Lesser the Shots, Higher the Hallucinations: Exploration of Genetic Information Extraction using Generative Large Language Models

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

[1]School of Computing Technologies, RMIT University, Melbourne, Australia
[2]School of Computing and Information Systems
The University of Melbourne, Melbourne, Australia
{milindi.kodikara2, karin.verspoor}@rmit.edu.au

Abstract

Organisation of information about genes, genetic variants, and associated diseases from vast quantities of scientific literature texts through automated information extraction (IE) strategies can facilitate progress in personalised medicine.

We systematically evaluate the performance of generative large language models (LLMs) on the extraction of specialised genetic information, focusing on end-to-end IE encompassing both named entity recognition and relation extraction. We experiment across multilingual datasets with a range of instruction strategies, including zero-shot and few-shot prompting along with providing an annotation guideline. Optimal results are obtained with few-shot prompting. However, we also identify that generative LLMs failed to adhere to the instructions provided, leading to over-generation of entities and relations. We therefore carefully examine the effect of learning paradigms on the extent to which genetic entities are fabricated, and the limitations of exact matching to determine performance of the model.

1 Introduction

There is a persistent need for organised genetic information to support advancements in scientific discovery and personalised healthcare (Putman et al., 2023; Dagdelen et al., 2024). 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 (Goel et al., 2023; Chang et al., 2024). 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, saving time and effort (Xu et al., 2024; Singhal et al., 2016; Khordad and Mercer, 2017; Goel et al., 2023).

Early solutions for automation explored rule-based, machine learning, and/or statistical methods for text mining of biomedical literature (Sekimizu et al., 1998; Temkin and Gilder, 2003; Coulet et al., 2010). Most such approaches failed to reach adequate accuracy levels to be used practically for biocuration, one of the key limitations being the weak generalisation of models (Elangovan et al., 2022). Despite that, certain approaches, for example (Khordad and Mercer, 2017; Verspoor et al., 2016), provided good results showing that automated methods have good potential to extract information from biomedical literature (Singhal et al., 2016; Dagdelen et al., 2024).

The natural language processing (NLP) task of *information extraction (IE)* addresses extraction of structured knowledge from natural language texts (Xu et al., 2024). This process is pivotal for automating curation of biomedical information.

In this work, our focus is on the IE tasks of Named Entity Recognition (NER) where entity spans are identified and annotated with a type, Relation Extraction (RE) where specified entity types are identified and the relation type between the identified entities is classified, and end-to-end encompassing both NER and RE steps, NERRE. We target entities related to disease-associated genetic variation, including genes, mutations, and the diseases themselves.

Recently, methods based on generative AI have shown promising results for biomedical IE (Xu et al., 2024; Goel et al., 2023; Dagdelen et al., 2024). 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 LLMs that utilise 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 (Yu et al., 2023; Sainz et al., 2024).

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 (Sahoo et al., 2024). 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.

We explore the effectiveness of utilising a general generative LLM for end-to-end IE of genetic information. Our key contributions are:

- Experimentation with a range of instruction strategies, including zero-shot and few-shot prompting, across three genetic variant literature datasets, including one Spanish-language corpus.
- A detailed exploration of the limitations of using generative technologies for extraction of highly domain-specialised information.

This expands prior work on genetic IE both in breadth and depth, providing insight into the most effective use of generative LLMs for these tasks.

2 Methods

Our experiment involved an end-to-end IE pipeline with a manually crafted library of prompts for each IE task. We explored the impact of these prompts with the inclusion of examples under various in-context learning paradigms and the addition of an annotation guideline.

After pre-processing, prompts were sent to GPT-3.5 Turbo via OpenAI API calls to perform the specified task. The results were then post-processed to conform to the brat format (Stenetorp et al., 2012) for evaluation. This involved mapping each entity presented in the system output to a specific span of text where the entity appears. We processed each entity/relation in order, so that the first entity term in the output was mapped to the first occurrence of the term in the text, etc.

During post-processing of the results, hallucinated instances – defined here as entities or relations that could not be projected into the relevant text – were identified and discarded. These hallucinated instances were classified into two types, namely, over-generated hallucinations and fabricated hallucinations. *Over-generated hallucinations* are instances containing one or more entities that were found in the accompanied text but could not be mapped to any position on the text, after previous entities were mapped. *Fabricated instances* included one or more entities and/or relations that were not found in the text at all.

A method overview appears in Figure 1. Code is available at `https://github.com/Milindi-Kodikara/RMIT-READ-BioMed/releases/tag/v2.0`.

2.1 Data

Three annotated genetic variation corpora, GenoVarDis for NER (Agüero, 2024), TBGA for RE (Marchesin and Silvello, 2022) and Variome for NER+RE (Verspoor et al., 2013), were utilised.

Distribution of data in these three datasets is shown in Table 1. More details are provided in the Appendix; the schema of each dataset is outlined in Table A1 and the entity and relation types are summarised in Table A2.

2.1.1 GenoVarDis (Agüero, 2024)

We utilised the dataset provided for the GenoVarDis challenge (Agüero-Torales et al., 2024; Chiruzzo et al., 2024) consisting of Spanish-language texts manually translated from 497 English-language biomedical texts (titles and abstracts), and 136 Spanish-language biomedical texts (titles and abstracts) directly available from PubMed[1]. The data was split 70%-10%-20% for training, development (not used here) and test sets. We present results for experiments utilising both Spanish and English language prompts (cross-linguistic setting, following (Kodikara and Verspoor, 2024)).

2.1.2 TBGA (Marchesin and Silvello, 2022)

TBGA dataset was specifically created for biomedical RE using the DisGeNET database, which is one of the largest collections of genes and variants involved in human diseases (González et al., 2019). TBGA dataset is one of the largest publicly available English-language datasets created for genetic RE, with 700K publications with 200K instances and 100K gene-disease pairs annotated semi-automatically.

[1] `https://pubmed.ncbi.nlm.nih.gov/`

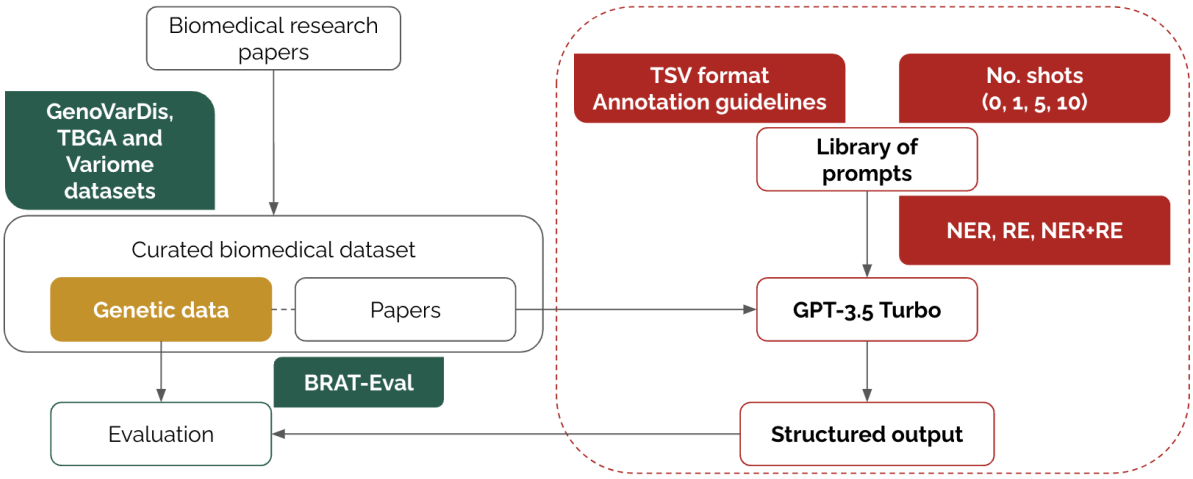

Figure 1: Overview of method

Table 1: Dataset statistics

Dataset	Train set			Test set			Total		Avg text length
	No. Texts	Gold entities	Gold relations	No. Texts	Gold entities	Gold relations	No. Texts	Gold	
GenoVarDis	427	8199	0	136	2101	0	563	10300	248
TBGA	178264	356528	178264	5	41032	20516	178269	596340	25
Variome	10	710	355	110	8590	4295	120	13950	331

2.1.3 Variome corpus (Verspoor et al., 2013)

The small Variome dataset of English-language inherited colorectal cancer texts is richly annotated for genetic variants, diseases and relations, relevant for cataloguing and interpreting human generic variation and its relationship to disease.

2.2 Model

Open AI's GPT model `gpt-35-turbo-16k` was utilised to perform the IE tasks. This model was selected as it has been shown to be effective for various IE tasks across domains (see Section 4).

Requests were sent to the Chat Completions API, containing prompts and our API key, using Azure Open AI to receive the responses containing the extracted tuples and triplets in the requested format.

2.3 Prompts

Each manually crafted prompt contains attributes as shown below.

- `prompt_id`: A unique identifier for the prompts. The `prompt_id` is a combination of the prompt index and the number of examples in the prompt. For cross-linguistic prompts, for NER, the `prompt_id` has "en" and "es" appended to the tail to distinguish between English and Spanish language instructions.
- `instruction`: Outline of the task for the model. (Example in Section A.1).
- `guideline`: Task annotation guidelines. This attribute varies between tasks as the relevant entities and relations to extract, as well as their definitions, differ. (Example in Section A.2.)

  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 (Sainz et al., 2024).
- `examples`: Number of examples to be embedded depending on the learning paradigm. Experimented values: {0, 1, 5, 10}.

  Each `example` consists of a text and associated annotations sampled randomly from the training datasets.
- `expected_output_format`: Defines the expected output structure and format. This attribute is a fixed string value and varies based on the task. The aim is to provide further

```
"prompt_id": "p4_ten_shot_es",

"instruction": "Encuentre las entidades en el siguiente texto en español. La
cantidad de entidades encontradas debe coincidir con la cantidad de veces que se
menciona la entidad en el texto.",

"guideline": "Una entidad es una variante en la secuencia de ADN ('DNAMutation'),
número RS ('SNP'), mutación CÓSMICA ('SNP'), alelo en la secuencia de ADN
('DNAAllele'), tipo salvaje y mutaciones ( 'NucleotideChange-BaseChange'), entidades
variantes con información insuficiente ('OtherMutation'), gen ('Gene'), entidades
patológicas ('Disease') o ID de transcripción ('Transcript').",

"examples": 10,

"expected_output": "Muestra los resultados en formato tsv con los encabezados
'label' para anotar la entidad como una de 'DNAMutation', 'SNP', 'DNAAllele',
'NucleotideChange-BaseChange', 'OtherMutation', 'Gene', 'Disease', 'Transcript' y
'span' para la entidad identificada Proporcione cada etiqueta y intervalo en una
nueva línea.",

"text": "Text: ..."
```

Figure 2: Example prompt

clarity on the task, thereby improving performance (Jiao et al., 2023). (Example in A.3).

All results are requested in tab separated vector (TSV) format. We further specify the headers for the extracted tuples and triplets.

- `text`: The embedded text from biomedical literature.

The prompt library consisted of 16 prompts with RE and NER+RE each being explored using 4 prompts and NER being explored using 8 prompts, 4 prompts for each language. The prompt library was manually crafted and refined iteratively based on trial and error with training instances.

An example from the prompt library is shown in Figure 2.

2.4 Evaluation

Industry standard metrics of Precision, Recall, and F1 score are used to evaluate performance.

The brateval[2] tool tailored for evaluation of data in the BRAT format[3], is used to compare extracted entities and/or relations against the gold standard data (Albahem et al., 2013).

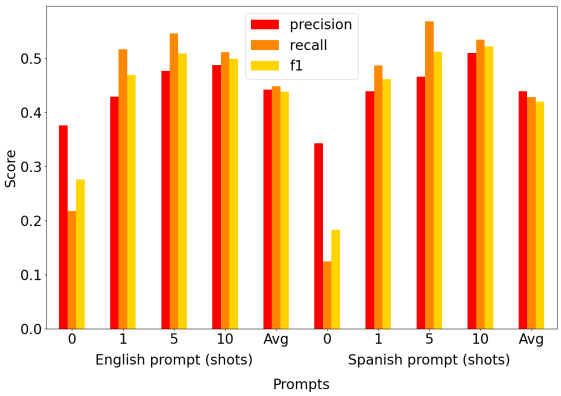

Figure 3: Results for varying number of shots for GenoVarDis (NER), grouped by prompt language

3 Findings

3.1 Few shot prompting leads to higher entity recognition

Optimal performance was obtained utilising prompts with five to ten examples for GenoVarDis (NER) and Variome (NER+RE) as shown in Figures 3 and 5. Worst performance for both datasets was observed for prompts with no examples (zero shot). In contrast, best performance for TBGA on RE was obtained through zero-shot prompting (Figure 4).

[2] https://github.com/READ-BioMed/brateval
[3] https://brat.nlplab.org/

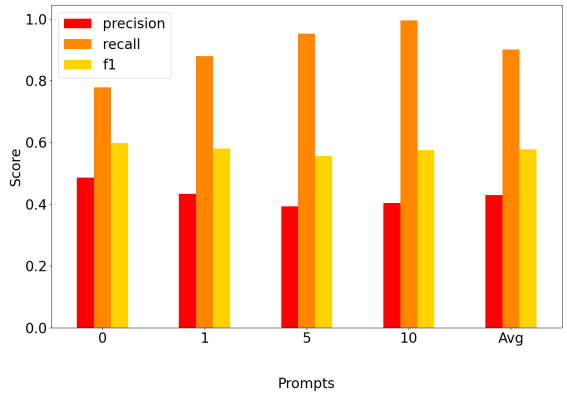

Figure 4: Results for varying number of shots for TBGA (RE)

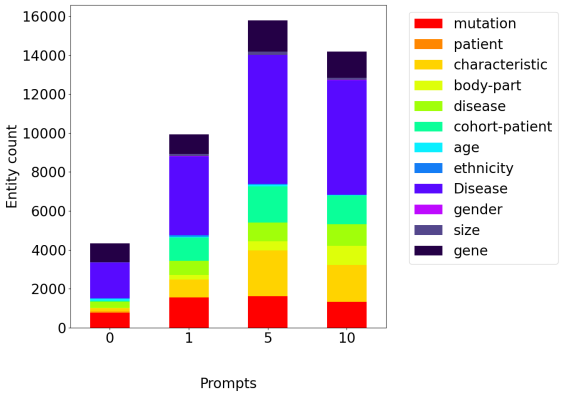

Figure 6: Extracted entity types for varying number of shots for Variome (NER+RE)

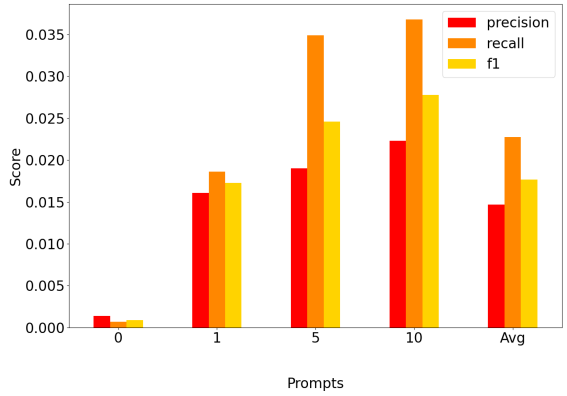

Figure 5: Results for varying number of shots for Variome (NER+RE)

A significant improvement in the F1 score could be observed for entity recognition with the incremental addition of examples in the prompts whereas little variation was observed for RE. Despite that, it should be noted that highest recall can be observed for the prompt with ten examples for RE showing that this addition has led to a better identification of the entities and relations. Moreover, more variation in types extracted could be observed with the increase of examples, for example, Figure 6 shows types such as 'cohort-patient' and 'body-part' being extracted for Variome.

The increased performance utilising few-shot prompting can be attributed to the ability of generative LLMs to learn in-context which was achieved with the addition of examples of texts, extracted genetic entities and identified relations, and their associated labels (Brown et al., 2020).

3.2 High recall, low precision across tasks for few-shot prompting

It can be seen across all three IE tasks that recall is higher than precision for few-shot prompting. This leads us to infer that a significant amount of correct entities matching the ground truth were captured despite generating false positive entities (see further detail in Figures A10 and A11).

This could be attributed to the generative nature of these models leading to over-generation, thereby extracting a large number of truly correct entities while also producing many false positives.

3.3 Low recall, high precision across tasks for zero-shot prompting

It can be observed across tasks that recall is lower and precision is higher for zero-shot prompting. For example, for NER, one of the reasons was the model over-generating tuples with the misaligned entity position in place of the extracted span, for example for the label 'Disease' the model would state '0-29' instead of the span name 'Glioblastoma multiforme congenito infratentorial' which was found at '11-59'.

This could be deduced to be due to the model being unable to learn in-context due to the lack of examples, leading to identification of a limited number of correct entities and over-generating false negative entities (Brown et al., 2020).

3.4 Lesser the shots, higher the hallucinations

One of the key failures observed was the inability of the model to adhere to the task outlined in the prompt leading to hallucinations and incorrect extraction of entities and relations. Hallucinations were entities that were discarded as fabrications or

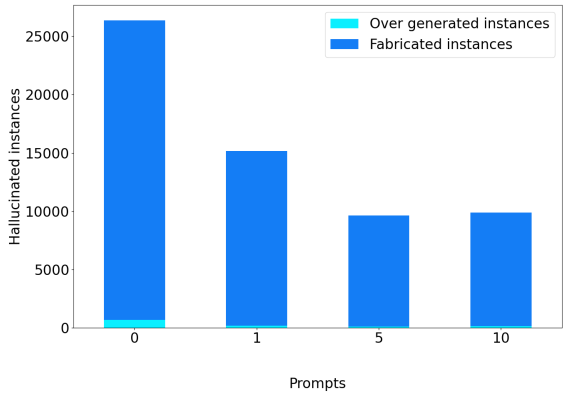

Figure 7: Hallucinations by type for varying number of shots for TBGA (RE)

over-generations (see definitions, Section 2).

Extracting named entities from the GenoVarDis dataset resulted in a majority of over-generated and a minute amount of fabricated hallucinations for all prompts, with the exception of the Spanish-language prompt adhering to zero-shot prompting which resulted in an extensive amount of fabrications. Extracting relations from the TBGA dataset resulted mainly in fabricated instances while end-to-end NER+RE utilising the Variome dataset showed a mix of both hallucination types.

A decrease in the amount of hallucinated instances was observed with the addition of examples in the prompts. A gradual increase in the number of matching instances extracted can be observed with the increase in the number of examples (see detail in Figures A1, A2, A3).

Overall, these hallucinations may be due to various factors, including the complexity of the IE tasks, limitations in the prompts with regard to providing context for the tasks, the generative nature of the model used, and limitations due to the LLM not being specifically trained on biomedical data. Further breakdown of hallucination types can be found in Figure 7, or Appendix Figures A4-A5.

It should be noted that issues such as fabrication and over-generation are a result of the generative nature of the model explored in this paper. Such issues are not encountered with traditional information extraction and classification approaches.

3.4.1 Fabrications

Upon manual inspection of the extracted data, hallucinated data and the gold standard data, the following reasons for the fabrications were identified.

1. Letter case of the entity not matching the entity in the text, for example hallucinated entity 'Carcinomas basocelulares' being stated in all lower case in the associated text.
2. Spans containing the desired entity with fabricated words or characters before or after the identified entity, for example, the entity 'dipeptidyl peptidase IV' in a TBGA dataset text is extracted by the model as 'dipeptidyl peptidase-4 inhibitor'.
3. Entity spans being produced instead of the entity string being extracted. This phenomenon was mainly observed for the Spanish language dataset, GenoVarDis, when using zero-shot prompting. Based on an analysis of the breakdown of types of the entities impacted by this, a majority of these positions were annotated as type 'Gene' (Figure 8).
4. Complete fabrications which could not be mapped to any position in the text, for example, the extracted relation 'Gene: SIVA Disease: NA' was discarded as a hallucination due to 'NA' not appearing in the corresponding text from the TBGA dataset, *'Thus, the role of SIVA in tumorigenesis remains unclear.'*.
5. The model would not adhere to the outlined output structure.

3.4.2 Over-generation

Entity recognition resulted in the majority of the over-generated instances observed. While most of these instances could be mapped to a position in the relevant text, the output included more entity mentions than were actually stated in the text. As such, these entity tuples being marked as hallucinations. For example, in one text in the GenoVarDis dataset, the gene 'PMP22' is mentioned in seven locations while the model hallucinated an additional 34 instances.

3.5 Exact matching leading to high amounts of false positives

Extracted tuples and triplets were neither manually manipulated nor normalised during post-processing, as our goal was to explore the direct performance results, based on exact matching of the extracted entities and the identified relations relations with the gold standard data.

One of the contributing factors to the inadequate performance of the tasks can be attributed to the model labelling entities with fabricated labels, for

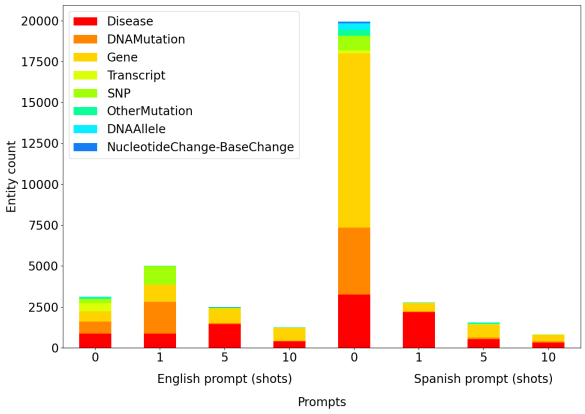

Figure 8: Hallucinations by entity type for varying number of shots for GenoVarDis (NER)

example, a large portion of the extracted entities for the Variome corpus reference the label 'Disease' for what should have been 'disease' entities. This led to a high number of false positives, Figure 6. While easily resolved through case-insensitive matching, this illustrates how the LLM did not strictly follow instructions; the annotation guideline only specifies 'disease' as a label option for the entities.

We have observed that the model mislabels entities and misidentifies relations, especially when the entities and the relations are highly specific to the biomedical domain.

Furthermore, one of the key issues that arises due to exact matching is when the model extracts the entities and identifies the correct relations whilst adding further information to the span, introducing false positives. For example, the target Variome corpus entity 'characteristic microsatellite instable' was extracted as 'characteristic microsatellite instable **tumours**'.

Issues seen with exact matching could be avoided with normalisation, error correction or changes to the evaluation settings such as relaxed/overlap matching and considering multiple plausible annotations, similar to the methodologies outlined by Dagdelen et al. (Dagdelen et al., 2024).

4 Related Work

The availability of generative AI and LLMs has driven substantial developments in NLP. These LLMs are being used for IE due to their capabilities related to text generation, understanding and generalisation.

4.1 IE tasks utilising LLMs

Research has explored joint IE tasks, NER and RE, utilising LLMs for successful IE using scientific datasets specifically designed to test IE of biomedical data (Dagdelen et al., 2024; Goel et al., 2023). Research shows that general-domain LLMs show great performance when various learning paradigms were utilised in their methods for IE from biomedical text, regardless of not being trained specifically for specific domains, specifically (Wadhwa et al., 2023; Agrawal et al., 2022). Moreover, LLMs have been shown to provide great results for medical NLP, which closely relates to biomedical NLP (Agrawal et al., 2022; Goel et al., 2023).

While not a generative LLM, BERT pre-trained and fine-tuned for biomedical data has shown great performance for task-specific NLP models compared to general-domain LLMs (Gu et al., 2020). The general-domain LLM, GPT-3, has been shown to perform close to fully supervised models and outperform existing solutions for IE of biomedical data for the task of RE (Wadhwa et al., 2023; Agrawal et al., 2022). When exploring gene set summarisation using zero-shot learning, it was found that the new GPT models performed well and were free of hallucinations but were unable to generalise missing key terms along the way (Joachimiak et al., 2023).

Inspired by the above research, this project utilised a general-domain generative LLM, GPT-3.5 Turbo, to conduct experiments on IE tasks to determine the performance of various prompting strategies and undertake a comprehensive analysis on how effectively genetic entities and relations can be extracted from scientific literature.

4.2 Prompt engineering for domain specific IE tasks

Variation in prompt strategies for IE has been shown to have a great impact on results with LLMs (Peng et al., 2023; Xu et al., 2024). There are many ways to design prompts under various learning paradigms and methods such as few-shot, zero-shot, chain-of-thought and question answering.

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 (Ashok and Lipton, 2023). Few-shot prompting has resulted in great

performance for both IE tasks, NER and RE, across various domains (Wadhwa et al., 2023; Goel et al., 2023). For example, performance achieved was found to be close to fully supervised models utilising 10 examples, which was found to be the optimal number of examples, adhering to the few-shot learning paradigm (Wadhwa et al., 2023). 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 (Agrawal et al., 2022). With the provision of annotated guidelines in the prompt along with fine-tuning, zero-shot results have shown to improve IE tasks (Sainz et al., 2024; Marchesin and Silvello, 2022). The above research indicates providing more context to the prompts provided to the models lead 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, it has not been compared with other learning paradigms for NER, RE tasks. Findings from above literature influences our research where we test the performance of NER, RE and joint NER and RE (NER+RE) with complex prompts with annotation guidelines under various paradigms.

Across various domains there are many investigations of the effect the output structure has on the performance of IE tasks. It was discovered that requesting the output from the model to be in a specific structure leads to an increase in accuracy of information extracted. Requesting the output to be in a table format via the prompt, where the table headers were either specified by the user or inferred using context by the models (Jiao et al., 2023); extracted text being output as a summary (Chang et al., 2024); structured output requested in the YAML format (Goel et al., 2023); output summarised into a natural sentence according to a predefined pattern and then extracted into an end-to-end (E2E) output template which has placeholders for the expected triggers and arguments (Hsu et al., 2021) are examples of different output formats which impacted performance of IE tasks. Inspired by the aforementioned research, in our approach we request the output to be structured in the tab separated vector (TSV) format along with the expected headers for the tuples and triplets extracted specified in order to obtain results with higher accuracy.

Upon exploration and evaluation of RE by looking at token-level annotation, phase level annotation and end-to-end relation extraction by Agrawal et al., it was found that it is difficult to guide LLMs to match exact schema (Agrawal et al., 2022). Moreover, it was discovered that there was bias in the results where the LLM was outputting a non-trivial answer even when none existed. This paper further highlighted the importance of crafting prompts for IE tasks to avoid such issues by, for example chaining multiple prompts and using an output structure such as sequence tagging. Findings from this influences our research greatly with relation to including more complexity and specificity when undertaking prompt engineering.

With various LLMs explored for exact word matching for joint NER and RE tasks, performance was shown to be negatively affected when the LLMs slightly change the phrasing or notion of the output when extracting entities and relations due to the ambiguity of the real-world IE tasks. Some of the solutions proposed to correct this issue include performing manual scoring of the results to assess correctedness of core information by looking at entity normalisation, error correction and multiple plausible annotations (Dagdelen et al., 2024).

According to Goel et al., it is clear that LLMs can significantly accelerate IE, with baseline accuracy compared to a trained NLP annotator (Goel et al., 2023). It was discovered that there was superior recall at the expense of precision when utilising LLMs. These results were stated to be mainly due to prompt engineering with few-shot paradigm without any parameter tuning directly. This was shown to save time and cost as it resulted in generating human expert-level annotations.

Based on the above, it can be observed that there has been a lack of a comprehensive investigation of the effectiveness of the prompt structure on an end-to-end IE process for genetic information extraction – particularly across NER, RE, and NER+RE – which was explored in this paper.

4.3 Biomedical literature and datasets

There exist limited datasets to test IE tasks in the biomedical domain. Some of the available datasets include GENIA (Kim et al., 2003), TBGA (Marchesin and Silvello, 2022), and UniProt (Bairoch and Apweiler, 1997), where data has been curated from English language literature. The lack of resources in the biomedical domain can be attributed to high level of expertise required for detailed annotation, lack of publicly available datasets, and restrictions on the usage of some existing datasets

with LLMs. For example, A Agrawal et al. (2022) utilise a dataset which was a modification of the English-language annotated CASI dataset (Moon et al., 2014) as it is publicly available to support NLP tasks. It is also worth noting the costliness in the curation of databases by experts in the biomedical field contributing to the lack of research in RE (Marchesin and Silvello, 2022). This leads to annotated corpora being limited in size, which prevents models from scaling effectively to large amounts of data (Elangovan et al., 2022). It was also found that general purpose LLMs find it difficult to provide good results for domain-specific extraction of information with datasets containing limited information (Park et al., 2023). It can be observed that in an already resource poor domain for IE, finding a publicly available dataset to support NLP research across languages in the biomedical domain is difficult. While there exists limited datasets trained on models to encourage multilingual IE, there is room to explore whether general-domain generative LLMs could be utilised to create robust datasets to improve IE tasks (Carrino et al., 2022).

4.4 IE tasks on non-English literature

It is worth noting that information from literature conducted in non-English domains has the potential to provide a diverse perspective to the biomedical knowledge built using English-language only datasets and aid in advancements in medical research (Rezaeian, 2015; AlShuweihi et al., 2020). The effectiveness of generative LLMs on the extraction of genetic information in a cross-linguistic setting using a Spanish-language dataset showed that on average English-language prompts provide higher performance agnostic of the language of the dataset (Kodikara and Verspoor, 2024). This was attributed to the fact that LLMs were predominantly trained on English-language data. In order to move towards creating solutions for non-English language literature, our research included an investigation of the limitations of NER using Spanish language scientific literature in GenoVarDis.

5 Conclusion

We explored the use of a generative LLM for end-to-end genetic information extraction across several tasks and datasets. We additionally explored limitations of using a generative model by analysing hallucinated instances generated for each IE task.

Through our evaluation of prompting strategies we show that few-shot prompting provides optimal performance for tasks involving named entity recognition. We further show that there is minimal effect of learning paradigms for identification of relations between genetic entities.

Key limitations of a generative model include over-generation and fabrication of entities demonstrating that generative models struggle to adhere to the task outlined in the instructions.

Further research needs to be conducted to explore ways in which performance can be further improved along with minimising the negative impacts of using generative models for IE in the biomedical domain before using them practically.

Acknowledgments

We thank the RACE Hub of RMIT University for providing access to the Azure Open AI API service.

References

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David A. Sontag. 2022. Large language models are few-shot clinical information extractors. In *Conference on Empirical Methods in Natural Language Processing*.

Marvin M. Agüero. 2024. GenoVarDis. Accessed on May 20, 2024.

Marvin M. Agüero-Torales, Carlos Rodríguez Abellán, Marta Carcajona Mata, Juan Ignacio Díaz Hernández, Mario Solís López, Antonio Miranda-Escalada, Sergio López-Alvárez, Jorge Mira Prats, Carlos Castaño Moraga, David Vilares, and Luis Chiruzzo. 2024. Overview of GenoVarDis at IberLEF 2024: NER of Genomic Variants and Related Diseases in Spanish. *Procesamiento del Lenguaje Natural*, 73.

Ameer Albahem, Karin Verspoor, and Antonio Jose Jimeno Yepes. 2013. BRAT-Eval v0.3.2.

Mohamed AlShuweihi, Said A. Salloum, and Khaled F. Shaalan. 2020. 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*.

Dhananjay Ashok and Zachary Chase Lipton. 2023. PromptNER: Prompting for named entity recognition. *arXiv*, abs/2305.15444.

Amos Bairoch and Rolf Apweiler. 1997. The SWISS-PROT protein sequence data bank and its supplement trembl. *Nucleic acids research*, 25 1:31–6.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda

Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Casimiro Pio Carrino, Joan Llop, Marc Pàmies, Asier Gutiérrez-Fandiño, Jordi Armengol-Estap'e, Joaquín Silveira-Ocampo, Alfonso Valencia, Aitor Gonzalez-Agirre, and Marta Villegas. 2022. Pretrained biomedical language models for clinical NLP in spanish. In *Workshop on Biomedical Natural Language Processing*.

Jiayu Chang, Shiyu Wang, Chen Ling, Zhaohui Qin, and Liang Zhao. 2024. Gene-associated disease discovery powered by large language models. volume abs/2401.09490.

Luis Chiruzzo, Salud María Jiménez-Zafra, and Francisco Rangel. 2024. 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*.

Adrien Coulet, Nigam H. Shah, Yael Garten, Mark A. Musen, and Russ B. Altman. 2010. Using text to build semantic networks for pharmacogenomics. *Journal of Biomedical Informatics*, 43 6:1009–19.

John Dagdelen, Alex Dunn, Sanghoon Lee, Nicholas Walker, Andrew S. Rosen, Gerbrand Ceder, Kristin A. Persson, and Anubhav Jain. 2024. Structured information extraction from scientific text with large language models. *Nature Communications*, 15.

Aparna Elangovan, Yuan Li, Douglas EV Pires, Melissa J Davis, and Karin Verspoor. 2022. Large-scale protein-protein post-translational modification extraction with distant supervision and confidence calibrated biobert. *BMC Bioinformatics*, 23:1–23.

Akshay Goel, Almog Gueta, Omry Gilon, Chang Liu, Sofia Erell, Lan Huong Nguyen, Xiaohong Hao, Bolous Jaber, Shashir Reddy, Rupesh Kartha, Jean Steiner, Itay Laish, and Amir Feder. 2023. Llms accelerate annotation for medical information extraction. In *Machine Learning for Health (ML4H) Symposium*.

Janet Piñero González, Juan Manuel Ramírez-Anguita, Josep Saüch-Pitarch, Francesco Ronzano, Emilio Centeno, Ferran Sanz, and Laura Inés Furlong. 2019. The disgenet knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research*, 48:D845 – D855.

Yu Gu, Robert Tinn, Hao Cheng, Michael R. Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3:1 – 23.

I-Hung Hsu, Kuan-Hao Huang, Elizabeth Boschee, Scott Miller, Premkumar Natarajan, Kai-Wei Chang, and Nanyun Peng. 2021. Degree: A data-efficient generation-based event extraction model. In *North American Chapter of the Association for Computational Linguistics*.

Yizhu Jiao, Ming Zhong, Sha Li, Ruining Zhao, Siru Ouyang, Heng Ji, and Jiawei Han. 2023. Instruct and extract: Instruction tuning for on-demand information extraction. In *Conference on Empirical Methods in Natural Language Processing*.

Marcin P. Joachimiak, John Harry Caufield, Nomi L. Harris, and Chris J. Mungall. 2023. Gene set summarization using large language models. *arXiv*.

Maryam Khordad and Robert E. Mercer. 2017. Identifying genotype-phenotype relationships in biomedical text. *Journal of Biomedical Semantics*, 8.

Jin-Dong Kim, Tomoko Ohta, Yuka Tateisi, and Junichi Tsujii. 2003. Genia corpus - a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19 Suppl 1:i180–2.

Milindi Kodikara and Karin M. Verspoor. 2024. Effectiveness of cross-linguistic extraction of genetic information using generative large language models.

Stefano Marchesin and G. Silvello. 2022. TBGA: a large-scale gene-disease association dataset for biomedical relation extraction. *BMC Bioinformatics*, 23.

Sungrim Moon, Serguei V. S. Pakhomov, Nathan Liu, James O. Ryan, and Genevieve B. Melton. 2014. A sense inventory for clinical abbreviations and acronyms created using clinical notes and medical dictionary resources. *Journal of the American Medical Informatics Association*, 21 2:299–307.

Gilchan Park, Byung-Jun Yoon, Xihaier Luo, Vanessa Lpez-Marrero, Patrick Johnstone, Shinjae Yoo, and Francis J. Alexander. 2023. Automated extraction of molecular interactions and pathway knowledge using large language model, galactica: Opportunities and challenges. In *Workshop on Biomedical Natural Language Processing*.

C.A.I. Peng, Xi Yang, Kaleb E. Smith, Zehao Yu, Aokun Chen, Jiang Bian, and Yonghui Wu. 2023. Model tuning or prompt tuning? a study of large language models for clinical concept and relation extraction. *Journal of Biomedical Informatics*, page 104630.

Tim E. Putman, Kevin Schaper, Nicolas Matentzoglu, Vincent Rubinetti, Faisal S Alquaddoomi, Corey Cox, John Harry Caufield, Glass Elsarboukh, Sarah

Gehrke, Harshad B. Hegde, Justin T. Reese, Ian Braun, Richard M. Bruskiewich, Luca Cappelletti, Seth Carbon, Anita R. Caron, Lauren E. Chan, Christopher G. Chute, Katherina G Cortes, Vinícius De Souza, Tommaso Fontana, Nomi L. Harris, Emily L Hartley, Eric Hurwitz, Julius O. B. Jacobsen, Madan Krishnamurthy, Bryan Laraway, James A McLaughlin, Julie A. McMurry, Sierra A T Moxon, Kathleen R Mullen, Shawn T O'Neil, Kent A. Shefchek, Ray Stefancsik, Sabrina Toro, Nicole A. Vasilevsky, Ramona L Walls, Patricia L. Whetzel, David Osumi-Sutherland, Damian Smedley, Peter N. Robinson, Christopher J. Mungall, Melissa A. Haendel, and Monica C. Munoz-Torres. 2023. The Monarch Initiative in 2024: An analytic platform integrating phenotypes, genes and diseases across species. *Nucleic Acids Research*, 52:D938 – D949.

Mohsen Rezaeian. 2015. Disadvantages of publishing biomedical research articles in English for non-native speakers of English. *Epidemiology and Health*, 37.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Sohel Mondal, and Aman Chadha. 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv*, abs/2402.07927.

Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, and Eneko Agirre. 2024. GoLLIE: Annotation guidelines improve zero-shot information-extraction. In *The Twelfth International Conference on Learning Representations*.

Takeshi Sekimizu, Hyun Seok Park, Hyun Seok Park, Junichi Tsujii, and Junichi Tsujii. 1998. Identifying the interaction between genes and gene products based on frequently seen verbs in medline abstracts. *Genome informatics. Workshop on Genome Informatics*, 9:62–71.

Ayush Singhal, Michael Simmons, and Zhiyong Lu. 2016. Text mining genotype-phenotype relationships from biomedical literature for database curation and precision medicine. *PLoS Computational Biology*, 12.

Pontus Stenetorp, Sampo Pyysalo, Goran Topic, Tomoko Ohta, Sophia Ananiadou, and Junichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Conference of the European Chapter of the Association for Computational Linguistics*.

Joshua M. Temkin and Mark R. Gilder. 2003. Extraction of protein interaction information from unstructured text using a context-free grammar. *Bioinformatics*, 19 16:2046–53.

Karin Verspoor, Antonio Jimeno Yepes, Lawrence Cavedon, Tara McIntosh, Asha Herten-Crabb, Zoë Thomas, and John-Paul Plazzer. 2013. Annotating the biomedical literature for the human variome. *Database*, 2013:bat019.

Karin M. Verspoor, Go Eun Heo, Keun Young Kang, and Min Song. 2016. Establishing a baseline for literature mining human genetic variants and their relationships to disease cohorts. *BMC Medical Informatics and Decision Making*, 16.

Somin Wadhwa, Silvio Amir, and Byron C. Wallace. 2023. Revisiting relation extraction in the era of large language models. *Proceedings of the conference. Association for Computational Linguistics. Meeting*, 2023:15566–15589.

Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, and Enhong Chen. 2024. Large language models for generative information extraction: A survey. *Frontiers of Computer Science*.

Ping Yu, Hua Xu, Xia Hu, and Chao Deng. 2023. Leveraging generative ai and large language models: A comprehensive roadmap for healthcare integration. *Healthcare*, 11.

A Appendix

A.1 Example Spanish prompt for NER

"Encuentre las entidades en el siguiente texto en español. La cantidad de entidades encontradas debe coincidir con la cantidad de veces que se menciona la entidad en el texto."

A.2 Example guideline for NER

"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')."

A.3 Example expected output format for RE

"Display results in the tsv format with the column headers 'Gene', 'Disease', 'Relation' to annotate the entities. Provide each triplet in a new line."

A.4 Further analysis of results

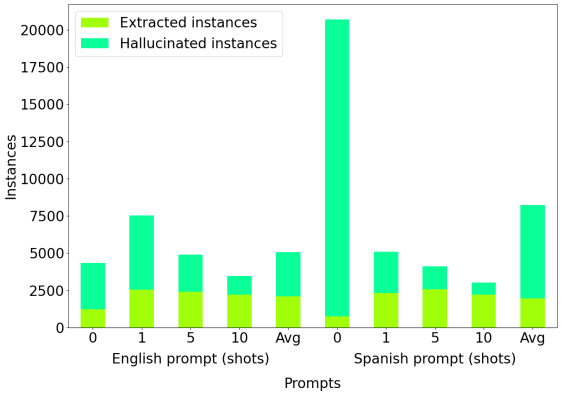

Figure A1: Instances for varying number of shots for GenoVarDis (NER)

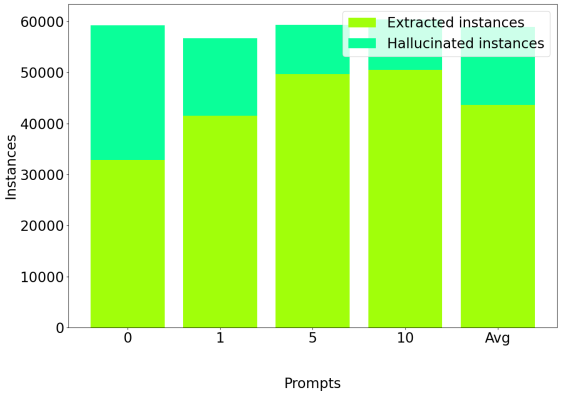

Figure A2: Instances for varying number of shots for TBGA (RE)

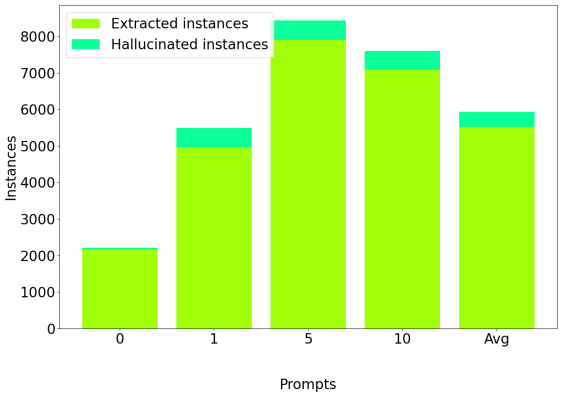

Figure A3: Instances for varying number of shots for Variome (NER+RE)

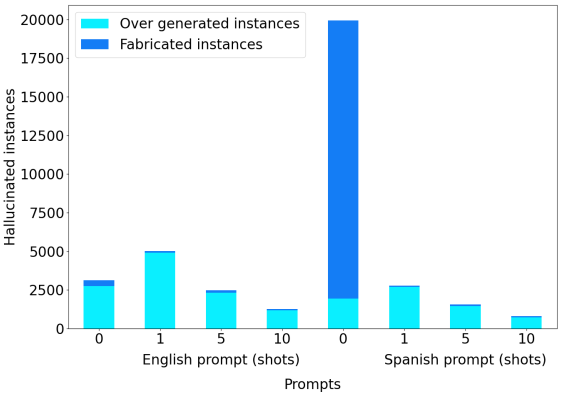

Figure A4: Hallucinations by type for varying number of shots for GenoVarDis (NER)

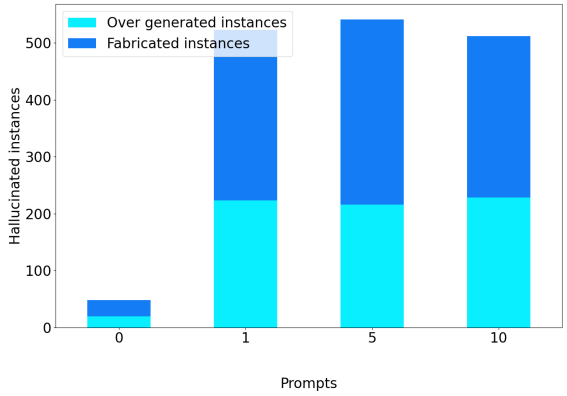

Figure A5: Hallucinations by type for varying number of shots for Variome (NER+RE)

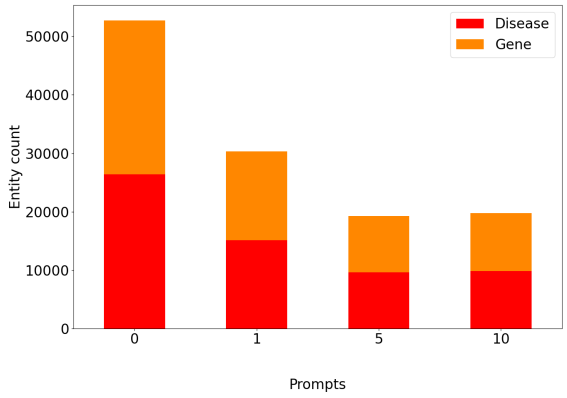

Figure A6: Hallucinations by entity type for varying number of shots for TBGA (RE)

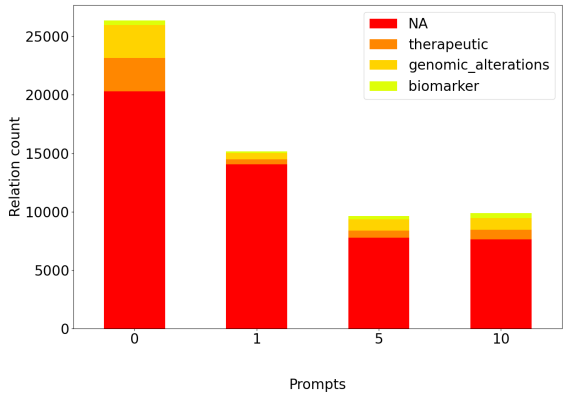

Figure A7: Hallucinations by relation type for varying number of shots for TBGA (RE)

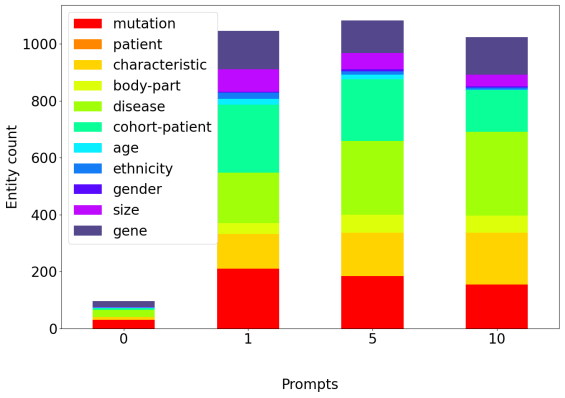

Figure A8: Hallucinations by entity type for varying number of shots for Variome (NER+RE)

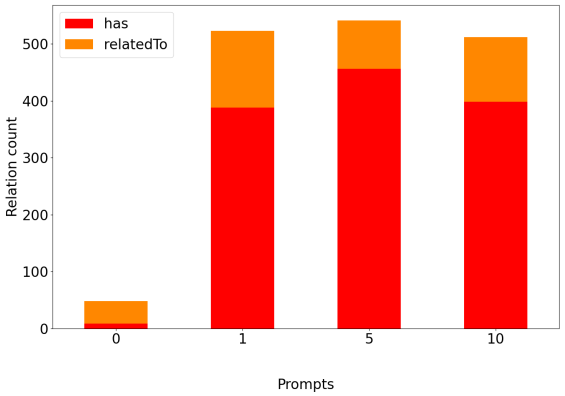

Figure A9: Hallucinations by relation type for varying number of shots for Variome (NER+RE)

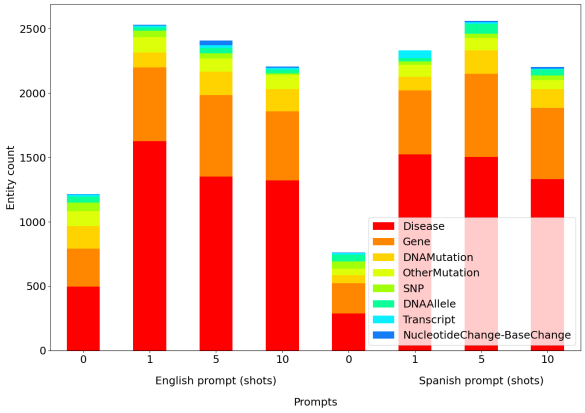

Figure A10: Extracted entity types for varying number of shots for GenoVarDis (NER)

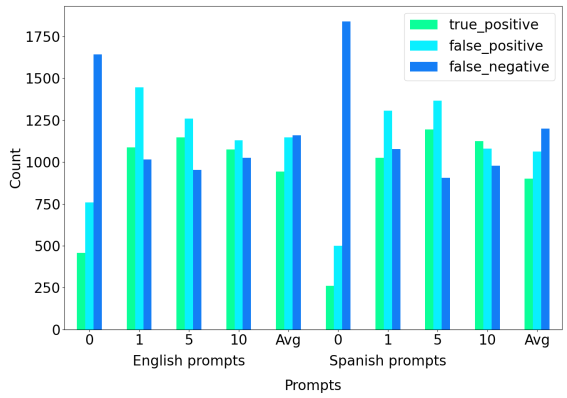

Figure A11: Entity division for varying number of shots for GenoVarDis (NER) grouped by prompt language

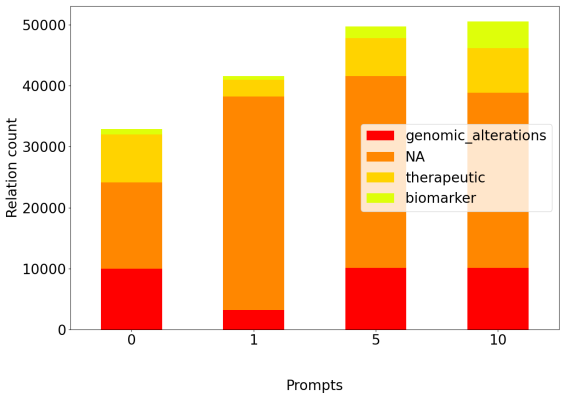

Figure A14: Extracted relation types for varying number of shots for TBGA (RE)

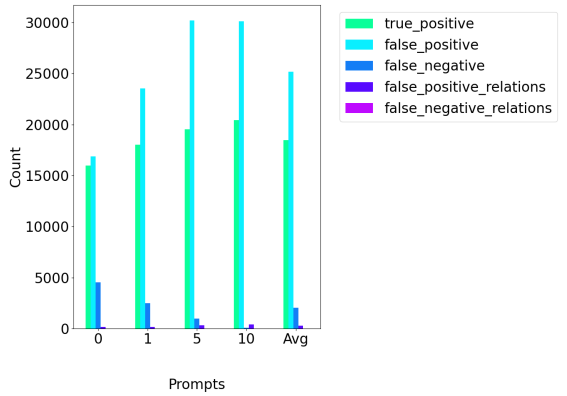

Figure A12: Entity and relation division for varying number of shots for TBGA (RE)

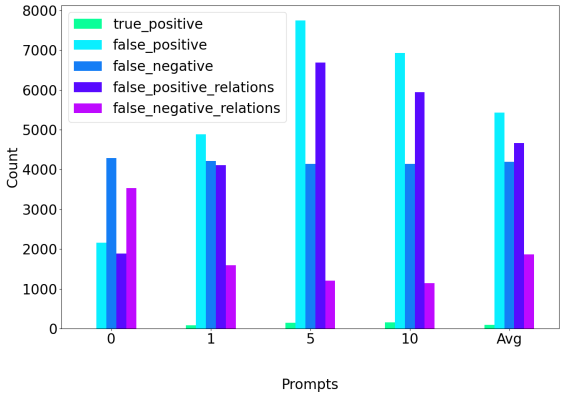

Figure A15: Entity and relation division for varying number of shots for Variome (NER+RE)

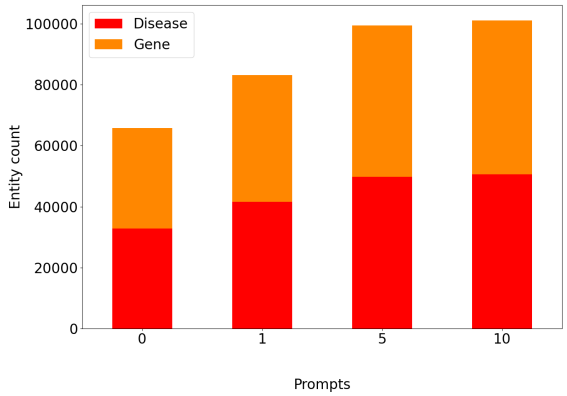

Figure A13: Extracted entity types for varying number of shots for TBGA (RE)

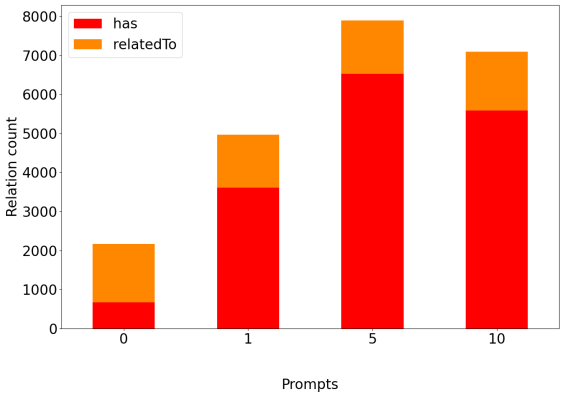

Figure A16: Extracted relation types for varying number of shots for Variome (NER+RE)

Table A1: Dataset annotation schema

Dataset	Annotation type	Label	Description
GenoVarDis	Entity	DNAAllele	Allele on DNA sequence
	Entity	DNAMutation	Variant on DNA sequence
	Entity	Disease	Disease
	Entity	Gene	Gene
	Entity	NucleotideChange-BaseChange	Wild type and mutant
	Entity	OtherMutation	Variant with insufficient information
	Entity	SNP	RS number, COSMIC mutation
	Entity	Transcript	Transcript
TBGA	Entity	Disease	Disease
	Entity	Gene	Gene
	Relation	biomarker	Gene is a biomarker for the disease
	Relation	genomic _alterations	Genomic alteration is linked to the gene associated with the disease phenotype
	Relation	therapeutic	Drug associated with disease
	Relation	NA	False association
Variome	Entity	characteristic	Characteristic of disease or tumour
	Entity	age	Number or range indicating how old a person/group of people is
	Entity	body-part	An organ or anatomical location in a person
	Entity	cohort-patient	patient - Individual with a disease; cohort - A group of people
	Entity	disease	An abnormal condition affecting the body of an organism.
	Entity	ethnicity	Where a person/group of people comes from, either based on ethnic origin or where they live
	Entity	gender	Terms indicating whether someone is male or female
	Entity	gene	Segment of DNA that codes for a protein
	Entity	mutation	Alteration of nucleotides or amino acids
	Entity	size	Number of people in a cohort, or mutation frequency
	Relation	has	X–has–Y
	Relation	relatedTo	X–relatedTo–Y

Label descriptions taken directly from the associated papers.

Table A2: Breakdown of dataset entity and relation types

Dataset	Label	Training set count	Test set count
GenoVarDis	DNAAllele	139	15
	DNAMutation	496	73
	Disease	4028	1433
	Gene	3093	514
	NucleotideChange-BaseChange	51	1
	OtherMutation	271	271
	SNP	120	120
	Transcript	1	1
TBGA	Disease	178264	20516
	Gene	178264	20516
	biomarker	20145	2315
	genomic _alterations	32831	2209
	therapeutic	3139	384
	NA	122149	15608
Variome	characteristic	136	1363
	age	10	79
	body-part	37	454
	cohort-patient	133	2016
	disease	237	2137
	ethnicity	7	38
	gender	2	78
	gene	15	825
	mutation	81	945
	size	52	655
	has	293	3714
	relatedTo	62	581