 Results

This section first presents the results of the experiments conducted to evaluate the performance of distinct LLMs in the tasks of NER and NED, comparing the NER results with those of I^2E full-text search. Next, we evaluate the impact of the proposed heuristics by comparing the number of EL candidates processed by the LLMs before and after their application.

5.1 NER Results

Table 5 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 medicine and the active ingredient. GPT-4o achieved the highest accuracy for commercial medicine 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. 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 5. NER Accuracy (%)

Model	Medicine 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

Table 6 shows how often the active ingredient was correctly identified when the medicine commercial name was not, and the combined coverage. GPT-4o and Claude 3.7 demonstrated the highest retrieval 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. 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 6. Residual NER Coverage (%)

Model	Active Ingredient	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

5.2 NED Results

Table 7 presents the NED accuracy of each model on the candidate entities generated by the I^2E framework. All models receive the same candidate lists, ensuring a fair comparison of their reasoning and selection capabilities.

Table 7. NED Accuracy Comparison

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

An additional experiment was conducted under varying levels of ambiguity to further assess the robustness of the entity linking process. As shown in Figure 3, the evaluated 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. GPT-4o and Claude 3.7 consistently outperformed the other models across all levels of ambiguity, maintaining high accuracy even when the number of 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

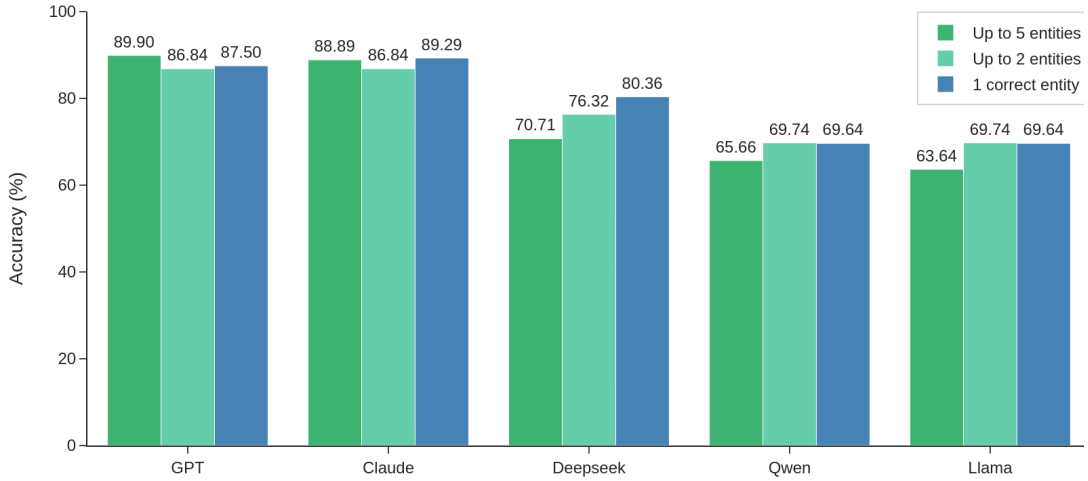

Figure 3. LLMs' NED accuracy for different ambiguity levels.

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.

5.3 Candidate Filtering Heuristics Impacts

Figure 4 shows the reduction in the candidate list size achieved by heuristic 1. These results demonstrate its filtering effectiveness, particularly for outliers. The most severe outlier, which originally had more than 3,500 candidates, had its list reduced to fewer than 2,000 candidates.

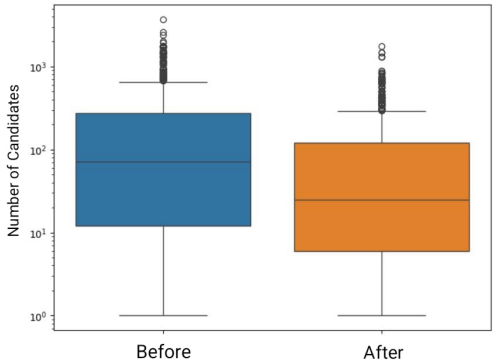

Figure 4. Number of candidates before and after of applying heuristics 1.

Figure 5 presents the reduction achieved by heuristic 2. Although this heuristic produced a noticeable decrease in median values and upper quartiles, its performance was less effective for outliers. In particular, the most extreme cases did not benefit from this strategy: the two most severe outliers showed no reduction at all in their number of candidates, indicating that this heuristic is less sensitive to atypical or irregular descriptions.

Figure 6 compares the overall reduction in candidate list size achieved by the two proposed heuristics. Note that

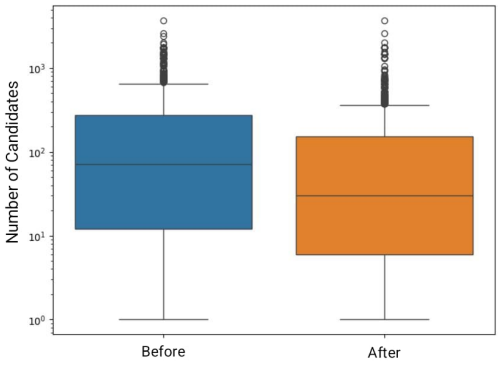

Figure 5. Number of candidates before and after of applying heuristics 2.

heuristic 2 reduced the average number of candidates per record by only 26%, whereas heuristic 1 achieved a substantially higher reduction of approximately 40%. Moreover, heuristic 2 exhibited poorer performance in the most critical cases, particularly for outlier records. This behavior can be explained by the fact that records associated with larger candidate sets often correspond to medicines composed of mixtures of two different active ingredients, whose presentation patterns are more complex and heterogeneous. Such characteristics are harder to capture using highly specific regular expressions, which limits the effectiveness of the measurement unit-based heuristics. However, even the quite modest reductions achieved through these simple heuristics were enough to yield proportional decreases in the number of tokens submitted to language models for NED. This translated into noise reduction, performance gains, and lower costs.

Figure 7 shows the accuracy gains of distinct LLMs on NED after applying the heuristics. We remind the reader that these results of further experiments to assess the effectiveness of the candidate filtering heuristics were obtained using an extended KG and data in which the number of EL candidates per NF-e item description is higher than in previously reported experiments, making the NED problem much more challenging. Therefore, these accuracies are not comparable

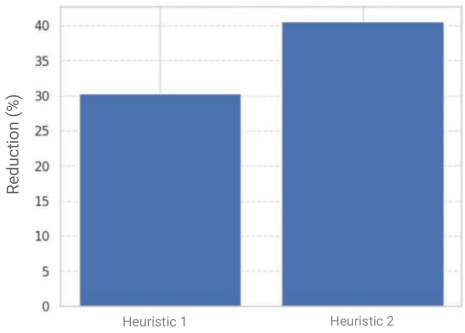

Figure 6. Comparison of candidate reduction between heuristics.

to those reported in Section 5.2, which refer to experiments with more limited KG and data.

Notice in Figure 7 that all evaluated models showed significant gains in accuracy when operating on a reduced candidate set, demonstrating the effectiveness of the filtering stage in enhancing NED results. Notably, gemma3:27b and llama3.3:70b achieved the highest accuracy scores among all evaluated models. It is also important to highlight the strong performance of Gemma, which attained superior results despite having less than half the number of parameters of LLaMA and Qwen, indicating high efficiency for this task.

Conversely, mistral:7b, the smallest model evaluated in terms of parameter count, showed lower performance than the other ones, underscoring the impact of model capacity on disambiguation quality even after candidate filtering.

5.4 Computational Costs

We evaluated I^2E NER costs on an Intel i7-13700 server (64GB RAM). Processing 2k invoices items via Apache Solr took 337.57 seconds. The time expenditure for this approach stems primarily from checking a static dictionary of 34,202 medicines. LLMs (e.g., Claude 3.7, GPT-4o) averaged 385 seconds for the same task. Financially, I^2E requires only an ordinary local server. In contrast, processing our annual volume of 10 million invoices using commercial APIs (Claude 3.7 or GPT-4o) which match I^2E's accuracy, would cost between USD 22,435.90 and USD 30,128.20. Open-source LLMs, on the other hand, have lower performance than required by our applications.

The EL stage increases computational demand due to larger prompts containing candidate lists, taking 20–30 minutes and costing approximately USD 1.30 for 156 descriptions. However, our filtering heuristics reduced candidate lists by up to 40, cutting total execution time by 8 to 12 minutes. This substantial reduction demonstrates the filtering stage's critical role in enabling scalable, cost-effective hybrid EL deployments.

5.5 Discussion

The competitive NER performance of I^2E demonstrates that cost-effective solutions can effectively substitute the extensive use of LLMs. The lower performance of most LLMs in recognizing active ingredients likely occurs because these are frequently omitted in invoice descriptions, appearing mostly in generic medicine purchases. Conversely, our EL results confirm the LLMs' strength in handling variability and semantic ambiguity (e.g., dosage and packaging) during disambiguation. The substantial performance gap between frontier models (GPT-4o, Claude) and open-weight alternatives ($\sim$70B versions of DeepSeek, Qwen, and LLaMA) suggests that massive scale and advanced reasoning are crucial for interpreting noisy real-world texts.

To explicitly connect our scientific claims with the experimental findings, we highlight three key points. First, our claim that our hybrid pipeline offers an efficient and much cheaper solution than pure LLMs is supported by the execution time and cost analysis presented in subsection 5.4. Second, the claim that index-based candidate generation maintains effectiveness is validated by our NER evaluation, showing I^2E (88.46% accuracy) performing comparably to GPT-4o (91.03%). Finally, our claim that domain-specific heuristics improve efficiency without sacrificing accuracy is corroborated by the filtering experiments, which demonstrated up to a 40% reduction in candidate lists.

5.6 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 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 Entity Linking (EL) approach that combines efficient full-text indexes and domain knowledge for candidate generation, augmented with regex-based heuristics for filtering, delegating only disambiguation to LLMs. The framework is adaptable to other domains by adjusting the KB, heuristics, and prompt templates. Evaluations on public medicine purchases showed that I^2E achieves NER accuracy comparable to GPT-4o. Moreover, various LLMs efficiently disambiguated these pre-filtered candidates with accuracies approaching 90%.

Ultimately, our hybrid pipeline can leverage LLMs more effectively for EL. While our index-based search provides an efficient alternative for the initial NER stage, our core contribution is the overall EL framework. By offloading candidate generation to transparent symbolic methods, we achieve state-of-the-art effectiveness while drastically reducing the computational costs of pure LLM-based solutions.

Future work includes extending I^2E to other domains (e.g., vehicles, clinical documents) and exploring alternative indexing strategies to better capture lexical variations. We also plan to refine our heuristic-filtering mechanisms

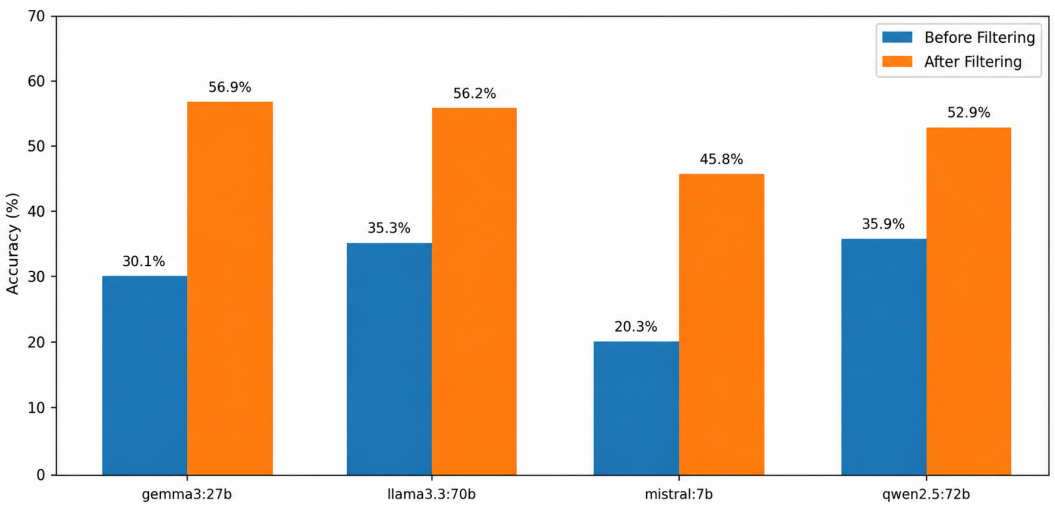

Figure 7. LLM Accuracies on EL before and after applying heuristics.

[Azevedo et al., 2026], investigate deductive reasoning in knowledge graphs, and incorporate few-shot or in-context learning to enhance LLM disambiguation in highly noisy scenarios.

Acknowledgements

This work has been supported by FAPESC grant 2021TR1510, the Céos project financed by the Public Ministry of Santa Catarina State (MPSC), and the National Institute of Science and Technology in Artificial Intelligence Responsible for Computational Linguistics, Information Processing and Dissemination (INCT-TILD-IAR).

Competing interests

The authors declare that they have no competing financial interests or personal relationships that could influence the work reported.

Availability of data and materials

The datasets generated and/or analyzed during the current study are not available due to a non-disclosure agreement regarding the electronic invoices of Santa Catarina State public purchases. However, the data from Anvisa and CMED are public and we are currently preparing datasets for similar experiments using disclosed data of federal public purchases.

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*.

Albuquerque, D. L., Santos, V. S., Nack, P., Fileto, R., and Dorneles, C. F. (2025). Language models are not a panacea: Combining them with domain knowledge and efficient indexes for entity linking. In *Simp. Brasileiro de Banco de Dados (SBBD)*, pages 479–492. SBC.

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 Inter. Conf. 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.

de Almeida, V. A., da Silva, S. S., Chire, J., Vicenzi, L., Borges, N. H., Kociolek, H., de Castro Rocha, S. M., Gomes, F. N., Ferreira, J. C., Marques, O., and e Oliveira, L. E. S. (2026). Clinical named entity recognition in the portuguese language: a benchmark of modern bert models and llms.

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.

Hepp, M. (2008). Goodrelations: An ontology for describing products and services offers on the web. In *Knowledge Engineering: Practice and Patterns: 16th International Conference, EKAW 2008, Acitrezza, Italy, September 29-October 2, 2008. Proc.16*, pages 329–346. Springer.

Islam, K. S., Nipu, A. S., Wu, J., and Madiraju, P. (2025). Llm-based prompt ensemble for reliable medical entity recognition from ehrs. In *2025 IEEE International Conference on Information Reuse and Integration and Data Science (IRI)*, pages 162–167. IEEE.

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.

Musen, M. A. (2015). The protégé project: a look back and a look forward. *AI matters*, 1(4):4–12.

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.

Oliveira, L. E. S. e., Peters, A. C., da Silva, A. M. P., Gebeluca, C. P., Gumiela, Y. B., Cintho, L. M. M., Carvalho, D. R., and Hasan, S. A. (2022). Semclinbr–a multi-institutional and multi-specialty semantically annotated corpus for portuguese clinical nlp tasks. *Journal of Biomedical Semantics*, 13(1):11. DOI: 10.1186/s13326-022-00269-1.

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.

Richter-Pechanski, P., Seiferling, M., Kiriakou, C., Schwab, D. M., Geis, N. A., Dieterich, C., and Frank, A. (2025). Medication information extraction using local large language models. *Journal of biomedical informatics*, page 104898.

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.

Ruas, P., Conceição, S., and Couto, F. M. (2026). Systematic review of named entity linking and knowledge organisation systems in biomedical and clinical domains. *ACM Comput. Surv.*, 58(10). DOI: 10.1145/3796222.

Santos, V. and Dorneles, C. (2024). Unveiling the segmentation power of llms: Zero-shot invoice item description analysis. In *Anais do XXXIX Simp. Brasileiro de Bancos de Dados*, pages 549–561, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/sbbd.2024.240820.

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. DOI: 10.1109/TKDE.2021.3117715.

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, L., Ma, Y., Bi, W., Lv, H., and Li, Y. (2024). An entity extraction pipeline for medical text records using large language models: analytical study. *Journal of Medical Internet Research*, 26:e54580.

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 Inter. Conf. on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Xiao, G., Lanti, D., Kontchakov, R., Komla-Ebri, S., Güzel-Kalaycı, E., Ding, L., Corman, J., Cogrel, B., Calvanese, D., and Botoeva, E. (2020). The virtual knowledge graph system ontop. In *Inter. Semantic Web Conference*, pages 259–277. Springer.

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*.

Yu, A., Weile, J., and Courtot, M. (2026). Benchmarking llm-based information extraction tools for medical documents. *medRxiv*, pages 2026–01.