REAL: A Retrieval-Augmented Entity Linking Approach for Biomedical Concept Recognition

Darya Shlyk[1]*, Tudor Groza[2], Stefano Montanelli[1], Emanuele Cavalleri[1] and Marco Mesiti [1]
[1]Università degli Studi di Milano, Via Giovanni Celoria 19, 20133 Milan, Italy
[2]School of Electrical Engineering, Computing and Mathematical Sciences,
Curtin University, Kent St, Bentley WA 6102, Australia

Abstract

Large Language Models (LLMs) offer an appealing alternative to training dedicated models for many Natural Language Processing (NLP) tasks. However, outdated knowledge and hallucination issues can be major obstacles in their application in knowledge-intensive biomedical scenarios. In this study, we consider the task of biomedical concept recognition (CR) from unstructured scientific literature and explore the use of Retrieval Augmented Generation (RAG) to improve accuracy and reliability of the LLM-based biomedical CR. Our approach, named REAL (Retrieval Augmented Entity Linking), combines the generative capabilities of LLMs with curated knowledge bases to automatically annotate natural language texts with concepts from bio-ontologies. By applying REAL to benchmark corpora on phenotype concept recognition, we show its effectiveness in improving LLM-based CR performance. This research highlights the potential of combining LLMs with external knowledge sources to advance biomedical text processing. Source code is available at: `https://github.com/dash-ka/REAL-BioCR`.

1 Introduction

Biomedical Concept Recognition (CR) aims to identify and link textual mentions of biomedical concepts to entries in expert-curated knowledge bases and ontologies. CR combines two subtasks from the standard information extraction pipeline: entity recognition (NER) and entity linking (EL), sometimes referred to as named entity disambiguation (NED) or grounding. NER aims to detect strings in text that refer to classes of biomedical entities, such as phenotypes, diseases, or genes. EL maps those strings to terms in an ontology, such as Human Phenotype Ontology (HPO) (Köhler et al., 2014) for phenotypic features, Mondo (Vasilevsky et al., 2020) for disease terms, and HGNC (Eyre et al., 2006) for human genes. Automated CR methods represent an active research area and are essential for a range of downstream biomedical applications. In genomic medicine, for instance, accurately recognizing phenotype concepts from free-text medical notes is the starting point to improve genetic disease diagnostics (Labbé et al., 2023).

State-of-the-art CR systems rely on fine-tuning transformer-based language models pretrained on biomedical texts, such as BioBERT (Lee et al., 2020), and have restricted scope, targeting a single or few application domains (Feng et al., 2022; Luo et al., 2021). The major limitation of these approaches is the need for domain-specific training with expert-labeled corpora, which is not always feasible due to the scarcity of annotated data in the biomedical field (Fries et al., 2022). On the other hand, general-purpose Large Language Models (LLM), such as OpenAI's Generative-Pretrained Transformer (GPT), have demonstrated remarkable zero and few-shot learning abilities, offering significant potential for biomedical NLP. Recent studies have shown promising results when using LLMs in clinical information extraction without domain-specific training (Agrawal et al., 2022; Meoni et al., 2023). However, challenges persist regarding factual accuracy in generated responses, hindering their usability for knowledge-intensive tasks in specialized domains (Gao et al., 2023; Reese et al., 2023). To address these challenges, Retrieval-Augmented-Generation (RAG) (Lewis et al., 2020) has been recently proposed as a technique to enhance LLMs with relevant information retrieved from external knowledge bases through semantic similarity calculation.

Our study aims to explore the application of the RAG paradigm in the context of biomedical CR. To this end, we developed REAL, a Retrieval-Augmented Entity Linking approach for ontology-based CR. To overcome the limitation of training dedicated NER and EL models, our approach lever-

*Corresponding author: darya.shlyk@unimi.it

Proceedings of the 23rd Workshop on Biomedical Language Processing, pages 380–389
August 16, 2024. ©2024 Association for Computational Linguistics

ages prompting techniques with general purpose LLMs to handle both tasks in a unified pipeline. Given a text, REAL first identifies mentions of concepts belonging to some target biomedical domain through a zero-shot NER, and then associates these mentions to terms in the domain ontology using retrieval-enhanced Entity Linking. By embedding the mention and ontology concepts into a common dense space, the retrieval mechanism provides the LLM with a selection of candidates from a bio-ontology identified through nearest neighbor search. By synergistically combining the retrieval mechanism with prompt-engineering, REAL aims to leverage up-to-date knowledge with existing knowledge bases, thereby improving the accuracy and reliability of LLM-based CR.

We summarize our contributions as follows:

- We propose a novel RAG-based approach that leverages general-purpose LLMs for automatic annotation of unstructured scientific literature with concepts from bio-ontologies. Our approach is versatile and can be easily adopted in various application domains without requiring domain-specific training.

- We conduct experiments with two benchmark corpora, studying the effectiveness of our approach on the phenotype concept recognition task. The results show that REAL can achieve competitive performance, indicating a great promise for the RAG paradigm in the context of biomedical concept recognition.

2 Related Work

2.1 Biomedical Concept Recognition

Biomedical CR tools predominantly rely on dictionary-based methods, using lexical matching with lookup tables. The OBO annotator (Taboada et al., 2014), the NCBO annotator (Jonquet et al., 2009), and the Monarch Initiative platform (Putman et al., 2023) are examples of tools that achieve high precision, but often suffer from low recall.

To overcome the limitations of dictionary-based methods, the recent research explored the use of neural-based models, with significant performance improvements. State-of-the-art approaches leverage pretrained BERT (Bidirectional Encoder Representations from Transformers) architectures. For instance, PhenoBert (Feng et al., 2022) implements a complex pipeline exploiting convolutional neural networks (CNNs) with BERT to automatically recognize HPO terms from free text. Phenotagger (Luo et al., 2021) is a hybrid approach that combines dictionary and deep learning methods. Specifically, Phenotagger fine-tunes a pretrained BioBERT model on weakly supervised datasets. These solutions necessitate task-specific training, requiring extensive computational resources and significant human effort for the manual annotation of large training corpora.

With the the advent of ChatGPT, researchers started to consider prompt-based approaches that leverage impressive language understanding capabilities of instruction-based generative models to address a wide spectrum of NLP tasks with no domain or task-specific training. One of the most prominent examples is SPIRES (Structured Prompt Interrogation and Recursive Extraction of Semantics) (Caufield et al., 2024) that leverages LLMs to assist the automatic construction of knowledge bases. Given an input text and a user-defined conceptual schema, the method recursively prompts an LLM to extract structured knowledge conforming with the schema's classes relevant for a given domain. The schema guides the LLM in extracting named entities that meet specific property constraints. To map extracted entities to ontology identifiers, SPIRES adopts the Ontology Access Kit library (OAKlib), which provides interfaces for external annotation tools, including the OBO annotator, and the Ontology Lookup Service.

2.2 Prompt-based Phenotyping

Several recent studies have employed prompt engineering techniques with LLMs to evaluate their capability in performing end-to-end phenotype concept recognition. Labbé et al. (2023) prompt GPT3.5 model to directly extract HPO term labels alongside corresponding IDs from medical texts. Their study highlights the limitations associated with purely prompt-based concept recognition, suggesting that potential improvements could be achieved by integrating factual knowledge from reference resources to aid in the generation process.

Groza et al. (2024) evaluated the OpenAI GPT-3.5 and GPT-4.0 models on phenotype concept recognition by testing alternative prompting strategies, including pipelined and in-context learning approaches. The former involves two sequential prompts: one for phenotype extraction and another for linking to HPO IDs. The latter approach incorporates the target subset of HPO label - ID pairs from the reference ontology inside the prompt as

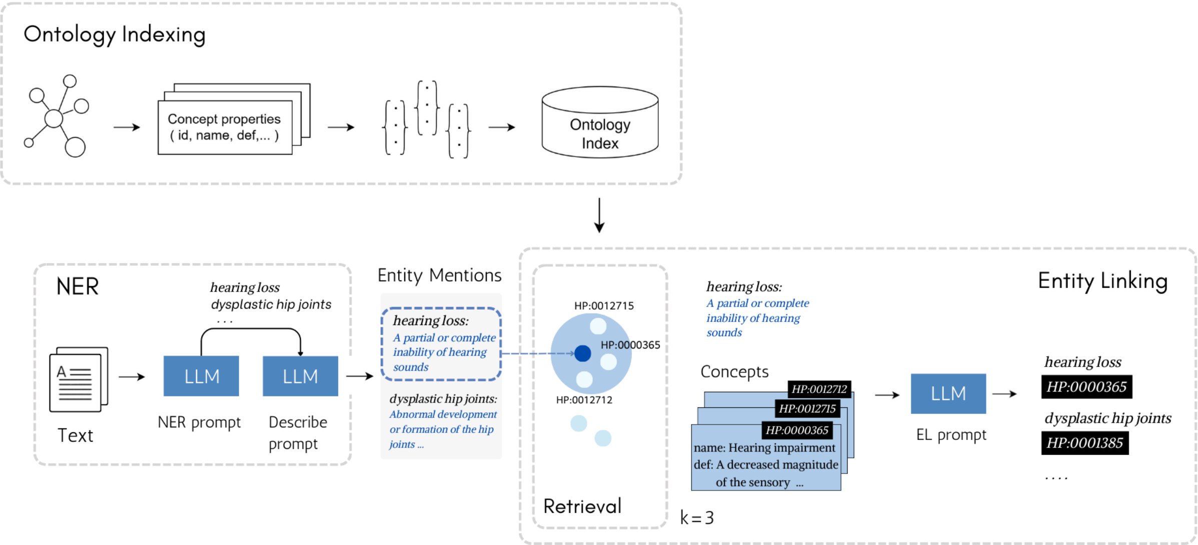


Figure 1: A high-level overview of the REAL approach.

context. Their findings demonstrate that in-context learning with pre-filtered ontology terms has the potential to surpass state-of-the-art CR systems.

The idea to couple parametric knowledge of an LLM with vast external knowledge repositories to improve the factuality and accuracy of the LLM responses forms the foundation of RAG. This technique involves chunking and embedding the knowledge resource into a set of vectors, followed by retrieving *top-k* relevant chunks based on semantic similarity with the user query, which are then incorporated into the LLM prompt. To the best of our knowledge, ours is the first work to explore the application of RAG in the field of biomedical concept recognition (CR). In REAL, we employ RAG to assist the LLM in linking textual mentions of biomedical entities to terms in domain ontologies.

3 Methodology

The ontology-based CR problem can be formally presented as two consecutive tasks, NER and EL, as follows. Let $\mathcal{O}$ denote a set of concepts $\{C_1, \ldots, C_n\}$ defined in the domain ontology used for text annotation. Given a text T, the NER task identifies textual mentions of biomedical entities from the target domain, $m_1, \ldots, m_h$. Then, the EL task consists in assigning each entity mention m_i to a concept $C \in \mathcal{O}$ that best represents it.

As shown in Figure 1, REAL implements CR in a pipeline consisting of three main phases: *Ontology Indexing*, *NER*, and *Retrieval-enhanced Entity Linking*. The ontology indexing is executed only once during the pre-processing to convert the concepts in $\mathcal{O}$ into a searchable index. The main workflow starts with the zero-shot NER (in the left bottom of the figure), where we prompt the LLM to extract instances of a specified entity type from T and generate a short definition for each of them. In the Retrieval-enhanced Entity Linking phase (right bottom part of the figure), we search in the ontology index the *top-k* most similar concepts to the embedding of m. They are the candidates for entity linking and the best matching is identified by properly instructing an LLM prompt. Details of the approach are provided in the remainder.

3.1 Ontology Indexing

To implement RAG for CR with the domain ontology $\mathcal{O}$ as a reference knowledge resource, we create vector embeddings for concepts in $\mathcal{O}$ and index them inside a vector store. This process creates an ontology index $\mathcal{I}$, that we can query to retrieve ontology concepts with the most similar embedding vector to the embedding of a given query.

In this study, we used ChromaDB[1], an open-source vector database, to store concept embeddings and perform semantic similarity search in the embedding space using the cosine similarity function. However, the proposed method can employ any database that enables efficient vector search capabilities. Unlike other vector stores, ChromaDB provides interfaces to popular LLM providers, and automatically computes the embedding from text using the specified embedding model. Specifically, this study uses the OpenAI

[1] https://www.trychroma.com/

Concept prompt

```
As a clinical expert, write a single sentence
definition that explains the meaning of the
concept: { concept name }
```

Figure 2: Prompt template for generating a definition for a given concept name.

`text-embedding-ada-002` model for concept embedding. We store the computed embeddings alongside the concept properties inside the index $\mathcal{I}$. To construct input texts for the embedding model, we employ the concept name and its definition provided among concept properties in a given ontology. This design choice stems from the need to create a vector representation that captures the meaning of a concept, with the concept name and definition providing the minimal necessary information to achieve this goal. Whenever a textual definition is not available in the concept properties, we automatically generate one from the concept name by prompting an LLM using the prompt in Figure 2 (the variable part of the prompt is colored in green).

Example 1 *Suppose we are interested in creating an ontology index for the HPO. A vectorial representation for each HPO concept is generated by concatenating its name and definition and stored into ChromaDB with other concept properties. For instance, for the HPO term "Breast carcinoma" (id:* `HP:0003002`*), the following text has been used for generating the concept embedding:*

```
name: Breast carcinoma
definition: Presence of carcinoma in breast
```

The prompt template in Figure 2 is used to generate a single sentence definition for 2,586 HPO terms that do not have a definition in the ontology.

3.2 NER

Given the input text T and the specification of the target biomedical entities to be extracted, the NER step produces a set of pairs $\mathcal{P} = \{(m_i, d_i) \mid 0 \leq i \leq h\}$, where m_i represents a mention of the target entity extracted from T, and d_i denotes a concise definition for that entity mention. This step employs zero-shot prompting with LLMs using two consecutive prompts: the *NER prompt* for entity extraction and the *Describe prompt* for definition generation. The NER prompt in Figure 3 incorporates the input text T and directs the LLM to extract spans in T that represent instances of the target entity type defined for a given application domain. The domain adaptation of the NER task is performed by changing the type of the target biomedical entity specified in the prompt, e.g., genes, phenotypes. The Describe prompt in Figure 4 operates on the list of entity mentions $m_1, \ldots, m_h$, produced with the NER prompt, and tasks the LLM to generate a definition for each extracted mention. Besides the list of entity mentions, this prompt also includes the original text T to help the LLM to compose contextually informed entity definitions.

NER prompt

```
From the text below, extract all mentions of
the following entities in the following format:
entities : < a semicolon-separated list of noun
phrases containing a mention of { target entities }
It must be semicolon-separated.
Split entities containing words "and", "," into
separate entities.>

Text:
{ input text }
```

Figure 3: Prompt template for NER.

Describe prompt

```
Given a text and a semicolon-separated list of
entities from that text, write a definition for
each entity in the following format:
<entity>: < a detailed definition that explains
the meaning of the entity in one sentence >

Here is the text:
{ input text }

Here are the entities:
{ entity mentions }
```

Figure 4: Prompt template for entity description.

Example 2 *Let T be the following text:*

```
The combination of either the skin tumours
or multiple odontogenic keratocysts.
```

Suppose we are interested in extracting mentions of human phenotypes from T. Then, in the NER prompt, we specify the following target entities: "human phenotypes, including physical abnormalities, symptoms of disease, and inherited disorders". Given T, the NER prompt extracts the entities: `skin tumors` *and* `odontogenic keratocysts`*.*

For each of them (and considering T) a definition is generated using the prompt in Figure 4 as follow:

skin tumors: Abnormal growths or masses that occur in the skin and can be benign or malignant.

odontogenic keratocysts: Cysts that develop in the jawbones and are derived from the remnants of dental tissue.

3.3 Retrieval Augmented Entity Linking

Given the Ontology index $\mathcal{I}$, and a set $\mathcal{P}$ of pairs (m, d) obtained as a result of the NER phase, for each element in $\mathcal{P}$, the EL phase is realized in two steps: *Candidate Retrieval* and *Entity Linking*.

3.3.1 Candidate Retrieval

Given the entity mention m with its definition d and a user-defined parameter k, we first embed (m, d) into the same embedding space with ontology concepts, and then retrieve *top-k* semantically similar concepts $\{C'_1, \ldots C'_k\}$ from $\mathcal{I}$ by approximate k-nearest neighbor search. We adopt the same embedding strategy for ontology concepts to enable consistent representation of the entity mentions in the common vector space (see Section 3.1).

Example 3 *Consider the first entity mention identified in Example 2. The following text is used to compute the mention embedding q:*

name: skin tumors
definition: Abnormal growths or masses that occur in the skin and can be benign or malignant.

By querying $\mathcal{I}$ with q, we can retrieve the top-3 concepts in HPO with the highest similarity score, according to cosine similarity function:

- ID: HP:0008069
  name: Neoplasm of the skin
  definition: A tumor (abnormal growth of tissue) of the skin.
  score: 0.9329
- ID: HP:0000951
  name: Abnormality of the skin
  definition: An abnormality of the skin.
  score: 0.8974
- ID: HP:0012056,
  name: Cutaneous melanoma
  definition:The presence of a melanoma of skin.
  score: 0.8937

```
EL prompt

As an expert clinician, your task is to accurately
identify the { domain ontology } concept mentioned
in the provided text using the concepts listed below.
Accuracy is paramount. If the text does not precisely
refer to any of the concepts listed below, please
return "None"; otherwise, return the corresponing
concept ID in the following format:
answer: ‹ concept ID or None ›
confidence: ‹ one of: HIGH, LOW, MEDIUM ›

Here are some examples:
{ examples }

Below are the concepts:
{ candidate concepts }

Text:
{ entity description }
```

Figure 5: Prompt template for EL.

3.3.2 Entity Linking

To ground m using the retrieved candidate set $\{C'_1, \ldots C'_k\} \subset \mathcal{O}$, we re-frame the EL task as a multiple-choice selection and prompt the LLM to identify the ontology concept among the provided candidates that best matches an entity description (m, d). The candidate concepts are provided as part of the prompt with their properties, including concept ID, name and definition. When selecting a concept for a given mention m, the LLM is instructed to associate a confidence level with its answer (a value in $\{\texttt{HIGH}, \texttt{MEDIUM}, \texttt{LOW}\}$), which we use as a filtering mechanism when parsing the EL results. The EL prompt is generated according to the template in Figure 5, and adopts a few-shot learning technique, where the LLM learns to perform the EL task in-context by following a set of examples provided as part of the prompt.

Examples are ontology-specific and present the following structure: i) a list of concepts with the ID, name, and definition; ii) a text describing the entity mention to be grounded; iii) the expected answer; and iv) the associated confidence level. In the case of the HPO, Figure 6 shows the example that can be included in the EL prompt for one-shot Entity Linking. This negative example serves the purpose of instructing the LLM to be conservative and refrain from mapping any entity mention extracted with NER unless the matching concept belongs to the provided candidates.

```
[Concept A]
ID: HP:0034057
name: Fetal anomaly
definition: Structural or functional abnormalities of the fetus.

[Concept B]
ID: HP:0034058
name: Abnormal fetal morphology
definition: Any structural anomaly of the fetus.

[Concept C]
ID: HP:0034059
name: Abnormal fetal physiology
definition: Any functional anomaly of the fetus.

Text:
name: physical abnormalities
definition: Structural or functional abnormalities in
the body that can be observed or measured.


answer: None
confidence: HIGH
```

Figure 6: An HPO-specific example for the EL task.

Example 4 *Consider the mention and the set of retrieved candidates in Example 3. The template in Figure 5 is filled with: the name of the domain ontology (HPO); the HPO-specific example in Figure 6; the retrieved candidates and their properties; the description of the entity to be grounded. Invoking the LLM with EL prompt, the following result is returned:*

```
answer: HP:0008069
confidence: HIGH
```

4 Experiments

4.1 Benchmark Corpora

To validate our approach, we evaluate the performance of REAL for clinical phenotyping and phenotype annotation using two publicly available benchmark datasets: the HPO GSC+ (Lobo et al., 2017) and the dev component of the corpus published by BioCreative VIII Track 3 (Weissenbacher et al., 2023), referred to as BIOC-GS hereafter. HPO GSC+ consists of 228 manually annotated PubMed abstracts, with a total of 1933 annotations that cover 497 unique HPO IDs. The BIOC GS consists of 454 clinical observations manually annotated for phenotypes identified during dysmorphology physical examinations, that cover a total of 358 unique HPO IDs. As a reference resource for grounding, we use HPO, that provides a standardized vocabulary of phenotypic abnormalities associated with human hereditary and other diseases (Köhler et al., 2019). After preprocessing the ontology file, we indexed a total of 18.536 HPO concepts (See Section 3.1).

4.2 Experimental Setting

Currently, the REAL implementation relies on the OpenAI GPT models and feeds the prompts to the LLM by calling the OpenAI API. For evaluation, we use the `gpt-3.5-turbo-16k` model accessed through the GPT-3 completion endpoint, with default settings for temperature and `max tokens`. The number of LLM calls per document is estimated as follows: 2 requests sent to the OpenAI completion API endpoint[2] in the NER step, one for entity extraction and one for definition generation. Followed by h calls in the EL step, one call for each extracted mention. Additionally, for candidate retrieval, each entity mention requires a call to the OpenAI embedding API endpoint[3], which is handled automatically by the ChromaDB vector store. Due to constraints on the context window size (16,385 tokens for `gpt-3.5-turbo-16k` model), we limit the retrieved candidate set to a small number. In our experiments, we set $k = 3$, and included three candidate concepts in the EL prompt, as we observed no substantial improvements when using a larger number of candidates (see Section 4.4 for further discussion). Moreover, to ensure precise results in the EL phase, we opt to consider only mention/concept pairs associated with a `HIGH` confidence level, discarding less confident answers generated by the LLM.

To evaluate the effectiveness of the RAG paradigm in the context of the LLM-based biomedical concept recognition, we benchmark against a base case, where we directly instruct the GPT-3.5 model to extract and align HPO concepts from the input text using a single instructional prompt. The baseline prompt used in the experiments is adopted from Groza et al. and reported in Appendix 8.

In assessing the role of the LLM component in the entity linking step, our evaluation involves two distinct grounding strategies: one relies on the LLM to select the appropriate candidate concept for a given entity mention, while the other always selects the first matching concept retrieved by the embedding-based search. We refer to this latter strategy, which does not utilize the LLM in the linking phase, as *REAL-1st HIT* to differentiate it from the strategy using GPT3.5 for grounding, which we denote as *REAL-GPT3.5*. For a fair comparison with existing unsupervised methods for concept recognition, our evaluation in-

[2] `https://api.openai.com/v1/chat/completions`

[3] `https://api.openai.com/v1/embeddings`

	Document level			Mention level		
System	P	R	F1	P	R	F1
GPT-3.5	0.12	0.28	0.16	0.07	0.17	0.10
SPIRES	**0.84**	0.31	0.45	**0.84**	0.19	0.31
REAL-1st hit	0.40	**0.49**	0.44	0.33	**0.36**	0.39
REAL-GPT3.5	0.68	0.48	**0.56**	0.67	0.32	**0.43**

Table 1: Evaluation results on HPO GSC+

	Document level			Mention level		
System	P	R	F1	P	R	F1
GPT-3.5	0.26	0.33	0.29	0.22	0.29	0.25
SPIRES	**0.93**	0.31	0.47	**0.93**	0.19	0.47
REAL-1st hit	0.59	0.49	0.42	0.59	0.48	0.41
REAL-GPT3.5	0.69	**0.67**	**0.66**	0.68	**0.66**	**0.65**

Table 2: Evaluation results on BIOC GS

cludes SPIRES, a close prompt-based alternative to the REAL approach, accessible with local installation of the OntoGPT Python package [4]. For entity linking, SPIRES uses the OBO annotator (Taboada et al., 2014), a state-of-the-art dictionary-based method, designed for automatic annotation of biomedical literature with HPO terms. To execute HPO concept recognition with SPIRES, we utilize a predefined template for extracting human phenotypes, which is provided with the OntoGPT installation.[5] For phenotype extraction, SPIRES uses the `gpt-3.5-turbo-16k` model.

We evaluate the results by computing standard metrics for the concept recognition task: precision (P), recall (R) and F1-score (F1). The evaluation is performed at both document and mention levels. At the document level, we compute true positives as a set of target concepts that were found at least once in a given document and assigned a correct HPO identifier. At the mention-level, we account for all occurrences of a target concept within a document.

4.3 Results

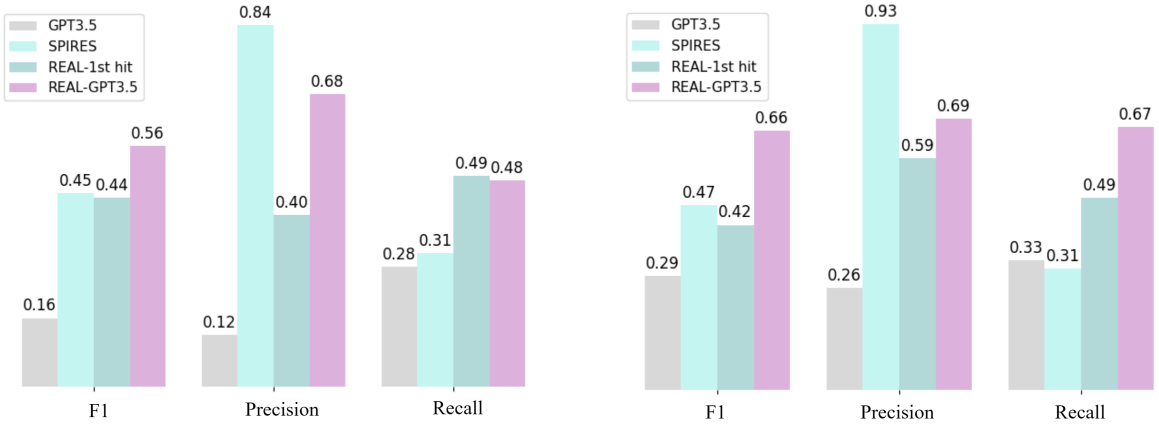
Tables 1 and 2 present the evaluation results for the HPO concept recognition on HPO GSC+ and BIOC GS datasets, respectively. To facilitate the performance comparison across systems, Figure 7 illustrates the precision, recall, and F1 score values considering the document level evaluation, which closely reflects the pattern observable at the mention level. Consistent results are also observed when conducting testing on the two datasets, as discussed in this section. Among other methods, *REAL-GPT3.5* can correctly recognize more HPO concepts, achieving the best F1 scores at both mention and document level. The dictionary matching method used with the OBO annotator, allows SPIRES to achieve the highest precision, which does not compensate for poor recall rates. The results show that the retrieval mechanism integrated in REAL significantly improves the recall, compared to other methods. In fact, *REAL-1st hit*, that uses the 1st retrieved concept for entity linking, achieves similar F1 score as SPIRES while balancing the precision and recall rates. Comparing the two grounding strategies, we observe that *REAL-GPT3.5* improves the precision over *REAL-1st hit* at both mention and document level. Leveraging the LLM for entity linking produces more precise results as it enables reasoning over the best match through multiple-choice selection and effectively filters out spurious extractions, that is, entity mentions erroneously identified as phenotypes by NER. In summary, the results on the GSC+ and BIOC GS datasets demonstrate the effectiveness of the REAL approach for phenotype concept recognition. Our experiments here were limited to GPT-3.5, but it is likely that GPT-4 will yield even better results.

4.4 Error analysis and discussions

The formulation of the NER prompt represents one of the critical aspects for the success of the approach. Poor results on NER propagate down the pipeline affecting the usefulness of the entity linking step. We assess the completeness of the NER results through a manual analysis of the generated extractions. Our evaluation suggests that the NER prompt achieves pertinent extractions providing a comprehensive coverage of the phenotypic features in both corpora. Some extractions from the HPO GSC+ include concepts not covered in the HPO ontology, such as mentions of diseases (*Prader-Willi syndrome, Angelman syndrome*), or generic phenotype-related concepts (*human anomaly*, *genetic abnormalities*). Additionally, we observe that a number of HPO terms extracted by REAL lack annotation in HPO GSC+. For instance, the phenotype *"Uniparental disomy"* is recognized 17 times in the corpus, but it is not present in the gold standard annotations, despite the existence of the exact match in the HPO: *"Uniparental disomy"* (`HP:0032382`). Such extractions represent the main cause of false positives and frequently in-

[4] `https://github.com/monarch-initiative`

[5] `https://github.com/monarch-initiative/ontogpt/blob/main/src/ontogpt/templates/`



Document level - HPO GSC+

Document level - BIOC GS

Figure 7: Document level evaluation results on HPO GSC+ and BIOC GS.

volve HPO terms close to the root of the taxonomic tree, such as *phenotypic abnormality (*`HP:0000118`*)* and *mode of inheritance* (`HP:0000005`).

Upon examining the frequently missed HPO terms, we identified two main causes of false negatives. The first issue is specific to HPO GSC+ and stems from the overlapping concepts, where phrases contain multiple nested HPO terms. For instance, the phrase *"skin tumors"* is annotated with both *"Neoplasm of the skin"* (`HP:0008069`) and *"Neoplasm"* (`HP:0002664`). By design, REAL, extracts mentions of entities as a whole and annotates to the most specific HPO term, failing to produce identifiers for nested concepts. This explains a high number of omissions for generic terms such as, *"Nurofibroma"*, *"Schwannoma"*, and *"Meningoma"*, usually nested within more specific HPO concepts.

The second issue involves complex entity mentions, frequently found in clinical notes, that have a form of compound and prepositional phrases, such as, *"scarring between 2 and 3"* and *"2,3 syndactyly bilaterally in feet"*. These extractions may produce definitions where the meaning of the entity is altered with respect to the target HPO term, yielding a poor set of retrieved candidates. For instance, the extracted mention, *"scars on axillary lines bilaterally"* produces a definition (*"Permanent marks or blemishes that have formed on the skin in the areas of the armpits, appearing on both sides of the body."*), that shifts the entity's meaning away from the target concept , *"Scarring"* (`HP:0100699`), towards related HPO terms, such as *"Axillary freckling"*, and *"Axillary lymphadenopathy"*. Moreover, due to the high level of granularity of the extracted mentions in the BIOC GS dataset, entities are often grounded in HPO terms that are more specific than those provided in the annotations. For example, the mention *"skins on the right foot feet thickend"* is mapped to *"Hypertrophy of skin of soles"*(`HP:0007403`) (i.e., *"Thick skin of soles"*), instead of the target *"Thickend skin"* (`HP:0001072`). These and similar issues can arise as a consequence of annotation idiosyncrasies that vary across benchmarks and could be addressed via additional post-processing of the NER results.

By analyzing the retrieval results, we found that around 65% of target concepts in HPO GCS+ (78% in BIOC GS) were effectively retrieved among candidates using approximate k-nearest neighbor search with $k = 3$. Preliminary experiments with greater values of k show no significant improvement, suggesting that the effectiveness of the candidate retrieval step mostly depends on the ability of the LLM to produce entity descriptions that are semantically close to target HPO terms. Our approach relies on the LLM to produce factual definitions for extracted mentions. However, future research might explore alternative strategies to ensure the factuality of the generated definitions. (Remy and Demeester, 2023).

It is important to stress that the domain expertise requirements vary across different phases of the concept recognition pipeline. In the grounding step using RAG, the domain knowledge is provided from outside, significantly reducing the expertise required by the LLM for entity normalization. In contrast, the biomedical NER task relies on the domain knowledge encoded within the model's parameters, demanding greater familiarity with the target domain to accurately recognize and define entities. This makes the NER task more knowledge-intensive and crucial for the overall success of the approach.

5 Concluding remarks

In this work, we introduced a novel approach for ontology-based concept recognition, that leverages RAG to harness general-purpose LLMs for automatic annotation of biomedical texts with classes from domain ontologies. The approach does not require domain specific training, but relies on prompt-engineering for both NER and EL tasks, integrating a retrieval mechanism to dynamically source domain knowledge from biomedical ontologies. We discussed the effectiveness of our approach on clinical phenotyping and phenotype annotation with experiments conducted on HPO GSC+ and BIOC GS benchmark corpora. Ongoing efforts focus on refining the prompt design to enhance performance and consider the integration with other GenAI providers. Using GPT models through OpenAI's API hinders the reproducibility of the results, which represents the main limitation of the current implementation. We plan to address this issue using a local installation of open-source LLMs. Furthermore, future research activities include conducting a comprehensive cross-domain evaluation to assess the generalizability of the proposed solution to diverse application domains.

Acknowledgments

This research was supported by: the "National Center for Gene Therapy and Drugs based on RNA Technology", PNRR-NextGeneration EU program [G43C22001320007]; the MUSA - Multilayered Urban Sustainability Action - Project, funded by the PNRR-NextGeneration EU program ([G43C22001370007], Code ECS00000037).

References

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. *arXiv preprint arXiv:2205.12689*.

J Harry Caufield et al. 2024. Structured Prompt Interrogation and Recursive Extraction of Semantics (SPIRES): a method for populating knowledge bases using zero-shot learning. *Bioinformatics*, 40(3):btae104.

Tina A Eyre, Fabrice Ducluzeau, Tam P Sneddon, Sue Povey, Elspeth A Bruford, and Michael J Lush. 2006. The HUGO gene nomenclature database, 2006 updates. *Nucleic acids research*, 34(suppl_1):D319–D321.

Yuhao Feng, Lei Qi, and Weidong Tian. 2022. PhenoBERT: a combined deep learning method for automated recognition of human phenotype ontology. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 20(2):1269–1277.

Jason Alan Fries, Natasha Seelam, Gabriel Altay, Leon Weber, Myungsun Kang, Debajyoti Datta, Ruisi Su, Samuele Garda, Bo Wang, Simon Ott, et al. 2022. Dataset debt in biomedical language modeling. In *Challenges & Perspectives in Creating Large Language Models*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Tudor Groza, Harry Caufield, Dylan Gration, Gareth Baynam, Melissa A Haendel, Peter N Robinson, Christopher J Mungall, and Justin T Reese. 2024. An evaluation of GPT models for phenotype concept recognition. *BMC Medical Informatics and Decision Making*, 24(1):30.

Clement Jonquet, Nigam H Shah, Cherie H Youn, Mark A Musen, Chris Callendar, and Margaret-Anne Storey. 2009. NCBO annotator: semantic annotation of biomedical data. In *Int'l Semantic Web Conf., Poster and Demo Session (ISWC 2009)*, 171.

Sebastian Köhler, Leigh Carmody, Nicole Vasilevsky, Julius O B Jacobsen, Daniel Danis, Jean-Philippe Gourdine, Michael Gargano, Nomi L Harris, Nicolas Matentzoglu, Julie A McMurry, et al. 2019. Expansion of the human phenotype ontology (HPO) knowledge base and resources. *Nucleic acids research*, 47(D1):D1018–D1027.

Sebastian Köhler, Sandra C Doelken, Christopher J Mungall, Sebastian Bauer, Helen V Firth, Isabelle Bailleul-Forestier, Graeme CM Black, Danielle L Brown, Michael Brudno, Jennifer Campbell, et al. 2014. The human phenotype ontology project: linking molecular biology and disease through phenotype data. *Nucleic acids research*, 42(D1):D966–D974.

Thomas Labbé, Pierre Castel, Jean-Michel Sanner, and Majd Saleh. 2023. ChatGPT for phenotypes extraction: one model to rule them all? In *45th Annual Int'l Conf. of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1–4. IEEE.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Manuel Lobo, Andre Lamurias, Francisco M Couto, et al. 2017. Identifying human phenotype terms by combining machine learning and validation rules. *BioMed Research International*, 2017.

Ling Luo, Shankai Yan, Po-Ting Lai, Daniel Veltri, Andrew Oler, Sandhya Xirasagar, Rajarshi Ghosh, Morgan Similuk, Peter N Robinson, and Zhiyong Lu. 2021. PhenoTagger: a hybrid method for phenotype concept recognition using human phenotype ontology. *Bioinformatics*, 37(13):1884–1890.

Simon Meoni, Eric De la Clergerie, and Theo Ryffel. 2023. Large language models as instructors: A study on multilingual clinical entity extraction. In *Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 178–190.

Tim E Putman, Kevin Schaper, Nicolas Matentzoglu, Vincent P Rubinetti, Faisal S Alquaddoomi, Corey Cox, J Harry Caufield, et al. 2023. The Monarch Initiative in 2024: an analytic platform integrating phenotypes, genes and diseases across species. *Nucleic Acids Research*, 52(D1):D938–D949.

Justin T Reese, Daniel Danis, J Harry Caulfied, Elena Casiraghi, Giorgio Valentini, Christopher J Mungall, and Peter N Robinson. 2023. On the limitations of large language models in clinical diagnosis. *medRxiv*.

François Remy and Thomas Demeester. 2023. Automatic glossary of clinical terminology: a large-scale dictionary of biomedical definitions generated from ontological knowledge. *arXiv preprint arXiv:2306.00665*.

Maria Taboada, Hadriana Rodríguez, Diego Martínez, María Pardo, and María Jesús Sobrido. 2014. Automated semantic annotation of rare disease cases: a case study. *Database*, 2014:bau045.

Nicole Vasilevsky, Shahim Essaid, Nico Matentzoglu, Nomi L Harris, Melissa Haendel, Peter Robinson, and Christopher J Mungall. 2020. Mondo disease ontology: harmonizing disease concepts across the world. In *CEUR Workshop Proceedings, CEUR-WS*, volume 2807.

Davy Weissenbacher, Siddharth Rawal, Xinwei Zhao, Jessica RC Priestley, et al. 2023. PhenoID, a language model normalizer of physical examinations from genetics clinical notes. *medRxiv*.

A Appendix

Baseline prompt

~~~
You will be provided with a text in triple backticks.
The task is to perform automated concept recognition
using the Human Phenotype Ontology and extract all
Human Phenotype Ontology concepts found in the text.
Include the HPO ID of the concepts you find in the
result.

``` { input text } ```
~~~

Figure 8: Baseline prompt for HPO CR.