AutoPCR: Automated Phenotype Concept Recognition by Prompting

Anonymous ACL submission

Abstract

Phenotype concept recognition (CR) is a fundamental task in biomedical text mining, enabling applications such as clinical diagnostics and knowledge graph construction. However, existing methods often require ontology-specific training and struggle to generalize across diverse text types and evolving biomedical terminology. We present AutoPCR, a prompt-based phenotype CR method that does not require ontology-specific training. AutoPCR performs CR in three stages: entity extraction using a hybrid of rule-based and neural tagging strategies, candidate retrieval via SapBERT, and entity linking through prompting a large language model. Experiments on four benchmark datasets show that AutoPCR achieves the best average and most robust performance across both mention-level and document-level evaluations, surpassing prior state-of-the-art methods. Further ablation and transfer studies demonstrate its inductive capability and generalizability to new ontologies.

1 Introduction

Biomedical text mining plays a key role in unlocking clinical and scientific knowledge from unstructured data sources such as clinical notes and research articles. A fundamental step in this process is ontology-based concept recognition (CR), which aims to identify textual mentions of concepts defined in a given ontology from input text. An ontology is a formal, structured representation of domain-specific knowledge curated by experts, consisting of standardized concepts with associated names, definitions, synonyms, and hierarchical relationships. Phenotype CR, a specific instance of ontology-based CR using the Human Phenotype Ontology (HPO) (Robinson et al., 2008; Köhler et al., 2019), has become a central research focus. This is largely due to the availability of richly annotated datasets (Weissenbacher et al., 2023; Lobo et al., 2017; Anazi et al., 2017), yet the task remains challenging because of the specialized and rapidly evolving nature of biomedical terminology and ontologies. Phenotype CR plays a critical role in downstream biomedical applications. For example, genetic disease diagnostics require accurate identification of phenotype concepts in clinical notes (Labbé et al., 2023), and biomedical knowledge graph construction relies on robust CR from scientific literature to support integrative data analysis and knowledge discovery (Soman et al., 2024).

Traditionally, CR can be divided into two stages: (1) entity extraction, which generates spans from the input text for further linking to the ontology; (2) entity linking, which links entities to semantically similar concepts from the ontology. Different CR methods adopt varying designs for these two stages, resulting in distinct strengths and limitations.

Early CR methods are primarily dictionary-based, which rely on lookup tables and string-matching techniques to identify concept mentions. While these approaches offer high precision, they suffer from low recall due to limited vocabulary coverage and inability to handle linguistic variations (Jonquet et al., 2009; Taboada et al., 2014). In recent years, researchers have increasingly adopted machine learning models that either employ biomedical named entity recognition (bioNER) methods for accurate entity extraction or domain-specific language models to enhance language understanding in biology. However, as these models are trained against a fixed ontology, they must be re-trained to recognize new concepts, limiting their usability for frequently updated ontologies like HPO (updated monthly). In contrast, large language models (LLMs), such as GPT-4 (Achiam et al., 2023), have demonstrated strong zero-shot learning capabilities, offering new possibilities for phenotype CR. Recent studies have shown that LLMs can effectively extract clinical information without domain-specific fine-tuning (Agrawal et al., 2022; Meoni et al., 2023). Despite their potential,

LLMs face challenges related to factual consistency and reliability in knowledge-intensive tasks (Chen et al., 2023; Reese et al., 2024). To address these challenges, retrieval-augmented generation (Lewis et al., 2020, RAG) has emerged as an effective technique to improve LLM's CR performance by incorporating relevant information retrieved from the ontology through semantic similarity-based retrieval mechanisms. However, to improve recall, a large number of entities need to be generated, which conflicts with the limited throughput of OpenAI's API. This highlights *the need for efficient semantic similarity models to filter noisy entities and prioritize high-quality candidate concepts*.

In this study, we propose **AutoPCR**[1], an automated phenotype CR method based on prompting, which consists of three sequential steps. First, to ensure that extracted entities are biologically meaningful while maintaining high recall, AutoPCR employs two complementary strategies for entity extraction, namely a rule-based strategy for shorter, free-form text, and a neural tagging approach for longer, standardized content. In the second step, AutoPCR retrieves candidate concepts using a semantic similarity model. Although phenotype CR has been extensively studied, prior work has largely overlooked advances in distantly-supervised biomedical entity linking. As a popular choice, AutoPCR adopts SapBERT (Liu et al., 2021a), a fine-tuned PubMedBERT (Gu et al., 2021) on the UMLS ontology (Bodenreider, 2004, a superset of HPO), to identify semantically relevant concepts. In the final step, AutoPCR performs entity linking by prompting an LLM. Each entity and its associated candidate set are encoded into a structured prompt that includes the entity string along with detailed information of each candidate concept. This approach enables accurate entity disambiguation without ontology-specific fine-tuning. Being model-agnostic and easily adaptable, it is well-suited for deployment in low-resource or rapidly evolving biomedical domains.

Our key contributions are as follows.

- **Superior and robust performance.** AutoPCR achieves the best average and most robust ranking on both mention-level and document-level evaluations across three benchmark datasets, changing the fact that prompt-based methods have lagged behind dictionary-based or neural approaches.
- **Inductive capability.** AutoPCR maintains superior performance even without prior exposure to HPO, making it adaptable to the frequent updates of the HPO ontology.
- **Generalizability.** AutoPCR transfers well to a new ontology without reconfiguration and supports rapid deployment within minutes, offering greater potential for real-world application in other biomedical domains.

[1]Our code is available at https://anonymous.4open.science/r/AutoPCR-3520.

2 Related Work

2.1 Phenotype concept recognition

Phenotype CR has grown into multiple methodological paradigms, including dictionary-based methods, neural methods, and prompt-based methods using LLMs.

Dictionary-based methods identify entities by exhaustively matching input spans against ontology concepts using lookup tables or string similarity measures. Tools such as the NCBO annotator (Jonquet et al., 2009), the OBO annotator (Taboada et al., 2014), Doc2HPO (Liu et al., 2019), and ClinPhen (Deisseroth et al., 2019) exemplify this strategy. More recently, FastHPOCR (Groza et al., 2024b) utilizes groups of morphologically equivalent words generated by GPT-4 to address lexical variations and represents the state of the art.

Neural methods typically take advantage of deep learning architectures, such as convolutional neural networks (CNNs) and BERT (Devlin et al., 2019), with supervised training on ontology concepts to improve phenotype CR performance. There are two common entity extraction strategies adopted by neural methods, serving as the foundation for subsequent entity linking. One line of methods, e.g., PhenoTagger (Luo et al., 2021) and PhenoTagger++ (Qi et al., 2024), adopts a rule-based strategy that exhaustively generates entity spans after removing punctuation and function words. PhenoTagger predicts concepts from extracted entities with a trainable BioBERT (Lee et al., 2020) coupled with dictionary-based methods. PhenoTagger++ further integrates concept embeddings derived from the ontology structure using TransR (Lin et al., 2015) into the BioBERT model. The other line of methods adopts a neural tagging approach that leverages off-the-shelf bioNER tools, e.g., Stanza (Zhang et al., 2021), to identify biomedical textual segments and then exhaustively extract entities from these segments. NCR (Arbabi et al., 2019) matches CNN-encoded entities with ontol-

ogy concepts represented by hierarchically aggregated fastText (Bojanowski et al., 2017) embeddings. PhenoBERT (Feng et al., 2022) improves NCR by using multiple CNNs to retrieve top candidate concepts for each phenotype subcategory and then re-ranks them with a trainable BERT that scores each entity-concept pair. Recently, PhenoBCBERT and PhenoGPT (Yang et al., 2024) train Bio+ClinicalBERT (Alsentzer et al., 2019) and GPT-3 (Brown et al., 2020) on manually labeled data to improve rare disease recognition.

LLMs, such as GPT-4, have enabled phenotype CR without the need for ontology-specific training, allowing for rapid adaptation to evolving biomedical vocabularies. Prompt-based methods, such as Labbé et al. (2023) and Groza et al. (2024a), utilize prompt engineering to guide LLMs in directly extracting biomedical concepts from text. On the other hand, the RAG-based method REAL (Shlyk et al., 2024) relies on LLMs to extract biomedical entities, generate their definitions, and perform entity linking. Candidate concepts are first retrieved using a generalist embedding model based on names and definitions, and then passed to a LLM for final linking using concept-level information.

2.2 Distantly-supervised biomedical entity linking

Distant supervision has emerged as a practical solution for biomedical entity linking, where manual annotation is costly and often infeasible. These methods leverage ontology-derived supervision, enabling large-scale training without labeled corpora.

BioSyn (Sung et al., 2020) aligns entities with concept aliases using a combination of character-level features and dense embeddings trained from UMLS alias tables. SapBERT (Liu et al., 2021a) further improves generalizability by fine-tuning PubMedBERT through contrastive learning over concept name pairs to align synonyms. KrissBERT (Zhang et al., 2022) incorporates context into contrastive learning by using PubMed-derived concept contexts and employs a cross-attention encoder to re-rank candidate concepts. Its inference strategy retains multiple context embeddings per concept, enabling context-aware matching. Sasse et al.'s (2024) recent work shows that synthetic entities generated by LLMs like Llama-2 (Touvron et al., 2023) can further enhance normalization performance, especially under distribution shifts, highlighting the complementary role of generative methods in distant supervision frameworks.

3 Method

In this section, we present AutoPCR, an automated method for phenotype CR based on prompting. We first provide a formal definition of the CR task and describe how it is decomposed into three sequential sub-tasks. We then define each sub-task rigorously and detail how AutoPCR addresses them through an integrated and modular design. The architecture of AutoPCR is shown in Figure 1.

3.1 Problem formulation

Definition 3.1 (Concept Recognition) *Given a domain ontology $O{=}(C, I)$ containing concepts $C{=}\{c_1, \ldots, c_n\}$ and related concept-level information I (e.g., definition and synonyms) and a piece of input text T, the task is to find f satisfying*

$$f(O,T) = \{(i,j,c) \mid i,j \in [0,|T|], \; T[i{:}j] \sim c \in C\} \quad (1)$$

which extracts entities from the input text with start and end offsets that can be mapped to semantically similar concepts in the ontology.

We decompose the CR task f into three sequential sub-tasks, namely entity extraction f_{EE}, candidate concept retrieval f_{CCR}, and entity linking f_{EL}, i.e., $f = f_{\text{EL}} \circ f_{\text{CCR}} \circ f_{\text{EE}}$, as detailed respectively in the next three sections.

3.2 Entity extraction

Definition 3.2 (Entity Extraction) *Given an ontology $O{=}(C, I)$ and a piece of input text T, the task is to find f_{EE} satisfying*

$$f_{EE}(O,T) \supset \{(i,j) \mid i,j \in [0,|T|], \; T[i{:}j] \sim \exists c \in C\} \quad (2)$$

which extracts entity spans from the text that may correspond to ontology concepts.

We adopt two complementary strategies for extracting entities, tailored to different types of input text. For shorter, free-form text such as clinical notes, we follow a rule-based strategy inspired by PhenoTagger. The input text is split into sentences, tokenized, POS-tagged using NLTK (Bird et al., 2009), and converted to lowercase. All n-gram spans ($n{\in}[2, 10]$) are then enumerated as candidate entities, excluding unigrams due to their limited variability and tendency to be misclassified as false positives. A part-of-speech filter is applied to eliminate spans that begin or end with punctuation or

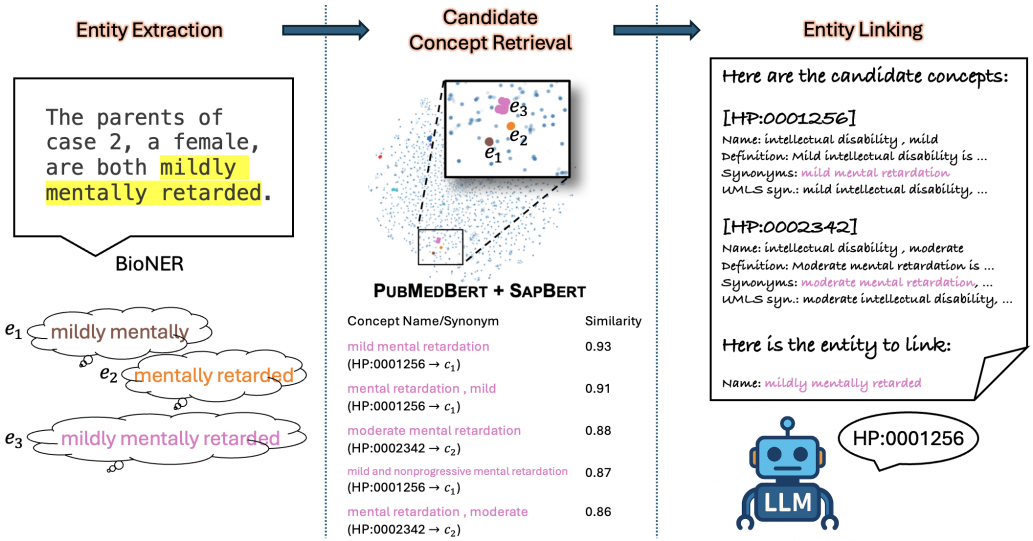

Figure 1: Architecture of AutoPCR. It performs concept recognition in three stages: entity extraction using a hybrid of rule-based and neural tagging strategies (e.g., extracted entities e_1, e_2, and e_3), candidate concept retrieval via SapBERT initialized from PubMedBERT (e.g., retrieved concepts c_1 and c_2 for e_3), and entity linking through prompting an LLM (e.g., linked concept c_1). The detailed prompt template is shown in Section 3.4.

function words, including prepositions, conjunctions, and determiners.

For longer, standardized content such as scientific abstracts, we adopt a neural tagging approach inspired by PhenoBERT. Each sentence is first processed using Stanza with the "ner-i2b2" processor and the "mimic" package to extract clinically relevant segments. To improve coverage, additional segments are generated by further splitting the sentences on punctuation and conjunctions. All resulting segments are then used as windows for n-gram extraction ($n \in [2, 10]$) to generate entities.

3.3 Candidate concept retrieval

Definition 3.3 (Candidate Concept Retrieval) *Given an ontology $O=(C, I)$, an entity $e=T[i{:}j]$ from text T, and a maximum number of candidates k, the task is to find f_{CCR} satisfying*

$$f_{CCR}(O, e, k) = C_{cand} \subset C \quad s.t. \quad |C_{cand}| \leq k, \quad e \sim \exists c \in C_{cand} \quad \vee \quad C_{cand} = \emptyset, \quad e \not\sim \forall c \in C \tag{3}$$

which retrieves a small set of potentially matching concepts from the ontology for each extracted entity.

To retrieve candidate concepts for each extracted entity e from the input text, we adopt an embedding-based retrieval strategy using SapBERT, which encodes semantically similar ontology terms into neighboring vectors. All ontology concepts and their synonyms are pre-encoded and indexed as dense vectors. Each extracted entity is also encoded with SapBERT and then compared against the index using cosine similarity. If the top similarity exceeds a high-confidence threshold τ_1, the entity is directly linked to the most similar concept. Otherwise, if the similarity falls within a relaxed interval $(\tau_2, \tau_1]$, we retrieve the top-k most similar concepts to form the candidate set C_{cand} for downstream entity linking. This two-stage thresholding allows us to balance precision and recall while avoiding unnecessary disambiguation when the match is unambiguous.

3.4 Entity linking

Definition 3.4 (Entity Linking) *Given an ontology $O=(C, I)$, an entity $e=T[i{:}j]$ from text T, and candidate concepts C_{cand}, the task is to find f_{EL} satisfying*

$$f_{EL}(O, e, C_{cand}) = c \in C_{cand} \sim e \tag{4}$$

which selects the most semantically similar concept from the retrieved candidate concepts.

Inspired by the prompt-based method REAL, we perform entity linking by prompting GPT-4o-mini. For each extracted entity e and its retrieved candidate concept set C_{cand}, we construct a structured prompt that includes the entity name along with the list of candidate concepts. Each concept is represented by its ID, name, definition, synonyms, and

cross-referenced UMLS synonyms. Specifically, the following prompt template is used.

> **System prompt:**
> As an expert clinician, your task is to accurately link the entity using the concepts listed below. Accuracy is paramount. If the entity does not precisely refer to any of the concepts listed below, please return "None"; otherwise, return the corresponding concept ID in the following format:
> answer:<concept ID or None>
> confidence:<one of HIGH, LOW, MEDIUM>
>
> Here are the concepts:
> {*candidate concepts with ID, name, definition, synonyms, UMLS synonyms*}
>
> **User prompt:**
> Here is the entity to link:
> label: {*entity string*}

The output contains both a concept ID and a confidence level, and only predictions with "HIGH" confidence are retained to ensure precision and reduce false positives.

3.5 Post-processing

AutoPCR may generate overlapping entity spans, as different n-grams or segments can cover intersecting text regions. These overlapping entities may link to the same or different concepts and thus require conflict resolution. If they are linked to different concepts, we retain all of them. If they are linked to the same concept, we apply resolution strategies corresponding to the used entity extraction strategy (Luo et al., 2021; Feng et al., 2022). When using the rule-based strategy for entity extraction, we keep the span with the highest matching score. When using the neural tagging approach, we retain the longest span, based on the intuition that longer spans typically convey more precise semantic meaning in standardized texts.

4 Experiments

We carry out extensive experiments to answer three research questions regarding our AutoPCR model. (**RQ-1**) How does AutoPCR perform against other baselines on various datasets? (**RQ-2**) How much does each module of AutoPCR contribute to its performance? (**RQ-3**) Can AutoPCR generalize effectively and efficiently to a different ontology other than HPO? Section 4.1 introduces the CR benchmark datasets. Section 4.2 describes implementation details. Section 4.3 presents the baseline systems. Sections 4.4, 4.5, and 4.6 correspond to our answers to the three research questions.

4.1 Concept recognition benchmark

To evaluate the performance of AutoPCR, we conduct experiments on *four* widely-used benchmark datasets: (i) **BIOC-GS**, the development set from BioCreative VIII Track 3 (Weissenbacher et al., 2023), consisting of 382 clinical observation records from dysmorphology physical examinations, with an average length of 8.5 words. It includes 607 phenotype mentions covering 315 unique HPO concepts. An example record is "ABDOMEN: Small umbilical hernia. Mild distention. Soft." (ii) **GSC-2024**, a refined GSC+ (Lobo et al., 2017) dataset given by FastHPOCR, which comprises 228 scientific abstracts from PubMed. We follow PhenoTagger's data split, using 22 abstracts for development. The test split contains 2,034 phenotype mentions linked to 451 unique HPO concepts, with an average of 150 words. (iii) **ID-68**, which contains 68 real clinical notes from families with intellectual disabilities (Anazi et al., 2017), and is manually annotated by PhenoBERT in the same way as GSC+. It includes 857 phenotype mentions covering 433 unique HPO concepts, with an average length of 157 words. (iv) The **NCBI** disease corpus (Doğan et al., 2014), consisting of 100 PubMed abstracts, with an average length of 205 words. It includes 960 phenotype mentions covering 198 unique MEDIC concepts. These datasets involve both free-form and standardized phenotype mentions, covering a wide range of use cases.

We follow the evaluation pipeline of PhenoTagger, a popular baseline. BIOC-GS, GSC-2024, and ID-68 are grounded on HPO release "20240208" under the root node "phenotypic abnormality." NCBI is grounded on MEDIC, a curated disease ontology integrating MeSH (Lipscomb, 2000) and OMIM (Amberger et al., 2019), under the root node "diseases." Mention-level and document-level precision (P), recall (R), and F1 scores are employed as evaluation metrics. For mention-level evaluation, a predicted entity is considered correct only if there exists a ground-truth entity with the same linked concept and overlapping offsets. For document-level evaluation, linked concepts are treated as a set per document, and metrics are computed by comparing these sets, followed by micro-averaging across all documents.

4.2 Implementation details

We adopt different entity extraction strategies for each dataset, as described in Section 3.2. For BIOC-GS, we use the rule-based strategy from PhenoTagger, which suits shorter, free-form clinical text. For GSC-2024, ID-68, and NCBI, we apply the neural tagging approach from PhenoBERT, as these datasets consist of longer, standardized content.

For candidate concept retrieval, we set τ_1=0.95 $\in [0.9, 0.92, 0.95, 0.98]$ and τ_2=0.85 $\in [0.8, 0.82, 0.85, 0.88, 0.9]$ to distinguish between high- and low-confidence matches. For similarity scores in the range $(\tau_2, \tau_1]$, we retain up to k=5 $\in [3, 5, 10]$ candidates. All hyperparameters are tuned on the GSC-2024 development set.

All experiments are run once with fixed seeds and zero temperature on a Linux server with an Intel Xeon Gold 6242R CPU (16 cores, 3.1GHz), 128 GB RAM, and one NVIDIA Tesla V100 (32 GB). The total cost of running AutoPCR is $< \$1$.

4.3 Baseline systems

We compare AutoPCR against various phenotype CR baselines across three categories. Dictionary-based methods include NCBO (Jonquet et al., 2009), OBO (Taboada et al., 2014), ClinPhen (Deisseroth et al., 2019), and FastHPOCR (Groza et al., 2024b). Neural methods include NCR (Arbabi et al., 2019), PhenoTagger (Luo et al., 2021), PhenoBERT (Feng et al., 2022), and PhenoTagger++ (Qi et al., 2024). Prompt-based methods, based on GPT-4o-mini, include direct prompting with Groza et al.'s (2024a) template 4 and REAL (Shlyk et al., 2024). All baselines are evaluated with their default parameter settings. Mention-level results are omitted for ClinPhen, since it does not provide offsets for extracted entities.

4.4 Main results

We report *mention-level* and *document-level* results for all methods across three benchmark datasets grounded on HPO in Tables 1 and 2. AutoPCR consistently achieves top-tier performance, ranking second in five out of six settings and third in the remaining one, resulting in the best average mention-level rank of 2.33 and document-level rank of 2.00. Its performance is also the most stable, with the smallest rank variance among all methods. The second-best method is FastHPOCR, a dictionary-based approach that ranks third on average for both evaluation levels. However, its performance drops notably on BIOC-GS, indicating reduced stability across datasets. The third-best performers are PhenoTagger and PhenoBERT, two neural models each with an average rank of 4.00. PhenoTagger performs better on BIOC-GS but worse on ID-68, whereas PhenoBERT shows the opposite trend, with stronger results on ID-68 and weaker performance on BIOC-GS. Both models exhibit higher variability compared to AutoPCR. Traditional dictionary-based methods such as NCBO, OBO, and ClinPhen consistently achieve high precision, but suffer from low recall, leading to lower F1 score rankings. The early neural model NCR is outperformed by its successor PhenoBERT as expected, while PhenoTagger++, intended as an improved version of PhenoTagger, does not yield any notable performance gain. The naive prompting method using GPT-4o-mini performs poorly compared to all other baselines. In contrast, the retrieval-augmented method REAL achieves the best result on BIOC-GS, but ranks near the bottom on the other datasets, highlighting the limitations of its model design. These results comprehensively address **RQ-1**, demonstrating the performance superiority and robustness of AutoPCR. We further analyze these outcomes in detail, showing how the design of AutoPCR contributes to its performance under varying dataset characteristics.

On BIOC-GS, which consists of noisy clinical observations, we observe a clear performance trend across method categories: prompt-based methods outperform neural methods, which in turn outperform dictionary-based ones, consistent with prior findings (Groza et al., 2024a; Qi et al., 2024; Shlyk et al., 2024). This pattern reflects their respective modeling capabilities. Prompt-based methods rely on LLMs with strong language understanding abilities for entity linking, enabling them to better handle free-form and lexically diverse clinical text. Neural methods, typically fine-tuned on structured, ontology-aligned corpora, struggle to generalize to such noisy input. Dictionary-based methods perform worst, as they lack semantic understanding and rely solely on exact or approximate string matching. Within the prompt-based category, for the same reason, REAL performs better than AutoPCR due to its use of LLM-generated definition-similarity-based concept retrieval, which relies less on the surface forms of the entities.

On GSC-2024 and ID-68, which feature longer and more standardized text with less surface-level variation in entities, AutoPCR remains highly com-

Method	Performace (%)	BIOC-GS				GSC-2024				ID-68				Avg. Rk.
		P	R	F1 (Rk.)		P	R	F1 (Rk.)		P	R	F1 (Rk.)		
Dictionary -based	NCBO	71.80	45.80	55.92	9	**96.62**	51.57	67.25	7	86.53	65.23	74.39	7	7.67
	OBO	73.89	45.31	56.17	8	87.41	52.66	65.72	8	82.55	63.01	71.47	8	8.00
	ClinPhen	–	–	–	11	–	–	–	11	–	–	–	11	11.00
	FastHPOCR	60.53	60.46	60.50	5	91.66	**79.25**	**85.01**	**1**	87.23	71.76	78.75	3	3.00
Neural	NCR	55.90	60.96	58.32	7	81.14	74.88	77.88	6	78.64	78.18	78.41	4	5.67
	PhenoTagger	56.56	**70.35**	62.70	3	86.16	78.12	81.95	4	83.49	73.75	78.32	5	4.00
	PhenoBERT	64.50	55.85	59.86	6	88.04	74.98	80.99	5	**94.33**	**78.76**	**85.85**	**1**	4.00
	PhenoTagger++	56.87	69.52	62.57	4	87.71	78.27	82.72	2	79.52	73.86	76.59	6	4.00
Prompt -based	GPT-4o-mini	2.89	2.47	2.66	10	15.31	7.62	10.18	10	18.00	15.52	16.67	10	10.00
	REAL	**75.20**	62.93	**68.52**	**1**	76.27	47.59	58.61	9	76.33	65.46	70.48	9	6.33
	AutoPCR (ours)	62.54	67.38	64.87	2	91.17	75.42	82.55	3	85.54	73.16	78.87	2	**2.33**

Table 1: Mention-level performance of all methods on three benchmark datasets. AutoPCR achieves the best average and most robust ranking across datasets and methods.

Method	Performace (%)	BIOC-GS				GSC-2024				ID-68				Avg. Rk.
		P	R	F1 (Rk.)		P	R	F1 (Rk.)		P	R	F1 (Rk.)		
Dictionary -based	NCBO	72.90	45.86	56.30	10	**99.33**	52.26	68.49	7	89.06	64.69	74.95	7	8.00
	OBO	74.46	45.36	56.38	9	85.98	52.40	65.12	8	83.36	62.55	71.47	8	8.33
	ClinPhen	64.78	51.16	57.17	8	86.20	44.99	59.12	10	74.96	61.92	67.82	10	9.33
	FastHPOCR	60.56	60.76	60.66	5	95.15	**77.61**	**85.49**	**1**	87.75	71.38	78.72	3	3.00
Neural	NCR	55.16	61.09	57.97	7	81.88	73.09	77.24	6	79.18	77.68	78.42	4	5.67
	PhenoTagger	57.78	**70.70**	63.59	3	87.57	76.62	81.73	4	83.29	73.52	78.10	5	4.00
	PhenoBERT	65.07	56.13	60.27	6	90.33	73.87	81.27	5	**94.11**	**78.56**	**85.64**	**1**	4.00
	PhenoTagger++	58.21	69.87	63.51	4	89.38	76.70	82.55	3	79.40	73.90	76.55	6	4.33
Prompt -based	GPT-4o-mini	50.77	44.15	47.23	11	18.25	9.75	12.71	11	19.53	16.77	18.05	11	11.00
	REAL	**75.94**	63.25	**69.02**	**1**	80.21	54.66	65.02	9	76.64	66.20	71.04	9	6.33
	AutoPCR (ours)	64.11	67.72	65.86	2	93.91	74.01	82.78	2	85.65	73.01	78.83	2	**2.00**

Table 2: Document-level performance of all methods on three benchmark datasets. AutoPCR achieves the best average and most robust ranking across datasets and methods.

petitive, even though the advantage of language understanding becomes less pronounced. The top-performing baselines on these two datasets, FastHPOCR and PhenoBERT, are both evaluated on data partially labeled by their respective research teams, potentially introducing bias in their favor. Also, neither method achieves top performance on the other's dataset. AutoPCR, by contrast, consistently ranks just behind these baselines and achieves the (second) best average rank across mention-level and document-level evaluations. Unlike in BIOC-GS, AutoPCR significantly outperforms REAL in this setting. This advantage stems from its more comprehensive entity extraction strategy, which improves recall, and a candidate concept retrieval module that efficiently models semantic similarity by filtering noisy entities and prioritizing high-quality candidates based on surface forms.

4.5 Ablation studies

To assess the contribution of each module in AutoPCR and answer **RQ-2**, we conduct ablation experiments, as summarized in Table 3. We evaluate the following three variants of AutoPCR.

- **Variant 1.** This variant simulates a scenario where the target ontology is newly introduced and unseen during model training. We re-train SapBERT (initialized from PubMedBERT) on UMLS excluding all HPO-related concepts.
- **Variant 2.** This variant examines whether domain-specific alignment in SapBERT is essential for candidate concept retrieval. We replace SapBERT with PubMedBERT for embedding, without any contrastive fine-tuning on UMLS.
- **Variant 3.** This variant evaluates the importance of the entity linking module. We remove the linking module entirely and rely solely on candidate concept retrieval. If the top candidate exceeds the similarity threshold τ_1, it is directly assigned as the final prediction, bypassing the GPT-based disambiguation step.

The full AutoPCR model overall outperforms all ablated variants across datasets and metrics, confirming the importance of each module. Variant 2 exhibits the most severe performance drop, especially on GSC-2024 and ID-68, highlighting the importance of domain-specific alignment in

	Prior Knowledge	Candidate Retrieval	Entity Linking	Mention-level F1%			Document-level F1%		
				BIOC-GS	GSC-2024	ID-68	BIOC-GS	GSC-2024	ID-68
Variant 1		√	√	64.77	81.79	78.28	65.76	81.95	78.22
Variant 2	√		√	58.03	67.29	71.29	58.99	67.31	71.34
Variant 3	√	√		61.89	81.07	**79.34**	62.91	80.83	**79.14**
AutoPCR (ours)	√	√	√	**64.87**	**82.55**	78.87	**65.86**	**82.78**	78.83

Table 3: Ablation studies on the impact of prior ontology knowledge, SapBERT-based candidate concept retrieval, and entity linking. Removing any single module leads to overall performance degradation.

	NCBI		
Method	M-level F1 (%)	D-level F1 (%)	Deployment Time
FastHPOCR	43.08	52.25	24m10s
NCR	36.29	29.95	∼28h
PhenoTagger	47.75	64.03	15h8m2s
AutoPCR (ours)	**51.47**	**66.08**	**2m38s**

Table 4: Performance and deployment time of selected phenotype concept recognition methods on the NCBI dataset using the MEDIC ontology.

embedding-based retrieval. Variant 3 performs reasonably well on standardized abstracts such as GSC-2024 and ID-68, suggesting that surface-form matching alone may suffice for low-ambiguity datasets. However, it underperforms notably on BIOC-GS, which contains noisier clinical narratives with higher ambiguity, underscoring the value of LLM-based reasoning in resolving fine-grained semantic distinctions. Notably, Variant 1 maintains relatively strong performance despite the complete removal of HPO-specific knowledge. This finding demonstrates that AutoPCR can perform inductive inference over entirely new or rapidly evolving ontologies, such as HPO.

4.6 Generalizability

To evaluate the generalizability and deployment efficiency of AutoPCR, we conduct an additional experiment on the NCBI dataset using the MEDIC ontology. Table 4 reports both mention-level and document-level results, along with the time required to deploy each method. We define *deployment time* as the total time needed to prepare a method for inference on a new ontology, including ontology-specific index construction or re-training. We compare AutoPCR with three strong and representative baselines: FastHPOCR, NCR, and PhenoTagger. PhenoTagger++ and PhenoBERT are excluded due to the lack of publicly available training pipelines for adapting to new ontologies.

AutoPCR achieves the highest F1 scores on both mention-level and document-level, while also being the most efficient to deploy, requiring only 2 minutes and 38 seconds to build the ontology index of embedded concepts for retrieval. In contrast, neural methods such as PhenoTagger and NCR require several hours of retraining and still fall short in performance. FastHPOCR deploys more quickly than neural methods but performs substantially worse than both PhenoTagger and AutoPCR, indicating limited generalizability. These results answer **RQ-3** and demonstrate that AutoPCR can generalize effectively and efficiently to a new ontology without reconfiguration. Its ability to support rapid deployment makes it well suited for real-world, off-the-shelf use in other biomedical domains.

5 Conclusion and Future Work

In this work, we introduce AutoPCR, an automated method for phenotype concept recognition. It extracts entities using a hybrid of rule-based and neural tagging strategies, retrieves candidate concepts via biomedical pre-trained embeddings, and links entities to ontology concepts through large language models. Our experiments demonstrate that AutoPCR achieves not only superior accuracy and robustness in phenotype recognition, but also remarkable inductiveness and generalizability to unseen ontologies–hence the name "automated."

Future work will explore several directions to further enhance the adaptability and scalability of AutoPCR. First, we plan to support multilingual ontologies and cross-lingual concept recognition by incorporating cross-lingual variants of SapBERT (Liu et al., 2021b), enabling broader application to non-English biomedical corpora. Second, we aim to leverage richer contextual signals to improve entity linking in complex scenarios. Third, we plan to replace the API-based LLM with a locally deployed model fine-tuned to reject unlinkable entities, eliminating the need for manual threshold tuning and offering a more efficient and cost-effective alternative. Finally, we are interested in integrating AutoPCR into downstream pipelines such as biomedical knowledge graph construction and phenotype-driven disease diagnosis, in order to assess its utility in real-world biomedical applications.

Limitations

The current implementation of AutoPCR relies on the OpenAI API for entity linking, which introduces latency due to external service calls and incurs usage costs. While this setup enables strong zero-shot generalization, it may hinder scalability in large-scale or time-sensitive applications. Additionally, AutoPCR assumes access to well-structured biomedical ontologies with informative definitions and synonyms; performance may be affected when such resources are sparse or inconsistently curated. Finally, although AutoPCR is ontology-agnostic by design, the candidate concept retrieval step depends on embedding-based similarity, which may be less effective for highly abstract or semantically overloaded concepts.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

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*.

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. *arXiv preprint arXiv:1904.03323*.

Joanna S Amberger, Carol A Bocchini, Alan F Scott, and Ada Hamosh. 2019. OMIM.org: leveraging knowledge across phenotype–gene relationships. *Nucleic Acids Research*, 47(D1):D1038–D1043.

Shams Anazi, Sateesh Maddirevula, Vincenzo Salpietro, Yasmine T Asi, Saud Alsahli, Amal Alhashem, Hanan E Shamseldin, Fatema AlZahrani, Nisha Patel, Niema Ibrahim, and 1 others. 2017. Expanding the genetic heterogeneity of intellectual disability. *Human Genetics*, 136:1419–1429.

Aryan Arbabi, David R Adams, Sanja Fidler, Michael Brudno, and 1 others. 2019. Identifying clinical terms in medical text using ontology-guided machine learning. *JMIR Medical Informatics*, 7(2):e12596.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl_1):D267–D270.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Shiqi Chen, Siyang Gao, and Junxian He. 2023. Evaluating factual consistency of summaries with large language models. *arXiv preprint arXiv:2305.14069*.

Cole A Deisseroth, Johannes Birgmeier, Ethan E Bodle, Jennefer N Kohler, Dena R Matalon, Yelena Nazarenko, Casie A Genetti, Catherine A Brownstein, Klaus Schmitz-Abe, Kelly Schoch, and 1 others. 2019. ClinPhen extracts and prioritizes patient phenotypes directly from medical records to expedite genetic disease diagnosis. *Genetics in Medicine*, 21(7):1585–1593.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. NCBI disease corpus: a resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, 47:1–10.

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.

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

Tudor Groza, Dylan Gration, Gareth Baynam, and Peter N Robinson. 2024b. FastHPOCR: pragmatic, fast, and accurate concept recognition using the human phenotype ontology. *Bioinformatics*, 40(7):btae406.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

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 *ISWC 2009-8th International Semantic Web Conference, Poster and Demo Session*, 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, and 1 others. 2019. Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Research*, 47(D1):D1018–D1027.

Thomas Labbé, Pierre Castel, Jean-Michel Sanner, and Majd Saleh. 2023. ChatGPT for phenotypes extraction: one model to rule them all? In *2023 45th Annual International Conference 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, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

Carolyn E Lipscomb. 2000. Medical subject headings (MeSH). *Bulletin of the Medical Library Association*, 88(3):265.

Cong Liu, Fabricio Sampaio Peres Kury, Ziran Li, Casey Ta, Kai Wang, and Chunhua Weng. 2019. Doc2Hpo: a web application for efficient and accurate HPO concept curation. *Nucleic Acids Research*, 47(W1):W566–W570.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021a. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4228–4238, Online. Association for Computational Linguistics.

Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2021b. Learning domain-specialised representations for cross-lingual biomedical entity linking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 565–574, Online. Association for Computational Linguistics.

Manuel Lobo, Andre Lamurias, and Francisco M Couto. 2017. Identifying human phenotype terms by combining machine learning and validation rules. *BioMed Research International*, 2017(1):8565739.

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, Theo Ryffel, and Eric Villemonte de La Clergerie. 2023. Large language models as instructors: a study on multilingual clinical entity extraction. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 178–190. Association for Computational Linguistics.

Jiewei Qi, Ling Luo, Zhihao Yang, Jian Wang, Huiwei Zhou, and Hongfei Lin. 2024. An improved method for phenotype concept recognition using rich HPO information. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1135–1140. IEEE.

Justin T Reese, Daniel Danis, J Harry Caufield, Tudor Groza, Elena Casiraghi, Giorgio Valentini, Christopher J Mungall, and Peter N Robinson. 2024. On the limitations of large language models in clinical diagnosis. *medRxiv*, pages 2023–07.

Peter N Robinson, Sebastian Köhler, Sebastian Bauer, Dominik Seelow, Denise Horn, and Stefan Mundlos. 2008. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *The American Journal of Human Genetics*, 83(5):610–615.

Kuleen Sasse, Shinjitha Vadlakonda, Richard E Kennedy, and John D Osborne. 2024. Disease entity recognition and normalization is improved with large language model derived synthetic normalized mentions. *arXiv preprint arXiv:2410.07951*.

Darya Shlyk, Tudor Groza, Stefano Montanelli, Emanuele Cavalleri, Marco Mesiti, and 1 others. 2024. REAL: a retrieval-augmented entity linking approach for biomedical concept recognition. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 380–389. Association for Computational Linguistics.

Karthik Soman, Peter W Rose, John H Morris, Rabia E Akbas, Brett Smith, Braian Peetoom, Catalina Villouta-Reyes, Gabriel Cerono, Yongmei Shi, Angela Rizk-Jackson, and 1 others. 2024. Biomedical knowledge graph-optimized prompt generation for large language models. *Bioinformatics*, 40(9):btae560.

Mujeen Sung, Hwisang Jeon, Jinhyuk Lee, and Jaewoo Kang. 2020. Biomedical entity representations with synonym marginalization. In *Proceedings of the 58th Annual Meeting of the Association for Computational*

Linguistics, pages 3641–3650, Online. Association for Computational Linguistics.

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.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Davy Weissenbacher, Siddharth Rawal, Xinwei Zhao, Jessica RC Priestley, Katherine M Szigety, Sarah F Schmidt, Mary J Higgins, Arjun Magge, Karen O'Connor, Graciela Gonzalez-Hernandez, and 1 others. 2023. PhenoID, a language model normalizer of physical examinations from genetics clinical notes. *medRxiv*, pages 2023–10.

Jingye Yang, Cong Liu, Wendy Deng, Da Wu, Chunhua Weng, Yunyun Zhou, and Kai Wang. 2024. Enhancing phenotype recognition in clinical notes using large language models: PhenoBCBERT and PhenoGPT. *Patterns*, 5(1).

Sheng Zhang, Hao Cheng, Shikhar Vashishth, Cliff Wong, Jinfeng Xiao, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. Knowledge-rich self-supervision for biomedical entity linking. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 868–880, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yuhao Zhang, Yuhui Zhang, Peng Qi, Christopher D Manning, and Curtis P Langlotz. 2021. Biomedical and clinical English model packages for the Stanza Python NLP library. *Journal of the American Medical Informatics Association*, 28(9):1892–1899.