Inferring Scientific Cross-Document Coreference and Hierarchy with Definition-Augmented Relational Reasoning

Lior Forer
The Hebrew University of Jerusalem
lior.forer@mail.huji.ac.il

Tom Hope
The Hebrew University of Jerusalem
The Allen Institute for AI
tomh@allenai.org

Abstract

We address the fundamental task of inferring cross-document coreference and hierarchy in scientific texts, which has important applications in knowledge graph construction, search, recommendation and discovery. Large Language Models (LLMs) can struggle when faced with many long-tail technical concepts with nuanced variations. We present a novel method which generates context-dependent definitions of concept mentions by retrieving full-text literature, and uses the definitions to enhance detection of cross-document relations. We further generate *relational* definitions, which describe how two concept mentions are related or different, and design an efficient reranking approach to address the combinatorial explosion involved in inferring links across papers. In both fine-tuning and in-context learning settings, we achieve large gains in performance on data subsets with high amount of different surfaces forms and ambiguity, that are challenging for models. We provide analysis of generated definitions, shedding light on the relational reasoning ability of LLMs over fine-grained scientific concepts.

1 Introduction

As the volume of scientific papers continues to explode, so does the intricate web of concepts mentioned across papers. Papers in computer science, for example, discuss fine-grained concepts such as algorithms, methods, tasks and problem areas (Luan et al., 2018). Being able to automatically identify, e.g., when papers are referring to the same problem while using very different language (*cross-document coreference*), or that a method used in one paper is part of a more general family of methods (*hierarchy*)—could help unlock a wide range of applications, including in scientific knowledge-base construction (Mondal et al., 2021; Hope et al., 2021), information retrieval (Wang et al., 2023; Kang et al., 2024; Wu et al., 2024), recommendation systems (Mysore et al., 2023), systems for paper reading and learning new topics (Murthy et al., 2022; Guo et al., 2024), and even in computational creativity for science (Hope et al., 2017; Portenoy et al., 2022; Wang et al., 2024; Srinivasan and Chan, 2024). However, detecting cross-document coreference and hierarchy relations in scientific papers poses significant challenges (Cattan et al., 2021b). Concepts manifest in long-tail, nuanced variations. Technical language is characterized by ambiguity, different surface forms and fine-grained levels of concept abstractions. As an example illustrating several of these challenges, consider *energy-based network embedding*, a specific method under the broader concept of *neural representation learning* despite no surface overlap or lexical cues that may suggest one concept subsumes the other; while *low energy embedded networks* in communications is unrelated, despite strong surface similarity (Figure

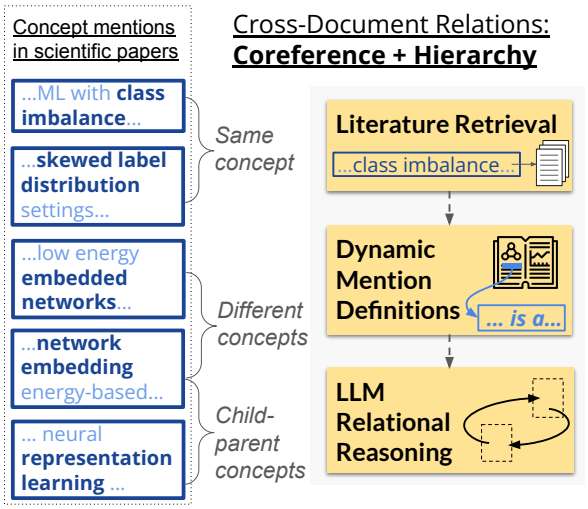

Figure 1: We detect cross-document coreference and hierarchy by augmenting original inputs from papers with context-sensitive definitions and relational reasoning.

Transactions of the Association for Computational Linguistics, vol. 14, pp. 802–828, 2026. https://doi.org/10.1162/TACL.a.685
Action Editor: Marie-Catherine de Marneffe. Submission batch: 4/2025; Revision batch: 10/2025; Published 6/2026.

1). Surface forms *network*, *energy*, *embedding/embedded* refer to different scientific concepts in different contexts (ambiguity). Another example is *class imbalance* which refers to a *skewed label distribution* (Figure 1)—the same underlying scientific concept with different terms.

We explore the ability of large language models (LLMs) to detect cross-document coreference and hierarchy, i.e., detecting when scientific concept mentions across two papers co-refer (e.g., *network embedding* and *neural representation of graphs*), and when one concept mention is a parent/child of another mention (*network embedding* and the parent concept *neural representation learning*).

We present a novel method for this challenging setting. Given a concept mention in a paper, our approach first generates a concept *definition* (e.g., *network embedding is a neural representation of graphs...*). Definitions are generated by retrieving literature relevant to the mention and synthesizing a concept description that takes into account both global literature information and the local context surrounding the mention. We augment original inputs with the definitions, and train LLMs to make use of them for detecting cross-document coreference and hierarchy. We further introduce *relational* definitions, which explicitly describe how two concept mentions are related or different. To avoid combinatorial explosion when generating relational definitions across all possible pairwise candidates, we create an efficient two-stage re-ranking approach that first generates *singleton* definitions and uses them to score and select candidates for which relational definitions will be computed.

We train and evaluate our approach on the large and challenging SciCo dataset (Cattan et al., 2021b), a benchmark for scientific cross-document coreference and hierarchy detection. The examples used above are real cases appearing in SciCo. In both fine-tuning (FT) and in-context learning (ICL) settings, our definition-based approach achieves substantial improvements in detection accuracy, particularly on challenging data subsets marked by a high amount of different surface forms and ambiguity. These hard cases account for 10–20 percent of the test data. In the FT setting, we obtain new state-of-the-art results, with singleton and relational definition augmentation leading to large gains in CoNLL F1 over models that do not use definitions, and particularly strong gains in hierarchy detection. In the ICL setting, we evaluate the ability of `GPT-4o-mini` equipped with the advanced prompt optimizer DSPy (Khattab et al., 2024); we find that when our definition augmentation is added, results substantially increase.

Finally, we conclude with an extensive analysis of the strengths and limitations of contextualized, relational, and retrieval-based definitions, shedding light on when and how they contribute to performance. This analysis opens up intriguing avenues for future research, particularly in refining the generation and utilization of dynamic concept definitions and in improving the relational reasoning (Alexander, 2016) ability of LLMs over nuanced, fine-grained technical concepts.[1]

2 Problem Setting

We focus on the fundamental problem of hierarchical cross document coreference resolution (H-CDCR), first introduced in Cattan et al. (2021b). In H-CDCR, our goal is to induce clusters of context-sensitive mentions that co-refer to the same concept, and to deduce a hierarchy among these concept clusters. Cattan et al. (2021b) created SciCo, a challenging dataset for training and evaluating H-CDCR models focusing on the complex scientific domain; in SciCo, the goal is to infer coreference clusters from sets of mentions extracted from computer science papers, and a hierarchy over the co-reference clusters. For example, as illustrated in Figure 1, the goal would involve detecting that *class imbalance* and *skewed label distribution* belong to the same underlying concept cluster, that *embedded networks* does not belong in the same cluster as *network embedding*, and that *network embedding* belongs to a concept cluster that is a child of the parent concept *representation learning*.

Formally, we are provided with a collection of documents D, each $d \in D$ annotated with concept $mentions$. Let $M_d = \{m_1, m_2, ..., m_n\}$ be the set of mentions in document d and M the set of mentions across all $d \in D$. Similar to cross-document co-reference resolution, the first goal is to cluster the mentions in M into disjoint clusters $C = \{C_1, C_2, ..., C_n\}$. Each cluster should contain mentions $\{m|m \in C_i\}$ that refer to the same underlying concept. The second goal is to determine a hierarchy over those clusters, specifically a hierarchical relation between cluster pairs $C_i \rightarrow C_j$ that reflects on a $referential\ hierarchy$; a rela-

[1] Code, models and generated definitions available at https://github.com/TomHopeLab/ScicoRadar.

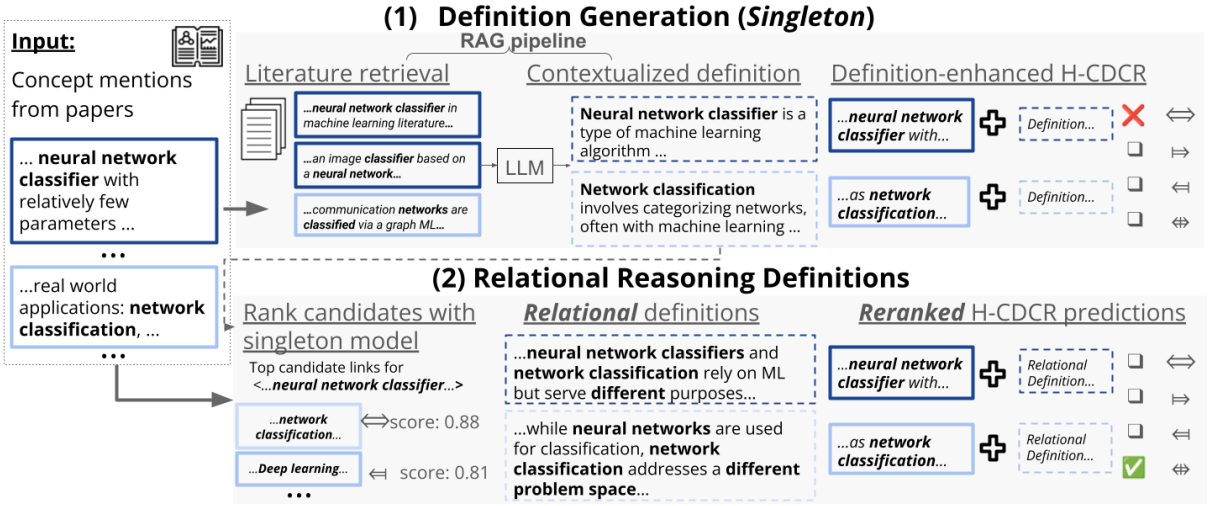

Figure 2: Overview of SCICO-RADAR. We are given as input papers with concept mentions (e.g., methods and tasks). **(1)** For each mention, we first create *singleton* definitions by retrieving relevant literature and using an LLM to generate context-dependent concept definitions. These definitions are used to augment the original inputs. We use the augmented input to train an LLM to detect cross-document coreference and hierarchy. **(2)** Using the model trained with singleton definitions, we rank promising top-K candidates, and create for them *relational* definitions that explicitly reason about pairwise concept relationships and further augment the input to enhance detection.

tion $C_i \rightarrow C_j$ exists when the concept underlying C_j entails a $reference$ to C_i. Given documents D and mentions M, the goal is to construct clusters C and a hierarchy graph $G_C = (C, E)$ by learning from a set of N examples $\{(D^k, M^k, G_c^k)\}_{k=1}^N$.

As we discuss next, we explore learning in both the fine-tuning (FT) and in-context learning (ICL) settings. Our novel approach involves augmenting mentions M with dynamic, contextual *definitions*. We aim to synthesize definitions for specific mentions—and *relational* definitions explicitly describing one or more relations between specific *pairs* of mentions—such that the generated definitions assist models in more accurately inferring the correct cross-document relations.

3 Methods

In this section, we present our SCICO-RADAR (Scientific Concept Induction with Relational Definition Augmented Reasoning) approach for the scientific H-CDCR task.

Overview At a high level, our approach consists of several components, as illustrated in Figure 2. We first retrieve global contexts from scientific papers corresponding to each mention and its local context. Next, we generate a definition for each mention based on the global and local context. We explore two types of definitions. The first type is a *singleton* definition which defines the mention given its context; for example, the definition for *kernel* may correspond to the operating systems or machine learning sense of the term, depending on the context. The second type we explore is a *relational* definition, where a mention is defined explicitly in relation to another mention. To avoid combinatorial explosion of generating relational definitions for all possible pairs of mentions and candidates, we further design a candidate re-ranking method inspired by IR research (Guo et al., 2020), in which only top-ranked candidates are routed to the relational definition generation module. Finally, we use the generated definitions to augment mention representations as part of training LLMs on SCICO, in both fine-tuning and in-context learning experiments. We now provide further details.

Adding definitions to H-CDCR models Following Cattan et al. (2021b), we consider a multi-class classification setting where each mention pair (m_1, m_2) can be assigned into four classes: (1) m_1 and m_2 co-refer (2) $m_1 \rightarrow m_2$ (3) $m_2 \rightarrow m_1$ and (0) none. We denote our definition augmentation model as a function $\mathcal{D}(\cdot, \cdot)$, where $\mathcal{D}$ is applied to a pair of candidate mentions m_1, m_2 and enriches each mention with a contextual definition. $\mathcal{D}$ may correspond either to singelton definitions, in which case the definitions for m_1 and m_2 are generated in

isolation of each other, or relational definitions in which the definition of m_1 depends on m_2 and vice versa. Equipped with $\mathcal{D}$ that enriches mention contexts, during training we use a set of mentions M to learn a pairwise scorer $f(*)$ (Yang et al., 2022) that minimizes

$$L = -\frac{1}{N} \sum_{\substack{m_1, m_2 \in \mathcal{M} \\ m_1 \neq m_2}} y \cdot \log(f(\mathcal{D}(m_1, m_2))), \quad (1)$$

where y represents one of the four possible classes, N denotes the total number of training pairs, $f(*)$ is a neural language model.

RAG-based definition generation We use a Retrieval-Augmented Generation (RAG) pipeline for definition generation. Given a specific mention and its context, our method first uses a retriever model to find relevant documents or passages from a large corpus; the retrieved contexts are then fed into a generative model which synthesizes contextualized definitions. Formally, for every mention $m_i \in M$, we denote the passage surrounding m_i as ϕ_{m_i} and $\psi(\phi_{m_i})$ is an embedding function that returns a vector representation which we use for retrieving relevant contexts from a large corpus $\mathcal{C}$ of passages obtained from scientific papers.[2]

We embed all passages in $\mathcal{C}$ using ψ. Then, for every mention m_i we retrieve a set of passages $\{c_1^i, c_2^i, ..., c_j^i\}$ with maximal dot product similarity to $\psi(\phi_{m_i})$. To reduce noise, we further follow common practice and apply a reranking and filtering step using a re-ranker LM ξ that returns a new re-ranked and filtered context set $\mathcal{R}_i = \{c_{1^*}, ..., c_{j^*}\} = \xi(\{c_1, c_2, \ldots, c_j\})$, where $\mathcal{R}_i$ is an ordered set of contexts for m_i, after reranking and filtering. Lastly, we feed $m_i; \phi_{m_i}; \mathcal{R}_i$ into an LLM that generates contextual definition d_i. In Figure 3, we show the prompt logic used for generating d_i (see technical details in Appendix A.1).

Relational definitions We introduce another method for generating definitions, which we name as *relational*. Relational definitions are designed to describe mentions in connection to another mention, thereby providing a more direct, explicit understanding of how two mentions may be related to each other. To create relational definitions, we feed into an LLM the *pairwise* contexts of two mentions, $m_i; \phi_{m_i}; \mathcal{R}_i$ and $m_j; \phi_{m_j}; \mathcal{R}_j$, with a custom prompt (Figure 3) for defining m_i in relation to m_j or vice versa. Note that relational definitions are asymmetric as they are centered around an "anchor" mention (see examples in Figure 2 and Table 5).

However, generating relational definitions for all pairs of mentions quickly explodes; for instance, in the SciCo test set this would entail 564,352 unique pairs for which we would need to generate definitions, where each mention contains multiple contexts from papers as well as the surrounding paragraph. On the other hand, singleton definitions may be computed offline for individual mentions and stored and retrieved efficiently. Thus, in order to make the relational definition generation process feasible, we create a two-stage method in which the first stage consists of using singleton definitions to efficiently filter the space of possible candidates, and in the second stage we create relational definitions only for a small subset of candidates. This approach broadly takes after the common retrieve-and-rerank pipeline often used in information retrieval systems (Lin et al., 2020) and also cascade approaches to link prediction (Safavi et al., 2022).

More formally, for each mention $m_i \in M$ and each candidate m_j we use singleton definition augmentation to compute likelihood scores $f(\mathcal{D}(m_i, m_j))$ (each prediction of the model is a logit of size 4). We first filter out pairs for which the model predicted *none* (no relationship). Then, for each $f(\mathcal{D}(m_i, m_j))$ we compute the softmax probabilities and rank the predictions based on the model's confidence in its assigned relation. We set a threshold value θ to select the top 25% pairs as a tradeoff between accuracy and computational efficiency; we compare different values of this threshold on the *Hard-10* subset in Appendix A.4). For each mention m_i, we generate relational definitions solely for the top pairs of mentions for which $f(\mathcal{D}(m_i, m_j)) \geq \theta$.

`GPT-4o` based definition generation In addition to the RAG-based approach, we explore the use of `GPT-4o` (OpenAI, 2024) for generating definitions. This method leverages the model's extensive parametric knowledge without incorporating external context retrieval. Specifically, for each mention m_i, we provide `GPT-4o` with the passage surrounding m_i denoted as ϕ_{m_i}. Unlike the RAG pipeline, this approach does not utilize additional context from external sources. While this method simplifies the definition generation process, it may reduce the quality of generated definitions, as we explore

[2] We use the corpus of full-text arXiv papers (§4).

further in Sections 4.2 and 5; it also substantially increases costs.

LLM training Finally, given definitions we integrate them into LLMs for training. The current state of the art reported by Cattan et al. (2021b) was based on fine-tuning a LONGFORMER model (Beltagy et al., 2020). We fine-tune recent state of the art LLM, and also experiment with both few-shot prompting and the more advanced DSPy (Khattab et al., 2024) prompt optimization framework (see more details in Appendix A.3).

4 Experiments

In this section we describe our experimental setup and results. We first outline the implementation details (§4.1), and then present the main results (§4.2), which introduce the baselines and report performance on both the hardest subsets and the overall dataset.

4.1 Implementation Details

Finetuning with definitions We train models both with the inclusion of definitions and without them. As base models, we report results for fine-tuning MISTRAL-7B[3] (Jiang et al., 2023), and LONGFORMER (Beltagy et al., 2020) which was used originally in Cattan et al. (2021b). We augment the original data to inject a definition for every mention. For MISTRAL-7B we construct a prompt (Figure 6) following the ChatML[4] format. For Longformer the input is processed as follows. For each mention $m1, m2$ and its corresponding paragraph, mention markers <m> and </m> surrounding the mention are used for obtaining a mention representation. We insert additional <def> and </def> markers surrounding each definition d_{m_1}, d_{m_2}. Finally we concatenate all the above with the separation tokens <s> and </s> separating each mention (See prompt in Figure 7).

We fine-tuned the MISTRAL-7B model from the HuggingFace (Wolf et al., 2020) library by adding a classification head, replacing the original head with a task-specific classification layer.[5] The classification head is a simple feed-forward layer with a softmax activation function (see Appendix A.2 for hyper parameters, hardware and implementation details). We also attempted an alternative generative fine-tuning approach with the same setup, where the model was tasked with generating the correct relationship for each pair. However, this approach led to worse results in our experiments.

ICL In our experiments with few-shot in-context learning (ICL) we use two types of prompts: one incorporating definitions and the other without them. We include one example per class (four total; DSPy further optimizes in-context examples). The prompt with definitions provides additional context for each example (see prompt in Figure 9), while the prompt without definitions relies solely on the examples (Figure 8). To account for variability in results due to few-shot selection, we report average results for 5 runs for each configuration, each time sampling a different set of few-shot examples and corresponding definitions (within each run, the same ICL examples are used for both with/without definition settings). We structure the output to extract classifications and reasoning automatically.[6]

Due to the combinatorial nature of our task with cross-document links across a potentially huge candidate space, we focus ICL evaluation on a strong yet comparatively cheap model, `GPT-4o-mini`, and applied only to a subset of the test set (§4.2). Our ICL experiments cost roughly $500, while a stronger `GPT-4o` model would have incurred a cost of over $16000 which is beyond our academic budget. Concurrent results (Lior et al., 2024) on SCICO with various open-source MISTRAL models in the few-shot setting led to very poor results in comparison to our results with `GPT-4o-mini` even without our definition-augmentation approach.

Evaluation Metrics For co-reference evaluation, we adopt the standard metrics MUC, B^3, CEAFe, and LEA, and report the widely used *CoNLL F1*, defined as the average of the MUC, B^3, and CEAFe F1 scores, as our primary measure. To evaluate the hierarchy, we use the following evaluation methods created in Cattan et al. (2021b):

- **Cluster-Level Hierarchy Score** checks whether each predicted parent–child link between mention clusters aligns with *at least one* mention pair in the gold hierarchy. This

[3] To check generalizablity of our results beyond Mistral, we also conduct a brief ablation experiment with another recent open source LLM (Abdin et al., 2024). We observe similar overall trends. See Tables 14–19 in Appendix.

[4] `github.com/openai/openai-python/blob/release-v0.28.0/chatml.md`

[5] `huggingface.co/transformers/v3.0.2/model_doc/auto.html#automodelforsequenceclassification`

[6] `dspy-docs.vercel.app/docs/deep-dive/signature/executing-signatures#inspecting-output`.

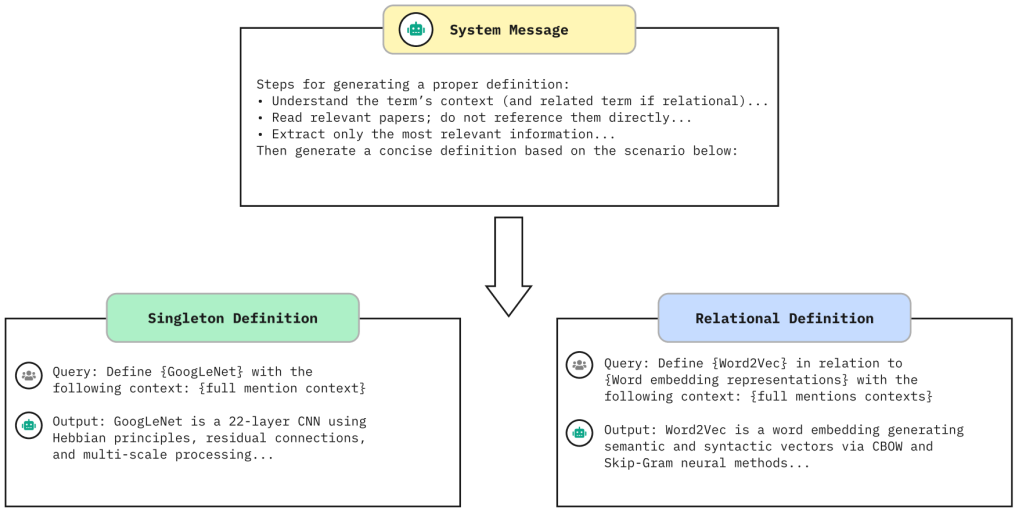

Figure 3: Logic and structure of the prompt used to generate singleton and relational definitions.

design avoids "double penalizing" minor coreference mismatches, For example, if the system predicts a parent–child relationship between clusters [Information Extraction] → [Definition Extraction, Pattern-based Extraction], and at least one pair of mentions (e.g., "Information Extraction" → "Definition Extraction") exists in the gold hierarchy, this counts as correct. We also consider a stricter version (50% Hierarchy Score) that requires *at least half* of the mentions in the predicted cluster pair to align with the gold parent–child link.

- **Path-Distance Score** compares distances between mentions (in terms of the number of edges separating them) in the predicted hierarchy vs. the gold. It provides partial credit when the distances are close, even if not perfectly matched. For instance, consider the mentions "Information Extraction" and "Pattern-based Definition Extraction": if they are separated by two edges in the gold hierarchy but only one edge in the predicted hierarchy, the metric gives partial credit proportional to the closeness of these path lengths.

We refer the reader to Cattan et al. (2021b) for full technical details on these hierarchy metrics.

4.2 Results

We next present the experimental results. We begin by establishing baselines and evaluating on the hardest subsets of the dataset, where the task is most challenging and the benefits of our definition based approach has the most impact. We then assess performance on the full dataset to measure overall improvements, and conclude with an ablation study that examines the specific contribution of different types of definitions.

Evaluation focus In addition to reporting overall performance (see Table 3), we focus on the "hard" data subsets provided by Cattan et al. (2021b), comprised of the 10% (*Hard-10*) or 20% (*Hard-20*) topics (collections of candidate mentions) with lowest scores as ranked according to CoNLL F1 by using Levenshtein distance in agglomerative clustering as a baseline. The subsets are particularly challenging, with a high amount of different surface forms and ambiguity, and are thus of most interest; e.g., the best model reported by Cattan et al. (2021b), achieved only 64.34 F1 for coreference CoNLL F1 on the *Hard-10* subset, while achieving 76.46 F1 on the full test set.

Baselines As a primary baseline, we use the model introduced in the original SciCo paper Cattan et al. (2021b), which defines the standard evaluation setting for this task. In addition, we include

Model	Coreference	Hierarchy		Path
	CoNLL F1	F1	F1-50%	Ratio
SciCo-Longformer(Cattan et al., 2021b)	64.34	45.46	28.07	34.26
GPT 4o-mini	51.28	40.22	16.95	31.04
GPT 4o-mini$_{Singleton}$	54.91	43.33	20.13	32.56
SciCo-Mistral-7B	69.68	50.25	37.36	49.55
SciCo-Radar$_{Singleton}$	70.30	54.70	43.40	50.00
SciCo-Radar$_{Relational}$	**71.90**	**56.98**	**45.93**	**53.85**

Table 1: Results on top 10% hardest topics (*Hard-10* subset). `GPT-4o-mini` uses few-shot examples (one per class). `GPT-4o-mini`$_{Singleton}$ adds our singleton definitions, which increase performance by several points. Fine-tuning Mistral-7B on SciCo leads to a large improvement. SciCo-Radar refers to Mistral-7B fine-tuned on SciCo along with our definition augmentation approach. SciCo-Radar leads to a further large boost in results, particularly in terms of hierarchy detection. SciCo-Radar with relational definitions achieves the best results.

Metric	SciCo-Longformer (Baseline)			SciCo-Longformer$_{\text{Singleton def}}$		
	All	***Hard-20***	***Hard-10***	**All**	***Hard-20***	***Hard-10***
CoNLL F1	76.46	68.18	64.34	78.18	71.10	65.28
Hierarchy F1	46.09	46.44	45.46	47.43	49.28	46.21
Hierarchy F1 50%	35.14	29.61	28.07	38.52	33.77	29.52
Path Ratio	44.91	36.21	34.26	48.81	44.07	35.36

Table 2: Results for SciCo-Longformer (Baseline) and SciCo-Longformer$_{\text{Singleton def}}$ across the full SCICO test set and two harder subsets (*Hard-20*, *Hard-10*). Metrics (§4.1) shown are CoNLL F1, cluster-level hierarchy F1 as well as 50% and path ratio.

our own fine-tuned models that do not use definitions, in order to isolate the contribution of definition based augmentation. We also report results for in-context learning (ICL) with GPT-4o-mini, to establish a baseline for definition augmentation without finetuning. While SEAM Lior et al. (2024) also reports results on SciCo and uses more recent models, its performance falls well below that of the original SciCo baseline.

Evaluation summary We report our results in Table 1 (*Hard-10*), Table 2 (Longformer all data) and Table 3 (all data) from the main paper, along with the full Tables [14, 15, 16, 17, 18, 19] in the Appendix (results in all subsets reflect similar model trends). ICL few-shot results are averaged across 5 runs as described earlier. We first observe that, not very surprisingly, fine-tuning of recent LLMs leads to large gains over the smaller Longformer model used originally in Cattan et al. (2021b). Fine-tuning also achieve considerable gains over concurrent results for SciCo CDCR in the few-shot setting with larger variants from the Mistral family (see Table 3 in Lior et al. (2024)), highlighting the value of fine-tuning in this task. More interestingly, our contextualized definitions led to strong improvements, for both LLM fine tuning and ICL. Our SciCo-Radar$_{\text{Relational}}$ model, which consists of fine-tuning Mistral-7B on SciCo augmented by relational definitions, achieved the highest results in every evaluation category, as exemplified in Table 1. Generally, relational definitions outperformed singleton definitions, by carrying more nuanced information about relationships that the model might miss with the more generic singleton definitions[7] (see Section 5 for more analysis). ICL with larger LLMs (OpenAI, 2024), including prompt optimization with DSPy, underperformed smaller fine-tuned models in our experiments, though future work may explore this further with more models and in-context examples.

As seen in Table 2, augmenting Longformer with our contextual definitions also resulted in substantial enhancements in both hierarchy and coreference scores, by up to 3.38 and 1.72 respectively.

[7] Paired bootstrap tests (Appendix A.5) confirm statistically reliable gains for hierarchy metrics across both hard subsets and for CoNLL F1 on *Hard-10*, with the remaining metrics reflecting positive upward trends.

Model	Coreference	Hierarchy		Path
	CoNLL F1	F1	F1-50%	Ratio
SciCo-Longformer (Baseline)	76.46	46.09	35.14	44.91
SciCo-Mistral-7B	81.82	56.84	49.70	57.21
SciCo-Radar$_{Singleton}$	81.95	57.33	50.58	57.27
SciCo-Radar$_{Relational\ def}$	**82.07**	**58.03**	**50.72**	**57.60**

Table 3: Results on the full SciCo test set. Fine-tuning Mistral-7B substantially outperforms the Longformer baseline, and adding singleton or relational definitions gives further improvements, with relational definitions achieving the best overall scores.

Even more substantial improvements were observed within the *Hard-20* subsets, particularly in hierarchy and hierarchy 50% scores, achieving improvements of 2.84 and 4.16. With Mistral-7B, incorporating definitions consistently improved all metrics. As seen In Table 1, incorporating definitions notably boosted Mistral's hierarchy F1 and F1-50% scores by up to 4.5 and 6.1 using singleton definitions, and by 6.78 and 8.63 with relational definitions.

We also report additional coreference and hierarchy precision-recall metrics in Tables 14, 15 and 17, 18 in the Appendix. For coreference, the F1 gains with relational definitions come from a substantial boost in recall with only a relatively small dip in precision. For hierarchy, we see a similar trend where both singleton and relational definitions improve recall while maintaining stable precision.

Finally, to assess the stability of ICL results with random few-shot examples or examples optimized by random search in DSPy, we report average results for five runs (Table 4).[8] We find that the effects are overall stable, with definitions helping boost results across both the few-shot and DSPy experiments;[9] interestingly, we do not observe gains from using DSPy in our experiments.

Overall vs. Hard Subsets Table 3 shows only modest overall improvements on the full SciCo test set, whereas Table 1 reveals much larger gains on the *Hard-10* and *Hard-20* subsets. This is expected, since many mentions can be resolved by simple lexical overlap or the model's prior knowledge, reducing the benefits of adding definitions. By contrast, the *Hard-10*/*Hard-20* cases feature ambiguous concepts and more variations of different surface forms, precisely where retrieval-based definitions can clarify those subtle distinctions or shared properties, thus offering substantial improvements. Additionally, these challenging cases make up only 10–20% of the data, so the strong gains do not dramatically affect the overall average. Nonetheless, linking these rare or nuanced mentions accurately is crucial for downstream scientific tasks, for example, knowledge graph construction or document retrieval, where mislinking can lead to incorrect representations of fine-grained concepts. A similar phenomenon was observed by Ravi et al. (2023) in event coreference, where the majority of mention pairs were straightforwardly matched by a model's surface or parametric knowledge. Substantial benefits from injecting external knowledge was reported primarily for the smaller more challenging fraction of mentions, in line with our findings on the *Hard-10*/*Hard-20* subsets. Notably, in table 2, we observe that the gains from adding definitions for the Longformer model are larger on the full and the *Hard-20* test sets compared to the *Hard-10* subset, which contrasts with the trends we observed for stronger models. We hypothesize that Longformer struggles more with the inherently ambiguous and challenging examples in the Hard 10 subset, limiting its ability to fully exploit the additional context provided by definitions. Consequently, while definition augmentation still improves performance, the relative gains are smaller on the hardest cases compared to simpler ones.

`GPT-4o` definitions As an additional baseline and ablation, we use the powerful and more costly `GPT-4o` to generate definitions by providing the model with the mention and its surrounding context but no additional retrieval, using only the model's

[8]The SciCo dataset contains 69,682 mention pairs in the *Hard-10* subset, along with extensive context and definitions per mention; running our ICL experiments on this subset cost $500. The full test set contains 564,352 mention pairs.

[9]We also experimented with several runs of fine-tuning, finding negligible variance across seeds.

Model	Coreference	Hierarchy		Path
	CoNLL F1	F1	F1-50%	Ratio
GPT 4O-MINI	51.28 ± 2.21	40.22 ± 1.70	16.95 ± 1.62	31.04 ± 2.23
GPT 4O-MINI$_{Singleton\ def}$	54.91 ± 0.61	43.33 ± 0.96	20.13 ± 0.61	32.56 ± 1.07
GPT 4O-MINI$_{DSPy}$	50.71 ± 0.66	42.23 ± 0.94	14.74 ± 1.36	27.41 ± 1.29
GPT 4O-MINI$_{DSPy\ Singleton\ def}$	54.90 ± 0.52	43.72 ± 1.16	19.69 ± 1.97	30.70 ± 0.56

Table 4: `GPT-4o-mini` ablation study. Co-reference and hierarchy scores with different configurations, average results and SE (five runs) on the *Hard-10* subset. We see the positive effects of definitions with and without DSPy optimization. We focus on `GPT-4o-mini` due to its relatively affordable cost ($500 to produce this table) and speed, important in our setting of inferring cross-document links in large-scale literature.

Context and Prediction	Singleton Def	Relational Def
1: ... WAN expansion regardless of network structures and to enhance the ***network cyber-security***...	**1**: Protecting networks and information from unauthorized access ... firewalls, antivirus software, encryption...	**1**: Protecting networks ... firewalls, encryption ... In contrast to cyber-safety awareness individual behavior and education ...
2: ... situation concerning ***cyber-safety awareness*** in schools and has adopted a short-term approach towards cyber-safety ... **Ground Truth:** $\not\Longleftrightarrow$ **No Def Model:** $\Longleftrightarrow$	**2**: Recognition of potential online threats ... educating individuals ... about protecting personal information ... **Correct Prediction:** $\not\Longleftrightarrow$	**2**: Recognition of potential online threats ... Unlike network cyber-security, which protects systems, cyber-safety awareness seeks to protect and inform users ... **Correct Prediction:** $\not\Longleftrightarrow$
1: ... they can use a technology for ***online marketing*** using e-commerce ...	**1:** Using digital channels to promote products or services... web advertising, e-commerce, targeted promotions based on customer data...	**1**: Using digital channels to promote products or services... closely related to targeted ads, which are a key component of online marketing strategies...
2: ... data gathering for purposes of personalization, ***targeted ads***... **Ground Truth:** $\Longrightarrow$ **No Def Model:** $\not\Longrightarrow$	**2:** Online advertisements tailored to individuals or groups... **Correct Prediction:** $\Longrightarrow$	**2**: A form of online marketing ... often employs sophisticated user profiling methods ... **Correct Prediction:** $\Longrightarrow$
1: ... These findings provide new possibilities of systematic F0 control for ***conversational speech synthesis*** ...	**1**: The generation of speech that mimics natural, everyday conversation, often utilizing machine learning techniques to control ...	**1**: the generation of speech that mimics natural, everyday conversation, often utilizing machine learning techniques ... It can be distinguished from traditional speech synthesis ...
2: ... SND and diarization can then be used ... achieving modern performance in ***conversational speech processing tasks*** ... **Ground Truth:** $\Longleftarrow$ **No Def Model:** $\not\Longleftarrow$	**2**: A series of computational procedures applied to spoken language during social interactions ... encompass Speech Separation, Speaker Diarization, Automatic Speech Recognition ... **Correct Prediction:** $\Longleftarrow$	**2**: Operations involved in analyzing and understanding spoken language, including Speech Separation, Diarization, Automatic Speech Recognition ... a precursor to applications such as conversational speech synthesis ... **Correct Prediction:** $\Longleftarrow$

Table 5: Examples where the model without definitions misidentifies the true relation. The singleton and relational definitions add important context enabling the model to accurately determine the correct relation. $\not\Longleftrightarrow$ - no relationship, $\Longleftrightarrow$ - co-reference, $\Longrightarrow$ - 1 $\rightarrow$ 2 parent-child relationship and $\Longleftarrow$ - 2 $\rightarrow$ 1 parent-child relation.

vast knowledge embedded in its parameter space (Ye et al., 2023; Bubeck et al., 2023). We find that after training MISTRAL-7B on these definitions, a big gap in performance emerged in favor of retrieval-based definitions in terms of hierarchy predictions (Table 12), along with a smaller but noticeable advantage in coreference scores, too (Table 19). Upon reviewing the generated definitions, we observed that `GPT-4o` often generated definitions that were more generic and tended to be more fixated on irrelevant context details, while retrieval-based definitions naturally assisted the definition generation to be both more detailed and specific while zooming out and abstracting-away irrelevant context. For example, in Table 11, the GPT-4o definition of "representation learning" is too focused on a specific application involving embodied agents and cognitive phenomena, while the singleton defi-

nition abstracts away from these details and instead highlights the broader machine learning context, the kind of generalization needed to resolve this hierarchical relationship (Table 11 in the Appendix; §5 for more analysis).

5 Qualitative Analysis

In this section, we explore the impact of definitions on model performance by presenting qualitative examples and analyzing their effects. We also perform error analysis, revealing different sources of error which may inform future improvements. Our focus is on understanding how different types of definitions can influence the output of the model, specifically in cases where the base model struggles. We use our best-performing fine-tuned SciCo-Radar model for these analyses.

5.1 When Definitions Generally Help

We begin with providing qualitative evidence to support the fundamental hypothesis in this paper, showing that incorporating dynamic definitions substantially improves model performance over the non-definition approach. We discuss two main ways in which we qualitatively observe definitions broadly help, before diving deeper into specific types of definitions and their various trade-offs.

Lexical ambiguity, different surface forms, and fine-grained hierarchy Lexical ambiguity, different surface forms, and fine-grained concept hierarchies are at the heart of the challenge for the H-CDCR task in complex technical domains. Our findings confirm our hypothesis that definitions provide useful context that helps models understand lexical and semantic nuances for resolving the true relationships of mentions, or lack there of, by detecting shared meanings and aspects that are not obvious from the text alone. For instance (see also Table 5, row 2), the definition of *online marketing operation* describes it as encompassing various digital strategies, including *targeted ads*, clarifying that the former subsumes the latter. Similarly, defining *targeted ads* as a specific type of online advertisement tailored to specific demographics makes the hierarchical relationship clear. As another example, demonstrating lexical ambiguity (see also Table 5, row 1), the definitions for *network cyber-security* and *cyber-safety awareness* make it clear that one mention focuses on technical measures to secure network infrastructure, while the other emphasizes educating individuals about online safety practices. These definitions help the model understand that, despite surface similarities, the mentions do not belong to the same conceptual cluster or share a hierarchical link.

Context lacking information & context interference We observe that, unsurprisingly, vague or ambiguous contexts can make it hard to clarify whether two different mentions are co-referential, hierarchical, or unrelated. Additionally, a mention might be a minor part of a broader discussion where the surrounding text focuses on other concepts, potentially injecting irrelevant context into the model's decision on the cross-document relation. We observe that definitions generally help mitigate these challenges by providing detailed explanations that highlight similarities or shared concepts that extend beyond the surface-level context. For example, *conversational speech synthesis* and *conversational speech processing tasks* (see Table 5, row 3) where the first mention is part of a broader discussion on F0 control and speaking attitudes, and the second mention involves a focus on privacy-preserving features in speech processing. We note that this type of context can affect the quality of definitions as well; however, the retrieved global literature information may help mitigate problems in the local context, such that the definition generation model can often successfully generate a useful enough definition to assist the classification model.

5.2 Impact of Relational Definitions

Following our exploration of how definitions generally improve model performance over non-definition approaches, we now turn to the impact of relational definitions compared to singleton definitions. We examine key cases where relational definitions outperform singleton definitions, particularly in capturing complex hierarchical or co-referential relationships, and discuss some trade-offs involved in both approaches.

Although singleton definitions often proved to be more effective than no definitions, we found they can often fall short (see also Table 10 in the Appendix). In some cases, definitions share overlapping concepts without adequately distinguishing the unique roles of each mention. Definitions may fail to capture the shared relationship between two mentions, especially when they focus on different aspects of the same underlying concept. This can lead the model to accentuate differences rather than recognize the commonality between the mentions.

Context and Prediction	Singleton Def	Relational Def
1: ... Then the RNN language models predict probability distribution by the following equation ... is a ***word embedding matrix network*** ...	**1**: A matrix in which each row corresponds to a word in a given vocabulary ... used in natural language processing models to project discrete word tokens into a continuous vector space ...	**1**: A matrix where each row corresponds to a word in a given vocabulary ... This matrix is used to project discrete word tokens into distributed representations of words, which are dense vectors ...
2: ... RNN can utilize ***distributed representations of words*** by first converting ... **Ground Truth:** $\iff$ **No Def Prediction:** $\not\iff$	**2**: The way words are represented as vectors in computational models ... These vectors capture semantic and syntactic properties of words ... **Wrong Prediction:** $\not\iff$	**2**: The way words are represented as vectors ... In the context of RNNs, these distributed representations are utilized to convert tokens into vectors, forming a matrix that captures ... **Correct Prediction:** $\iff$
1: ... the ***simulated-annealing-based approach*** is proposed to construct a better timing-constrained ...	**1**: A computational technique used to optimize a problem by iteratively improving an initial solution through a series of modifications ...	**1**: A type of meta-heuristic approach that utilizes the principles of annealing used to optimize a problem by iteratively improving an initial solution through ...
2: ... results of numerous implementations of well-known ***meta-heuristic approaches*** ... **Ground Truth:** $\Longleftarrow$ **No Def Prediction:** $\not\iff$	**2**: A category of optimization algorithms that include Evolutionary Algorithms, Genetic Algorithms, and Swarm Intelligence techniques among others ... **Wrong Prediction:** $\not\iff$	**2**: A class of high-level, general-purpose optimization algorithms that ... These approaches include Evolutionary Algorithms, Simulated Annealing, Swarm Intelligence, and others ... **Correct Prediction:** $\Longleftarrow$
1: ... ***neural network classifier*** employs relatively few parameters compared to other deep learning ...	**1**: A type of machine learning algorithm that uses artificial neural networks to categorize data into different classes ...	**1**: A type of machine learning algorithm that uses artificial neural networks to classify ... Although both neural network classifiers and network classification rely on machine learning techniques, they serve different purposes ...
2: ... real world applications such as ***network classification***, and anomaly detection ... **Ground Truth:** $\not\iff$ **No Def Prediction:** $\iff$	**2**: A process that involves categorizing networks or graphs based on their structure, often utilizing machine learning algorithms ... **Wrong Prediction:** $\iff$	**2**: The process of categorizing networks or their components based on ... While neural networks can be used for classification tasks, network classification addresses a different problem space, focusing on ... **Correct Prediction:** $\not\iff$

Table 6: Analysis of relational definitions. Relational definitions can help clarify mention relationships in cases where singleton definitions are unable. $\not\iff$ - no relationship, $\iff$ - co-reference, $\Longrightarrow$ - 1 $\rightarrow$ 2 parent-child relationship and $\Longleftarrow$ - 2 $\rightarrow$ 1 parent-child relationship.

For instance (see also Table 6, row 1), the relational definitions more explicitly detail the relationship between *word embedding matrix* and *distributed representations of words*. Similarly, in the case of *simulated-annealing-based approach* and *meta-heuristic approaches* (Table 6, row 2), the relational definitions highlight Simulated Annealing as an approach in the broader umbrella of meta-heuristic approaches. In addition, relational definitions help clarify when there is no connection between two mentions in cases where singleton definitions do not provide sufficient discriminative information, e.g., in the case of *neural network classifier* and *network classification* (Table 6, row 3).

Relational definitions: Error cases Relational definitions may sometimes emphasize similarities or overlaps irrelevant to the actual relation between mentions. For example, the relational definitions for *Transformer model* and *Equivariant Transformers (ETs)* (see Table 10 in the Appendix) incorrectly identifies a hierarchical relationship, while the mentions refer to different architecture families.

5.3 Retrieval definitions vs. GPT4o definitions

Finally, we qualitatively examine how retrieval definitions differ to `GPT4o`-based definitions. We find that generally, retrieval generates more detailed and comprehensive definitions covering more diverse concept aspects, that allow detecting when one concept is more general than the other; conversely, `GPT4o` tends to generate more vanilla, less rich definitions (e.g., see Table 11, row 1 in the Appendix). Additionally `GPT4o` definitions tend more often to fixate and over-rely on context, making the definitions focused on too narrow aspects of

the context which do not allow to zoom out and detect abstractions and subsumptions (see example in Table 11, row 2 in the Appendix). These findings help explain why retrieval definitions appear to add several points in Hierarchy F1 vs. GPT4o definitions (Table 19 in the Appendix).

6 Related Work

Coreference Resolution Our task is closely related to the cross-document co-reference resolution task (Yang et al., 2015; Cattan et al., 2021a). In scientific domains, CDCR presents unique challenges due to the abstract nature of technical concepts and the need to differentiate between closely related terms at varying levels of specificity (Cattan et al., 2021b). Ravenscroft et al. (2021) further extend this line of work by introducing a cross-document, cross-domain coreference resolution task, which addresses coreference resolution between scientific articles and related news reports. Their work emphasizes the linguistic complexity introduced by domain-specific variations, highlighting significant limitations of the existing CDCR methods. Finin et al. (2009) explored leveraging external structured knowledge through Wikitology, a knowledge base built from Wikipedia, DBpedia, and Freebase, demonstrating that incorporating structured semantic information substantially improves cross-document entity coreference performance. Recent advancements in coreference resolution have been reached with LLMs. As demonstrated by Le and Ritter (2023), Gan et al. (2024), LLMs like GPT-4, InstructGPT and Llama variants, when employed with prompt-based techniques, have shown strong performance in co-reference resolution tasks in the more popular within-document setting. Yang et al. (2022) analyzed GPT models ability to resolve coreference using QA-based prompting. They found that while those models can generate reasonable answers, their coreference resolution performance is inconsistent and highly sensitive to prompt phrasing, which highlights some limitations of LLMs in those domains. In our approach, we explore LLMs in the challenging cross-document setting, in the scientific domain, and include also cross-document *hierarchy* detection. We use both in-context learning and fine-tuning with dynamically generated (relational) definitions to improve CDCR performance.

Definitions in downstream NLP tasks Earlier works leveraged external knowledge to improve text understanding. Al-Harbi et al. (2017) addressed lexical ambiguity in QA by incorporating WordNet-based definitions and domain-specific ontologies to refine word meanings. By integrating context knowledge, their approach improved word sense disambiguation, ensuring more precise retrieval in question answering. Yin and Roth (2018) similarly used dictionary definitions to improve hypernymy detection, addressing the limitations of distributional embeddings, which infer meaning based on co-occurrence patterns in large corpora. Although the usage of static definitions can be effective, they are often insufficient in our cross-document coreference and hierarchy task. Terms may have context dependent meanings that can vary across different papers, research fields, or even within the same document. In our approach, instead of relying on predefined dictionary definitions, we incorporate contextual definitions, generated with retrieved external knowledge in order to capture the nuanced semantic distinctions or similarities.

More recent works includes Munnangi et al. (2024) who enhanced biomedical NER capabilities of LLMs in the few-shot setting by incorporating static definitions of canonicalized entities primarily from the UMLS knowledge base Bodenreider (2004). In our method, we generate *dynamic* singleton and relational definitions specifically tailored to the unique context of the mentions in scientific texts, by retrieving relevant literature. Our definitions are designed to enhance understanding of cross-document relationships. Xin et al. (2024) recently proposed a method that uses LLMs as context augmenters to enhance entity linking (EL) with mention descriptions which are integrated into traditional EL models to improve linkage with canonical entities in general-domain KBs. Silva et al. (2020) introduce XTE (Explainable Text Entailment), an approach that explicitly leverages structured lexical definitions to improve semantic reasoning in RTE (Recognizing Textual Entailment). While their approach addresses entailment rather than cross-document coreference, it similarly demonstrates the benefit of augmenting neural models with structured concept definitions to enhance performance in relational reasoning tasks.

Relational Descriptions generation There has been scarce work we are aware of on generating relational concept descriptions. Murthy et al. (2022) introduced a system named ACCoRD for generating diverse descriptions of how scientific concepts

are related in order to provide varied perspectives on scientific concepts. The focus of ACCoRD was to build a small-scale user-facing system demonstration; unlike ACCoRD, our work uses relational definitions to directly improve NLP tasks such as CDCR and hierarchy inference. Brahman et al. (2021) generate natural language definitions in the form of rationales to explain defeasible inference, a type of nonmonotonic reasoning where new information can strengthen or weaken prior conclusions. These definitions include both local rationales for individual concepts and relational explanations of how updates alter entailment. They are generated by prompting a pretrained language model with a premise, hypothesis, and update, and trained via distant supervision without retrieval. While focused on entailment rather than coreference, their work similarly highlights the value of textual definitions for modeling conceptual relationships.

7 Conclusion

We presented SciCo-Radar, a novel approach based on dynamic definition generation and relational reasoning for enhancing the ability of Large Language Models (LLMs) to infer cross-document coreference and hierarchy relationships in scientific papers. Our method employs context-dependent and relational definitions generated through literature retrieval, along with a re-ranking based scheme to avoid combinatorial explosion across all potential mention pairs. SciCo-Radar has shown consistent improvements across both fine-tuning and in-context learning settings, with particularly strong gains on challenging subsets, and moderate improvements on the overall dataset. We showed that with our finetuning and dynamic definition integration approach, smaller open models can outperform much larger models. We also conducted in-depth analyses into the quality of generated definitions and failure modes which could be valuable to explore in future work. In addition, paired bootstrap tests provide further evidence that the improvements, particularly on the hardest subsets, are consistent across resampling.

Future research may focus on improving the generation and utilization of dynamic concept definitions, as well as improving the relational reasoning of LLMs when handling nuanced, fine-grained scientific concepts. Generating relational definitions can be computationally intensive at scale; we thus do not apply relational definitions across all possible mention pairs. Exploring ways to further optimize the efficiency of our re-ranking approach, or other approaches for explicit relational reasoning that avoid combinatorial explosion with infeasible costs, may lead to more scalable solutions for large-scale scientific applications. Another interesting direction is to investigate how the addition of structured metadata such as publication venue, field of study or year can help. For example, the term *transformer* may be interpreted as a neural model in a machine learning conference but as power equipment in an electrical engineering journal. Metadata signals could enhance both the generation of contextual definitions and the prediction of cross-document relations, particularly in ambiguous or long-tail cases. In addition, while orthogonal to our contribution, we note that the SciCo dataset (which is over 3 times larger than the common CDCR benchmark ECB+ (Cybulska and Vossen, 2014)) could be expanded by collecting more annotations of coreference clusters and hierarchies across many papers from different domains, to better capture the breadth and scale of literature. While our approach improves performance, particularly on harder examples, it comes with additional computational cost due to relational definition generation and fine-tuning. This cost also carries environmental impact especially in large-scale settings. Developing more efficient alternatives for relational reasoning and knowledge integration is not only important for scalability but also for reducing environmental footprint. We leave exploration of lower-cost yet effective variants as a promising direction for future work. Finally, it would be exciting to consider further applications of our method, in more general cross-document tasks.

Acknowledgments

We thank the reviewers and the action editor for their valuable feedback and suggestions. This research was supported by the Azrieli Foundation.

References

Marah I Abdin, Sam Ade Jacobs, and Ammar Ahmad et al Awan. 2024. Phi-3 technical report: A highly capable language model locally on your phone. Technical Report MSR-TR-2024-12, Microsoft.

Omar Al-Harbi, Shaidah S. Jusoh, and Norita Md

Norwawi. 2017. Lexical disambiguation in natural language questions (nlqs). *ArXiv*, abs/1709.09250.

Patricia A Alexander. 2016. Relational thinking and relational reasoning: harnessing the power of patterning. *NPJ science of learning*, 1(1):1–7.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document Transformer. *arXiv:2004.05150*.

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

Faeze Brahman, Vered Shwartz, Rachel Rudinger, and Yejin Choi. 2021. Learning to rationalize for nonmonotonic reasoning with distant supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12592–12601.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, John A. Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuan-Fang Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. *ArXiv*, abs/2303.12712.

Arie Cattan, Alon Eirew, Gabriel Stanovsky, Mandar Joshi, and Ido Dagan. 2021a. Cross-document coreference resolution over predicted mentions. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5100–5107.

Arie Cattan, Sophie Johnson, Daniel S Weld, Ido Dagan, Iz Beltagy, Doug Downey, and Tom Hope. 2021b. SciCo: Hierarchical cross-document coreference for scientific concepts. In *3rd Conference on Automated Knowledge Base Construction*.

Agata Cybulska and Piek Vossen. 2014. Using a sledgehammer to crack a nut? lexical diversity and event coreference resolution. In *LREC*, pages 4545–4552.

Timothy W. Finin, Zareen Syed, James Mayfield, Paul McNamee, and Christine D. Piatko. 2009. Using wikitology for cross-document entity coreference resolution. In *AAAI Spring Symposium: Learning by Reading and Learning to Read*.

Yujian Gan, Massimo Poesio, and Juntao Yu. 2024. Assessing the capabilities of large language models in coreference: An evaluation. In *International Conference on Language Resources and Evaluation*.

Jiafeng Guo, Yixing Fan, Liang Pang, Liu Yang, Qingyao Ai, Hamed Zamani, Chen Wu, W. Bruce Croft, and Xueqi Cheng. 2020. A deep look into neural ranking models for information retrieval. *Information Processing & Management*, 57(6):102067.

Yue Guo, Joseph Chee Chang, Maria Antoniak, Erin Bransom, Trevor Cohen, Lucy Lu Wang, and Tal August. 2024. Personalized jargon identification for enhanced interdisciplinary communication. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4535–4550.

Tom Hope, Aida Amini, David Wadden, Madeleine van Zuylen, Sravanthi Parasa, Eric Horvitz, Daniel S Weld, Roy Schwartz, and Hannaneh Hajishirzi. 2021. Extracting a knowledge base of mechanisms from COVID-19 papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4489–4503.

Tom Hope, Joel Chan, Aniket Kittur, and Dafna Shahaf. 2017. Accelerating innovation through analogy mining. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 235–243.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego

de Las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, L'elio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *ArXiv*, abs/2401.04088.

Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *ArXiv*, abs/2310.06825.

SeongKu Kang, Shivam Agarwal, Bowen Jin, Dongha Lee, Hwanjo Yu, and Jiawei Han. 2024. Improving retrieval in theme-specific applications using a corpus topical taxonomy. In *Proceedings of the ACM on Web Conference 2024*, pages 1497–1508.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2024. DSPy: Compiling declarative language model calls into state-of-the-art pipelines. In *International Conference on Learning Representations*.

Nghia T. Le and Alan Ritter. 2023. Are large language models robust coreference resolvers?

Sean Lee, Aamir Shakir, Darius Koenig, and Julius Lipp. 2024. Open source strikes bread - new fluffy embeddings model.

Xianming Li and Jing Li. 2024. AnglE-optimized text embeddings. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1825–1839.

Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2020. Pretrained Transformers for text ranking: BERT and beyond.

Gili Lior, Avi Caciularu, Arie Cattan, Shahar Levy, Ori Shapira, and Gabriel Stanovsky. 2024. SEAM: A stochastic benchmark for multi-document tasks. *arXiv preprint arXiv:2406.16086*.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232.

Ishani Mondal, Yufang Hou, and Charles Jochim. 2021. End-to-end construction of NLP knowledge graph. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1885–1895.

Monica Munnangi, Sergey Feldman, Byron C. Wallace, Silvio Amir, Tom Hope, and Aakanksha Naik. 2024. On-the-fly definition augmentation of LLMs for biomedical NER. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3878–3893.

Sonia K. Murthy, Kyle Lo, Daniel King, Chandra Bhagavatula, Bailey Kuehl, Sophie Johnson, Jon Borchardt, Daniel S. Weld, Tom Hope, and Doug Downey. 2022. ACCoRD: A multi-document approach to generating diverse descriptions of scientific concepts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 200–213.

Sheshera Mysore, Mahmood Jasim, Andrew McCallum, and Hamed Zamani. 2023. Editable user profiles for controllable text recommendations. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 993–1003.

OpenAI. 2023. GPT-4 technical report.

OpenAI. 2024. GPT-4o system card.

Jason Portenoy, Marissa Radensky, Jevin D West, Eric Horvitz, Daniel S Weld, and Tom Hope. 2022. Bursting scientific filter bubbles: Boosting innovation via novel author discovery. In

Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, pages 1–13.

James Ravenscroft, Arie Cattan, Amanda Clare, Ido Dagan, and Maria Liakata. 2021. Cd^2cr: Co-reference resolution across documents and domains. In *Conference of the European Chapter of the Association for Computational Linguistics*.

Sahithya Ravi, Chris Tanner, Raymond Ng, and Vered Shwartz. 2023. What happens before and after: Multi-event commonsense in event coreference resolution. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1708–1724, Dubrovnik, Croatia. Association for Computational Linguistics.

Tara Safavi, Doug Downey, and Tom Hope. 2022. CascadER: Cross-modal cascading for knowledge graph link prediction. *AKBC*.

Vivian Dos Santos Silva, André Freitas, and Siegfried Handschuh. 2020. XTE: Explainable text entailment. *ArXiv*, abs/2009.12431.

Arvind Srinivasan and Joel Chan. 2024. Improving selection of analogical inspirations through chunking and recombination. In *Proceedings of the 16th Conference on Creativity & Cognition*, pages 374–397.

Jianyou Wang, Kaicheng Wang, Xiaoyue Wang, Prudhviraj Naidu, Leon Bergen, and Ramamohan Paturi. 2023. Doris-mae: scientific document retrieval using multi-level aspect-based queries. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 38404–38419.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. 2024. Scimon: Scientific inspiration machines optimized for novelty. In *ACL Anthology: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 279–299*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Junde Wu, Jiayuan Zhu, and Yunli Qi. 2024. Medical graph rag: Towards safe medical large language model via graph retrieval-augmented generation. *arXiv preprint arXiv:2408.04187*.

Amy Xin, Yunjia Qi, Zijun Yao, Fangwei Zhu, Kaisheng Zeng, Xu Bin, Lei Hou, and Juanzi Li. 2024. LLMAEL: Large language models are good context augmenters for entity linking. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 2634–2644.

Bishan Yang, Claire Cardie, and P. Frazier. 2015. A hierarchical distance-dependent bayesian model for event coreference resolution. *Transactions of the Association for Computational Linguistics*, 3:517–528.

Xiaohan Yang, Eduardo Peynetti, Vasco Meerman, and Chris Tanner. 2022. What GPT knows about who is who. In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, pages 75–81, Dublin, Ireland. Association for Computational Linguistics.

Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, Jie Zhou, Siming Chen, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. A comprehensive capability analysis of GPT-3 and GPT-3.5 series models. *ArXiv*, abs/2303.10420.

Wenpeng Yin and Dan Roth. 2018. Term definitions help hypernymy detection. In *International Workshop on Semantic Evaluation*.

A Appendix

A.1 Definition Generation technical details

For each definition generation iteration we retrieve the relevant papers, using `mxbai-embed-large-v1` (Lee et al., 2024; Li and Li, 2024) as an embedding model with state-of-the-art performance. We use the `Chroma DB`[10] open-source vector database for fast retrieval using approximate nearest neighbor search. We rerank retrieved results with a cross-encoder (`mxbai-rerank-large-v1`[11]. Results are then finally fed into the definition generation model, for which we use `Mixtral 8x7B`, a mixture of experts LLM (Jiang et al., 2024). We employ this method to produce singleton definitions for each individual mention and relational definitions (see Figures 4-5 for specific prompts and A.2 for hardware details and limitations).

A.2 Additional Hardware and LoRA Hyper-Parameter Details

Hardware Setup and Runtime We employed 8 NVIDIA RTX A6000 48GB GPUs to generate both singleton and relational definitions through our RAG pipeline. Generating relational definitions only for the top-ranked mention pairs (§3) took roughly two weeks, while to generate relational definitions for the entire test set, we would need roughly 40 days. For MISTRAL-7B, Fine-tuning was conducted on 3 NVIDIA RTX A6000 48GB GPUs over 2 epochs. The training was carried out with a batch size of 2, 4 gradient accumulation steps and a maximum sequence length of 1664 tokens (these choices were made due to our academic resources). We used a default learning rate of 2e-5 with a weight decay of 0.001.

LoRA Configuration and Hyper-Parameters We use Low-Rank Adaptation of LLM (LoRA), a parameter efficient fine-tuning method (Hu et al., 2022) with the following parameters selected according to our resource constraints: lora_alpha: 16, lora_dropout: 0.1, lora_r: 8, SEQ_CLS as the task type and modules_to_save=["score"] in order to update and save the weights of the classification head we added to the model.

A.3 ICL

DSpy We select 200 pairs for training and 60 pairs for development. To ensure balanced evaluation, we drew a random subset of examples, ensuring an equal number of labeled examples across the different relations. Within our DSPy program we incorporate chain-of-thought (CoT) reasoning and BootstrapFewShotWithRandomSearch optimizer.[12] In CoT, a model is instructed to reason step by step to get to its answer, which can improve performance (Wei et al., 2022). BootstrapFewShotWithRandomSearch is an enhanced version of the BootstrapFewShot teleprompter,[13] which uses a teacher (GPT4 OpenAI (2023)) - student (GPT4o-mini) approach to generate few-shot demonstrations by sampling from the training data. During the training phase, the teacher model simulates potential outputs and rationales for labeled instances. The teacher model iteratively generates few-shot examples by attempting to predict the correct outputs based on the training data, learning from its own successes and refining also the generated rationale over time. This process is repeated multiple times, with a random search conducted over the generated demonstrations. The chosen program is the one that achieves the highest performance on the dev set. During each iteration on the test set, the program receives few-shot examples along with their bootstrapped rationales, assisting it in forming better rationales, thereby resulting in improved predictions.

A.4 Additional Ablation study

Relational definitions Ablation To further examine the effect of relational definitions, we conducted an ablation study restricted to the *Hard-10* subset (Table 9). Generating relational definitions for all mention pairs across the full SCICO test set is computationally infeasible, as it would require producing hundreds of thousands of pairwise definitions, therefore we focused on the smallest subset. Interestingly, while coreference scores improved only modestly, the hierarchy metrics saw larger gains, strengthening our hypothesis that relational definitions particularly help the model capture hierarchical structure. To balance feasibility and performance, we set the candidate selection threshold at 25%, which we found to offer a practical trade-off between computational cost and accuracy.

[10] https://github.com/chroma-core/chroma

[11] mixedbread.ai/docs/reranking/models

[12] https://dspy-docs.vercel.app/docs/building-blocks/optimizers

[13] https://dspy-docs.vercel.app/docs/deep-dive/teleprompter/bootstrap-fewshot

A.5 Statistical Significance Analysis

Statistical Significance We evaluate whether improvements of our SCICO-RADAR$_{Relational}$ model over the strongest baseline (SCICO-MISTRAL-7B) are statistically reliable using one-tailed paired bootstrap tests over topics.

Method For each subset (*Hard-10*, *Hard-20*) and metric, we compute the observed difference Δ and generate $20{,}000$ bootstrap samples by resampling topics with replacement. For each sample, we recompute the metric and record the resampled difference Δ_b. The one-tailed p-value is the proportion of samples in which the relational model does not outperform the baseline.

Results On the *Hard-10* subset, improvements across all hierarchy metrics and CoNLL F1 are statistically significant ($p < 0.05$). On the *Hard-20* subset, the relational model yields statistically significant improvements in both Hierarchy F1 ($p = 0.050$) and Hierarchy F1-50% ($p = 0.048$), while gains in Path Ratio and CoNLL F1 remain positive but are not statistically significant. These patterns are consistent with our hypothesis that relational definitions most strongly benefit the most challenging cases.

Subset	**Metric**	Δ	p
Hard-10	Path Ratio	+4.30	0.047
Hard-10	Hier F1-50%	+8.57	0.044
Hard-10	Hierarchy F1	+6.72	0.047
Hard-20	Hierarchy F1	+4.77	0.050
Hard-20	Hier F1-50%	+4.99	0.048
Hard-20	Path Ratio	+0.48	0.113

Table 7: Paired bootstrap significance results (20k samples). Δ denotes improvement of RELATIONAL over the MISTRAL-7B baseline on hierarchy metrics.

Subset	**Model**	F1	Δ	p
Hard-10	Relational	71.89	+1.91	0.040
Hard-20	Relational	76.23	+0.64	0.086

Table 8: Paired bootstrap significance results (20k samples) for CoNLL F1. Positive improvements on *Hard-20* are not statistically significant.

Model	Coreference	Hierarchy		Path
	CoNLL F1	F1	F1-50%	Ratio
SciCo-Radar$_{\textit{Relational def 25\%}}$	71.90	56.98	45.93	53.85
SciCo-Radar$_{\textit{Relational def 60\%}}$	71.96	57.71	46.25	54.31
SciCo-Radar$_{\textit{Relational def 100\%}}$	**72.33**	**58.15**	**46.82**	**55.17**

Table 9: Ablation for relational definitions on *Hard-10* subset.

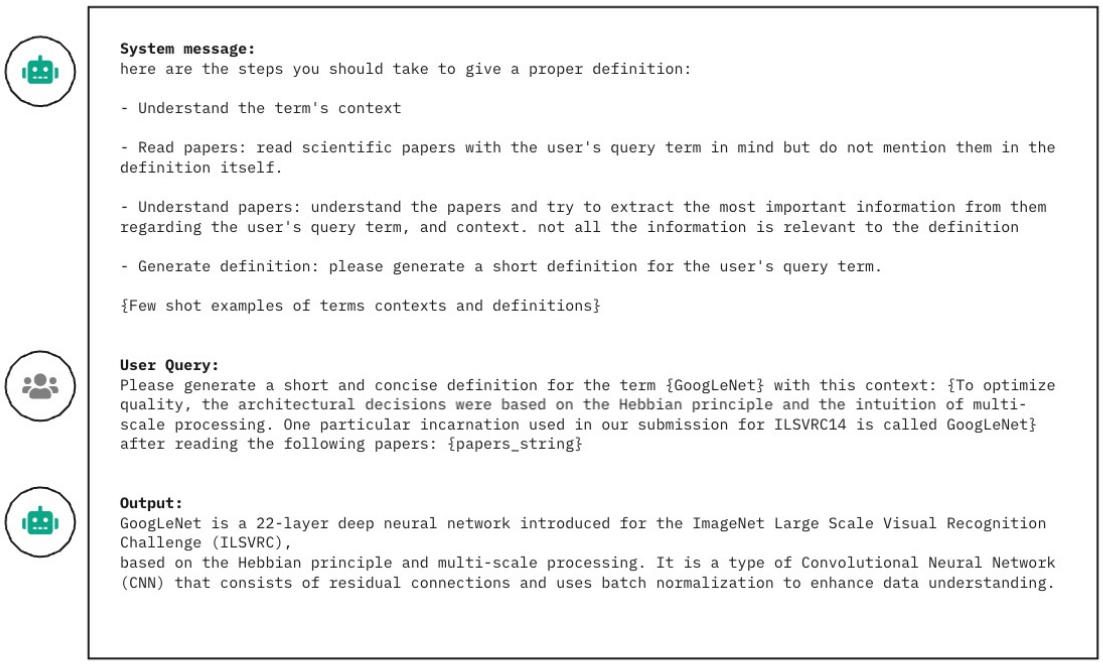

Figure 4: Singleton definition generation prompt.

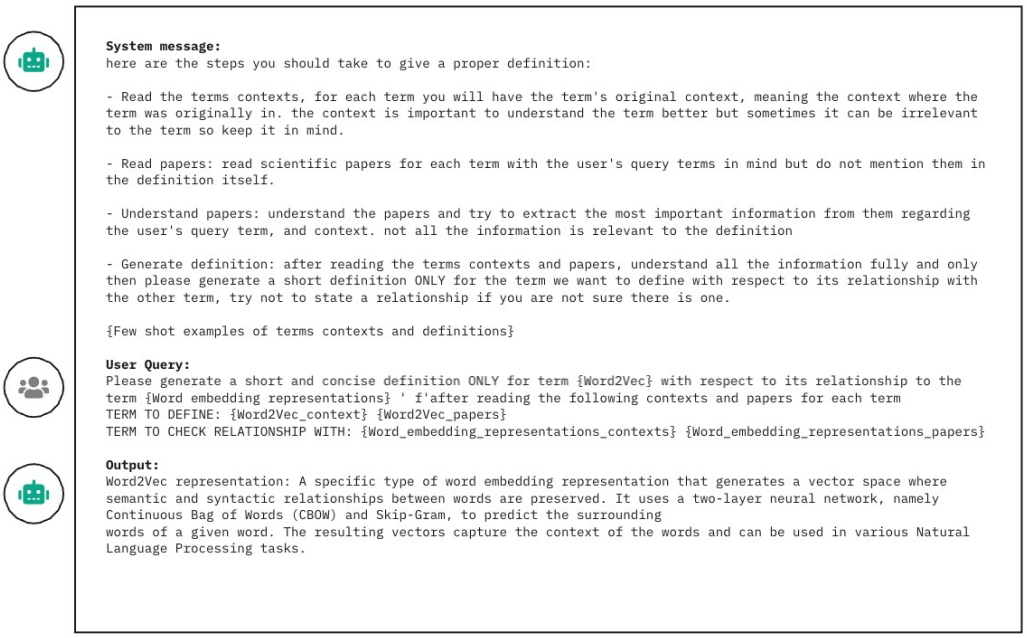

Figure 5: Relational definition generation prompt.

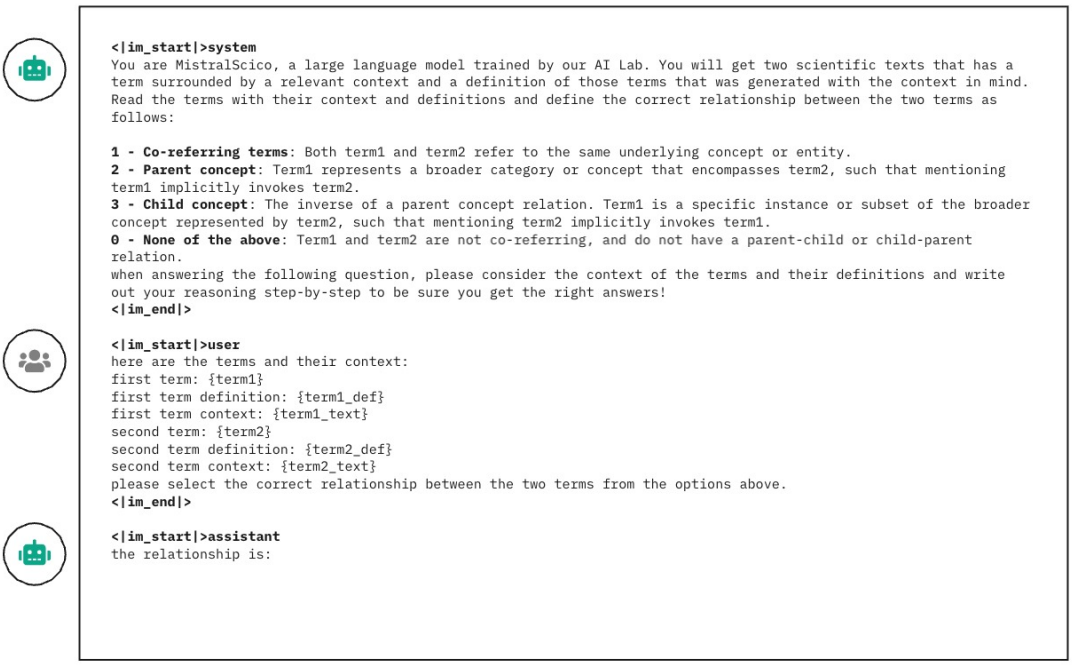

Figure 6: Chat ML finetuning prompt format.

Context and Prediction	Definition:
1: ... in various tasks such as ***semantic parser***, question-answering ...	**1:** A Natural Language Processing tool that maps natural language into ... enabling the interpretation and manipulation of meaning ...
2: ... modeled ***syntactic parsing*** and SRL jointly... **Ground Truth:** $\not\Longleftrightarrow$ **No Def Model:** $\not\Longleftrightarrow$	**2:** The process of analyzing a sentence to determine its grammatical structure ... provide structural information about sentences, which can aid in understanding meaning ... **Singleton Prediction:** $\Longleftrightarrow$
1: ... Three adapted deep neural networks : ***ResNet***, LSTM ...	**1:** A type of deep residual network, primarily used in CNNs, that consists of several stacked residual units. Each unit is a collection of convolutional layers, accompanied by a shortcut connection ...
2: ... which are extracted by a ***deep pre-trained ResNet*** ... **Ground Truth:** $\Longleftrightarrow$ **No Def Model:** $\Longleftrightarrow$	**2:** A Residual Network that has been previously trained on a large dataset, allowing it to extract meaningful features from images. It is often used as a component in various computer vision tasks ... **Singleton Prediction:** $\not\Longleftrightarrow$
1: ... extracted from DCT coefficients to achieve ***unsupervised segmentation of image objects*** ...	**1:** The process of dividing an image into distinct sections or clusters based on certain visual characteristics, without requiring any prior labeled data or manual intervention...
2: ... we introduce group saliency to achieve superior ***unsupervised salient object segmentation*** ... **Ground Truth:** $\Longrightarrow$ **No Def Model:** $\Longrightarrow$	**2:** The process of identifying and separating the most prominent objects in an image or image collection... utilizes intrinsic features and relationships among images to locate and segment salient objects that maximize inter-image similarities and intra-image distinctness ... **Singleton Prediction:** $\not\Longleftrightarrow$
1: ... the inference speed is twice that of sparsemax in ***Transformer model*** ...	**1:** A type of artificial neural network that uses an encoder-decoder structure ... the concept of Equivariant Transformers extends the Transformer model ...
2: ... We propose ***Equivariant Transformers (ETs)***, a family of ... **Ground Truth:** $\not\Longleftrightarrow$ **No Def Model:** $\Longrightarrow$	**2:** A type of Transformer model designed to incorporate prior knowledge on the transformation invariances of a domain ... **Relational Prediction:** $\Longrightarrow$

Table 10: Examples of Definitions with wrong predictions. We show cases where definitions can negatively impact the model predictions. $\not\Longleftrightarrow$ - no relationship, $\Longleftrightarrow$ - co-reference, $\Longrightarrow$ - 1 $\rightarrow$ 2 parent-child relationship and $\Longleftarrow$ - 2 $\rightarrow$ 1 parent-child relationship.

Context and Prediction	Singleton Def	GPT4o Def
1: ... embodied gameplay on ***representation learning*** in the context of perspective taking ...	**1**: A process in machine learning where machines automatically discover the representations needed for detection or classification from raw data. It involves developing algorithms to automatically learn and improve the feature representations ...	**1**: The process by which artificial agents autonomously discover and encode useful information and features from their observations of an environment, enabling them to understand concepts such as occlusion, object permanence, free space, and containment without needing large labeled datasets for each new task.
2: ... A line of research learn the ***embeddings of graph nodes*** through matrix factorization ... **Ground Truth:** $\Longrightarrow$	**2**: the transformation of nodes in a graph into a set of vectors, where each vector represents a node and aims to capture the graph topology ... These embeddings are often used as inputs for other machine learning tasks such as ... **Correct Prediction:** $\Longrightarrow$	**2**: the low-dimensional vector representations of nodes within a graph, learned through techniques like matrix factorization, which capture the structural and relational information of the nodes to facilitate various machine learning tasks ... **Wrong Prediction:** $\not\Longleftrightarrow$
1: ... we show how ***context-based word embeddings*** , vector representations of words ...	**1:** A type of word representation NLP that captures the variations of word meanings based on their context ... Context-based word embeddings differ from classic word embeddings as they provide word vector representations based on their context, thereby implicitly providing a model ...	**1**: Vector representations of words that capture their meanings based on the surrounding context, used to enhance search capabilities and aid in label prediction for incident reports.
2: ... models outperforms these models even though they use pre - trained ***word embeddings*** ... **Ground Truth:** $\Longleftarrow$	**2:** A type of word representation that allows words with similar meaning to have a similar representation. They are often trained on large amounts of text data and can capture ... **Correct Prediction:** $\Longleftarrow$	**2**: The representation of words in a continuous vector space where words with similar meaning have similar vectors. **Wrong Prediction:** $\not\Longleftrightarrow$

Table 11: Analysis of retrieval vs. GPT-4o definitions. Examples of GPT-4o-generated definitions that do not help accurately resolve relations, in contrast to retrieval-enhanced definitions. $\not\Longleftrightarrow$ - no relationship, $\Longleftrightarrow$ - co-reference, $\Longrightarrow$ - 1 $\rightarrow$ 2 parent-child relationship and $\Longleftarrow$ - 2 $\rightarrow$ 1 parent-child relationship.

Model	Hierarchy		Path Ratio
	F1	F1-50%	
SciCo-Radar$_{GPT\text{-}4o}$	51.2	41.6	50.9
SciCo-Radar $_{RAG}$	54.7	43.4	50.0

Table 12: Ablation: Comparison between definitions generated by RAG and GPT-4o. It is noticeable that the RAG definitions provide an edge with the more challenging *Hard-10* subset on the hierarchy scores.

Model	Coreference	Hierarchy		Path
	CoNLL F1	F1	F1-50%	Ratio
SciCo-Longformer (Baseline)	68.18	46.44	29.61	36.21
SciCo-Mistral-7B	75.59	54.59	44.53	54.79
SciCo-Radar$_{Singleton}$	75.88	57.61	47.41	53.31
SciCo-Radar$_{Relational\ def}$	**76.23**	**59.36**	**49.52**	**55.27**

Table 13: Model results on the *Hard-20* subset.

	Hierarchy			Hierarchy 50%		
	Recall	Precision	F1	Recall	Precision	F1
SciCo-Longformer *(Baseline)*	39.67	54.99	46.09	27.46	48.76	35.14
SciCo-Longformer$_{Singleton\ def}$	40.54	57.13	47.43	31.35	49.94	38.52
Phi-3-mini	34.16	70.96	46.12	29.58	64.00	40.46
Phi-3-mini$_{Singleton\ def}$	36.01	69.75	47.50	31.12	63.01	41.66
SciCo-Mistral-7B	47.62	70.49	56.84	42.32	60.22	49.70
SciCo-Radar$_{GPT\text{-}3.5\ def}$	44.35	70.30	54.39	39.46	62.79	48.46
SciCo-Radar$_{GPT\text{-}4o\ def}$	46.77	74.00	57.31	41.98	65.33	51.12
SciCo-Radar$_{Singleton\ def}$	49.76	67.61	57.33	43.86	59.74	50.58
SciCo-Radar$_{Relational\ def}$	49.41	70.25	58.03	42.75	62.34	50.72

Table 14: Recall, Precision and F1 according to the Cluster-level Hierarchy Score (§4.1) for all models on the full test set.

	Hierarchy			Hierarchy 50%		
	Recall	Precision	F1	Recall	Precision	F1
SciCo-Longformer *(Baseline)*	42.71	50.88	46.44	21.30	48.56	29.61
SciCo-Longformer$_{Singleton\ def}$	42.94	57.83	49.28	24.78	53.00	33.77
Phi-3-mini	31.84	68.64	43.50	24.66	61.02	35.13
Phi-3-mini$_{Singleton\ def}$	36.86	63.68	46.79	26.44	60.61	36.83
SciCo-Mistral-7B	44.73	70.02	54.59	35.09	60.93	44.53
SciCo-Radar$_{GPT\text{-}3.5\ def}$	40.81	68.81	51.23	32.62	62.38	42.84
SciCo-Radar$_{GPT\text{-}4o\ def}$	44.50	71.60	55.30	35.43	63.13	45.39
SciCo-Radar$_{Singleton\ def}$	53.34	67.33	57.61	39.35	59.63	47.41
SciCo-Radar$_{Relational\ def}$	50.90	71.20	59.36	40.02	64.92	49.52

Table 15: Recall, Precision and F1 according to the Cluster-level Hierarchy Score (§4.1) for all models on the *Hard-20* subset.

	Hierarchy			Hierarchy 50%		
	Recall	Precision	F1	Recall	Precision	F1
SciCo-Longformer *(Baseline)*	39.96	52.73	45.46	20.49	44.55	28.07
SciCo-Longformer$_{Singleton\ def}$	41.78	51.69	46.21	23.53	39.61	29.52
Phi-3-mini	26.57	64.46	37.63	18.26	56.02	27.54
Phi-3-mini$_{Singleton\ def}$	36.06	62.78	45.75	22.28	60.15	32.56
SciCo-Mistral-7B	39.96	67.69	50.25	28.60	53.85	37.36
SciCo-Radar$_{GPT\text{-}3.5\ def}$	35.09	66.92	46.04	24.14	57.79	34.05
SciCo-Radar$_{GPT\text{-}4o\ def}$	40.77	68.97	51.25	31.85	60.34	41.69
SciCo-Radar$_{Singleton\ def}$	46.45	66.67	54.75	34.48	58.61	43.42
SciCo-Radar$_{Relational\ def}$	48.07	69.93	56.98	36.51	61.89	45.93

Table 16: Recall, Precision and F1 according to the Cluster-level Hierarchy Score (§4.1) for all models on the *Hard-10* subset.

	MUC			B^3			CEAFe			LEA			CoNLL
	R	P	F1	R	P	F1	R	P	F1	R	P	F1	F1
SciCo-Longformer *(Baseline)*	85.19	85.84	85.52	74.26	75.77	75.01	69.37	68.36	68.86	71.75	73.19	72.46	76.46
SciCo-Longformer$_{Singleton\ def}$	86.59	86.68	86.64	77.52	76.23	76.87	69.25	72.91	71.03	75.22	73.77	74.49	78.18
Phi-3-mini	90.87	87.70	89.06	85.05	77.01	80.83	73.38	76.85	75.07	83.30	74.72	78.78	81.72
Phi-3-mini$_{Singleton\ def}$	90.99	88.06	89.50	85.09	78.35	81.58	73.86	77.88	75.81	83.36	76.17	79.60	82.30
SciCo-Mistral-7B	92.51	86.71	89.51	86.72	75.05	80.47	73.03	78.11	75.49	85.19	72.77	78.49	81.82
SciCo-Radar$_{GPT\text{-}3.5\ def}$	91.42	87.17	89.24	85.22	75.91	80.30	72.94	77.31	75.07	83.50	73.67	78.28	81.54
SciCo-Radar$_{GPT\text{-}4o\ def}$	91.91	86.63	89.19	86.07	74.22	80.10	71.68	77.08	75.58	84.40	71.84	77.62	81.62
SciCo-Radar$_{Singleton\ def}$	90.46	88.28	89.36	83.39	78.09	80.66	74.11	77.62	75.82	81.58	75.85	78.61	81.95
SciCo-Radar$_{Relational\ def}$	90.90	88.11	89.48	84.35	77.56	80.81	73.28	78.74	75.91	82.63	75.32	78.80	82.07

Table 17: Coreference results for all the models on the entire SCICO test set according to all coreference metrics: MUC, B^3 , CEAFe, LEA and CoNLL F1.

	MUC			B^3			CEAFe			LEA			CoNLL
	R	P	F1	R	P	F1	R	P	F1	R	P	F1	F1
SciCo-Longformer *(Baseline)*	78.38	86.40	82.19	58.04	75.71	65.71	60.03	53.61	56.64	55.08	73.26	62.88	68.18
SciCo-Longformer$_{Singleton\ def}$	81.95	85.49	83.69	66.37	73.20	69.62	60.12	59.86	59.99	63.46	70.54	66.81	71.10
Phi-3-mini	86.28	87.71	86.99	76.11	75.23	75.67	64.56	64.28	64.42	73.80	72.70	73.24	75.69
Phi-3-mini$_{Singleton\ def}$	86.09	88.08	87.07	74.12	78.13	76.07	66.33	62.84	64.54	71.68	76.78	74.13	76.12
SciCo-Mistral-7B	88.41	86.30	87.34	77.78	71.89	74.77	64.94	64.39	64.66	45.46	69.41	72.31	75.59
SciCo-Radar$_{GPT\text{-}3.5\ def}$	87.03	87.19	87.11	76.29	73.01	74.61	65.15	66.00	65.57	73.96	70.65	72.26	75.77
SciCo-Radar$_{GPT\text{-}4o\ def}$	87.59	86.03	86.95	77.52	70.18	74.85	63.41	63.95	65.35	75.13	67.51	71.12	75.72
SciCo-Radar$_{Singleton\ def}$	86.22	88.15	87.17	73.50	75.45	74.46	64.49	63.95	64.22	71.05	72.98	72.00	75.88
SciCo-Radar$_{Relational\ def}$	86.59	88.19	87.39	74.59	76.65	75.71	63.74	65.25	64.45	72.24	72.98	72.61	76.23

Table 18: Coreference results for all the models on the *Hard-20* subset of the SCICO test set according to all coreference metrics: MUC, B^3 , CEAFe, LEA and CoNLL F1.

	MUC			B^3			CEAFe			LEA			CoNLL
	R	P	F1	R	P	F1	R	P	F1	R	P	F1	F1
SciCo-Longformer *(Baseline)*	76.37	85.15	80.52	53.85	71.84	61.55	52.79	49.24	50.96	50.85	68.91	58.52	64.34
SciCo-Longformer$_{Singleton\ def}$	78.33	85.35	81.69	57.15	72.27	63.83	51.22	49.44	50.32	54.25	69.20	60.82	65.28
Phi-3-mini	82.25	86.78	84.45	68.65	72.20	70.38	56.38	56.38	56.38	65.92	69.15	67.50	70.40
Phi-3-mini$_{Singleton\ def}$	80.94	86.47	83.61	63.84	75.41	69.15	58.19	53.38	55.69	60.89	72.44	66.16	69.48
SciCo-Mistral-7B	85.51	84.63	85.06	71.63	65.95	68.97	54.92	57.51	56.19	68.78	62.72	65.61	69.98
SciCo-Radar$_{GPT\text{-}3.5\ def}$	83.42	85.43	84.41	68.79	69.18	68.98	56.45	57.48	56.96	66.09	66.12	66.11	70.12
SciCo-Radar$_{GPT\text{-}4o\ def}$	84.46	85.24	84.85	71.15	67.29	69.16	56.02	56.53	56.27	68.43	64.01	66.15	70.10
SciCo-Radar$_{Singleton\ def}$	82.90	86.87	84.84	67.44	70.96	69.15	56.40	57.43	56.91	64.78	67.98	66.34	70.30
SciCo-Radar$_{Relational\ def}$	85.90	84.47	85.18	71.92	68.52	70.18	60.86	59.78	60.31	69.08	65.56	67.28	71.89

Table 19: Coreference results for all the models on the *Hard-10* subset of the SCICO test set according to all coreference metrics: MUC, B^3 , CEAFe, LEA and CoNLL F1.

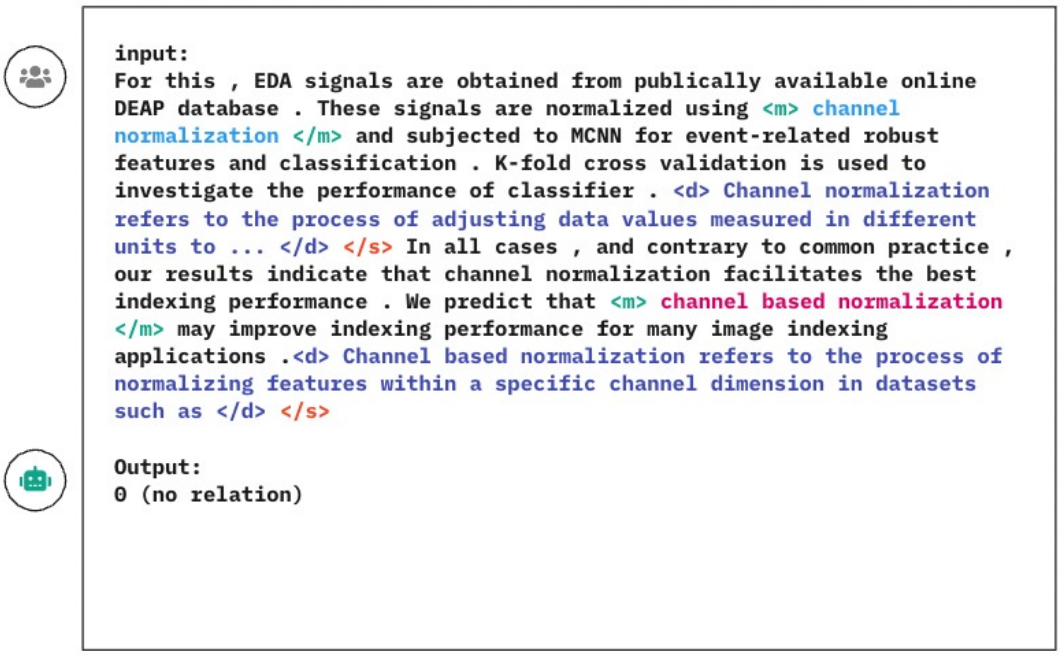

Figure 7: Longformer format.

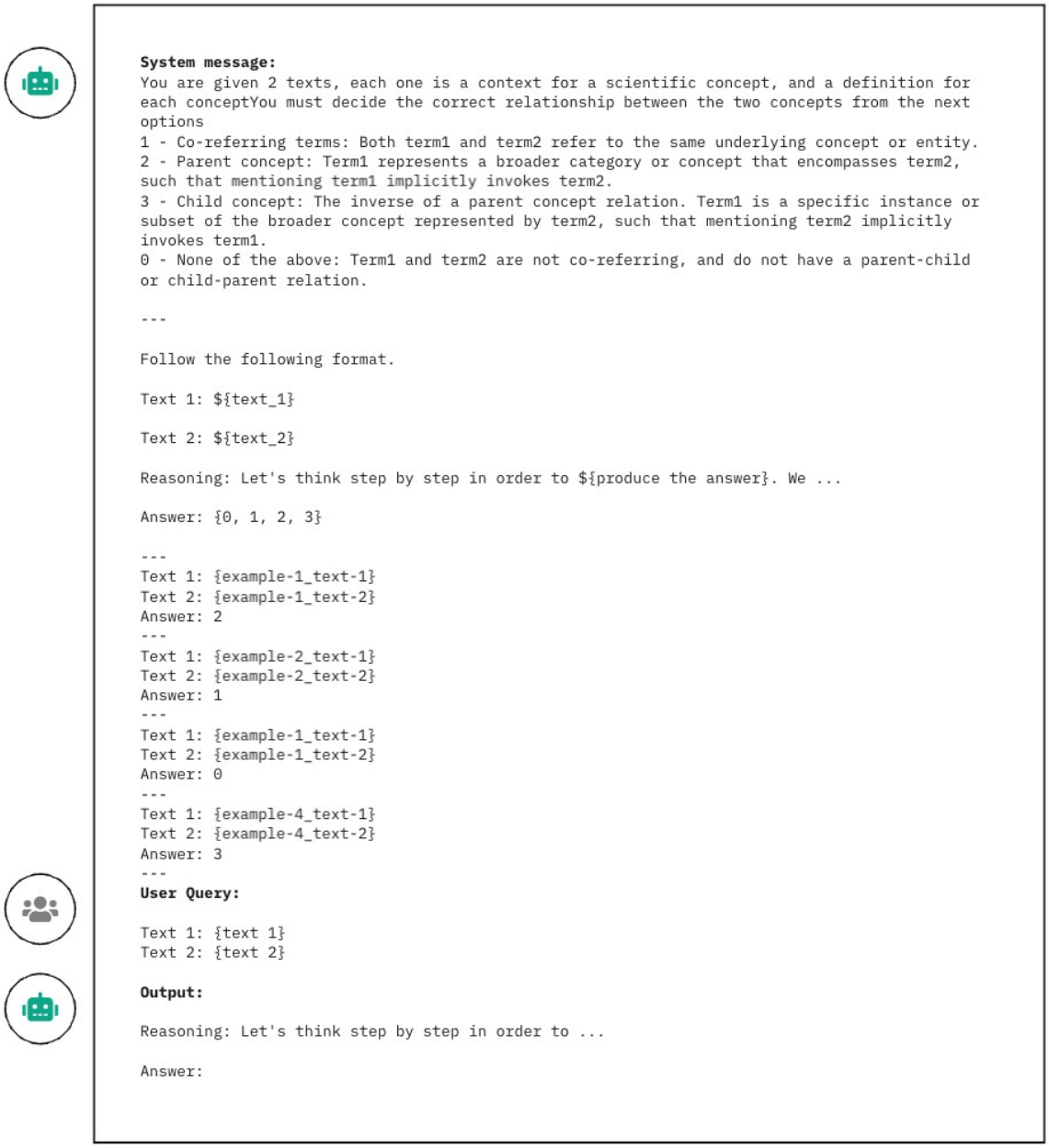

Figure 8: Few shots prompt format.

System message:
You are given 2 texts, each one is a context for a scientific concept, and a definition for each conceptYou must decide the correct relationship between the two concepts from the next options
1 - Co-referring terms: Both term1 and term2 refer to the same underlying concept or entity.
2 - Parent concept: Term1 represents a broader category or concept that encompasses term2, such that mentioning term1 implicitly invokes term2.
3 - Child concept: The inverse of a parent concept relation. Term1 is a specific instance or subset of the broader concept represented by term2, such that mentioning term2 implicitly invokes term1.
0 - None of the above: Term1 and term2 are not co-referring, and do not have a parent-child or child-parent relation.

Follow the following format.

Text 1: ${text_1}

Definition 1: a contextual definition for the first concept

Text 2: ${text_2}

Definition 2: a contextual definition for the second concept

Reasoning: Let's think step by step in order to ${produce the answer}. We ...

Answer: {0, 1, 2, 3.}

Text 1: {example-1_text-1}
Definition 1: {def_1}
Text 2: {example-1_text-2}
Definition 2: {def_2}
Answer: 2

Text 1: {example-2_text-1}
Definition 1: {def_1}
Text 2: {example-2_text-2}
Definition 2: {def_2}
Answer: 1

Text 1: {example-3_text-1}
Definition 1: {def_1}
Text 2: {example-3_text-2}
Definition 2: {def_2}
Answer: 0

Text 1: {example-4_text-1}
Definition 1: {def_1}
Text 2: {example-4_text-2}
Definition 2: {def_2}
Answer: 3

User Query:

Text 1: {text-1}
Definition 1: {def_1}
Text 2: {text-2}
Definition 2: {def_2}

Output:

Reasoning: Let's think step by step in order to ...

Answer:

Figure 9: Few shots with definitions prompt format.