Journal of Information and Data Management, 2025, 16:1, doi: 10.5753/jidm.2025.4298

PromptNER: An Automatic Prompt-Learning Data Labeling Approach for Named Entity Recognition on Sensitive Data

Claudio M. V. de Andrade [**Universidade Federal de Minas Gerais** | ***mailto:claudio.valiense@dcc.ufmg.br***]
Fabiano Muniz Belém [**Universidade Federal de Minas Gerais** | ***mailto:fmuniz@dcc.ufmg.br***]
Celso França [**Universidade Federal de Minas Gerais** | ***mailto:celsofranca@dcc.ufmg.br***]
Marcos Carvalho [**Universidade Federal de Minas Gerais** | ***mailto:marcoscarvalho@dcc.ufmg.br***]
Marcelo Ganem [**Universidade Federal de Minas Gerais** | ***mailto:marceloganem@dcc.ufmg.br***]
Gabriel Teixeira [**Universidade Federal de Minas Gerais** | ***mailto:gabrielmedeiros@dcc.ufmg.br***]
Gabriel Jallais [**Universidade Federal de Minas Gerais** | ***mailto:gabrieljallais@dcc.ufmg.br***]
Alberto H. F. Laender [**Universidade Federal de Minas Gerais** | ***mailto:laender@dcc.ufmg.br***]
Marcos A. Gonçalves [**Universidade Federal de Minas Gerais** | ***mailto:mgoncalv@dcc.ufmg.br***]

Departamento de Ciência da Computação, Universidade Federal de Minas Gerais, Av. Antônio Carlos, 6627 - Pampulha, Belo Horizonte, MG, 31270-901, Brazil.

Received: 23 March 2024 • **Published:** 20 January 2025

Abstract We address the task of Named Entity Recognition (NER) for entities of the types *Organization* and *Product/Service* found in textual complaints recorded on Web platforms. Due to the high inference power of Large Language Models (LLM's), there is a growing interest in applying them to distinct problems. However, they face issues of high infrastructure cost and privacy concerns when using external API's. Accordingly, in this article we propose **PromptNER**, an approach that uses LLM's for the recognition of entities in consumers' complaints and use them to locally train simpler models, such as *SpERT* (Span-based Entity and Relation Extraction Transformer), to address the task of entity and relation extraction, achieving scalabilty and privacy. Our PromptNER enhanced model achieves significant gains, between 41%-129% in F-score compared to the SpERT model trained with manually-labeled data and between 30%-268% over recent (zero-shot) Large Language Models (Llama 3.1).

Keywords: Automatic Labeling, Named Entity Recognition, Prompt Learning, Large Language Models, Sensitive Data.

1 Introduction

In Brazil, every day thousands of consumer reports are logged on platforms such as Consumidor.gov.br[1] and Procon[2], reporting complaints about products and services. Recognizing entities of specific types such as *Organizations* and *Products* or *Services* in these reports is essential for regulatory bodies to take appropriate measures to protect consumer rights. However, due to the large volume of data recorded on these platforms, manual extraction of such information is impractical. In this context, Named Entity Recognition (NER) is a task aimed at automatically identifying the entities that are mentioned in a text and classifying them in predefined categories (e.g., *Organization*, *Product* or *Service*) [Belém *et al*., 2023].

In this article, we address a particularly challenging real-world NER scenario coming from the Public Prosecutor's Office of Minas Gerais (Ministério Público de Minas Gerais or, in short, MPMG), where it is necessary to recognize named entities classified as Organization and Product or Service types from a set of consumer complaints without any previous manual labeling. Since such entities aggregate important information contained in texts, NER is a useful task in various applications such as record deduplication [de Carvalho *et al*., 2006; Silva *et al*., 2019; Mangaravite *et al*., 2022], data integration [Brunner and Stockinger, 2020], knowledge base construction [Niu *et al*., 2012], search in unstructured text collections [Caputo *et al*., 2009; Rodrigues *et al*., 2022] and text classification [Cunha *et al*., 2020; Constantino *et al*., 2022, 2023; de Andrade *et al*., 2023].

Transformer-based approaches such as Bidirectional Encoder Representations from Transformers (BERT) [Devlin *et al*., 2019] constitute the state-of-the-art in NER methods. Among these approaches, the discriminative model SpERT (Span-based Entity and Relation Transformer) stands out, thus being adopted in our solution. SpERT relies on semantic (i.e., contextual) aspects of word sequences, leading to high effectiveness across various reference datasets with relatively low computational cost.

However, one of the major challenges of discriminative methods is their dependency on a relatively large amount of labeled (training) data, which is absent in many real-cases, including the dataset provided by the MPMG. Obtaining labeled data is typically done manually, incurring on high financial/labor costs in terms of human resources (working hours) with the potential introduction of noise due to hard-to-label cases.

In this article we address this challenge by leveraging a

[1] `https://www.consumidor.gov.br`
[2] `https://www.procon.pr.gov.br`

[Texto]: Estou recebendo cobranças da Claro de produtos que não reconheço. Vide tela em anexo. CPF: xxx.xxx.xxx-xx
[Organização]: Claro
[Texto]: Estou a pouco tempo no mercado livre , minha primeira venda e já estou com problema . o mercado livre suspendeu a minha conta , sem motivos ! preciso que resolvam meu problema o mais rápido possível. indignada com a plataforma !
[Organização]: mercado livre

Figure 1. Example of a prompt provided and generated by BLOOM. The black text was provided to BLOOM, while the red text was generated by it.

machine learning approach based on prompt-learning [Luo *et al*., 2022; Ye *et al*., 2023]. This relatively new approach leverages the power of large-scale language models (LLM's), such as ChatGPT[3], BLOOM[4] and Llama[5]. Specifically, this approach employs a prompt (instruction, phrase, or initial text) presented to the LLM to initiate a conversation or perform a specific task.

In the context of the NER task, prompts can be used to identify entities in texts of complaints. For instance, when employing an LLM to identify organization names, we can provide some examples of labeled complaints (indicating where the entities are identified), followed by an unlabeled complaint. The LLM then infers the entities from the examples in the unlabeled complaint. Figure 1 presents a prompt request submitted to BLOOM in which the words "mercado" and "livre" (composing the name of a popular e-commerce Brazilian platform) are not manually labeled, thus ending the request at "[Organization]:". As a response, the generative model returns the complete text (in red) with the term "mercado livre", automatically labeling the given organization.

The proposed approach can be divided into three steps as depicted in Figure 2. The first step, "Complaints Crawling", aims to collect complaints related to the use of the Consumidor.gov.br platform, whose texts do not include any marking of the entities being complained about. The second step, "Prompt-Learning", utilizes a generative model to identify the entities mentioned in the complaint. Finally, the third and last step executes a Fine-Tuning process aimed at adjusting a state-of-the-art model in the NER task to the specific domain of complaints related to organizations and products or services. At the end of these three steps, we have a model capable of effectively labeling complaints, while avoiding additional costs and dangers of calling external LLMs with sensitive data.

In summary, this article aims to address the following research questions:

RQ1: What is the level of agreement among evaluators regarding the labeling of consumer reports using a generative model?

RQ2: What is the effectiveness of the automatic labeling obtained from LLM's such as BLOOM?

RQ3: How does our solution compare to state-of-the-art baselines, including SpERT tuned using only manually labeled data, and prompt-oriented (zero-shot) LLMs (in case Llama 3.1)?

[3]`https://chat.openai.com`
[4]`https://huggingface.co/bigscience/bloom`
[5]`https://llama.meta.com/`

To answer RQ1, we measure the amount of agreement among evaluators, as quantified by the Krippendorff's alpha coefficient [Gwet, 2011]. For RQ2, we compute precision, recall and F1-scores for each entity category, and discuss representative correctly and incorrectly classified examples. Finally, to answer RQ3, we compare two scenarios: one in which it is possible to manually label only a few examples (scenario 1), and another one in which data is labeled using a prompt-learning based approach (scenario 2). We compare with two baselines, both simulating a data scarcity scenario (scenario 1). In the first, we use SpERT with a few labeled examples to simulate data scarcity, and in the second, we use the latest LLM from META, Llama 3.1 to recognize the entities present in the dataset. Our solution differs from Llama 3.1 as it uses automatically labeled data obtained from the prompt-learning process, resulting in a fine-tuned model that can be applied in sensitive data environments. The LLM, in contrast, is used in a zero-shot mode, without tuning, using the promt only to guide the format of the output. Our final PromptNER solution is a discriminative model specifically tailored to recognize entities in manifestations.

For RQ1, we found that the agreement is 0.53 and is sufficient to validate the results. For RQ2, we noted that the LLM achieves better effectiveness in the organization entity due to a more restricted set of possibilities compared to product/services. For the final research question (RQ3), we compared two baselines that consider data scarcity with our solution and found that PromptNER achieves much higher effectiveness than the baselines, with improvements varying between 30%-268% in F-score.

In summary, the main contributions of this article are:

1. Proposal and evaluation of an automatic named entity labeling based on LLM's;
2. Proposal of more economical, scalable, and privacy-preserving models for NER through a fine-tuning process using automatically labeled data;
3. Application and evaluation of the proposed stages in the recognition of organizations, products, and services in consumer reports, with results demonstrating the benefits of the proposed methods.

The remainder of this article is organized as follows. Section 2 presents related work, while Section 3 describes the NER task addressed in this work. Section 4 describes the proposed strategies while 5 details the evaluation methodology. Experimental results are presented and discussed in Section 6. Finally, Section 7 concludes the article and points out some directions for future work.

2 Related Work

This section provides an overview of related work on the problem of Named Entity Recognition (NER), focusing on two specific types of approach: discriminative and generative. The main difference between these two types lies in the fact that generative approaches can generate texts from a request or an established pattern, while discriminative ones classify sequences of words or tokens in a sentence.

2.1 Discriminative Approaches

Discriminative approaches for Named Entity Recognition (NER) can be classified into two types: *token-based* and *span-based*. In the token-based approach, each word token in the text is classified according to the considered entity types, thus determining whether such a word occurs at the beginning, middle or end of the identified entity designation [Finkel *et al*., 2005; Patil *et al*., 2020]. Conversely, span-based strategies first identify all spans (sequences of tokens) smaller than a given limit and then classify each one of them [Eberts and Ulges, 2020; Fu *et al*., 2021; Liu *et al*., 2021]. Examples of token-based methods traditionally used in NER tasks are those based on Conditional Random Fields (CRF's) [Finkel *et al*., 2005; Patil *et al*., 2020]. CRF's are probabilistic models that use features of a particular token t in a text (such as patterns of uppercase and lowercase letters) and features of tokens adjacent to t to infer the category of each token.

Among span-based strategies, the Span-based Entity and Relation Transformer (SpERT) stands out as a state-of-the-art neural architecture for NER tasks [Eberts and Ulges, 2020]. SpERT encodes text spans into a vector representation based on pre-trained models such as Bidirectional Encoders from Transformers (BERT) [Devlin *et al*., 2019], classifying them into predefined categories of entities or as a "non-entity". The algorithm also represents pairs of spans as entities in the vector space and assigns them to predefined categories of relationships. Eberts and Ulges [2021] expanded the SpERT architecture to include clustering of mentions referring to the same entity in different segments of a text. Finally, Belém *et al*. [2022] proposed contextual reinforcement techniques and an entity delimitation strategy based on pre and post-processing of data in official documents such as judicial proceedings.

Finally, Ji *et al*. [2020] present an approach to jointly extract entities and relations from texts, which uses transformer-based models, specifically BERT, to encode text spans into vector representations used to classify entities and their relationships. They highlight the need for a span-based approach that considers the complete context in which entities appear to overcome the limitations of token-based approaches. Additionally, a span-specific attention model and contextual semantic representations are proposed, enabling a better understanding of the context of entities and their relationships.

2.2 Generative Approaches

Large Language Models (LLM's) have provided great advances in several NLP tasks, such as text classification [de Andrade *et al*., 2023] and text generation, specially in scenarios in which the amount of training data is very limited, known as zero-shot and few-shot learning [Brown *et al*., 2020]. The main reason of the predictive power of these models is their large scale when compared to previous ones. For example, GPT-3 presents 175 billion parameters, which is more than 100 times larger than its precursor GPT-2. GPT-4, the newest model created by OpenAI, goes even further and presents more than 1.75 trillion parameters [Liu *et al*., 2023]. Finally, ChatGPT has demonstrated its potential in various fields, including education, healthcare, reasoning, text generation, human-machine interaction, and scientific research. The reason for ChatGPT's success is that, while maintaining a lower number of parameters in comparison with GPT-3, it continually improves by exploiting Reinforcement Learning (RL) based on human feedbacks [Ouyang *et al*., 2022; Paiva *et al*., 2022; Christiano *et al*., 2023].

While the aforementioned examples of LLM's, particularly ChatGPT and GPT-4, currently produce top-notch results, they are not open-source. There are, however, many efforts to provide open source alternatives, such as LLaMA [Touvron *et al*., 2023], a collection of foundation language models ranging from 7B to 65B parameters, and BLOOM (BigScience Large Open-science Open-access Multilingual[6]), similar to GPT-3 in size.

All of the aforementioned models, open-source or not, can overcome the limitation of smaller traditional language models, which depends directly on the availability of a large amount of labeled data for the task at hand, including NER, object of the present work.

A possible strategy to employ LLM's to improve NER is to generate new instances of data that can be used to enhance discriminative learning. Among generative models, BLOOM (BigScience Large Open-science Open-access Multilingual), LLaMA (Open and Efficient Foundation Language Models), and ChatGPT (Chat Generative Pre-trained Transformer) emerge as possible alternatives for generating synthetic data for the NER task.

Given the increasing popularity of generative models and their effectiveness in various tasks, there is a growing interest in using these models to generate synthetic data quickly, cheaply, and possibly with good quality. However, generative models like ChatGPT and BLOOM are not sequence labelers like the best known models for the NER task and may produce hallucinations, i.e., label entities that are not present in the text. To address these limitations, studies such as those by Wang *et al*. [2023] propose techniques to leverage the power and size of LLM's as alternatives to discriminative models. Conversely, Tang *et al*. [2023] analyze the use of generative models as auxiliary tools rather than reliable sources for labeling clinical texts for NER tasks, questioning these models´ reliability regarding the quality of the generated responses and the protection of sensitive data.

In a recent work, Belém *et al*. [2022, 2023] proposed a data augmentation strategy based on generative AI to expand a training dataset by generating synthetic data containing entities and relations. In the present article, we tackle this problem differently by proposing an automatic labeling

[6] https://arxiv.org/abs/2211.05100

strategy of real unlabeled training data as a first step and then using the automatically labeled data for training discriminative approaches. Furthermore, in comparison to Belém *et al*. [2022, 2023], which dealt with official diaries, here we focus on a different domain (organization, products and services in consumer complaints), while we also perform an analysis of the agreement level of human evaluators.

These studies indicate a trend in exploring the potential of generative models as alternatives to discriminative models, as well as a way to perform data augmentation, allowing tasks using domain-specific NER models with scarce labeled data to be enhanced at a relatively low cost. Despite their promise, generative approaches do not yet constitute substitutes for discriminative approaches due to their high infrastructure cost and the privacy concerns discussed earlier [Zhang *et al*., 2022].

In sum, previous work has tackled the NER problem either using large generative models to directly extract the entities or by using smaller discriminative models [Tang *et al*., 2023]. Isolated, the former strategy presents high infra-structure costs or privacy concerns, while the latter requires a large amount of labeled data. Here, we join the best of each world, by first exploiting the generative approach to automatically label training data and then providing the resulting labeled data as training for discriminative approaches.

3 The Named Entity Recognition Task

The Named Entity Recognition (NER) task involves extracting and classifying entities mentioned in texts, allowing them to be distinguished among categories such as Person, Location, Organization, CPF (Individual Taxpayer Registry), CNPJ (National Registry of Legal Entities), and Phone Number, among others. Typically, the NER task processes unstructured text [Belém *et al*., 2011], i.e., text that does not have explicit indications of which of its tokens are named entities or not [Yadav and Bethard, 2018].

An example of the NER task, within the domain of consumer complaints focusing on the categories Organization and Product/Service, would be to identify, given the text "Eu comprei um celular da empresa A Ltd., mas a cobertura do serviço é terrível. Isto nunca aconteceu com meus outros celulares!" (in english, "I bought a mobile phone from Company A Ltd., but the service coverage is terrible. This never happened to me with other phones!"), the term "celular" as an entity of type Product/Service and the term "empresa A Ltd." as an entity of type Organization. The task of relation extraction, although addressed by the SpERT method, will not be covered within the scope of this article.

Table 1 presents examples of real complaints made on the **Consumidor.gov.br** platform. The Complaint column lists the complaint texts, and the Entity column indicates the complained entities. The first two rows of the table show complaints regarding entities of type Organization, while in the last two rows, the examples are related to entities of type Product/Service.

4 Proposed Architecture

This section presents the proposed solution for the NER problem addressed in this article. Based on the objectives and research questions outlined before, the following strategy was developed for addressing it. First, we describe the data collection process that will be used in subsequent steps. Next, we define a prompt structure for data labeling by BLOOM. An automation process was then developed to extract labels generated by BLOOM from a given set of database samples, formatted to match the input requirements for SpERT model training sets. The labels produced by BLOOM were then sample-evaluated, followed by training and evaluation of the metrics obtained with the SpERT model using the automatically labeled data.

Figure 2 presents an overview of the proposed PromptNER approach for named entity recognition (NER). The next subsections detail each stage of its architecture.

4.1 Data Collection

The collection of data (complaints) was performed using a web crawling process on the **consumidor.gov.br** platform, a public service that enables direct communication between consumers and companies to resolve consumer disputes through the Internet. In this direct communication, there is no intervention by the Public Power in individual dealings, making it a faster and more practical process. It is worth noting that all complaint data feed a public database, including information on best solution rates and satisfaction by company.

Through the **Consumidor.gov.br** platform, it was possible to obtain a set of complaints without any marking of the entities complained about in the text. The unstructured data can be found in the Consumer Report section of the platform. A total of 378,574 consumer reports linked to the state of Minas Gerais were collected.

4.2 Prompt-Learning

The second part of the architecture is similar to the pattern described in Section 1, where a small set of manually labeled examples is included in the prompt, followed by a new example for which the label produced by the generative model is sought. For the data presented here, four labeled examples were included, followed by the unlabeled example for which the entities are to be extracted.

In this stage, we used the generative model BLOOM, through the inference API provided by Hugging Face[7]. Most LLMs are developed by resource-rich organizations and are often kept out of public reach. We used BLOOM, an openly accessible language model with 176 billion parameters, designed and built thanks to the collaboration of hundreds of researchers. Specifically, we sent a request to BLOOM similar to that described in Figure 1, obtaining the completed data with the label, which is then saved for use in the next stage. It is worth noting that a filtering process occurs: labels generated by BLOOM that are not contained in the original text of the presented sample are discarded, and the sample is

[7] `https://huggingface.co/bigscience/bloom`

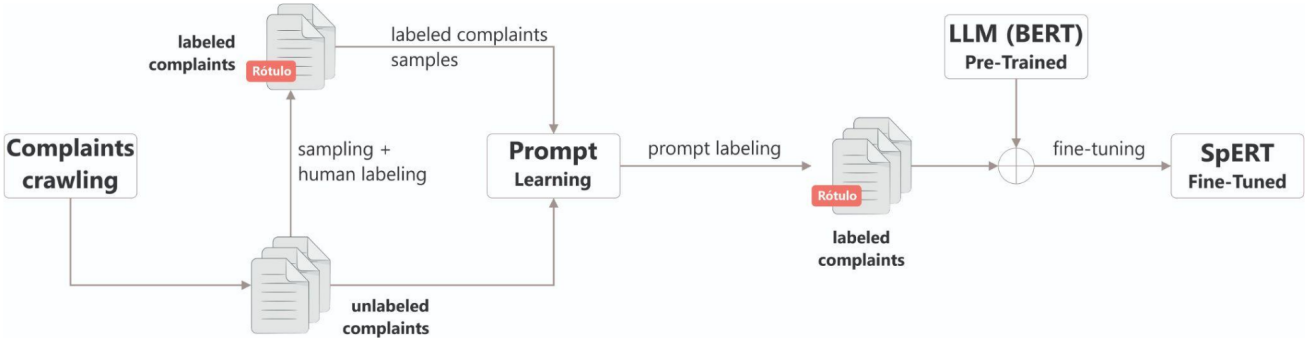

Figure 2. Proposed approach.

Table 1. Examples of complaints in the consumer.gov.br platform. In bold the claimed entity.

Complaint	Entity
I've only been in the **mercado livre** for a short time, my first sale and I'm already having a problem. Mercado Livre suspended my account for no reason! I need you to resolve my problem as quickly as possible. outraged by the **mercado livre** platform.	mercado livre
I'm receiving charges from **Claro** for products I don't recognize. See attached screen. CPF: xxx.xxx.xxx-xx	Claro
I purchased a **built-in electric oven** directly on the company's website, on XXXX, order no. For months I've been trying to get a visit from technical support to identify and repair the defect in the product, people call me, but never come. I am very disappointed, my loss and annoyance is incalculable.	Built-in electric oven
I had a **tim plan** and they offered me another plan, then they informed me that they would exempt me from the June bill, which was for me to call to cancel my plan and inform me of the discount they would give me on the June bill, when I called cancel, they informed me that this request would not be made.	TIM Plan

not used. This is a disadvantage of the generative model, as it does not allow us to limit the generation of tokens present in the input text.

A sample of the collected data is presented to evaluators to validate the quality of the labels produced by BLOOM and filtered by automation. Once validated, the data are provided to the SpERT model as a training set, and after training (a fine-tuning process of the model), evaluation metrics are calculated to assess the effectiveness of the proposed flow.

4.3 Fine-Tuning

The SpERT model is a popular and widely used model in the task of identifying entities in unstructured texts. Similar to other transformers, it allows for the process of adjusting its weights (fine-tuning), allowing the model to be adapted to the domain of the problem, in our case, recognizing organizations and products/services from consumer reports. The trained model is saved and can be used for entity recognition in new data sets without the need to send data (potentially confidential) to external environments.

Finally, but not least, the fine-tuning process of the SpERT can be done locally without transferring data, which is important for the privacy and security of sensitive data, such as those received through public agencies such as Procon.

5 Evaluation Methodology

This section describes the evaluation methodology adopted in this article. We present the data collections used (Subsection 5.1), how human evaluation was conducted (Subsection 5.2), the metrics used in the experimental evaluation (Subsection 5.3), and the parameterization of the SpERT method (Subsection 5.4).

5.1 Data Collections

From the collection of 378,574 complaints (Section 4.1), 7,858 of them were labeled by the generative model BLOOM. Table 2 presents the quantity of complaints for the entities of types *Organization* and *Product or Service* were labeled by BLOOM. After the automatic labeling process, the labeled data were provided as input for SpERT.

Table 2. Statistics of the Consumidor.gov.br dataset

Entity category	Amount of complaints
Organization	3,129
Product or Service	4,729

For evaluation, we used the 5-fold cross-validation process, where in each fold the data are divided into 5 partitions. Three of these partitions constitute the training set, which contains both the few examples that need to be manually labeled and the examples labeled automatically through the prompt in the LLM. These labeled data are used to adjust the model weights (fine-tuning). Another partition is used as the validation set for model parameterization. The last partition, the test set, consists of the data on which the NER is applied for evaluation.

Our methodology compares scenarios where only a few data points are manually labeled (Scenario-1) with the scenario where we have data labeled by prompt learning (Scenario-2). We used 100 manifestations (the total number of manually labeled data available) as training data in Scenario-1 and 7,858 in Scenario-2. In both scenarios, we kept the same test set.

5.2 Human Evaluation

From the complaints labeled through our prompt-based approach with the LLM BLOOM, we conducted a manual inspection with three evaluators analyzing a sample of 100 complaints, with 50 related to *Organization* entities and 50

related to *Product/Service* entities. We provided each evaluator with a spreadsheet containing the 100 complaints to assess whether the generative model was able to identify the complained entity, so that one evaluator did not have access to the responses of the others. It is worth noting that this is a costly task for a small work team, especially given the freeform writing that consumers use.

5.3 Evaluation Metrics

For evaluating the results, we used Precision, Recall, and F1-score, which capture different aspects of entity recognition effectiveness. Considering x as an entity type (e.g., x = Person or x = Organization), the precision of the algorithm to recognize entities of type x is calculated by the number of correct predictions divided by the total number of times the algorithm recognized type x. Recall shows how much the algorithm managed to cover mentions of entities of type x. Finally, the F1 score for x is defined as the harmonic mean between Precision and Recall. In our results, we present the F1 measure for each entity category and the average among them (Macro-F1).

5.4 SpERT Parameterization

For parameterizing the SpERT strategy, we used the values recommended by its authors [Eberts and Ulges, 2020]: a learning rate defined as $l_r = 5 \times 10^{-5}$, number of epochs $t = 20$, number of negative examples per sentence n^- = 100 (both for entities and relations), and batch size b_s = 2.

6 Experimental Results

This section presents the experimental results aimed at addressing the three research questions posed.

6.1 RQ1: What is the level of agreement among annotators regarding complaint labeling using LLM's?

In a sample of 50 complaints regarding the entity "Organization" the three evaluators agreed that the *BLOOM* successfully identified the respective organization in 43 of them. Regarding the recognition of the entity "Product/Service" out of 50 complaints, the *BLOOM* successfully identified the product or service in the text for 28 of them. In Section 6.2, we further discuss the differences in complexity between the two entity labels (Organization and Product/Service).

We employed the Krippendorff's alpha coefficient, which measures the level of agreement among annotators and is widely used when the number of annotators is greater than two [Akter and Wamba, 2016; Fabbri *et al*., 2021]. The scale of this metric ranges from -1 (maximum disagreement) to 1 (unanimity).

The Krippendorff's alpha coefficient obtained in our analysis was 0.53, suggesting an agreement among annotators sufficient to validate the results but still far from unanimity (represented by the value 1), indicating that manual labeling is a complex task prone to errors [Zhu *et al*., 2023]. This motivates solutions that rely on fewer manually labeled data, such as the proposed PromptNER approach.

In sum, when addressing RQ1, we found that generative approaches, represented by BLOOM, showed a relatively high level of agreement among annotators. The agreement was measublue using Krippendorff's alpha coefficient, which indicated a moderate level of agreement among annotators. This suggests that while the process of manual labeling presents a higher cost and remains prone to errors, the agreement achieved on automatically labeled samples was sufficient to validate the results obtained from the system.

6.2 RQ2: What is the effectiveness of labeling obtained from LLM's like BLOOM?

Table 3 presents the effectiveness of BLOOM in terms of precision, recall, and F-score per entity type from a sample containing 100 complaints and considering the assessments of three evaluators. BLOOM achieves an F-score of 0.83 for the "Organization" entity and 0.56 for "Product/Service," resulting in a Macro-F1 of 0.695.

We can observe superior effectiveness regarding the "Organization" entity compared to "Product/Service". This is probably due to the fact that terms associated with the organization occur in a more defined context followed by prepositions such as "to" "of" and "in". Another argument is that identifying "Product/Service" in some cases requires knowledge of the associated organization to know the services it provides. In the example prompt shown in Figure 3, knowledge of the organization **Correios** is necessary to identify the term "logística reversa" (reverse logistics) as a service offeblue. Another point is that the set of entities of the "Organization" type is more restricted than that of "Product/Service," leading the BLOOM model to have a larger coverage of examples of type "Organization" in its training process.

Finally, the complexity of identifying and correctly delimiting products and services in a text, even for humans, is higher than performing the same task for entities of type "Organization". For example, in the product description "ASUS Zenbook Pro 14 Duo OLED (UX8402) 16GB RAM", it is arguably acceptable to include or not additional attributes of the product such as the model and the memory as part of the identified entity. This also causes a lower level of agreement by human annotators as discussed in Section 6.1.

Table 3. Precision, Recall and F-Score in BLOOM sample

Type	Precision	Recall	F-Score
Organization	0.86	0.80	0.83
Product / Service	0.56	0.56	0.56

In sum, regarding RQ2, our results indicated that BLOOM achieved a higher level of effectiveness in identifying entities related to "Organization" compared to "Product/Service" This difference in effectiveness can be attributed to several factors, including the context in which the entities appear in complaints and the complexity of identifying products or services without knowledge of the associated organization, and intrinsic higher difficulty of labeling "Product/Service" entities. Furthermore, the results suggested that the BLOOM model benefited from a more restricted set of entities for

[Texto]: Fui ao Correios e não consegui utilizar a logística reversa para devolver meu Iphone-X de volta à Apple
[Produto ou Serviço]: logística reversa
[Produto ou Serviço]: Iphone-X
[Organização]: Correios
[Organização]: Apple

Figure 3. Example of a prompt with organization, product, and service.

"Organization" compared to "Product/Service" which contributed to its effectiveness.

6.3 Comparison with Baselines

Remind that the question we aim to answer in this section is "RQ3: How does our solution compare to state-of-the-art baselines, including SpERT tuned using only manually labeled data, and prompt-oriented (zero-shot) LLM's?´´

As baselines, we used the SpERT and Llama 3.1 models. SpERT training utilizes 100 manually inspected examples, representing the typical situation of a scarcity of manually labeled data, without employing automatic labeling based on LLM's. In LLaMA 3.1 baseline, we used a prompt similar to Figure 4 to provide information on the task and the desiblue output, but we provide no training examples. In other words, Llamma 3.1 is used in a zero-shot mode to simulate the situation in which we do not want the NER to have access to sensitive data. As these Large Language Models are trained with huge amounts of data, we want to assess the extent to which they perform without domain knowledge.

Our PromptNER architecture, on the other hand, utilizes labeling from the generative model, which obtains a larger amount of automatically labeled data at a much lower cost than manual labeling. All models use the same test partition of the 5-fold cross-validation to ensure they are evaluated on the same complaint set, with the difference lying in the training partition.

We present the results of both baselines and PromptNER in Table 4. We start by comparing SpERT and PromptNER in recognizing the "Organization" and "Product/Service" entity. The PromptNER architecture achieved an improvement in Precision of 41.0% (0.39 vs. 0.55), 37% in Recall (0.51 vs. 0.70), and 41% in F1-score (0.44 vs. 0.62) compared to SpERT, highlighting the importance of using the generative model for automatic labeling. In recognizing "Product/Service," the results were more impactful, achieving an improvement in precision, recall, and F-score of 168.7% (0.16 vs. 0.43), 80.0% (0.30 vs. 0.54), and 128.6% (0.21 vs. 0.48), respectively.

In a comparison between Llama 3.1 and PromptNER, the former is more precise (0.73 vs 0.55) in recognizing "Organizations" but at the cost of a very recall (0.36 vs 0.70), resulting in a lower F-score (0.48 vs 0.62) when compared to PromptNER. One possible explanation for Llama 3.1's superior precision for the "Organization" entity is its extensive pre-training. The training data for these large models is considerable, including the whole Wikipedia as well as data crawled from the Open Web, which often contain specific patterns associated with organizational names. This abundance of data about organizations likely contributes to the model's higher effectiveness in recognizing such entities.

For the "product/service" entity, PromptNER achieves higher effectiveness across all three metrics, demonstrating the benefits of fine-tuning the discriminative model using training data. Our PromptNER enhanced model achieves significant gains, between 30%-268% (F-Score 0.48 vs 0.13) over Llama 3.1. Another interesting point to note from Table 4 is that the confidence intervals for PromptNER results are small relative to the mean, indicating that the result was stable across partitions (folds), and therefore the model is highly generalizable (i.e., robust) to different training and testing partitions.

Table 5 presents examples of SpERT and PromptNER successes and errors regarding the recognition of Products/Services. In lines 1 and 2, only PromptNER correctly identified the complained product (highlighted in blue). Particularly, in line 1, SpERT partially suggested the organization's name "bahia" (related to Casas Bahia), which is not the focus of the product recognition task. In line 2, SpERT suggested the term "celular" (cell phone), which is not the complained product. In line 3, both models make errors —- both suggest the organization in blue—while the correct complained product is highlighted in blue. The complexity of this complaint is apparent— the models would need to relate the term "contestação" (dispute) to "faturas em aberto" (unpaid bills) for proper recognition to occur.

Table 4. Result of Precision, Recall, and F-Score with 95% confidence interval of the SpERT model and PromptNER architecture

Method	Entity	Precision	Recall	F-Score
SpERT	Organization	0.39 ± 0.09	0.51 ± 0.04	0.44 ± 0.05
Llama 3.1	Organization	0.73 ± 0.05	0.36 ± 0.04	0.48 ± 0.05
PromptNER	Organization	0.55 ± 0.03	0.70 ± 0.02	0.62 ± 0.02
SpERT	Product or Service	0.16 ± 0.04	0.30 ± 0.04	0.21 ± 0.03
Llama 3.1	Product or Service	0.25 ± 0.03	0.08 ± 0.02	0.13 ± 0.02
PromptNER	Product or Service	0.43 ± 0.02	0.54 ± 0.03	0.48 ± 0.02

In sum, answering RQ3, we found that PromptNER outperformed SpERT and LLama 3.1 in most metrics for both "Organization" and "Product/Service" entities. These improvements demonstrated the significance of utilizing generative models for data labeling, resulting in a larger amount of labeled data available for domain adaptation at a lower cost compared to manual labeling. The analyzed examples illustrated the successes and errors of the models in recognizing products or services in complaints, highlighting the challenges associated with entity recognition in this domain.

Overall, the experimental results underscore the potential of Language Model-based methods in automating labeling tasks and improving the effectiveness of downstream NER models. They also emphasize the importance of addressing challenges such as context understanding and entity disambiguation [Ferreira *et al.*, 2014] to further enhance effectiveness.

Table 5. Examples of SpERT and PromptNER successes and errors.

Complaint
Good morning, I made a purchase at Casas Bahia of 2 products, a set of blender, mixer, and orange juicer, and a rotating brush, and I canceled the order due to the delay in delivery because they gave me a date and it was not met... I asked for the refund of the amount, they refunded me the amount of the kit and not the rotating brush, and on the website, it states that I received the brush. I contacted them several times, even asked for proof that I had received, the signature of who received the product. They say they will refund the amount and until today nothing. I've already paid all installments.
In 2018 I contracted a postpaid plan for my cell phone number xxx, for a value of 49.90. For some months now, the plan's value has been increasing, currently reaching the amount of R$ 67.00. In the current invoices, they have been charging for SVAs (additional services in the plan) that I did not hire, do not use, and were not informed to me at the time of hiring, and that drastically increase the invoice amount. I called on 10/03/2021 to TIM and spoke to an attendant, who could not explain to me why the invoice amounts are increasing and did not solve anything about my situation.
I went to a store of the telecom operator TIM to contract a plan, but the attendant informed me that there were two unpaid bills under my CPF. I informed her that I do not recognize the number, filled out a handwritten letter and sent it to the operator requesting the dispute of the open amounts and informing that I do not know the line, but until today, nothing has been resolved.

7 Conclusions and Future Work

In this article, we addressed the issue of Named Entity Recognition (NER) on real complaint data collected from the website **Consumidor.gov.br**, aiming at identifying Organizations and Products/Services mentioned in the text. To achieve this, we designed and employed the PromptNER approach, which comprises two stages: (1) automatic labeling of public data based on Large Language Models (LLM's), and (2) application of the labeled data as training data for scalable and secure local models that can be maintained on a simpler and more private infrastructure, without the need for data exchange, thereby eliminating costs and concerns regarding the submission of confidential data to external LLM environments. Our model achieved promising results, with gains ranging from 41% to 129% in F-Score compared to the state-of-the-art model (SpERT) trained with manually labeled data as well as Large Language Models (Llamma 3.1) applied to NER (gains ranging 30 - 268%), with the additional advantage of dispensing with the costly process of manual labeling.

As future work we plan to test with other LLMs such as ChaptGPT and LLamma itself as automatic data labelers. We also plan to evaluate the proposed strategies in other domains, covering not only other entity types, but also addressing the Relation Extraction (RE) task. Finally, we intend to study the impact of the temporal dynamics [Mourão *et al*., 2008; Salles *et al*., 2010, 2017] of complaints in the NER task as seasonal and temporal issues (e.g., major events such as Easter and Christmas) may affect the probability of an organization or product entity to appear in a specific complaint.

Acknowledgements

We would like to thank the Public Ministry of the State of Minas Gerais for its support under the *Capacidades Analíticas* project.

Funding

This work was partially supported by grants from CAPES, CNPq, FAPEMIG, FAPESP and AWS.

Authors' Contributions

Cláudio Valiense, Fabiano Belém and Celso França: Conceptualization, Writing (review & editing), Methodology, Validation. *Marcos Carvalho, Marcelo Ganem, Gabriel Teixeira and Gabriel Jallais*: Writing (review), Methodology, Validation. *Alberto H. F. Laender and Marcos A. Gonçalves*: Conceptualization, Writing (review & editing), Validation, Project Management, and Supervision.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The datasets and the source code of the current study are available at `https://github.com/MPMG-DCC-UFMG/M01`.

References

Akter, S. and Wamba, S. F. (2016). Big data analytics in E-commerce: a systematic review and agenda for future research. *Electronic Markets*, 26(2):173–194. DOI: 10.1007/s12525-016-0219-0.

Belém, F., Martins, E., Pontes, T., Almeida, J., and Gonçalves, M. (2011). Associative tag recommendation exploiting multiple textual features. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, page 1033–1042, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/2009916.2010053.

Belém, F. M., de Andrade, C. M. V., França, C., Carvalho, M., Ganem, M. A. S., Teixeira, G., Jallais, G., Laender, A. H. F., and Gonçalves, M. A. (2023). Contextual reinforcement, entity delimitation and generative data augmentation for entity recognition and relation extraction in official documents. *Journal of Information and Data Management*, 14(1). DOI: 10.5753/JIDM.2023.3180.

Belém, F. M., Ganem, M. A. S., França, C., Carvalho, M., Laender, A. H. F., and Gonçalves, M. A. (2022). Reforço e delimitação contextual para reconhecimento de entidades e relações em documentos oficiais. In *Proceedings of the 37th Brazilian Symposium on Databases, SBBD 2022, Buzios, Brazil, September 19-23, 2022*, pages 292–303. SBC. DOI: 10.5753/SBBD.2022.224650.

Belém, F., Ganem, M., França, C., Carvalho, M., Laender, A. H. F., and Gonçalves, M. (2022). Reforço e Delimitação Contextual para Reconhecimento de Entidades e Relações em Documentos Oficiais. In *Anais do XXXVII Simpósio Brasileiro de Bancos de Dados*, pages 292–303.

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.

Brunner, U. and Stockinger, K. (2020). Entity Matching with Transformer Architectures - A Step Forward in Data Integration. In *Proceedings of the International Conference on Extending Database Technology*, pages 463–473.

Caputo, A., Basile, P., and Semeraro, G. (2009). Boosting a Semantic Search Engine by Named Entities. In *Foundations of Intelligent Systems*, pages 241–250.

Christiano, P., Leike, J., Brown, T. B., Martic, M., Legg, S., and Amodei, D. (2023). Deep reinforcement learning from human preferences.

Constantino, K., Cruz, V., Zucheratto, O., França, C., Carvalho, M., Silva, T. H., Laender, A., and Gonçalves, M. (2022). Segmentação e classificação semântica de trechos de diários oficiais usando aprendizado ativo. In *Anais do XXXVII Simpósio Brasileiro de Bancos de Dados*, pages 304–316, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/sbbd.2022.224656.

Constantino, K., H. P. Silva, T., B. Silva, J. V., L. Cruz, V. A., M. M. Zucheratto, O., Carvalho, M., Santos, W., França, C., M. V. de Andrade, C., H. F. Laender, A., and Gonçalves, M. A. (2023). Using active learning for segmentation and semantic classification of legal acts extracted from official diaries. *Journal of Information and Data Management*, (1).

Cunha, W., Canuto, S. D., Viegas, F., Salles, T., Gomes, C., Mangaravite, V., Resende, E., Rosa, T., Gonçalves, M. A., and Rocha, L. (2020). Extended pre-processing pipeline for text classification: On the role of meta-feature representations, sparsification and selective sampling. *Inf. Process. Manag.*, 57(4):102263. DOI: 10.1016/J.IPM.2020.102263.

de Andrade, C. M., Belém, F. M., Cunha, W., França, C., Viegas, F., Rocha, L., and Gonçalves, M. A. (2023). On the class separability of contextual embeddings representations – or "the classifier does not matter when the (text) representation is so good!". *Information Processing & Management*, 60(4):103336. DOI: https://doi.org/10.1016/j.ipm.2023.103336.

de Carvalho, M. G., Gonçalves, M. A., Laender, A. H. F., and da Silva, A. S. (2006). Learning to deduplicate. In Marchionini, G., Nelson, M. L., and Marshall, C. C., editors, *ACM/IEEE Joint Conference on Digital Libraries, JCDL 2006, Chapel Hill, NC, USA, June 11-15, 2006, Proceedings*, pages 41–50. ACM. DOI: 10.1145/1141753.1141760.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Eberts, M. and Ulges, A. (2020). Span-based Joint Entity and Relation Extraction with Transformer Pre-training. In *Proceedings of the 24th European Conference on Artificial Intelligence*, pages 2006–2013.

Eberts, M. and Ulges, A. (2021). An End-to-end Model for Entity-level Relation Extraction using Multi-instance Learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3650–3660.

Fabbri, A. R., Kryscinski, W., McCann, B., Xiong, C., Socher, R., and Radev, D. R. (2021). Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409. DOI: 10.1162/tacl_a_00373.

Ferreira, A. A., Veloso, A., Gonçalves, M. A., and Laender, A. H. F. (2014). Self-training author name disambiguation for information scarce scenarios. *J. Assoc. Inf. Sci. Technol.*, 65(6):1257–1278. DOI: 10.1002/ASI.22992.

Finkel, J. R., Grenager, T., and Manning, C. (2005). Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370.

Fu, J., Huang, X., and Liu, P. (2021). SpanNER: Named Entity Re-/Recognition as Span Prediction. In *Annual Meeting of the Association for Computational Linguistics*, pages 7183–7195.

Gwet, K. L. (2011). On the krippendorff's alpha coefficient.

Ji, B., Yu, J., Li, S., Ma, J., Wu, Q., Tan, Y., and Liu, H. (2020). Span-based Joint Entity and Relation Extraction with Attention-based Span-specific and Contextual Semantic Representations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 88–99.

Liu, C., Fan, H., and Liu, J. (2021). Span-Based Nested Named Entity Recognition with Pretrained Language Model. In Jensen, C. S., Lim, E.-P., Yang, D.-N., Lee, W.-C., Tseng, V. S., Kalogeraki, V., Huang, J.-W., and Shen, C.-Y., editors, *In Processing of the 26th International Conference Database Systems for Advanced Applications*, pages 620–628.

Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., He, H., Li, A., He, M., Liu, Z., Wu, Z., Zhao, L., Zhu, D., Li, X., Qiang, N., Shen, D., Liu, T., and Ge, B. (2023). Summary of chatgpt-related research and perspective towards the future of large language models. *Meta-Radiology*, 1(2):100017. DOI: 10.1016/j.metrad.2023.100017.

Luo, X., Xue, Y., Xing, Z., and Sun, J. (2022). PRCBERT: Prompt Learning for Requirement Classification using BERT-based Pretrained Language Models. In *Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering*, pages 1–13.

Mangaravite, V., Carvalho, M., Cantelli, L., Ponce, L. M., Campoi, B., Nunes, G., Laender, A. H. F., and Gonçalves, M. A. (2022). DedupeGov: Uma Plataforma para Integração de Grandes Volumes de Dados de Pessoas Físicas e Jurídicas em Âmbito Governamental. In *Anais do XXXVII Simpósio Brasileiro de Bancos de Dados*, pages 90–102. DOI: 10.5753/sbbd.2022.224655.

Mourão, F., Rocha, L., Araújo, R. B., Couto, T., Gonçalves, M. A., and Jr., W. M. (2008). Understanding temporal aspects in document classification. In Najork, M., Broder, A. Z., and Chakrabarti, S., editors, *Proceedings of the International Conference on Web Search and Web Data Mining, WSDM 2008, Palo Alto, California,*

USA, February 11-12, 2008, pages 159–170. ACM. DOI: 10.1145/1341531.1341554.

Niu, F., Zhang, C., Ré, C., and Shavlik, J. W. (2012). DeepDive: Web-scale Knowledge-base Construction using Statistical Learning and Inference. In *Proceedings of the Second International Workshop on Searching and Integrating New Web Data Sources, Istanbul, Turkey, August 31, 2012*, pages 25–28.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. (2022). Training language models to follow instructions with human feedback.

Paiva, B. B. M., Nascimento, E. R., Gonçalves, M. A., and Belém, F. (2022). A reinforcement learning approach for single redundant view co-training text classification. *Inf. Sci.*, 615:24–38. DOI: 10.1016/J.INS.2022.09.065.

Patil, N., Patil, A., and Pawar, B. (2020). Named entity recognition using conditional random fields. *Procedia Computer Science*, 167:1181–1188. International Conference on Computational Intelligence and Data Science. DOI: https://doi.org/10.1016/j.procs.2020.03.431.

Rodrigues, P. H. S., de Sousa, D. X., Rosa, T. C., and Gonçalves, M. A. (2022). Risk-sensitive deep neural learning to rank. In Amigó, E., Castells, P., Gonzalo, J., Carterette, B., Culpepper, J. S., and Kazai, G., editors, *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 803–813. ACM. DOI: 10.1145/3477495.3532056.

Salles, T., Rocha, L., Mourão, F., Gonçalves, M. A., Viegas, F., and Jr., W. M. (2017). A two-stage machine learning approach for temporally-robust text classification. *Inf. Syst.*, 69:40–58. DOI: 10.1016/J.IS.2017.04.004.

Salles, T., Rocha, L., Pappa, G. L., Mourão, F., Jr., W. M., and Gonçalves, M. A. (2010). Temporally-aware algorithms for document classification. In Crestani, F., Marchand-Maillet, S., Chen, H., Efthimiadis, E. N., and Savoy, J., editors, *Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2010, Geneva, Switzerland, July 19-23, 2010*, pages 307–314. ACM. DOI: 10.1145/1835449.1835502.

Silva, L., Canalle, G. K., Salgado, A. C., Lóscio, B., and Moro, M. (2019). Uma Análise Experimental do Impacto da Seleção de Atributos em Processos de Resolução de Entidades. In *Anais do XXXIV Simpósio Brasileiro de Bancos de Dados*, pages 37–48.

Tang, R., Han, X., Jiang, X., and Hu, X. (2023). Does synthetic data generation of llms help clinical text mining? *Computer Science Archive*, abs/2303.04360. DOI: 10.48550/arXiv.2303.04360.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023). Llama: Open and efficient foundation language models.

Wang, S., Sun, X., Li, X., Ouyang, R., Wu, F., Zhang, T., Li, J., and Wang, G. (2023). GPT-NER: Named Entity Recognition via Large Language Models. *Computer Science Archive*, abs/2304.10428. DOI: 10.48550/arXiv.2304.10428.

Yadav, V. and Bethard, S. (2018). A survey on recent advances in named entity recognition from deep learning models. In Bender, E. M., Derczynski, L., and Isabelle, P., editors, *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2145–2158, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ye, F., Huang, L., Liang, S., and Chi, K. (2023). Decomposed Two-Stage Prompt Learning for Few-Shot Named Entity Recognition. *Information*, 14(5). DOI: 10.3390/info14050262.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., *et al*. (2022). Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Zhu, Y., Ye, Y., Li, M., Zhang, J., and Wu, O. (2023). Investigating annotation noise for named entity recognition. *Neural Computing and Applications*, 35(1):993–1007. DOI: 10.1007/s00521-022-07733-0.