MA-COIR: Leveraging Semantic Search Index and Generative Models for Ontology-Driven Biomedical Concept Recognition

Shanshan Liu[1,2], Noriki Nishida[1], Rumana Ferdous Munne[1], Narumi Tokunaga[1], Yuki Yamagata[3,4], Kouji Kozaki[5], Yuji Matsumoto[1],
[1]RIKEN AIP [2]University of Tsukuba [3]RIKEN R-IH [4]RIKEN BRC
[5]Osaka Electro-Communication University
{shanshan.liu, noriki.nishida, rumanaferdous.munne, narumi.tokunaga, yuki.yamagata, yuji.matsumoto}@riken.jp
kozaki@osakac.ac.jp

Abstract

Recognizing biomedical concepts in the text is vital for ontology refinement, knowledge graph construction, and concept relationship discovery. However, traditional concept recognition methods, relying on explicit mention identification, often fail to capture complex concepts not explicitly stated in the text. To overcome this limitation, we introduce MA-COIR, a framework that reformulates concept recognition as an indexing-recognition task. By assigning semantic search indexes (ssIDs) to concepts, MA-COIR resolves ambiguities in ontology entries and enhances recognition efficiency. Using a pretrained BART-based model fine-tuned on small datasets, our approach reduces computational requirements to facilitate adoption by domain experts. Furthermore, we incorporate large language models (LLMs)-generated queries and synthetic data to improve recognition in low-resource settings. Experimental results on three scenarios (CDR, HPO, and HOIP) highlight the effectiveness of MA-COIR in recognizing both explicit and implicit concepts without the need for mention-level annotations during inference, advancing ontology-driven concept recognition in biomedical domain applications. Our code and constructed data are available at https://github.com/sl-633/macoir-master.

1 Introduction

Automatic recognition of biological concepts in the text aids experts in refining ontologies and consolidating domain knowledge. As structured knowledge evolves to include increasingly complex concepts (Gargano, 2023; Yamagata et al., 2024), identifying concepts often requires significant expert analysis. Traditional Concept Recognition (CR) methods are inadequate for supporting tasks such as ontology-driven knowledge graph construction, efficient literature retrieval for specific concepts, and the discovery of novel relationships between concepts.

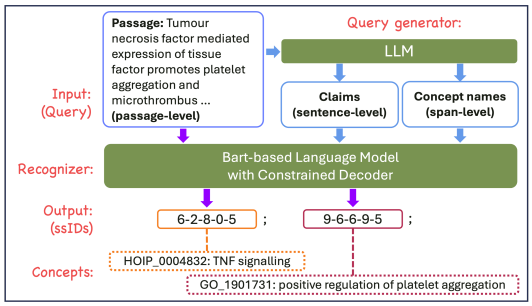

Figure 1: Concept recognition by MA-COIR follows the default workflow indicated by purple arrows. When an LLM generates simplified queries from a given passage, additional processes, denoted by blue arrows, are incorporated. When "6-2-8-0-5" is generated, "HOIP_0004832: TNF signalling" is predicted as a concept within the query.

Typically, recognizing ontology concepts in passages or sentences relies on identifying mentions - text spans where concepts appear. When mentions are provided, Entity Disambiguation (ED) can be applied to match each mention to a single entity or none at all (Wu et al., 2020; Jiang et al., 2024; Wang et al., 2023; OAKlib, 2023). When mentions are unknown, recognition may be achieved through a pipeline beginning with Named Entity Recognition (NER) to identify mentions, followed by ED to resolve these predictions (Shlyk et al., 2024; Caufield et al., 2024). Alternatively, end-to-end Entity Linking (EL) approaches can yield a series of (mention, entity) pairs (Kolitsas et al., 2018; Cao et al., 2020; Luo et al., 2021).

With advancements in Large Language Models (LLMs), several LLM-based pipeline methods for NER and ED have been introduced (Shlyk et al., 2024; Caufield et al., 2024). In-context learning (ICL) techniques reduce annotation requirements; however, a substantial performance gap remains between ICL and fully supervised methods (Shlyk

et al., 2024). While mention-based queries are typically generated to retrieve concepts, the limitation of this approach becomes evident when complex concepts do not appear explicitly as "mentions" within the text, rendering aforementioned mention-based recognition methods ineffective in real-world applications.

We propose **MA-COIR** (**M**ention-**A**gnostic **Co**ncept Recognition through an **I**ndexing-**R**ecognition Framework), a framework for recognizing biomedical concepts explicitly or implicitly mentioned in the text. Inspired by prior works (Tay et al., 2022; Jiang et al., 2024), we reformulate the concept recognition (CR) task into an indexing-recognition paradigm. This approach assigns each concept a semantic search index (ssID) and trains a neural model to predict ssIDs corresponding to concepts described in the input text (see Fig. 1).

By generating ssIDs instead of literal concept names, the framework resolves ambiguities caused by identical concept names within ontologies (e.g., concepts sharing preferred names but differing definitions). Additionally, the semantic alignment between concepts and their assigned indexes enhances model learning, enabling more powerful recognition.

Our method leverages a pretrained BART-based language model fine-tuned on a small dataset, thereby reducing computational demands and improving accessibility for domain experts. Furthermore, we explore LLM-generated queries and synthetic data, demonstrating the framework's utility in low-resource settings for real-world concept extraction tasks. Results across datasets (CDR, HPO, and HOIP) demonstrate the effectiveness of our framework.

Our contributions are:

- We propose MA-COIR, a novel framework for recognizing both explicit and implicit biomedical concepts without the need for prior identification of specific mentions, thereby reducing reliance on labor-intensive annotations needed for entity recognizer training.
- To the best of our knowledge, we are the first to integrate a semantic search index into biomedical concept recognition, improving generative model learning and enabling more efficient recognition.
- We demonstrate the utility of query and training data generated by an LLM in concept recognition tasks, providing a reference framework for efficient recognition in low-resource settings.

2 Related work

Biomedical Concept Recognition. In recent years, biomedical CR methods have largely followed two main approaches. The first approach involves fully-, weakly-, or self-supervised learning methods based on pretrained language models, such as domain-specific BERT or BART models (Liu et al., 2021; Lee et al., 2019; Yuan et al., 2022; Zhang et al., 2022), and fine-tuned these models on small annotated datasets (Luo et al., 2021). The second approach leverages the strong generalization capabilities of LLMs to perform NER and ED tasks in zero- or few-shot settings (Wang et al., 2023).Existing biomedical CR methods that operate without mention annotations are LLM-based. For instance, (Caufield et al., 2024) explored a schema guiding LLMs to perform NER with specified constraints, using (OAKlib, 2023) for subsequent ED tasks. (Shlyk et al., 2024) proposed the REAL framework, which combines LLM-based zero-shot NER with an ED method using retrieval-augmented generation (RAG). (El Khettari et al., 2024) developed an ICL demonstration selection strategy to generate concept names closely aligned with ontology terms, subsequently linking them based on the similarity between generated names and ontology terms.

Hierarchical Indexing. Hierarchical indexing has proven effective in handling large output spaces, as seen in applications like extreme multi-label classification (Zhang et al., 2021; Kharbanda et al., 2022) and document retrieval (Tay et al., 2022). By organizing labels or documents into tree-structured hierarchies based on semantic relationships, these methods improve computational efficiency and prediction performance. Notably, in the context of biomedical CR, well-defined concept taxonomies already exist through ontologies, offering a natural foundation for hierarchical organization. However, the application of hierarchical indexing in this field remains relatively unexplored despite its potential benefits.

3 Methodology

3.1 Task formulation

Let O represent a set of concepts $\{C_1, ..., C_n\}$ defined within a domain ontology. Given a query text

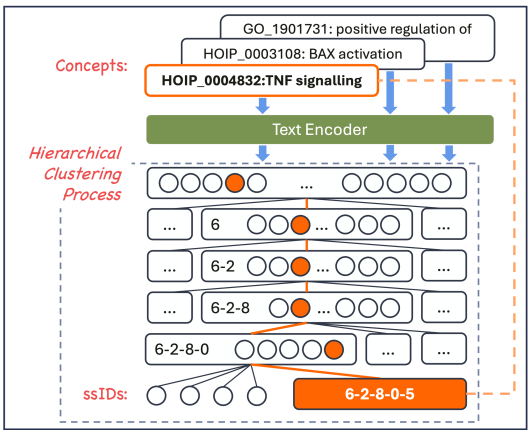

Figure 2: Indexing Phase in MA-COIR: A semantic search index (ssID) is assigned based on a label tree derived from the domain ontology. Through hierarchical clustering, the ssID for the concept "HOIP_0004832: TNF signaling" is "6-2-8-0-5".

Q, the CR task aims to identify a subset of concepts $\{C'_1, ..., C'_p\}$ from the ontology that are referenced in the text.

We approach the CR task as an end-to-end generative process. First, we assign each concept C a unique semantic search index (ssID). Then, our model generates one or more ssIDs for the input text Q, thereby retrieving the concepts are presented in the text.

3.2 Concept Indexing

As illustrated in Fig. 2, each concept C is represented as a vector E_C, obtained by encoding its canonical name $Name_C$ using a text encoder. Given our focus on the biomedical domain, we select SapBERT (Liu et al., 2021) as the text encoder.[1] The representation E_C is derived by averaging the last hidden states for the tokens in $Name_C$.

$$X_C = TextEncoder(Name_C) \in \mathbb{R}^{l \times H} \quad (1)$$

$$E_C = avg(X_C) \in \mathbb{R}^H \quad (2)$$

where l is the token length, and H is the dimension of each token's embedding.

Starting with the ROOT node that encompasses all concepts in the target ontology, we construct a label tree using a top-down **hierarchical clustering process**. Specifically, if a node contains more than g elements, we divide it into $\leq m$ categories until each leaf node corresponds to a single concept (with $g = 10, m = 10$ in this study)[2] by K-means algorithm implemented with Scikit-learn (Pedregosa et al., 2011). Each node is assigned an index based on its category, forming a sequence of "semantic search indexes" (ssIDs) that encode semantic information from each concept's representation.

3.3 Concept Recognition

During recognition phase following the indexing process, the input may consist of a passage (e.g., a paragraph of one PubMed article), a sentence, or a span (mention or concept name), while the output is a text sequence listing ssIDs (e.g., "6-2-8-0-5; 9-6-6-9-5;"). Each ssID is separated by a semicolon (";"), as illustrated in Fig. 1.

To effectively map natural language text to a formatted sequence, we selected a BART-based pretrained language model (facebook/bart-large) (Lewis et al., 2019). This model, with its encoder-decoder architecture and cross-attention mechanism, is well-suited for our tasks.

To ensure the BART-based model generates valid ssID sequences, we apply a constrained decoder that filters the output to retain only valid ssIDs. The decoder's vocabulary T is restricted to ssID tokens. The token embedding e_t for each token $t \subset T$ is obtained from the embedding layer $LmEmbedding$ of the language model LM:

$$e_t = LmEmbedding(t) \in \mathbb{R}^H \quad (3)$$

where H is the dimension of a token's embedding.

At the i-th time step, the decoder selects the token with the highest score based on the token embedding e_t and the last hidden state h_i. One feature $h_{i,t}$ is computed using a one-layer linear classifier:

$$h_i = LM(\hat{y}_{i-1}) \in \mathbb{R}^H \quad (4)$$

$$h_{i,t} = W_t^o h_i + b^o \quad (5)$$

where W^o is the weight and b^o is the bias of the classifier.

Another feature $e_{i,t}$ is the dot product of e_t and h_i, representing the relevance between the token t and h_i:

$$e_{i,t} = e_t h_i \quad (6)$$

[1]Through preliminary experiments, we observed that using the average of token embeddings yields better performance than the [CLS] token. We evaluated several pretrained language models, including BioBERT v1.1, PubMedBERT, SapBERT, and SciBERT, with SapBERT achieving the best results.

[2]We initially adopted DSI's setting (g=10, m=100) (Tay et al., 2022) but observed better performance with a smaller m. The choice of g=10 aligns naturally with our use of digits (0–9) to label clusters, forming an intuitive decimal tree.

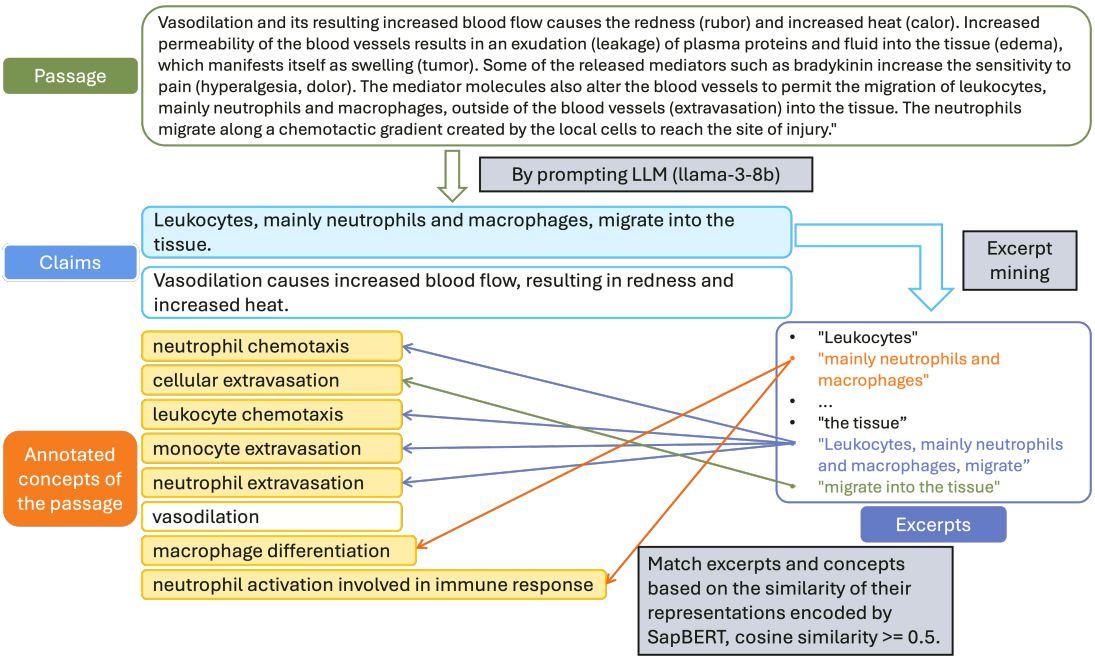

Figure 3: An example of constructing a claim-concept instance is as follows: Given a passage, we prompt the LLM to breakdown the passage into several claims. For one claim, we then perform excerpt mining. Next, we match these mined excerpts to the passage's annotated concepts by assessing semantic similarity. If an excerpt closely aligns with an annotated concept, we pair the concept with the claim. In this example, seven concepts are paired with a single claim, forming a claim-concept instance.

The final score of the token t is the average of two features:

$$z_{i,t} = avg(e_{i,t}, h_{i,t})) \tag{7}$$

$$\hat{y}_i = \arg\max_t(\sigma(z_{i,t})) \tag{8}$$

where $h_{i,t}, e_{i,t}, z_{i,t} \in \mathbb{R}^1$, σ is the Softmax function. The model parameters are optimized by minimizing the $CrossEntropyLoss(y, \hat{y})$.

Our preliminary experiments revealed that using only one canonical name-ssID pair to introduce a concept into the model did not provide strong performance. It is crucial to incorporate synonym-, mention-, and passage-ssID pairs for model improvement if they are available. Therefore, our model is trained on various input-output pairs. When the input is a span and the output is the ssID of a single concept, the model learns "indexing". When the input is a longer text and the output includes multiple ssIDs for the concepts are presented in the input, the model is trained for "recognition".

3.4 Multi-level queries generated by LLMs

Biomedical concepts are more challenging to recognize when the query is a passage compared to a sentence or span. By extracting shorter segments (e.g., sentences, phrases) from a passage, the model can better identify concepts that are difficult to capture when the query is a passage. Our framework, MA-COIR, is trained to process multiple levels of queries, enabling the integration of results from various query types derived from a passage into the final predictions.

In this study, we employ an open-source LLM - llama-3-8b (AI@Meta, 2024), to generate simplified queries from passages. For the CDR and HPO datasets, where concepts are associated with specific "mentions", the model generates concept names to serve as queries. Given that HOIP concepts are not consistently expressed as phrases, we use the model to transform passages into sentence-level claims and span-level concept names.

Claims are prioritized over segmented sentences because they encapsulate the passage's meaning in a coherent and self-contained manner, facilitating comprehension and recognition. In contrast, segmented sentences often lack sufficient context, leading to ambiguity. Claims provide the necessary abstraction and semantic synthesis, aligning more effectively with downstream tasks that rely on conceptual understanding.

The concept name generation is performed under a 10-shot ICL setting. For a given passage in the test set, we randomly select 10 passage-concept

Split	Data	Passage	Claim	Concept	Mention
Train	CDR	500	-	1,328	2,672
	HPO	182	-	416	926
	HOIP	225	682	337	-
Test	CDR	500	-	2,778	4,600
	HPO	23	-	159	237
	HOIP	37	165	265	-

Table 1: Statistics of instances.

pairs from the training set as demonstrations of the prompt.[3] Claim generation is done in a zero-shot setting due to the lack of annotated passage-claim pairs. Prompts we used are provided in Appendix Fig. 5.

3.5 Data augmentation

After breaking down the passage into claims using an LLM on the HOIP dataset, we generate claim-ssID pairs from the training set for semi-supervised learning. This data construction follows a common weakly supervised NER approach, consisting of two steps:

- Excerpt mining: Identify noun phrases and excerpts consisting of "a noun phrase and a verb linked to that noun phrase" using the dependency tree of a generated claim. We use spaCy (Honnibal and Montani, 2017) as the dependency parser.
- Labeling function: Represent each excerpt similarly to how a concept or query is represented, then compute the cosine similarity between the excerpt and annotated concepts from the passage. If any excerpt in the claim has a cosine similarity ≥ 0.5 to a gold concept, that concept is assigned to the claim.

Many matched excerpts only capture part of the meaning of the corresponding concept. Pairing the entire claim (which the excerpt appears) with the concept reduces noise compared to pairing the excerpt alone with the concept. An example of constructing a claim-concept instance is shown in Fig. 3.

4 Experiments

4.1 Datasets

Target concepts in an ontology are expressed frequently either as mentions or not. The motivation for proposing MA-COIR is to apply a pragmatic approach for the latter. To evaluate the framework's effectiveness in both cases, we conduct experiments on the three datasets.

CDR The pair of the MeSH [4] and BC5CDR dataset (Li et al., 2016). The 2015 version of the MeSH vocabulary includes 258K terms and BC5CDR comprises 1,500 passages annotated with MeSH terms based on entity mentions. MeSH is not a formally defined ontology, evaluating performance on this scenario establishes a reference for the lower bound of ontological content.

HPO The pair of Human Phenotype Ontology (HPO) (Gargano, 2023)[5] and HPO GSC+ dataset published by Lobo et al. (2017). The latest version of the HPO ontology includes over 19,000 concepts. The HPO GSC+ dataset comprises 228 PubMed abstracts and 1,933 mention annotations, each mention linked to a concept.

HOIP The pair of Homeostasis Imbalance Process (HOIP) ontology (Yamagata et al., 2024) and HOIP dataset (El Khettari et al., 2024).[6] The ontology includes over 60,000 concepts related to homeostasis imbalance processes, of which 44,439 biological process concepts are target concepts.

The dataset consists of 362 passages extracted from PubMed papers. Each passage is annotated with biological process concepts from the HOIP ontology. Mention annotations of concepts are not provided. Notably, a concept may be annotated based on its relevance to a process mentioned in the passage, even if the concept is not stated in the passage (this relevance may depend on the annotator's background knowledge).

We conduct training with the original train/dev set, and evaluation with a refined test set containing only explicitly mentioned concepts.

4.2 Comparison system

XR-Transformer. Prior to MA-COIR, no supervised biomedical CR model directly generated a list of ontology concepts from free text. By treating concepts as labels, CR task can be naturally framed as an instance of extreme multi-label text classification (XMC), where a passage is assigned multiple relevant ontology terms. We adopt XR-Transformer (Zhang et al., 2021), a state-of-the-art

[3] Preliminary experiments using n-shot settings ($n = 0, 1, 3, 5, 10$) for LLM prompting on the HOIP dataset showed that the best results were achieved with a 10-shot setting.

[4] https://www.ncbi.nlm.nih.gov/mesh/

[5] https://hpo.jax.org/

[6] https://github.com/norikinishida/HOIP-dataset

XMC model with top-tier performance across multiple public benchmarks, as a strong baseline.

kNN-searcher. Given the lack of existing approaches that do not use mentions for CR, we selected a straightforward baseline method: the top-k Nearest Neighbor (kNN) search, which can retrieve candidate concepts based on a given query. As the way we represent a concept E_C that described in Section 3.2, we get the representation of the query E_Q by the $TextEncoder$:

$$X_Q = TextEncoder(Q) \in \mathbb{R}^{l \times H} \quad (9)$$

$$E_Q = avg(X_Q) \in \mathbb{R}^{H} \quad (10)$$

where l is the token length of the query, and H is the dimension of a token's embedding.

With E_Q and representations of all concepts $\{E_{C_1}, ..., E_{C_n}\}$ as input vectors, we implemented Faiss (Douze et al., 2024) for a fast vector search of E_Q among large-scale concept spaces, by calculated similarity based on Euclidean distance. The kNN-searcher may return a candidate even if its distance from the query is large, when no other concepts closer to the query exceed the distance of the candidate. To mitigate false positives, we classify retrieved concepts with a similarity score < 0.6 as non-predictions.

Additionally, we conduct a comparative analysis of our approach against (Shlyk et al., 2024) and (El Khettari et al., 2024) under a controlled setup. Details are described in Section 6.4.

4.3 Setups

We trained MA-COIR and XR-Transformer using passage-, name-, and synonym-ssID pairs for all three datasets. When annotated mentions or generated claims were available, the model was trained with mention- and claim-ssID pairs. The models trained with synthetic claim-ssID pairs is referred to as **MA-COIR-a** and **XR-Transformer-a**. For checkpoint selection, we used only passage-ssID pairs from the development set. Evaluation involved testing the model with various types of queries, including passages, gold mentions (for CDR and HPO), generated claims (only for HOIP), and generated concept names. The statistics for the instances are provided in Table 1. Hyperparameters are listed in Appendix A.1.

4.4 Evaluation metrics

We evaluate all models using precision (Pre), recall (Rec), and micro F1-score (F1), measured across different query levels. For MA-COIR, we use beam search to generate top-k concept sequences per query. Each sequence is segmented into ssID-like spans using semicolons as delimiters. Spans not matching any defined ssID are discarded. All valid spans across k sequences are then merged and deduplicated to form the final prediction set. When multiple queries are derived from a single passage, their predictions are aggregated and compared against the gold annotations for that passage.

To ensure a fair comparison, passage-level input for the kNN-searcher is the same full-text passage used by MA-COIR, rather than shorter fragments obtained via "excerpt mining" we described in Section 3.5.

5 Results

Tables 2 and 3 summarize model performance across three biomedical concepts. On both CDR and HPO, MA-COIR consistently achieves the best F1 scores with passage-level inputs (47.6 and 60.0, respectively), while kNN-searcher and XR-Transformer perform best with span-level inputs. In the more challenging HoIP setting, MA-COIR-a and XR-Transformer-a outperform kNN-searcher, with XR-Transformer-a achieving the highest F1 for passage- and claim-level inputs ((19.8 and 23.4), and MA-COIR leading in the span-level setting (26.8). We analyze results from three complementary perspectives: concept type, input granularity, and real-world applicability.

Concept Type. The three datasets involve concept spaces of increasing complexity—from chemical and drug names (CDR), to phenotype abnormalities (HPO), and finally to abstract homeostasis imbalance processes (HoIP).

In CDR, most gold concepts are explicitly mentioned in text or have close surface-level synonyms, making the kNN-searcher highly effective. However, HPO concepts such as "Abnormality of body height" or "Abnormal platelet morphology" are semantically richer and less likely to appear verbatim. Here, supervised models like MA-COIR and XR-Transformer gain a clear edge by leveraging learned task-specific information.

HoIP presents the greatest challenge: many target concepts are abstract, fine-grained, and rarely expressed via identifiable mentions, challenging to recognize even for experts (e.g., "dysregulation of matrix metalloproteinase secretion"). In addition, HoIP lacks mention-ssID training pairs, limiting

Dataset	k	Query	MA-COIR			XR-Transformer			kNN-searcher		
			Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
CDR	1	Passage	51.0	**44.6**	**47.6**	**79.6**	11.6	20.3	13.3	0.1	0.1
		Mention	67.2	72.0	69.5	67.1	71.4	69.1	**75.5**	**82.5**	**78.9**
		Concept	57.2	41.2	47.9	57.2	41.5	48.1	**63.5**	**48.2**	**54.8**
	5	Passage	36.5	**49.6**	**42.0**	**45.3**	33.1	38.3	12.5	0.1	0.2
		Mention	17.1	74.8	27.9	13.8	73.6	23.3	**18.9**	**92.0**	**31.3**
		Concept	15.2	44.2	22.6	12.4	44.4	19.4	**16.5**	**56.1**	**25.5**
	10	Passage	**29.9**	**52.0**	**37.9**	26.7	39.0	31.7	10.5	0.1	0.2
		Mention	9.2	75.5	16.4	7.1	74.1	13.0	**11.4**	**93.1**	**20.3**
		Concept	8.3	45.4	14.0	6.4	44.8	11.2	**9.9**	**57.3**	**16.9**
HPO	1	Passage	67.7	**53.8**	**60.0**	**91.3**	13.5	23.5	33.3	0.6	1.3
		Mention	85.6	80.1	82.8	**88.1**	**85.3**	**86.6**	70.7	71.2	70.9
		Concept	**65.9**	57.1	61.2	65.2	**57.7**	**61.2**	58.5	50.6	54.3
	5	Passage	60.8	**57.7**	**59.2**	**61.7**	45.5	52.4	11.1	0.6	1.2
		Mention	21.2	84.0	33.8	19.2	**87.8**	31.5	**21.3**	**87.8**	**34.3**
		Concept	**18.5**	**66.7**	**29.0**	15.4	66.0	25.0	18.1	**66.7**	28.4
	10	Passage	**54.1**	59.6	**56.7**	43.9	**64.7**	52.3	7.7	0.6	1.2
		Mention	12.4	87.2	21.7	9.9	87.8	17.7	**13.9**	**89.1**	**24.0**
		Concept	**11.0**	**73.7**	**19.2**	8.2	67.9	14.6	**11.0**	67.9	18.9

Table 2: Results of the top-k generated sequences by MA-COIR and the top-k retrieved concepts by the XR-transformer and kNN-searcher on the CDR and the HPO. "mention" are gold annotated mentions of a passage. "concept" are generated concepts by the LLM given a passage. Red values indicate the highest F1 score achieved for each query type on a given dataset.

k	Query	MA-COIR			MA-COIR-a			XR-Transformer-a			kNN-searcher		
		Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
1	Passage	11.1	25.0	15.4	13.0	**27.3**	17.6	**32.4**	13.6	**19.2**	6.7	2.3	3.4
	Claim	8.2	21.6	11.9	14.1	**30.7**	19.3	**19.8**	28.4	**23.4**	6.7	8.0	7.3
	Concept	18.2	46.6	26.2	**18.5**	**48.9**	**26.8**	17.8	45.5	25.6	13.0	35.2	19.0
5	Passage	8.6	34.1	13.8	11.0	**39.8**	17.2	**14.6**	30.7	**19.8**	2.1	3.4	2.6
	Claim	6.0	45.5	10.7	**7.4**	**47.7**	**12.8**	6.5	45.5	11.4	3.8	17.0	6.3
	Concept	6.4	64.8	11.6	**6.7**	**68.2**	**12.1**	5.5	64.8	10.1	5.0	56.8	9.1
10	Passage	7.2	36.4	12.0	9.8	**45.5**	**16.2**	**10.0**	**42.0**	16.2	2.4	6.8	3.6
	Claim	4.7	54.5	8.7	**5.9**	**59.1**	**10.7**	4.2	55.7	7.8	2.6	17.0	4.4
	Concept	3.9	69.3	7.4	**4.4**	**78.4**	**8.4**	3.0	69.3	5.7	3.3	62.5	6.2

Table 3: Results of the top-k generated sequences by MA-COIR and the top-k retrieved concepts by the XR-Transformer and kNN-searcher on the HOIP dataset. "claim" and "concept" refer to generated claims and concepts, produced by the LLM given a passage. Red values indicate the highest F1 score achieved for each query type.

supervised grounding.[7] As a result, all models struggle, but the gap between supervised and unsupervised methods widens. This underscores a key insight: concept complexity and the mentioned way are critical determinants of method suitability.

Input Granularity. MA-COIR excels with passage-level inputs, outperforming XR-Transformer by large margins on CDR (47.6 vs. 38.3) and HPO (60.0 vs. 52.4), and achieving stronger recall on HoIP. The kNN-searcher, by contrast, underperforms in this setting due to poor alignment between full passages and span-based embeddings.

At the span-level, performance varies: MA-COIR outperforms XR-Transformer when given gold mentions on CDR, but lags slightly on HPO. When using concept names generated by LLMs, MA-COIR matches or exceeds XR-Transformer. This reflects the robustness of MA-COIR to input variation and highlights a key practical strength: in real applications, gold mentions are unavailable, and LLM-generated spans often differ in granularity from ontology entries, making retrieval harder. MA-COIR's adaptability makes it better suited for such realistic, mention-free scenarios.

Practical Considerations. On CDR and HPO, MA-COIR demonstrates strong and consistent performance, proving its effectiveness for real-world biomedical CR. On HoIP, XR-Transformer-a achieves slightly higher F1 than MA-COIR-a (19.8 vs. 17.6). This is largely due to the dataset's statistics: each passage contains, on average, 7.2 gold

[7] A study examining the impact of mention information on MA-COIR, conducted on CDR, revealed a significant difference with and without mention-ssID pairs as training data, as detailed in Appendix A.4.

concepts. XR-Transformer-a's fixed-k retrieval (with $k = 5$) benefits from limiting false positives, whereas MA-COIR-a uses beam search to generate unbounded concept sequences, trading off precision for recall. In practice, however, concept density varies across documents, and setting an optimal k is non-trivial, limiting the robustness of fixed-k methods like XR-Transformer.

On span-level CDR tasks, MA-COIR and XR-Transformer perform comparably, but both fall short of kNN-searcher when provided with gold mentions. On HPO, kNN-searcher is only competitive when given gold mentions and big k values (e.g., $k = 5$ or 10). Further analysis (Appendix A.3) reveals that MA-COIR struggles to recognize unseen concepts lacking training exposure—an issue shared with XR-Transformer. In contrast, kNN-searcher remains unaffected. Nonetheless, we believe this limitation can be mitigated via data synthesis strategies: our preliminary experiments confirm the feasibility of using synthetic samples to improve MA-COIR's generalization.

Summary. MA-COIR delivers strong performance across diverse concept types and input settings. While training data coverage remains a limitation, this can be addressed with scalable augmentation techniques. Given its flexibility, robustness to input variation, and effectiveness even without gold mentions, MA-COIR offers a practical and reliable solution for biomedical CR.

6 Analysis

6.1 Effectiveness of ssID

To verify the effectiveness of ssID, we compared it with other types of indexes can be used for the recognition on the HOIP.

- Random ID: Randomly assign a number to each concept as an index. The index ranges from 0 to the number of all ontology concepts.
- Ontology ID: The unique ID of each concept in the ontology is used as the index. Like "HOIP_0004832" is the ontology ID of "TNF signaling", and the index for generation.
- ssID (name): As described in Section 3.2.
- ssID (+hypernyms): The indexes are based on constructing a label tree using the concatenation of the representation of a name of each concept, and the average of the representations of its hypernyms. The hypernymy and hyponymy relations is known from the ontology. Let U_C denote a set of concepts that are hypernyms of concept C defined in the ontology. The representation of the concept C used for label tree construction changed from eq. 2 to eq. 4.

$$E_{U_{C_i}} = avg(X_{U_{C_i}}) \in \mathbb{R}^H \quad (11)$$

$$E_C = [avg(X_C) : avg(E_{U_C})] \quad (12)$$

where ":" is the concatenation operation, H is the dimension of a token's embedding, $E_C \in \mathbb{R}^{2\times H}$.

Index type	Pre	Rec	F1
Random ID	7.8	31.8	12.5
Ontology ID	6.7	**47.7**	11.8
ssID (name)	**11.1**	25.0	**15.4**
ssID (+hypernyms)	9.7	20.5	13.1

Table 4: Results of the top-1 generated sequence using various index types with the passage queries on the HOIP dataset by MA-COIR.

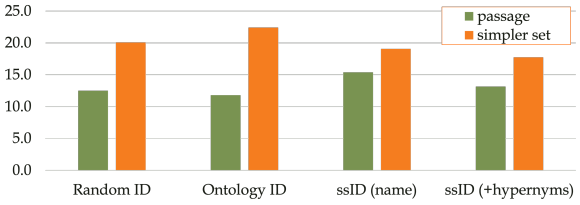

Figure 4: F1 scores by MA-COIR between complex query (passage) and the average of the simpler set of queries (claim/concept) from top-1 generated sequence using different indexes on the HOIP.

The experimental results are summarized in Table 4. Both Random ID and Ontology ID performed well on span-level queries, providing higher recall compared to ssIDs. On the other hand, using ssID (name) achieved the highest precision and F1 scores for passage-level queries. As shown in Fig. 4, ssID-based indexing demonstrates robustness across both complex and simple queries, whereas Random ID and Ontology ID perform optimally only on shorter queries. In the absence of tools to retrieve non-passage level information, ssID is clearly the superior choice.

6.2 Effectiveness of data augmentation

The results for the MA-COIR-a are presented in Table 3. Incorporating claim-ssID pairs, as described

Query	Pre	Rec	F1
passage	13.0	27.3	17.6
+ claim	12.5	45.5	19.7
+ concept	12.3	**64.8**	20.7
+ concept	**14.7**	61.4	**23.7**

Table 5: Results of the top-1 generated sequence by MA-COIR-a on HOIP.

Dataset	Method	Pre	Rec	F1
HPO	REAL-1st hit	40.0	49.0	44.0
	REAL-GPT3.5	**68.0**	48.0	56.0
	kNN-searcher	58.5	50.6	54.3
	MA-COIR	63.4	**54.5**	**58.6**
HOIP-o	ICL-Llama	**43.1**	11.8	18.6
	kNN-searcher	42.0	13.9	20.9
	MA-COIR	23.7	**19.6**	**21.5**

Table 6: Comparison between our methods and previous works. "HOIP-o" refers to the original test set.

in Section 3.5, leads to improvements across all metrics for all query types. F1 scores for claim-queries increase by 4.6 points compared to MA-COIR. Across all query types, the improvement in recall exceeds that in precision, indicating that the added data is both accurate (with minimal noise, which helps maintain precision) and diverse, benefiting all query types.

6.3 Combination of different-level queries

The results of combining predictions of various types of queries are presented in Table 5. While the accuracy of decomposing full passages into shorter units is low, MA-COIR captures additional concepts that are difficult to detect from full-length inputs alone. The predictions from different query levels exhibit partial but non-trivial overlap, revealing their complementary strengths.

Each query type offers distinct advantages. Aggregating predictions across all levels yields substantial gains. Recall improves significantly from $(27.3 \rightarrow 45.5 \rightarrow 64.8)$ when integrating all three, underscoring the value of multi-level querying.

6.4 More comparisons

Our framework operates under different setups compared to previous studies that were validated on the same dataset. We provide results using a more comparable setting to ensure fair evaluation (see Table 6).

For HPO dataset, REAL (Shlyk et al., 2024) reports results for two approaches: for an LLM generated mention, selecting the top-1 candidate from three candidates provided to GPT-3.5 (REAL-GPT3.5) or taking the top-1 concept retrieved by kNN searching (REAL-1st hit). For comparison, we report the results by MA-COIR trained without mention-ssID pairs and the kNN-searcher we implemented using concept queries with $k = 1$.

For HOIP dataset, El Khettari et al. (2024) report the results of a similarity-based kNN search for concepts generated by llama-3-8b in its few-shot setting (ICL-Llama). After retrieval, they filtered out out-of-dataset predictions. We replicated their approach by using their generated concepts as queries and applying the same filter with kNN-searcher and setting $k = 1$.

From the results of REAL-1st hit and kNN-searcher on HPO (F1: 44.0/54.3), as well as kNN-searcher on concepts from ICL-Llama and our generated concepts (F1: 18.6/20.9) on HOIP-o, we can infer that the quality of our generated concepts and the representation of concepts/queries is consistent with previous methods.

The removal of out-of-dataset concepts significantly reduced false positives in similarity-based methods, improving precision to over 40.0 on the HOIP-o. In contrast, MA-COIR does not predict concepts never appeared in the training phase, such post-processing does not provide benefits.

Overall, our supervised recognizer, MA-COIR, outperforms unsupervised LLM-based solutions like REAL-GPT3.5 and ICL-Llama.

7 Conclusion

We present the MA-COIR framework, a flexible and implementable solution for recognizing both simple and complex biomedical concepts explicitly or implicitly appeared in scientific texts, without requiring specific mention information. The framework meets the needs of domain experts, as demonstrated by experiments on three vocabulary/ontology-dataset pairs. We introduce efficient methods for obtaining queries at various levels and data augmentation using an LLM and proving their efficacy in low-resource scenarios. MA-COIR's adaptability to multi-level queries enhances its practical utility. We further provide an in-depth analysis of biomedical concept recognition and potential directions for future improvement.

Limitations

Although we would like MA-COIR to generate ssIDs for unseen concepts based on semantic similarities with seen concepts, results indicate that it lacks this capability. This restricts the model's applicability to the available dataset. Given that the annotated dataset contains significantly fewer concepts than the full ontology, further framework refinement is needed to allow comprehensive processing across different input levels and consistent mapping of all ontology concepts and their indexes.

It is essential to develop validation datasets that align with the needs of domain experts. In the HPO and HOIP test sets, the low proportion of unseen concepts limits the evaluation of the model's generalization to out-of-dataset concepts. Without observing MA-COIR's performance decline on the CDR dataset, this limitation might have gone unrecognized.

Last but not least, the performance of MA-COIR also depends on query quality. There is a substantial gap between results for concept names generated by an LLM and those derived from gold annotated mentions. Although we have not fully explored LLM-based query generation, it is unrealistic to expect consistent query quality across specialized biomedical domains. Thus, it is critical to both improve the model's robustness to lower-quality queries and identify ways to generate high-quality queries.

References

AI@Meta. 2024. Llama 3 model card.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2020. Autoregressive entity retrieval. *CoRR*, abs/2010.00904.

J Harry Caufield, Harshad Hegde, Vincent Emonet, Nomi L Harris, Marcin P Joachimiak, Nicolas Matentzoglu, HyeongSik Kim, Sierra Moxon, Justin T Reese, Melissa A Haendel, Peter N Robinson, and Christopher J Mungall. 2024. Structured Prompt Interrogation and Recursive Extraction of Semantics (SPIRES): a method for populating knowledge bases using zero-shot learning. *Bioinformatics*, 40(3):btae104.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library.

Oumaima El Khettari, Noriki Nishida, Shanshan Liu, Rumana Ferdous Munne, Yuki Yamagata, Solen Quiniou, Samuel Chaffron, and Yuji Matsumoto. 2024. Mention-agnostic information extraction for ontological annotation of biomedical articles. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 457–473, Bangkok, Thailand. Association for Computational Linguistics.

Michael A et al. Gargano. 2023. The human phenotype ontology in 2024: phenotypes around the world. *Nucleic Acids Research*, 52(D1):D1333–D1346.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Jyun-Yu Jiang, Wei-Cheng Chang, Jiong Zhang, Cho-Jui Hsieh, and Hsiang-Fu Yu. 2024. Entity disambiguation with extreme multi-label ranking. In *Proceedings of the ACM on Web Conference 2024*, pages 4172–4180.

Siddhant Kharbanda, Atmadeep Banerjee, Erik Schultheis, and Rohit Babbar. 2022. Cascadexml: Rethinking transformers for end-to-end multi-resolution training in extreme multi-label classification. *Advances in neural information processing systems*, 35:2074–2087.

Nikolaos Kolitsas, Octavian-Eugen Ganea, and Thomas Hofmann. 2018. End-to-end neural entity linking. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 519–529, Brussels, Belgium. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016. Biocreative V CDR task corpus: a resource for chemical disease relation extraction. *Database J. Biol. Databases Curation*, 2016.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. 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.

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.

OAKlib. 2023. Ontology access kit (oak).

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Darya Shlyk, Tudor Groza, Marco Mesiti, Stefano Montanelli, and Emanuele Cavalleri. 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, Bangkok, Thailand. Association for Computational Linguistics.

Yi Tay, Vinh Q. Tran, Mostafa Dehghani, Jianmo Ni, Dara Bahri, Harsh Mehta, Zhen Qin, Kai Hui, Zhe Zhao, Jai Gupta, Tal Schuster, William W. Cohen, and Donald Metzler. 2022. Transformer memory as a differentiable search index. *Preprint*, arXiv:2202.06991.

Qinyong Wang, Zhenxiang Gao, and Rong Xu. 2023. Exploring the in-context learning ability of large language model for biomedical concept linking. *Preprint*, arXiv:2307.01137.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Zero-shot entity linking with dense entity retrieval. In *EMNLP*.

Yuki Yamagata, Tatsuya Kushida, Shuichi Onami, and Hiroshi Masuya. 2024. Homeostasis imbalance process ontology: a study on covid-19 infectious processes.

Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. 2022. BioBART: Pretraining and evaluation of a biomedical generative language model. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 97–109, Dublin, Ireland. Association for Computational Linguistics.

Jiong Zhang, Wei-Cheng Chang, Hsiang-Fu Yu, and Inderjit Dhillon. 2021. Fast multi-resolution transformer fine-tuning for extreme multi-label text classification. *Advances in Neural Information Processing Systems*, 34:7267–7280.

Item	Value
model_card	facebook/bart-large
learning_rate	1e-5
num_epoch	50
batch_size	4
max_length_of_tokens	1024

Table 7: Hyperparameters of the recognizer.

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.

A Appendix

A.1 Hyperparameters

The BART-based language model (facebook/bart-large) used in MA-COIR for recognition is trained with hyperparameters listed in the Table 7.

The hyperparameters of the K-Means clustering algorithm used for hierarchical clustering process, are g and m, while g is the maximum number of the elements covered by a node when we can stop further dividing the node into smaller clusters. m is the number of clusters when we divide the elements in a node. For example, when $g = 10, m = 10$, if there are 9 elements in the current node, we do not divide the elements in this node by clustering; if there are 18 elements in the current node, we will do a clustering for these elements, so that these elements will be categorized into $m = 10$ clusters.

In this work, we set $g = 10, m = 10$. Our choice is based on two main considerations: (1) Empirical evidence: Preliminary experiments using the DSI-inspired configuration ($g = 10$, $m = 100$) resulted in lower F1 scores on the HOIP validation set, compared to the current setting. (2) Structural consistency: Using decimal numbering (0–9) aligns naturally with our hierarchical "ssID" design, which organizes concepts into 10 branches per level, facilitating both interpretability and implementation.

For the training of XR-Transformer, we implement the model with the library pecos[8], setting the hyperparameters provided by the authors, as those have already been tuned. The architecture of the Transformers model we used in the experiments is BERT.

[8]https://pypi.org/project/libpecos/

BC5CDR Concept: Please list all concepts referring to a chemical concept or a disease concept in following input (make sure not to add any information), and return the output as a jsonl, where each line is {"Chemical":[CHEMICAL]} or {"Disease":[DISEASE]}. If there is no chemical or disease concept in the input, it is fine to only return {"Chemical":None} or {"Disease":None}. Directly return the jsonl with no explanation or other formatting. The input is: "{passage}"

HPO GSC+ Concept: Please list all concepts referring to a medically relevant human phenotype concept in following input (make sure not to add any information), and return the output as a jsonl, where each line is {"phenotype":[PHENOTYPE]}. If there is no human phenotype concept in the input, it is fine to only return {"phenotype":None}. Directly return the jsonl with no explanation or other formatting. The input is: "{passage}"

HOIP Concept: Please list all biological processes involved in the phenomenon described in the following input (make sure not to add any information), and return the output as a jsonl, where each line is {"process":[PROCESS]}. If there is no process in the input, it is fine to only return {"process":None}. Directly return the jsonl with no explanation or other formatting. The input is: "{passage}"

HOIP Claim: Please breakdown the following input into a set of small, independent claims (make sure not to add any information), and return the output as a jsonl, where each line is {"claim":[CLAIM], "score":[CONF]}. The confidence score [CONF] should represent your confidence in the claim, where a 1 is obvious facts and results like 'The earth is round' and '1+1=2'. A 0 is for claims that are very obscure or difficult for anyone to know, like the birthdays of non-notable people. If the input is short, it is fine to only return 1 claim. Directly return the jsonl with no explanation or other formatting. The input is: "{passage}"

Figure 5: Prompt template for generating concept names / claims for passage. A prompt consists of task instruction, output format instruction, several demonstrations and the query.

		CDR		HPO	
k	Query	Seen	Unseen	Seen	Unseen
1	passage	57.2	0.3	60.0	0.0
	mention	92.4	0.0	89.3	0.0
	concept	52.9	0.0	63.6	0.0
5	passage	63.6	0.3	64.3	0.0
	mention	95.2	2.9	92.1	12.5
	concept	56.3	1.5	74.3	0.0
10	passage	66.6	0.4	66.4	0.0
	mention	95.8	4.0	95.0	18.8
	concept	57.7	2.2	80.7	12.5

Table 8: Recalls on the seen and unseen concepts of the top-k generated sequences by MA-COIR.

Data	Query	Pre	Rec	F1
All	passage	51.0	44.6	47.6
	mention	67.2	72.0	69.5
	concept	57.2	41.2	47.9
- mention	passage	36.1	30.5	33.1
	mention	39.5	42.8	41.1
	concept	32.4	22.3	26.4
- synonym	passage	48.2	42.3	45.0
	mention	67.4	72.0	69.6
	concept	58.2	41.4	48.3
- mention - synonym	passage	36.0	30.5	33.0
	mention	41.9	44.8	43.3
	concept	37.6	24.8	29.9

Table 9: Results on CDR with different training data. "All" contains passage-ssIDs pairs, name-ssID pairs, synonym-ssID pairs and mention-ssID pairs constructed from the original training set.

A.2 LLM Application

We applied a large language model llama-3-8b for query generation. For all concept generation tasks, the prompt consists of "instruction", "n demonstrations" under the n-shot setting, and the passage. The prompts we used for concept name generation on CDR, HPO and HOIP are shown in Fig. 5. For claim generation, the prompt template we used for a passage on HOIP is shown in Fig. 5. The generation is conducted in a zero-shot scenario cause there is no annotated data for passage-claim pairs.

A.3 Performance on seen and unseen concepts

Upon examining MA-COIR's performance on both seen (concepts appeared in the training set) and unseen concepts (concepts only appeared in the test set), we found that the performance gap between it and the kNN-searcher is primarily due to its inability to recognize unseen concepts. As presented in the Table 8, when we evaluated the model on unseen concepts, MA-COIR achieved a recall of nearly 0.0 on both the CDR and the HPO.

A.4 Training data for "Indexing" capability of the recognizer

The indexing capability of the model refers to the model's ability to generate the correct ssID for the query when it is a span. On datasets labelled with mentions, in addition to the canonical names and synonyms of a concept in the ontology that can be used to train model indexing capabilities, mentions are also very effective data. We conducted an ablation study on the CDR dataset to confirm the impact of synonym- and mention-ssID information on the model's ability to recognize concepts. The results can be seen in Table 9.

After removing the mention-ssID data, the model's performance dropped significantly; removing the synonym-ssID data, the performance on the passage-level query dropped less and even improved on the span-level query. This illustrates that the way a concept is expressed within a particular application (passage) is important for capturing the relationship between the concept and the ssID. Not only the indexing capability are influenced by removing mention data, but also the recognition on the passage query ($\downarrow$ 14.5 F1 score). The slight improvement after removing synonym-ssID pairs indicates how different the common expressions written in scientific papers and the technical terms of a concept are. Using synonyms to enrich concept information makes the query and a concept further apart in representation.