# Effectiveness of Cross-linguistic Extraction of Genetic Information using Generative Large Language Models

Milindi Kodikara[1], Karin Verspoor[1,2,*]

[1]School of Computing Technologies, RMIT University, Melbourne, Victoria, 3000, Australia
[2]School of Computing and Information Systems, The University of Melbourne, Melbourne, Victoria, 3010, Australia

**Abstract**

This paper presents the RMIT University system (RMIT-READ-BioMed) developed for the GenoVarDis shared task at IberLEF 2024, focusing on the task of Named Entity Recognition (NER) of genes, genetic variants, and associated diseases from Spanish-language scientific literature texts.

The approach involves exploration of a general generative Large Language Model (LLM), GPT-3.5, for NER. We explore the impact of providing English-language instructions with the Spanish-language target text (cross-linguistic setting) as compared to a within-language setting where the instruction language matches the language of the text. We further experiment with a range of instruction strategies, including zero-shot and few-shot prompting under these two settings. Results indicate that the optimal results could be obtained with English-language instructions under the few-shot learning paradigm, resulting in an F1-score of 0.5. While this approach does not match the top results achieved for the shared task, our experiments provide insight into limitations associated with simple prompting of LLMs in languages other than English.

**Keywords**

Natural Language Processing, Generative Large Language Models, Computational Biology, Generative AI, Bioinformatics, Named Entity Recognition



## 1. Introduction

There is a persistent need for organised genetic information to support advancements in scientific discovery and personalised healthcare [1, 2]. Typically, this organisation process involves extraction and storage of key entities and their relationships from vast amounts of biomedical literature into databases by biocurators. This is an arduous, costly, time consuming and manual task, prone to errors due to fatigue and volume [3, 4]. With the exponential growth of literature, efforts have been directed towards automating this process with natural language processing techniques to streamline curation of biomedical literature and save time [5, 6, 7, 3].

Early solutions for automation explored rule-based, machine learning, and/or statistical methods for text mining of biomedical literature [8, 9, 10, 11]. Most such approaches failed to reach adequate accuracy levels to be used practically for biocuration. Key limitations included weak generalization of models and the impact of semantic constraints. Despite that, approaches that utilized small training datasets, for example [7, 12], provided good results showing that automated methods have good potential to extract information from biomedical literature [6, 2].

The natural language processing (NLP) task of *information extraction (IE)* involves the process of structured knowledge being extracted from plain text [5]. This process is pivotal for automating curation of biomedical information. In this work, our focus is on the IE task of Named Entity Recognition (NER) where entity spans are identified and annotated with a type. Specifically, we target entities related to disease-associated genetic variation, including genes, mutations, and the diseases themselves.

*IberLEF 2024, September 2024, Valladolid, Spain*
*Corresponding author.
†These authors contributed equally.
s3667779@student.rmit.edu.au (M. Kodikara); karin.verspoor@rmit.edu.au (K. Verspoor)
https://milindi-kodikara.github.io/ (M. Kodikara);
https://scholar.google.com/citations?hl=en&user=dUxHnbcAAAAJ&view_op=list_works&sortby=pubdate (K. Verspoor)
0009-0002-5976-9781 (M. Kodikara); 0000-0002-8661-1544 (K. Verspoor)

Recently, methods based on generative AI have shown promising results for biomedical IE [5, 3, 2]. Hence, in our approach we explore the use of *generative Large Language Models (generative LLMs)* through *prompt engineering*. Generative LLMs are a specific class of LLM that utilize decoder-only algorithms to generate content in response to a *prompt*, or instruction, on the basis of a pre-trained language model. We specifically consider the Generative Pre-trained Transformer (GPT) models [13, 14]. The output of a generative LLM depends directly on the prompt that is provided as input, and the task of developing a suitable prompt for a given task or information need is termed prompt engineering [15]. A prompt can be crafted adhering to in-context *learning paradigms*, such as zero-shot or few-shot instructions. This involves providing either no (zero) or a small number (few) examples of the solution to a task in the prompt itself, to guide the generative LLM to the desired output.

NER has been extensively investigated by researchers under learning paradigms such as few-shot learning, showing successful extraction of information across domains such as politics, literature, and natural sciences [16]. Few-shot prompting has resulted in great performance for IE tasks including NER and relation extraction (RE), across various domains [17, 3]. Performance has in some cases been found to come close to fully supervised models utilizing 10 examples under the few-shot learning paradigm (e.g. [17]). Both zero-shot and few-shot prompting for IE from clinical text (which closely relates to genetic text) has been shown to be effective, using handcrafted prompt templates provided to a general-domain GPT based LLM [18]. With the provision of annotated guidelines in the prompt along with fine-tuning, zero-shot results have shown to improve IE tasks [14]. PromptNER, which extracts information using few-shot learning with a set of defined entities with high accuracy is an example of successful use of generative AI and prompt engineering for IE [16]. This research indicates that providing more context to the prompts leads to higher performance of IE tasks. It can also be noted that prompt engineering has been conducted to explore few-shot learning on biomedical data but it has not been systematically compared with other learning paradigms for biomedical IE tasks.

It can be deduced through existing literature that the structure of prompts can lead to variation in performance [16, 14]. Research related to prompt engineering has been successfully conducted in cross-domain settings [16]. Moreover, it can be observed that there is limited research conducted specifically in the area of automated extraction of genetic information using generative LLMs from texts in languages other than English [19]. This is worth exploring as extracting information from non-English literature has the potential to contribute to the enrichment of existing biomedical knowledge bases and support in the advancement of research [20, 21, 22].

In this project, the focus is on the challenge of NER of genes, genomic variants, and associated diseases from Spanish-language scientific text, in the context of the GenoVarDis competition which is a part of the IberLEF 2024 campaign [23, 24]. This challenge is the first of its kind exploring this topic, due to limited datasets for NLP tasks in the genetic domain, in particular for languages other than English.

We have explored the effectiveness of utilising a general, primarily English-language LLM for cross-linguistic IE in this challenge. We examine the impact of providing English and Spanish-language instructions with the Spanish-language target text (cross-linguistic setting), matching the instruction language to the text, and experimenting with a range of instruction strategies, including zero-shot and few-shot prompting.

## 2. Method

Our method involves the creation of an IE pipeline with a manually crafted library of prompts.

We explore the impact of these prompts under various learning paradigms and the provision of annotated guidelines. These prompts are submitted automatically to a generative LLM (GPT-3.5) to perform the task of NER, and the outputs are post-processed to conform to the required format. An overview of the method is depicted in Figure 1.

**Figure 1:** Overview of the methodology

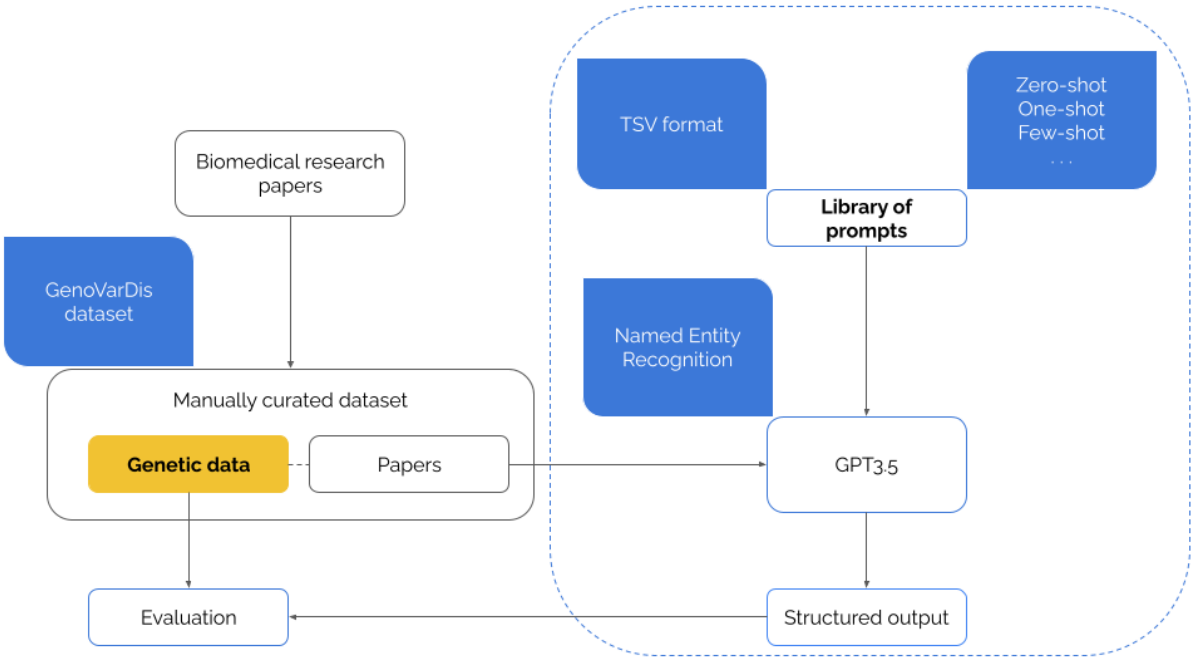



**Table 1**
Dataset statistics

| Data | Train | Evaluation | Test |
|---|---|---|---|
| Text | 427 | 70 | 136 |
| Gold annotations | 8199 | 1333 | 2102 |

## 2.1. Data

The dataset provided for the GenoVarDis challenge [24, 23] consisted of Spanish-language texts translated from 497 English-language biomedical texts (titles and abstracts) for the train and evaluation datasets, and 136 Spanish-language biomedical texts (titles and abstracts) originally derived from PubMed[1]. The data is split 70%-10%-20% for training, development (evaluation) and test sets as depicted in Table 1.

Along with the Spanish-language texts from the literature, the dataset included gold standard named entities for genes, genetic variants, and diseases, which were curated by human experts. The entity types were annotated according to label names as depicted in Table 2. This dataset was created with translations of English-language texts as there is a shortage of resources in other languages.

## 2.2. System Description

### 2.2.1. Platform

Our system, RMIT-READ-BioMed[2], is built using Jupyter notebooks in the Python programming language. Prompts were created as JSON objects for submission to the OpenAI GPT-3.5 API.

### 2.2.2. Prompt library

Each manually crafted prompt template contains the following attributes:

[1]https://pubmed.ncbi.nlm.nih.gov/
[2]https://github.com/Milindi-Kodikara/RMIT-READ-BioMed

**Table 2**
Entity types and their respective labels

| Entity type | Label |
|---|---|
| Variant on DNA sequence | DNAMutation |
| RS number | SNP |
| COSMIC mutation | SNP |
| Allele on DNA sequence | DNAAllele |
| Wild type and mutant | NucleotideChange-BaseChange |
| Variant with insufficient information | OtherMutation |
| Gene | Gene |
| Disease/Symptom | Disease |
| Transcript ID | Transcript |

**Figure 2:** Fixed guideline

> *"An entity is a variant on DNA sequence ('DNAMutation'), RS number ('SNP'), COSMIC mutation ('SNP'), Allele on DNA sequence ('DNAAllele'), wild type and mutations ('NucleotideChange-BaseChange'), variant entities with insufficient information ('OtherMutation'), gene ('Gene'), disease entities ('Disease') or Transcript ID ('Transcript')."*

- `prompt_id` : A unique identifier for each set of prompts, where a set contains two prompts, one written English and the other in Spanish. The id provides insight into the elements of the prompt, therefore, and is a combination of the IE task ("NER"), prompt language ("en" for English and "es" for Spanish), instruction index (eg: "`instruction_1`"), availability of guideline ("`guideline`" or omitted), number of examples, and output format.
- `instruction` : Clear and concise outline of the task for the model as depicted in Figure 4.
- `guideline` : Further clarity to the instruction by expanding on the entities to extract as depicted in Figure 2. This is a fixed string value. This attribute is set to be empty when no guideline required is provided to the prompt.
- `examples` : Number of examples to be embedded depending on the learning paradigm. Experimented values: {0, 1, 2, 5, 10, 20}
- `expected_output_format` : Describes the output structure and format, for example, Figure 3. Currently, this attribute is a fixed string value.
- `text` : The embedded text from biomedical literature.

Adding complexity and clarity to the task by providing an annotation guideline for the entities has been shown to increase performance. For example, provision of annotated guidelines in a prompt with no examples (zero-shot) has led to an improvement on the performance of LLMs on IE [14]. Therefore, we experiment with the inclusion of a fixed annotation guideline in the prompt library.

As including examples in few-shot learning paradigms affects the performance of the LLMs, we have explored the effect of prompts with 0, 1, 2, 5, 10 and 20 `examples`. These `examples`, which are embedded into the prompt, comprise of the Spanish-language texts and annotated data from the training dataset. The embedded examples are dynamically determined at run time depending on the number of examples required in the prompt.

There are various ways in which a result could be produced by a generative LLM, including JSON, tabulated, comma separated, and many more, that can impact the entities extracted. For example, requesting for the output to be depicted in table format leading to better usability and performance [25]. In this project we have specified that the output be separated by tabs. Moreover, we provide further detail by specifying the expected labels in the output for the identified spans.

The prompt library consisted of 9 prompts in English with varying values for the `instruction`, `guideline`, `examples`, and `expected_output` fields. Each prompt was also translated directly to

**Figure 3:** Example of the requested output format and structure

> *"Display results in the tsv format with the headers 'label' to annotate the entity as one of 'DNAMutation', 'SNP', 'DNAAllele', 'NucleotideChange-BaseChange', 'OtherMutation', 'Gene', 'Disease', 'Transcript' and 'span' for the identified entity. Provide each label and span in a new line."*

**Figure 4:** Prompt which resulted in the highest F1 score.

```
{
"prompt_id": "p_007_ner_en_instruction_1_20_few_shot_guideline_tsv",

"instruction": "Find the entities in the below Spanish language text. The number of
entities found should match the number of instances the entity is mentioned in the text.",

"guideline": "An entity is a variant on DNA sequence ('DNAMutation'), RS number ('SNP'),
COSMIC mutation ('SNP'), Allele on DNA sequence ('DNAAllele'), wild type and mutations
('NucleotideChange-BaseChange'), variant entities with insufficient information
('OtherMutation'), gene ('Gene'), disease entities ('Disease') or Transcript ID
('Transcript').",

"examples": 20,

"expected_output": "Display results in the tsv format with the headers 'label' to annotate
the entity as one of 'DNAMutation', 'SNP', 'SNP', 'DNAAllele',
'NucleotideChange-BaseChange', 'OtherMutation', 'Gene', 'Disease', 'Transcript' and 'span'
for the identified entity. Provide each label and span in a new line.",

"text": "Text: Síndrome de Gorlin en la edad pediátrica.
El sindrome de Gorlin (SG) es un trastorno de herencia autosomica dominante . . ."
}
```

the Spanish-language, resulting in an overall prompt count of 18. The prompt library was manually crafted and refined iteratively based on the performance of the model observed utilizing the training and evaluation datasets.

Each prompt id follows the format: "p_<prompt_number>_<IE task>_<language>_instruction_<instruction index>_<number of examples>_<learning paradigm>_<guideline>_<output type>_<output type index>". Certain fields are omitted from the prompt id when that field is not set, for example, if there is no guideline provided in the prompt, the prompt id will not contain the "`_guideline_`" field.

An example from the prompt library is depicted in Figure 4.

### 2.2.3. Entity extraction

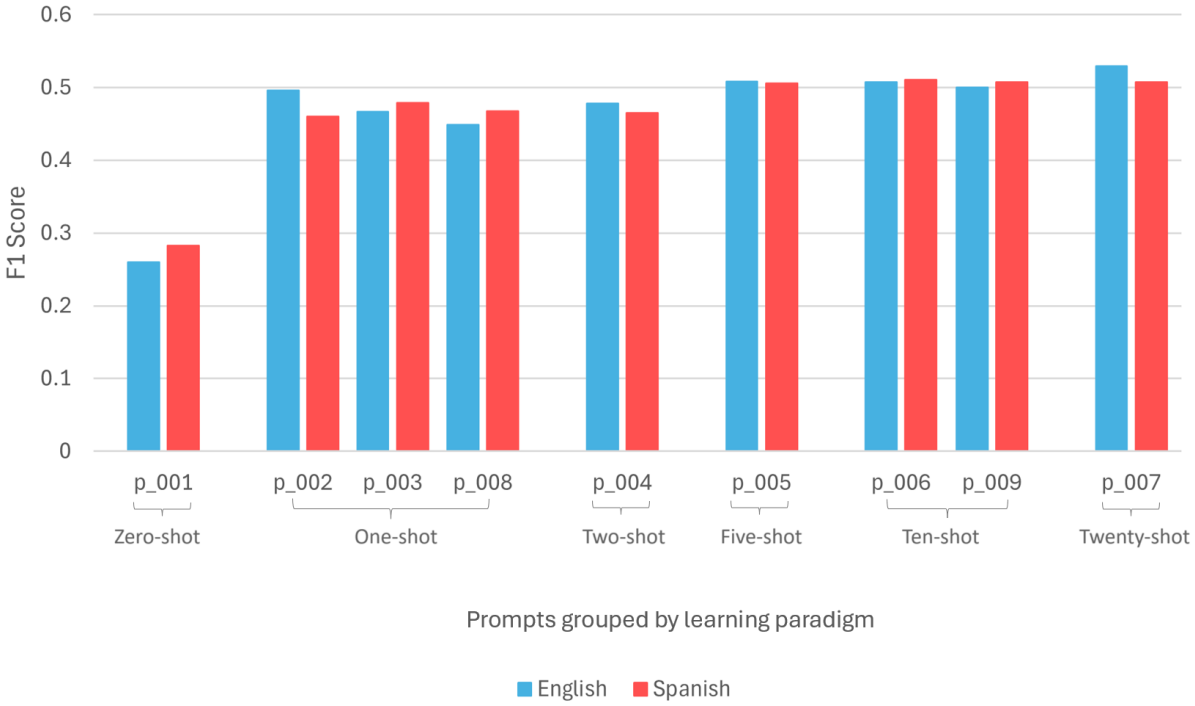
Open AI's GPT model `gpt-35-turbo-16k` was utilised to perform entity extractions from the provided text with a given prompt. This model was selected as it is optimized for chat and traditional completion tasks. A list of requests are sent to the Chat Completions API, containing prompts and our API key, using Azure Open AI receiving responses containing the model's outputs in the requested output formats.

### 2.2.4. Post-processing

The extracted list of tuples are processed to conform to the expected brat format [26] and further to the final single-file format requested from the competition.

Since the instructions to the GPT model only ask for a list of the extracted entities, while the

**Figure 5:** Impact of n-shot prompts on the F1 score, grouped by learning paradigm




submission format requires indicating the specific location in the text (text span) where the entity occurs, we attempt to locate each listed entity in the text by exact string match, annotating each matching text span as a mention of the corresponding entity type. Any span that cannot be matched to the text is filtered out from the submitted entities. We treat these unmatched entities as hallucinations erroneously produced by the generative LLM.

### 2.2.5. Evaluation

A tool that performs pairwise comparison of entities and relations in the BRAT format, brateval, is used for local evaluation of the extracted tuples against the gold standard data to get the overall statistics for true positives, false negatives, false positives, Precision, Recall, and F1 score [27].

## 3. Results

Table 3 demonstrates the performance of various prompts on the evaluation dataset.

The best F1 score was obtained utilizing prompt "`p_007_ner_en_20_few_shot_guideline_tsv`" which is a prompt written in the English-language with 20 examples provided under the few-shot learning paradigm, and including an annotation guideline providing further clarity to the prompt. Annotations identified based on this prompt were submitted as the final entry from our team into the GenoVarDis competition resulting in an F1 score of 0.548269, garnering us fourth place overall as shown in Table 4.

Table 5 depicts the performance of our system on the test data.

According to Figure 5, addition of examples have a correlation with the increase in the F1 score. Additionally, increasing the number of examples in the prompts have resulted in slight increases in performance. Spanish-language prompts seem to have performed better than English-language prompts in zero-shot prompting without an annotation guideline. Moreover, it can be seen that the result of the prompt with one example, no annotation guideline, and written in English-language outperforms their Spanish-language counterpart while the addition of the annotation guideline increases the performance

**Table 3**
System performance on development dataset

| prompt_id | true_positive | false_positive | false_negative | precision | recall | f1 |
|---|---|---|---|---|---|---|
| p_001_ner_en_instruction_1 | 313 | 1138 | 1020 | 0.2157 | 0.2348 | 0.2249 |
| p_001_ner_es_instruction_1 | 394 | 747 | 939 | 0.3453 | 0.2956 | 0.3185 |
| p_002_ner_en_instruction_1_one_shot_tsv | 634 | 721 | 699 | 0.4679 | 0.4756 | 0.4717 |
| p_002_ner_es_instruction_1_one_shot_tsv | 610 | 781 | 723 | 0.4385 | 0.4576 | 0.4479 |
| p_003_ner_en_instruction_1_one_shot_guideline_tsv | 603 | 730 | 730 | 0.4524 | 0.4524 | 0.4524 |
| p_003_ner_es_instruction_1_one_shot_guideline_tsv | 551 | 630 | 782 | 0.4666 | 0.4134 | 0.4383 |
| p_004_ner_en_instruction_1_2_shot_guideline_tsv | 605 | 711 | 728 | 0.4597 | 0.4539 | 0.4568 |
| p_004_ner_es_instruction_1_2_shot_guideline_tsv | 591 | 906 | 742 | 0.3948 | 0.4434 | 0.4177 |
| p_005_ner_en_instruction_1_5_few_shot_guideline_tsv | 685 | 814 | 648 | 0.457 | 0.5139 | 0.4838 |
| p_005_ner_es_instruction_1_5_few_shot_guideline_tsv | 666 | 873 | 667 | 0.4327 | 0.4996 | 0.4638 |
| p_006_ner_en_instruction_1_10_few_shot_guideline_tsv | 623 | 682 | 710 | 0.4774 | 0.4674 | 0.4723 |
| p_006_ner_es_instruction_1_10_few_shot_guideline_tsv | 622 | 757 | 711 | 0.4511 | 0.4666 | 0.4587 |
| p_007_ner_en_instruction_1_20_few_shot_guideline_tsv | 535 | 420 | 798 | 0.5602 | 0.4014 | 0.4677 |
| p_007_ner_es_instruction_1_20_few_shot_guideline_tsv | 581 | 412 | 752 | 0.5851 | 0.4359 | 0.4996 |
| p_008_ner_en_instruction_2_one_shot_guideline_tsv_2 | 577 | 687 | 756 | 0.4565 | 0.4329 | 0.4444 |
| p_008_ner_es_instruction_2_one_shot_guideline_tsv_2 | 625 | 722 | 708 | 0.464 | 0.4689 | 0.4664 |
| p_009_ner_en_instruction_2_10_few_shot_guideline_tsv_2 | 669 | 713 | 664 | 0.4841 | 0.5019 | 0.4928 |
| p_009_ner_es_instruction_2_10_few_shot_guideline_tsv_2 | 659 | 755 | 674 | 0.4661 | 0.4944 | 0.4798 |

**Table 4**
Official GenoVarDis challenge results

| competitor | f1 | precision | recall |
|---|---|---|---|
| ander.martinez | 0.820977 | 0.822350 | 0.819610 |
| VictorMov | 0.793455 | 0.790643 | 0.796287 |
| ELiRF-VRAIN | 0.734940 | 0.777483 | 0.696811 |
| RMIT-READ-BioMed (Milimeter98) | 0.548269 | 0.610754 | 0.497382 |
| orlandxrf | 0.530055 | 0.731769 | 0.415516 |
| GuillemGSubies | 0.428260 | 0.435531 | 0.421228 |
| Baseline | 0.319415 | 0.593790 | 0.218467 |
| Antares-Amazel | 0.300929 | 0.604017 | 0.200381 |

Note: Rows shaded in red and gray highlight the most performant model and the baseline model.

of these Spanish-language prompts. Similarly, 10-shot prompting follows the pattern where Spanish-language prompts outperform English-language prompts by a slight margin. Overall, prompt 007 with the highest number of examples, with an annotation guideline, written in English-language outperforms all the other prompts' results.

# 4. Discussion

## 4.1. Effect of prompting on the F1 score

This section aims to analyse the performance of our system on NER based on the results presented in Table 3 and Table 5. Upon inspection of the F1 scores, it can be observed that prompts containing only the instruction written in either English or Spanish under a zero-shot learning paradigm obtains the worst performance overall.

The provision of at least one example under one-shot learning paradigm to the prompt along with specification of the expected output structure results in a significant increase in performance compared with zero-shot learning. This increase in performance can be deduced to be due to the model's ability to learn in-context from the provided example. Additionally, providing more information with annotation guidelines and increasing the number of examples improve the F1 score slightly as depicted in Figure 5.

It can be observed that there are slight improvements in performance between when instructions are provided in English (cross-linguistic setting) vs. in Spanish (within-language setting), although the text to be analysed itself is in Spanish. This is likely due to the model being primarily trained on

**Table 5**
System performance on test dataset

| prompt_id | true_positive | false_positive | false_negative | precision | recall | f1 |
|---|---|---|---|---|---|---|
| p_001_ner_en_instruction_1 | 570 | 1716 | 1531 | 0.2493 | 0.2713 | 0.2599 |
| p_001_ner_es_instruction_1 | 550 | 1246 | 1551 | 0.3062 | 0.2618 | 0.2823 |
| p_002_ner_en_instruction_1_one_shot_tsv | 1057 | 1103 | 1044 | 0.4894 | 0.5031 | 0.4961 |
| p_002_ner_es_instruction_1_one_shot_tsv | 1039 | 1373 | 1062 | 0.4308 | 0.4945 | 0.4604 |
| p_003_ner_en_instruction_1_one_shot_guideline_tsv | 1095 | 1501 | 1006 | 0.4218 | 0.5212 | 0.4663 |
| p_003_ner_es_instruction_1_one_shot_guideline_tsv | 1016 | 1128 | 1085 | 0.4739 | 0.4836 | 0.4787 |
| p_004_ner_en_instruction_1_2_shot_guideline_tsv | 1049 | 1236 | 1052 | 0.4591 | 0.4993 | 0.4783 |
| p_004_ner_es_instruction_1_2_shot_guideline_tsv | 1122 | 1604 | 979 | 0.4116 | 0.5340 | 0.4649 |
| p_005_ner_en_instruction_1_5_few_shot_guideline_tsv | 1123 | 1199 | 978 | 0.4836 | 0.5345 | 0.5078 |
| p_005_ner_es_instruction_1_5_few_shot_guideline_tsv | 1191 | 1416 | 910 | 0.4568 | 0.5669 | 0.5059 |
| p_006_ner_en_instruction_1_10_few_shot_guideline_tsv | 1127 | 1214 | 974 | 0.4814 | 0.5364 | 0.5074 |
| p_006_ner_es_instruction_1_10_few_shot_guideline_tsv | 1186 | 1361 | 915 | 0.4656 | 0.5645 | 0.5103 |
| p_007_ner_en_instruction_1_20_few_shot_guideline_tsv | 1003 | 686 | 1098 | 0.5938 | 0.4774 | 0.5293 |
| p_007_ner_es_instruction_1_20_few_shot_guideline_tsv | 939 | 659 | 1162 | 0.5876 | 0.4469 | 0.5077 |
| p_008_ner_en_instruction_2_one_shot_guideline_tsv_2 | 986 | 1309 | 1115 | 0.4296 | 0.4693 | 0.4486 |
| p_008_ner_es_instruction_2_one_shot_guideline_tsv_2 | 1083 | 1448 | 1018 | 0.4279 | 0.5155 | 0.4676 |
| p_009_ner_en_instruction_2_10_few_shot_guideline_tsv_2 | 1117 | 1252 | 984 | 0.4715 | 0.5317 | 0.4998 |
| p_009_ner_es_instruction_2_10_few_shot_guideline_tsv_2 | 1147 | 1273 | 954 | 0.4740 | 0.5459 | 0.5074 |

English-language data as in most instances English-language prompts provide a higher performance.

## 4.2. Effect of entity matching

F1 scores of 0.526093 and 0.5293 were achieved from the GenoVarDis challenge evaluation and via brateval respectively for the same set of annotations from the test dataset using prompt `p_007_ner_en_20_few_shot_guideline_tsv`. As these F1 scores are closely similar, it can be deduced that the performance of the systems were measured based on exact matches of genetic entities.

## 4.3. Extracted entities and hallucinations

In this section a randomly selected example Spanish-language biomedical text from the test dataset, `PMID 24677153` (Figure 6), is analysed to gain further understanding of the performance. We examine differences between the gold standard entities (Table 8), entities extracted utilizing GPT 3.5 and the prompt library in our system (annotations from the final submission, Table 6), and hallucinated entities (Table 7).

As stated in Section 2, hallucinations were removed from the final annotations. The statistics of hallucinations discovered for each prompt for the test dataset and the development dataset are shown in Tables 9 and 10 respectively. The column "`matched_count`" reflects the final annotations.

The following can be noted upon observation:

1. Once our system identified certain spans, such as "`carcinomas basocelulares`" the model failed to find all occurrences of the span to be matched.
2. Fabricated spans such as "`fibromas ovaricos`" were discarded as hallucinations.
3. Spans that contain the expected entity but contains other words before or after the identified entity, for example "`gen CMT1A`" instead of "`CMT1A`", were discarded as hallucinations.
4. A reduction in hallucinated entities can be observed with the addition of examples in the prompt for the test dataset as depicted in Figure 7.
    - Compared to the number of hallucinated entities observed for zero-shot prompting, a reduction in the number of hallucinated entities can generally be observed with the addition of examples to prompts of either natural language. Prompt `p_004`, English-language version, and `p_002`, Spanish-language version, are exceptions to this.
    - Prompts `p_004` and `p_002`, written in English-language with examples produced the highest and lowest number of hallucinated entities respectively for prompts written in English-language.

**Figure 6:** Example Spanish text from test dataset

> 24677153|t|Síndrome de Gorlin en la edad pediátrica. 24677153|a|Introduccion. El sindrome de Gorlin (SG) es un trastorno de herencia autosomica dominante asociado a mutaciones en el gen PTCH1, cuya principal caracteristica es la aparicion de carcinomas basocelulares, unido a anomalias esqueleticas, queratoquistes odontogenicos y tumores intracraneales. Caso clinico. Niña de 3 años y 10 meses, ingresada por ataxia aguda. Destacan como antecedentes personales retraso psicomotor y como antecedentes familiares la sospecha de SG en la madre por quiste maxilar. En la exploracion, se aprecia macrocefalia con frente prominente e hipertelorismo, asi como nevo. Se solicita estudio genetico de SG, en el que se detecta la mutacion c.930delC en el exon 6 del gen PTCH1 en heterocigosis. Conclusiones. En el SG hay un aumento de la susceptibilidad al desarrollo de carcinomas basocelulares y es preciso un estrecho control dermatologico. Es necesario un seguimiento neurologico clinico y de imagen, mediante resonancia magnetica, para el diagnostico precoz de tumores intracraneales, fundamentalmente el meduloblastoma. Tambien son caracteristicos los queratoquistes odontogenicos, otras alteraciones cutaneas, fibromas cardiacos y ovaricos, asi como anomalias esqueleticas, que precisan controles clinicos y de imagen periodicos, y tratamiento en caso de ser necesarios, pero debe evitarse la radiacion. El SG es un trastorno poco frecuente, que se debe sospechar ante la presencia de alteraciones caracteristicas. Es necesario un seguimiento multidisciplinar, asi como establecer un protocolo de actuacion, para un temprano diagnostico y tratamiento de las complicaciones potencialmente graves derivadas de esta enfermedad.

- Zero-shot and two-shot prompting (prompt `p_004`) resulted in the highest values for total, matched and hallucinated entities for prompts written in English-language. Despite the similarity in entity counts, two-shot prompting shows a higher F1 score, closer to the optimal value, compared to zero-shot prompting with a difference in performance of 0.2184.
- Prompts `p_002` and `p_003`, written in Spanish-language with examples produced the highest and lowest number of hallucinated entities respectively for prompts written in Spanish-language. While prompt `p_002` resulted in extracting the highest number of hallucinated entities, it should be noted that the performance of prompt `p_002` shows a significant difference in F1 score of 0.1781 compared with the least performative Spanish-language prompt, `p_001`.
- Both versions of prompt `p_007`, extracted the least amount of total entities resulting in the least amount of matched entities. This prompt with twenty examples produced the best F1 scores for each of the respective natural languages with the English-language prompt depicting the optimal performance.
- Prompts with five or higher number of examples extracted similar amounts of hallucinated entities and are found to have resulted in similar F1 scores.

These hallucinations can be deduced to be due to the complexity of the cross-linguistic task of NER of biomedical entities, limitations in the prompts with regard to providing context for the task, generative nature of the model used, and limitations due to the LLM being trained predominantly on English-language data.

## 5. Conclusions and Future Work

This paper presents the system developed by RMIT University for the GenoVarDis shared task at IberLEF 2024, focusing on the task of Named Entity Recognition (NER) of genomic variants, genes, and its associated diseases from Spanish scientific literature.

Our approach involves exploring cross-linguistic NER of genetic information utilizing the generative LLM GPT-3.5 and a manually crafted library of prompts. We identified that few-shot learning paradigm works best for NER with an annotation guideline and the expected output structure outlined. Moreover, it is evident that the natural language of the prompt had only limited impact on the performance of the model for NER. We have demonstrated that cross-linguistic information extraction is feasible.

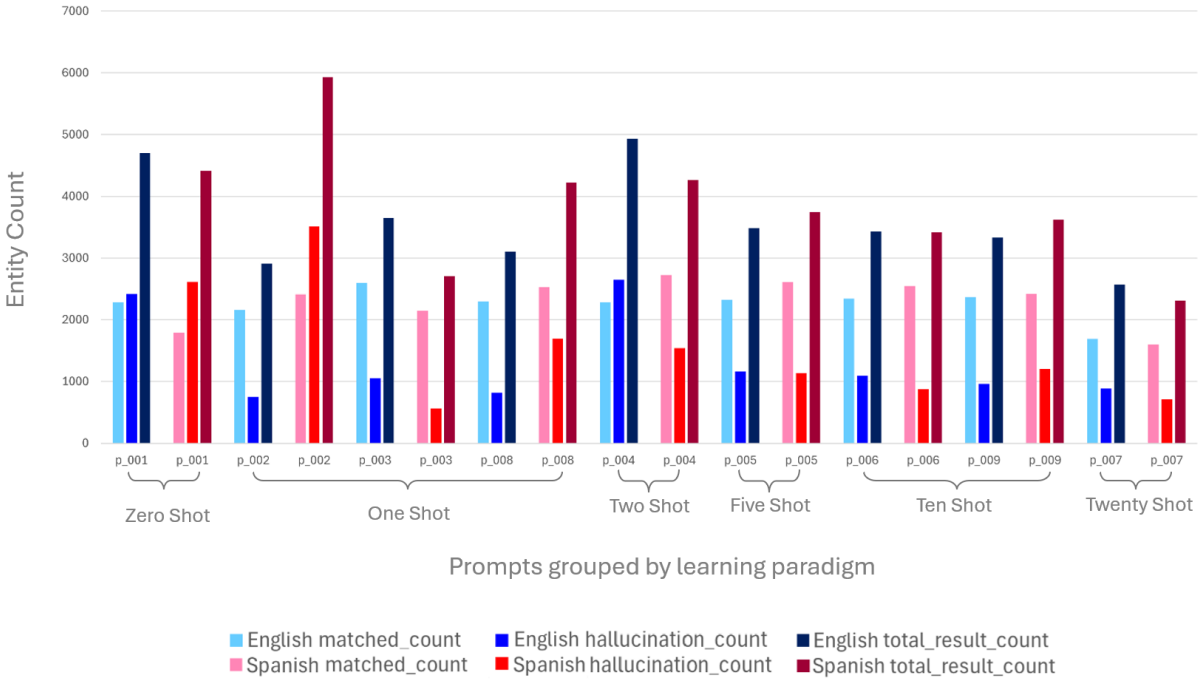
**Table 6**
Extracted entities from example Spanish text

| label | offset1 | offset2 | span |
|---|---|---|---|
| Disease | 11 | 29 | Síndrome de Gorlin |
| Gene | 186 | 191 | PTCH1 |
| DNAMutation | 729 | 738 | c.930delC |
| Gene | 760 | 765 | PTCH1 |
| Disease | 242 | 266 | carcinomas basocelulares |
| Disease | 300 | 328 | queratoquistes odontogenicos |
| Disease | 331 | 353 | tumores intracraneales |
| Disease | 1100 | 1114 | meduloblastoma |

**Table 7**
Hallucinated entities from example Spanish text

| label | span |
|---|---|
| Disease | Síndrome de Gorlin |
| Gene | PTCH1 |
| Gene | PTCH1 |
| Disease | fibromas ovaricos |
| Gene | PTCH1 |
| Disease | Síndrome de Gorlin |

**Figure 7:** Impact of n-shot prompts on extracted entities from test dataset texts, grouped by learning paradigm




In future work, we will be looking into improving performance and reducing hallucinations in various ways, such as structuring prompts to provide better context. We also plan to explore the effect of various output formats on the performance as it was evident that providing clear instructions on how the output needs to be structured provided better results. We can further consider adjusting GPT-3.5 temperature settings.

**Table 8**
Gold annotations for example Spanish text

| label | offset1 | offset2 | span |
|---|---|---|---|
| Disease | 11 | 29 | Síndrome de Gorlin |
| Disease | 81 | 99 | sindrome de Gorlin |
| Disease | 101 | 103 | SG |
| Gene | 186 | 191 | PTCH1 |
| Disease | 242 | 266 | carcinomas basocelulares |
| Disease | 276 | 298 | anomalias esqueleticas |
| Disease | 300 | 328 | queratoquistes odontogenicos |
| Disease | 331 | 353 | tumores intracraneales |
| Disease | 462 | 480 | retraso psicomotor |
| Disease | 527 | 529 | SG |
| Disease | 546 | 560 | quiste maxilar |
| Disease | 592 | 604 | macrocefalia |
| Disease | 629 | 643 | hipertelorismo |
| Disease | 654 | 658 | nevo |
| Disease | 692 | 694 | SG |
| DNAMutation | 729 | 738 | c.930delC |
| Gene | 760 | 765 | PTCH1 |
| Disease | 804 | 806 | SG |
| Disease | 861 | 885 | carcinomas basocelulares |
| Disease | 1056 | 1078 | tumores intracraneales |
| Disease | 1148 | 1176 | queratoquistes odontogenicos |
| Disease | 1184 | 1205 | alteraciones cutaneas |
| Disease | 1207 | 1236 | fibromas cardiacos y ovaricos |
| Disease | 1247 | 1269 | anomalias esqueleticas |
| Disease | 1404 | 1406 | SG |
| Disease | 410 | 422 | ataxia aguda |
| Disease | 1100 | 1114 | meduloblastoma |

## Acknowledgments

We thank the GenoVarDis challenge organisers for preparing the task and the datasets required for the challenge. We additionally thank the RACE Hub of RMIT University for providing access to the Azure Open AI API service.

**Table 9**
Hallucination statistics for test dataset

| prompt_id | f1 | matched_count | hallucination_count | total_count |
|---|---|---|---|---|
| p_001_ner_en_instruction_1 | 0.2599 | 2286 | 2414 | 4700 |
| p_001_ner_es_instruction_1 | 0.2823 | 1796 | 2614 | 4410 |
| p_002_ner_en_instruction_1_one_shot_tsv | 0.4961 | 2160 | 746 | 2906 |
| p_002_ner_es_instruction_1_one_shot_tsv | 0.4604 | 2412 | 3511 | 5923 |
| p_003_ner_en_instruction_1_one_shot_guideline_tsv | 0.4663 | 2596 | 1053 | 3649 |
| p_003_ner_es_instruction_1_one_shot_guideline_tsv | 0.4787 | 2144 | 564 | 2708 |
| p_004_ner_en_instruction_1_2_shot_guideline_tsv | 0.4783 | 2285 | 2647 | 4932 |
| p_004_ner_es_instruction_1_2_shot_guideline_tsv | 0.4649 | 2726 | 1536 | 4262 |
| p_005_ner_en_instruction_1_5_few_shot_guideline_tsv | 0.5078 | 2322 | 1161 | 3483 |
| p_005_ner_es_instruction_1_5_few_shot_guideline_tsv | 0.5059 | 2607 | 1131 | 3738 |
| p_006_ner_en_instruction_1_10_few_shot_guideline_tsv | 0.5074 | 2341 | 1091 | 3432 |
| p_006_ner_es_instruction_1_10_few_shot_guideline_tsv | 0.5103 | 2547 | 873 | 3420 |
| p_007_ner_en_instruction_1_20_few_shot_guideline_tsv | 0.5293 | 1689 | 883 | 2572 |
| p_007_ner_es_instruction_1_20_few_shot_guideline_tsv | 0.5077 | 1598 | 712 | 2310 |
| p_008_ner_en_instruction_2_one_shot_guideline_tsv_2 | 0.4486 | 2295 | 813 | 3108 |
| p_008_ner_es_instruction_2_one_shot_guideline_tsv_2 | 0.4676 | 2531 | 1687 | 4218 |
| p_009_ner_en_instruction_2_10_few_shot_guideline_tsv_2 | 0.4998 | 2369 | 963 | 3332 |
| p_009_ner_es_instruction_2_10_few_shot_guideline_tsv_2 | 0.5074 | 2420 | 1201 | 3621 |

Note: Compared with the provided 2102 gold annotations.

**Table 10**
Hallucination statistics for development dataset

| prompt_id | f1 | matched_count | hallucination_count | total_count |
|---|---|---|---|---|
| p_001_ner_en_instruction_1 | 0.2249 | 1451 | 780 | 2231 |
| p_001_ner_es_instruction_1 | 0.3185 | 1141 | 3288 | 4429 |
| p_002_ner_en_instruction_1_one_shot_tsv | 0.4717 | 1355 | 1357 | 2712 |
| p_002_ner_es_instruction_1_one_shot_tsv | 0.4479 | 1391 | 728 | 2119 |
| p_003_ner_en_instruction_1_one_shot_guideline_tsv | 0.4524 | 1333 | 1059 | 2392 |
| p_003_ner_es_instruction_1_one_shot_guideline_tsv | 0.4383 | 1181 | 441 | 1622 |
| p_004_ner_en_instruction_1_2_shot_guideline_tsv | 0.4568 | 1316 | 622 | 1938 |
| p_004_ner_es_instruction_1_2_shot_guideline_tsv | 0.4177 | 1497 | 1592 | 3089 |
| p_005_ner_en_instruction_1_5_few_shot_guideline_tsv | 0.4838 | 1499 | 569 | 2068 |
| p_005_ner_es_instruction_1_5_few_shot_guideline_tsv | 0.4638 | 1539 | 1681 | 3220 |
| p_006_ner_en_instruction_1_10_few_shot_guideline_tsv | 0.4723 | 1305 | 408 | 1713 |
| p_006_ner_es_instruction_1_10_few_shot_guideline_tsv | 0.4587 | 1379 | 463 | 1842 |
| p_007_ner_en_instruction_1_20_few_shot_guideline_tsv | 0.4677 | 955 | 600 | 1555 |
| p_007_ner_es_instruction_1_20_few_shot_guideline_tsv | 0.4996 | 993 | 387 | 1380 |
| p_008_ner_en_instruction_2_one_shot_guideline_tsv_2 | 0.4444 | 1264 | 1131 | 2395 |
| p_008_ner_es_instruction_2_one_shot_guideline_tsv_2 | 0.4664 | 1347 | 538 | 1885 |
| p_009_ner_en_instruction_2_10_few_shot_guideline_tsv_2 | 0.4928 | 1382 | 556 | 1938 |
| p_009_ner_es_instruction_2_10_few_shot_guideline_tsv_2 | 0.4798 | 1414 | 673 | 2087 |

Note: Compared with the provided 1333 gold annotations.

# References


[1] T. E. Putman, K. Schaper, N. Matentzoglu, V. Rubinetti, F. S. Alquaddoomi, C. Cox, J. H. Caufield, G. Elsarboukh, S. Gehrke, H. B. Hegde, J. T. Reese, I. Braun, R. M. Bruskiewich, L. Cappelletti, S. Carbon, A. R. Caron, L. E. Chan, C. G. Chute, K. G. Cortes, V. D. Souza, T. Fontana, N. L. Harris, E. L. Hartley, E. Hurwitz, J. O. B. Jacobsen, M. Krishnamurthy, B. Laraway, J. A. McLaughlin, J. A. McMurry, S. A. T. Moxon, K. R. Mullen, S. T. O'Neil, K. A. Shefchek, R. Stefancsik, S. Toro, N. A. Vasilevsky, R. L. Walls, P. L. Whetzel, D. Osumi-Sutherland, D. Smedley, P. N. Robinson, C. J. Mungall, M. A. Haendel, M. C. Munoz-Torres, The Monarch Initiative in 2024: an analytic platform integrating phenotypes, genes and diseases across species, Nucleic Acids Research 52 (2023) D938 – D949.

[2] J. Dagdelen, A. Dunn, S. Lee, N. Walker, A. S. Rosen, G. Ceder, K. A. Persson, A. Jain, Structured information extraction from scientific text with large language models, Nature Communications

15 (2024).
[3] A. Goel, A. Gueta, O. Gilon, C. Liu, S. Erell, L. H. Nguyen, X. Hao, B. Jaber, S. Reddy, R. Kartha, J. Steiner, I. Laish, A. Feder, LLMs accelerate annotation for medical information extraction, in: Machine Learning for Health (ML4H) Symposium, 2023.
[4] J. Chang, S. Wang, C. Ling, Z. Qin, L. Zhao, Gene-associated disease discovery powered by large language models, ArXiv abs/2401.09490 (2024).
[5] D. Xu, W. Chen, W. Peng, C. Zhang, T. Xu, X. Zhao, X. Wu, Y. Zheng, E. Chen, Large language models for generative information extraction: A survey, ArXiv abs/2312.17617 (2023).
[6] A. Singhal, M. Simmons, Z. Lu, Text mining genotype-phenotype relationships from biomedical literature for database curation and precision medicine, PLoS Computational Biology 12 (2016).
[7] M. Khordad, R. E. Mercer, Identifying genotype-phenotype relationships in biomedical text, Journal of Biomedical Semantics 8 (2017).
[8] T. Sekimizu, H. S. Park, H. S. Park, J. Tsujii, J. Tsujii, Identifying the interaction between genes and gene products based on frequently seen verbs in medline abstracts, Genome informatics. Workshop on Genome Informatics 9 (1998) 62–71.
[9] J. M. Temkin, M. R. Gilder, Extraction of protein interaction information from unstructured text using a context-free grammar, Bioinformatics 19 16 (2003) 2046–53.
[10] A. Coulet, N. H. Shah, Y. Garten, M. A. Musen, R. B. Altman, Using text to build semantic networks for pharmacogenomics, Journal of Biomedical Informatics 43 6 (2010) 1009–19.
[11] J. Tsujii, Natural Language Processing and Computational Linguistics, Computational Linguistics 47 (2021) 707–727. doi:10.1162/coli_a_00420.
[12] K. M. Verspoor, G. E. Heo, K. Y. Kang, M. Song, Establishing a baseline for literature mining human genetic variants and their relationships to disease cohorts, BMC Medical Informatics and Decision Making 16 (2016).
[13] P. Yu, H. Xu, X. Hu, C. Deng, Leveraging generative AI and large language models: A comprehensive roadmap for healthcare integration, Healthcare 11 (2023).
[14] O. Sainz, I. García-Ferrero, R. Agerri, O. L. de Lacalle, G. Rigau, E. Agirre, GoLLIE: Annotation guidelines improve zero-shot information-extraction, ArXiv abs/2310.03668 (2023).
[15] P. Sahoo, A. K. Singh, S. Saha, V. Jain, S. S. Mondal, A. Chadha, A systematic survey of prompt engineering in large language models: Techniques and applications, ArXiv abs/2402.07927 (2024).
[16] D. Ashok, Z. C. Lipton, PromptNER: Prompting for named entity recognition, ArXiv abs/2305.15444 (2023).
[17] S. Wadhwa, S. Amir, B. C. Wallace, Revisiting relation extraction in the era of large language models, Proceedings of the conference. Association for Computational Linguistics. Meeting 2023 (2023) 15566–15589.
[18] M. Agrawal, S. Hegselmann, H. Lang, Y. Kim, D. A. Sontag, Large language models are few-shot clinical information extractors, in: Conference on Empirical Methods in Natural Language Processing, 2022.
[19] C. P. Carrino, J. Llop, M. Pàmies, A. Gutiérrez-Fandiño, J. Armengol-Estap'e, J. Silveira-Ocampo, A. Valencia, A. Gonzalez-Agirre, M. Villegas, Pretrained biomedical language models for clinical nlp in spanish, in: Workshop on Biomedical Natural Language Processing, 2022.
[20] M. Rezaeian, Disadvantages of publishing biomedical research articles in english for non-native speakers of english, Epidemiology and Health 37 (2015).
[21] R. Bawden, K. Bretonnel Cohen, C. Grozea, A. Jimeno Yepes, M. Kittner, M. Krallinger, N. Mah, A. Neveol, M. Neves, F. Soares, A. Siu, K. Verspoor, M. Vicente Navarro, Findings of the WMT 2019 biomedical translation shared task: Evaluation for MEDLINE abstracts and biomedical terminologies, in: O. Bojar, R. Chatterjee, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, A. Martins, C. Monz, M. Negri, A. Névéol, M. Neves, M. Post, M. Turchi, K. Verspoor (Eds.), Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2), Association for Computational Linguistics, Florence, Italy, 2019, pp. 29–53. URL: https://aclanthology.org/W19-5403. doi:10.18653/v1/w19-5403.
[22] M. AlShuweihi, S. A. Salloum, K. F. Shaalan, Biomedical corpora and natural language processing

on clinical text in languages other than english: A systematic review, in: Recent Advances in Intelligent Systems and Smart Applications, 2020.
[23] L. Chiruzzo, S. M. Jiménez-Zafra, F. Rangel, Overview of IberLEF 2024: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), CEUR-WS.org, 2024.
[24] M. M. Agüero-Torales, C. Rodríguez Abellán, M. Carcajona Mata, J. I. Díaz Hernández, M. Solís López, A. Miranda-Escalada, S. López-Alvárez, J. Mira Prats, C. Castaño Moraga, D. Vilares, L. Chiruzzo, Overview of GenoVarDis at IberLEF 2024: NER of Genomic Variants and Related Diseases in Spanish, Procesamiento del Lenguaje Natural 73 (2024).
[25] Y. Jiao, M. Zhong, S. Li, R. Zhao, S. Ouyang, H. Ji, J. Han, Instruct and extract: Instruction tuning for on-demand information extraction, in: Conference on Empirical Methods in Natural Language Processing, 2023.
[26] P. Stenetorp, S. Pyysalo, G. Topic, T. Ohta, S. Ananiadou, J. Tsujii, BRAT: a web-based tool for nlp-assisted text annotation, in: Conference of the European Chapter of the Association for Computational Linguistics, 2012.
[27] A. Albahem, K. Verspoor, A. J. Jimeno Yepes, BRAT-Eval v0.3.2, 2013. URL: https://github.com/READ-BioMed/brateval.