Automating Knowledge Graph Construction from Literature: A Dual-Agent Approach with Progressive Ontology Prompting

Anonymous ACL submission

Abstract

To tackle the challenge of automating knowledge graph construction (KGC) from vast volumes of literature, in this paper, we propose a novel framework leveraging large language models (LLMs), which integrates a progressive ontology prompting (POP) algorithm with a dual-agent system, named LLM-Duo. On the one hand, the POP algorithm conducts a prioritized breadth-first search (BFS) across a predefined ontology, generating structured prompt templates and action sequences to guide the automatic annotation process. On the other hand, the LLM-Duo system features two specialized LLM agents, an explorer and an evaluator, working collaboratively and adversarially to continuously refine annotation quality. We showcase the real-world applicability of our framework through a case study focused on speech-language intervention discovery. Experimental results show that our approach surpasses advanced baselines, achieving more accurate and comprehensive annotations through a fully automated process. Our approach successfully identified 2,421 interventions from a corpus of 64,177 research articles in the speech-language pathology domain, culminating in the creation of a publicly accessible intervention knowledge base with great potential to benefit the speech-language pathology community.

1 Introduction

With millions of research articles published annually, the overwhelming amount of existing scientific knowledge presents both significant challenges and exciting opportunities for researchers, driving the need for advanced analytical tools and interdisciplinary approaches to access and utilize this wealth of information efficiently. Knowledge extraction from scientific literature enables researchers to keep abreast of the latest developments in their domain and facilitate valuable insights that can significantly enhance their work (Usai et al., 2018; Wang et al., 2023a). Yet, in such a vast ocean of data, only a very limited fraction of knowledge is collected and curated due to the low efficiency of the intensive human review process. In healthcare, for instance, evidence-based intervention refers to practices and treatments grounded in systematic research and proven effective through controlled studies (Rutten et al., 2021; Melnyk and Fineout-Overholt, 2022). It emphasizes the use of evidence from well-designed and well-conducted research as the foundation for healthcare decision-making (Sackett, 1997). One of the biggest challenges for healthcare providers is the efficient identification of relevant intervention evidence from an overwhelming body of research, highlighting the urgent need for automated knowledge extraction tools to streamline the process and enhance the accessibility of this valuable information.

In recent years, LLMs have been employed to categorize research papers, extract key findings, summarize complex studies, and create conversational assistants for question-answering and note generation, showing their impressive ability in understanding and extracting valuable insights from text (Goel et al., 2023; Achiam et al., 2023; Guo et al., 2023; Li et al., 2024a; Agrawal et al., 2022; Kojima et al., 2022; Qin et al., 2023; Zhong et al., 2023b). Many studies have utilized LLMs to streamline various subtasks involved in knowledge graph construction (KGC), such as named entity recognition (NER), relation extraction (RE), event extraction (EE), and entity linking (EL) (Wang et al., 2023b; Zhu et al., 2024; Li et al., 2024c; Sun et al., 2024; Li, 2024; Wadhwa et al., 2023). Some research has also explored the collaboration between LLMs and human annotators to improve annotation quality (Kim et al., 2024a; Wang et al., 2024; Tang et al., 2024). However, constructing knowledge graphs from long-range, domain-specific literature, such as scientific articles, poses a significant challenge. Developing annotated

datasets for training deep learning models in NER and RE tasks requires specialized domain expertise to accurately interpret the literature. In this context, LLMs offer a promising alternative through in-context learning, enabling scalable information extraction without the need for extensive labeled data. Despite these advancements, fully automated KGC remains a challenge, particularly when dealing with long-range documents. Most current KGC approaches focus on short texts, leaving significant potential for further development in handling more complex, lengthy content.

For automating KGC from literature with LLMs, one big challenge is the limited context window length of LLMs (An et al., 2024; Ding et al., 2024). This constraint restricts the amount of input text that the models can process at one time, potentially leading to incomplete analysis and missed connections between data points spread across larger documents. To address this issue, the Retrievel-Augmented Generation (RAG) technique (Lewis et al., 2020; Gao et al., 2023; Jiang et al., 2024) can be instrumental by combining a powerful retrieval component with a generative model, allowing the system to access a broader corpus of information beyond the immediate context window. Another challenge is coreference resolution (Lee et al., 2017), as LLMs often struggle with identifying coreferences consistently in long, complex documents due to inherent ambiguities. In the context of RAG, this difficulty can result in fragmented or contradictory responses, where the generated text fails to coherently integrate the retrieved information (Gan et al., 2024; Liu et al., 2024b).

In this paper, we address the challenge of automating knowledge graph construction via LLMs by formulating it as a prompt design and annotation scheduling problem with a predefined ontology graph structure and designing a framework leveraging two LLM agents to iteratively enhance the annotation quality. Specifically, we introduce a progressive ontology prompting (POP) algorithm that employs an outdegree-prioritized breadth-first search (BFS) across the ontology to create a series of prompt templates and action sequences to guide the annotation process conducted by LLMs. To enhance the annotation quality, we propose LLM-Duo, an interactive annotation framework by leveraging the power of LLMs while addressing the limitations of LLMs. Particularly, this framework integrates two LLM agents working both collaboratively and adversarially: 1) *Explorer*, a chatbot utilizing the RAG technique, works in a zero-shot setting to generate annotation results while engaging in debate with the *Evaluator* to justify its findings; and 2) *Evaluator*, an LLM responsible for critically assessing the *Explorer's* annotations and providing constructive feedback to help refine its outputs.

To showcase the practical impact of our approach, we apply our method in a case study of speech-language intervention discovery. We conduct experiments to compare our LLM-Duo annotation framework with several advanced LLM baselines including long context LLM (i.e., GPT-4-Turbo with 128k context window length), and RAG-based annotation chatbot with advanced prompting techniques including Chain-of-Thought (CoT) (Wei et al., 2022) and Self-Refine (Madaan et al., 2024). The experimental results demonstrate that our method not only delivers more accurate and comprehensive annotations over these strong baselines but also significantly accelerates the knowledge graph construction process. Furthermore, through our framework, we successfully curate a speech-language intervention knowledge base, providing a valuable resource for the speech-language pathology community. To our knowledge, this is the first intervention knowledge base in the speech-language pathology field.

2 Related Works

Mining structured knowledge from long-range documents is a great challenge due to the vast volume of content, the inherent ambiguity of natural language, and the individual bias of human interpretation (Ye et al., 2022; Jain et al., 2020; Tang et al., 2020). Particularly in domain-specific contexts, these challenges are further compounded by the need for specialized expertise, labor-intensive manual annotation, and difficulties in maintaining consistency and scalability (Zhao et al., 2021; Zheng et al., 2018). Automating knowledge extraction offers a promising solution to these issues, as advanced natural language processing (NLP) techniques can streamline the labor-intensive annotation process and enhance overall efficiency.

Traditional approaches to automated knowledge graph construction (KGC) typically rely on pipelines to handle various NLP tasks such as named entity recognition, relation extraction, coreference resolution, entity linking, and event detection (Luan et al., 2018; Martins et al., 2019; Wei

et al., 2019; Zhong et al., 2023a; Laurenzi et al., 2024). Recent advancements leverage LLMs to generate relational triplets in zero/few-shot settings for KGC, achieving promising results (Wei et al., 2023; Sun et al., 2024; He et al., 2024). Some studies (Zhang and Soh, 2024; Carta et al., 2023; Vamsi et al., 2024; Zhu et al., 2024) have further streamlined KGC by breaking it down into distinct phases, enabling LLMs to infer knowledge graph schemas without relying on predefined ontologies. However, these methods are often constrained to short texts or have only been validated on tasks like entity and relation extraction using human-annotated datasets, such as DuIE2.0 (Li et al., 2019) and DocRED (Yao et al., 2019), without being proven effective in real-world applications. As expert-crafted ontologies can guide LLMs to extract desired knowledge structure (Cauter and Yakovets, 2024). We introduce a novel method for constructing domain-specific knowledge graphs automatically via LLMs, which we detail in the following sections.

3 Preliminaries

A Knowledge Graph (KG) is a semantic network structured as an ontology, consisting of concepts and their relationships in a clear, interpretable format at scale (Ehrlinger and Wöß, 2016; Peng et al., 2023). For constructing KGs from literature, LLMs can enhance this process by leveraging their capabilities to understand long-range text. This allows for transforming unstructured data into structured formats, and finally, populating the ontology to create KG.

Focusing on automating knowledge graph construction from literature, we aim to develop a general framework that helps researchers and practitioners gather domain-specific knowledge in their fields. In our methodology, the KG ontology is crafted by domain experts, which can be represented by a directed acyclic graph (DAG) $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{F})$. Here $\mathcal{E}$, $\mathcal{R}$, and $\mathcal{F}$ are sets of concepts, relationships, and semantic triples respectively. $\mathcal{F}$ is a collection of triples (h, r, t) with a head concept $h \in \mathcal{E}$, a tail concept $t \in \mathcal{E}$, and a relation $r \in \mathcal{R}$ (Gruninger, 1995). To effectively instruct LLMs to construct knowledge graphs anchored to $\mathcal{G}$, the design of annotation prompts and the query sequences plays a crucial role. We thereby frame the problem of automated knowledge graph construction via LLMs as one of prompt design and scheduling, described by the following equation:

$$f(\mathcal{G}(\mathcal{E}, \mathcal{R}, \mathcal{F})) = \{(Prompt_i, Order_i) | i \in [1, N]\} \quad (1)$$

where f is a function that translates KG ontology into a set of annotation prompts and query sequences for the LLMs. A common case of f is directly prompting LLMs to generate triplets in a zero-shot/few-shot manner by including the KG schema within the prompt such as the annotation methods used in (Mihindukulasooriya et al., 2023; Kommineni et al., 2024; Mihindukulasooriya et al., 2023). However, those methods generate annotations in one shot and ignore the importance of contextual correlations between concepts within their surrounding neighborhood, resulting in incomplete annotations.

4 Methodology

In this section, we first introduce the POP algorithm that converts a KG ontology into a set of annotation prompt templates and query orders, then propose an interactive annotation framework based on two LLM agents to enable more convincing and accurate annotation generations.

4.1 Progressive Ontology Prompting

We develop a progressive ontology prompting (POP) algorithm that employs a prioritized BFS on the ontology graph $\mathcal{G}(\mathcal{E}, \mathcal{R}, \mathcal{F})$ to generate a set of annotation prompt templates and query sequences for LLMs. In our algorithm, the prompt formulation and scheduling follow a progressive manner. As illustrated in Figure 1, the annotation process begins at a source node (i.e., a concept node that only has outgoing edges) and continues by traversing its neighboring nodes in the order of a prioritized BFS. To allow for quick accessing a large portion of the graph, we enhance BFS by sorting neighboring nodes based on their out-to-in ratio $R(v)$, which is defined by:

$$R(v) = \frac{|\{(h, r, t) \in \mathcal{F}\}|h = v|}{|\{(h, r, t) \in \mathcal{F}\}|t = v|} \quad (2)$$

Our algorithm selects the neighboring node with the higher R value to visit in the next step. For instance, in the example of Figure 1, visiting the 'Patient' node before the 'Disorder' node can provide more context for the 'Disorder' concept annotation. For each concept node v, we use the visited nodes within its k-hop neighborhood as its context. The $Prompt_v$ for annotating concept v is crafted

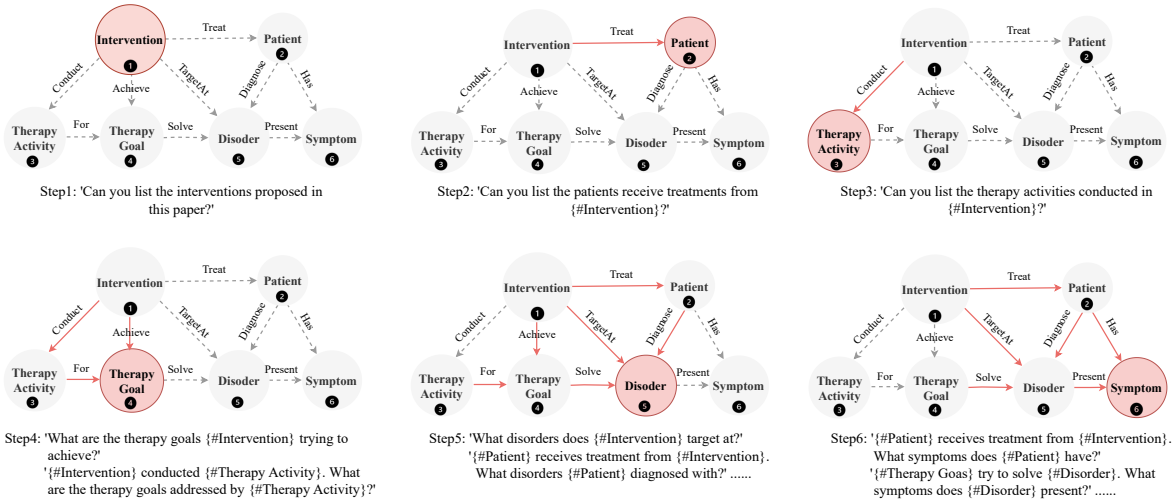


Figure 1: Illustration of prompt design and scheduling based on a prioritized BFS traversal over KG ontology.

based on its context and the completed annotations within that context. The action order $Order_v$ for $Prompt_v$ is determined by the sequence in which node v is visited during the prioritized BFS traversal.

The pseudocode of the POP algorithm is provided in Appendix A.1. Our algorithm first follows prioritized BFS traversal to capture the local k-hop context and visit order of a specific concept node v, then composes the annotation prompts $Prompt_v$ based on its ontology substructure $N_k(v)$ and completed annotations within its context. This process can be expressed as follows:

$$Prompt_v \leftarrow T_v(Annotation(N_k(v))) \tag{3}$$

$$T_v \leftarrow \{Prefix\,(N_{k-1}(u)) \oplus Question\,((v,e,u) \mid (v,e,u) \in F \mid u \in N_1(v)\} \tag{4}$$

, where $\oplus$ is the concatenation. $Prompt_v$ represents a set of annotation prompts for node v, generated by applying completed annotations to the prompt template T_v. As illustrated in Figure 1, the prompt template T_v consists of two parts: 1) *Question*, an annotation question derived from the relationship between node v and one of its neighboring nodes u; and 2) *Prefix*, a description based on the $k-1$ hop path of neighbor node u. We leverage few-shot learning to task LLMs in generating prompt templates. Example prompts and corresponding outputs are provided in Appendix A.2.

4.2 LLM-Duo Annotation Framework

To guarantee the integrity and reliability of LLM annotations, we propose LLM-Duo, a multi-agent annotation system. The central idea of a multi-agent system is to employ combinations of LLMs that can converse with each other to collaboratively accomplish tasks (Wu et al., 2023; Chen et al., 2024; Zhang et al., 2023). Drawing inspiration from the multi-agent debate idea in (Du et al., 2023; Kim et al., 2024b), we developed a framework where agents work both collaboratively and adversarially to enhance the quality of annotations.

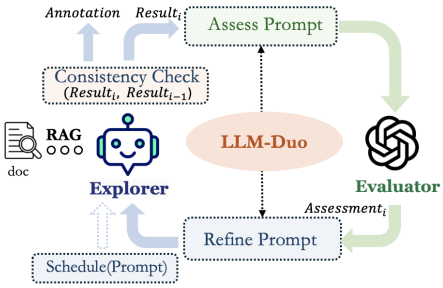


Figure 2: Iterative annotation with two LLM agents under the LLM-Duo framework.

The architecture of LLM-Duo is shown in Figure 2, featuring two LLM agents: the *explorer* and the *evaluator*. Specifically, the *explorer* is a chatbot performing annotation tasks using zero-shot question answering (QA). To ensure the *explorer's* outputs are grounded in the provided literature, RAG is employed to reference relevant sources, minimizing hallucinations. To further improve the accuracy and reliability of the annotations, the *evaluator* LLM is incorporated to review and validate the *explorer's* responses, ensuring higher-quality results.

Combined with the POP algorithm, LLM-Duo will be sequentially tasked with annotation prompts.

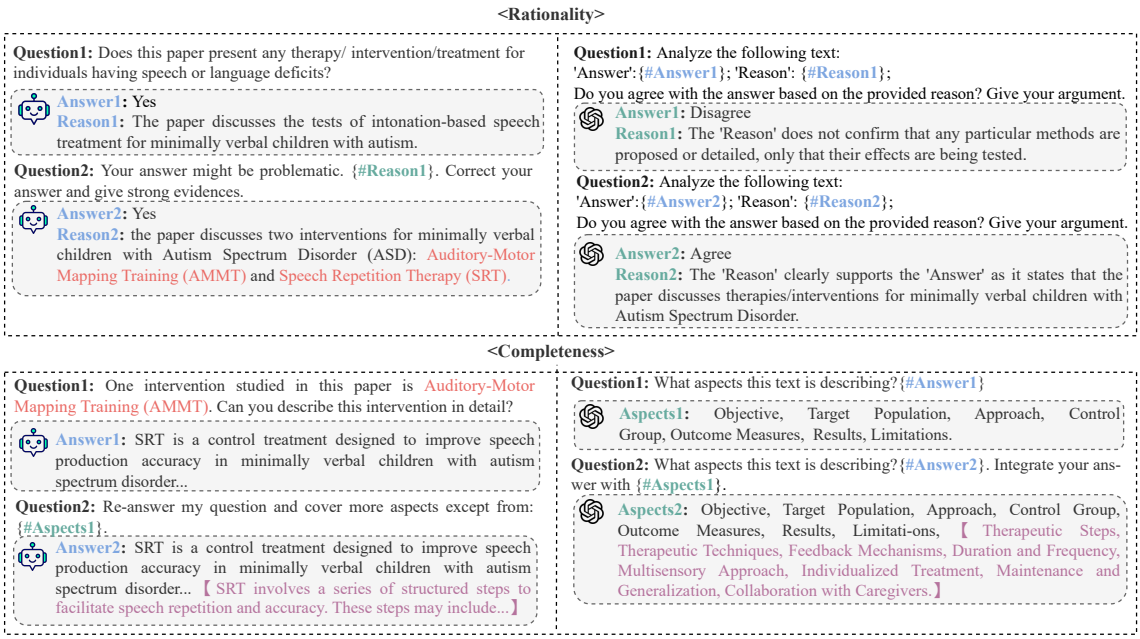


Figure 3: Annotation examples of speech-language intervention discovery using the LLM-Duo framework.

During each annotation cycle for a specific concept node, when focusing on concepts that emphasize rationality, the *explorer* provides an answer and an explanatory rationale to the *evaluator*. The *evaluator* then reviews the reasoning and offers feedback. Based on this feedback, the *explorer* either refines its answer or, if in disagreement, presents stronger evidence to defend its original answer and challenge the *evaluator's* judgment. For concepts that emphasize completeness, the *evaluator* extracts the aspects covered in each round of the *explorer's* answer, combines them with aspects from previous rounds, and prompts the *explorer* to expand further beyond the newly integrated aspect collection. This iterative process continues until the annotations reach a consistent and comprehensive state.

As the example shown in Figure 3, by facilitating interactive loops between two LLM agents, LLM-Duo enables more accurate and complete annotations. Although many multi-agent systems have been developed for various purposes (Chen and Si, 2024; Zhou et al., 2024; Zhu et al., 2024), LLM-Duo is the first multi-agent debate system specifically designed to enable rationality and completeness in KG annotation.

5 Experiments

5.1 Implementation

In the LLM-Duo framework, the *explorer* is a chatbot built on LLM with RAG, implemented with Llamaindex[1] framework. We use OpenAI 'text-embedding-3-large'[2] as the embedding model and set the chunk size to 256 tokens with an overlapping size of 128. Particularly, we use 'FastCoref' (Otmazgin et al., 2022) to process text chunks for coreference resolution before text embedding. Additionally, we include the document ID as metadata for chunks and apply a metadata filter in the chat engine to ensure that the *explorer* only answers based on the specific document being annotated. We use Chroma[3] as the vector database. We set the retrieval to be on the top 8 text chunks based on similarity scores reranked with SentenceTransformer-Rerank[4] employing the 'cross-encoder/ms-marco-MiniLM-L-2-v2'[5] model in Llamaindex. The evaluator is an external LLM who does not share any document context with the *explorer*.

5.2 Case Study: Speech-language Intervention Discovery

To evaluate the effectiveness of our method in a real-world context, we apply our automated KGC framework in the speech-language intervention discovery setting. Intervention discovery aims to systematically gather speech-language interventions from the literature to support evidence-based prac-

[1] https://www.llamaindex.ai
[2] https://platform.openai.com/docs/guides/embeddings/embedding-models
[3] https://github.com/chroma-core/chroma
[4] https://docs.llamaindex.ai/en/stable/examples/node_postprocessor/SentenceTransformerRerank
[5] https://huggingface.co/cross-encoder/ms-marco-MiniLM-L-2-v2

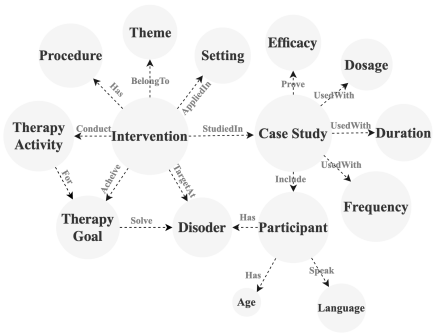


Figure 4: Ontology of speech-language intervention.

tice design. It involves identifying relevant studies and extracting key features of interventions, such as target disorder, procedure, efficacy, case study, therapy activity, etc., which are extremely labor-intensive for human reviewers, highlighting the potential of automated KGC based on LLMs to streamline the process. The intervention annotation ontology is shown in Figure 4. We further provide the concept definitions in Appendix B. To enable large-scale intervention discovery, we curated a literature corpus of 64,177 papers within the domain of speech-language pathology.

5.3 Annotation Baselines

The core idea behind LLM-Duo is to make LLM rectify its initial annotations through external feedback from another LLM. Instead of setting up another LLM for evaluation, recent studies demonstrate that LLMs can engage in self-correction to enhance their responses autonomously (Liu et al., 2024a; Li et al., 2024b). Notable examples of this include Chain-of-Thought (CoT) (Wei et al., 2022) and Self-Refine (Madaan et al., 2024). We separately equip the chatbot based on RAG with these two prompting methods for annotation and denote them as CoT-RAG and Self-Refine-RAG. The full prompts for these two methods can be found in Appendix C. Additionally, in LLM-Duo, a potential substitution of *explorer* is long-context LLM, which is capable of processing entire document tokens instead of chunking and retrieval with RAG. We refer to the LLM-Duo system as LLM-Duo-RAG when using *explorer* built on RAG, and as LLM-Duo-LongContext when utilizing long-context LLMs. Besides, we also compare with methods that directly input paper text to LLMs for zero-shot QA annotation without the evaluation feedback loop, including ShortContext LLM, LongContext LLM, OpenAI Assistant, and RAG.

5.4 Evaluation

In the experiment of comparing LLM-Duo with annotation baselines, we report six types of metrics: 1) Consistency Rounds (CR): the number of refine loops the method makes before achieving the annotation consistency; 2) Verbosity Index (VI): the number of covered aspects per 1k tokens in the annotations, which is an important metric for emphasizing content completeness; 3) Enumeration Quantity (EQ): the number of items listed in the annotations (i.e., therapy activities, therapy goals in our test case.); 4) Faithfulness (Faith): the extent of the annotation faithful to the provided literature document, which is measured by FaithfulnessEvaluator[6] of Llamaindex. 5) Accuracy (ACC): the percent of correct annotations in all LLM-generated annotations. 6) Cover: the percent of correct LLM-generated annotations to the total mentioned concept entities in the provided literature.

6 Results

In this section, we first provide a detailed evaluation of our progressive ontology prompting algorithm and the LLM-Duo annotation framework. Then, we showcase the results of speech-language intervention discoveries using our automated KGC framework.

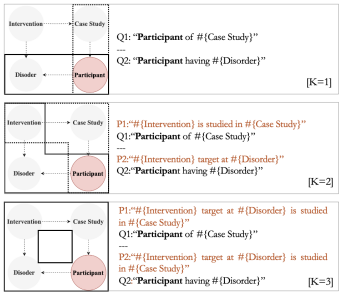
6.1 POP Algorithm Study

6.1.1 Context Size Analysis

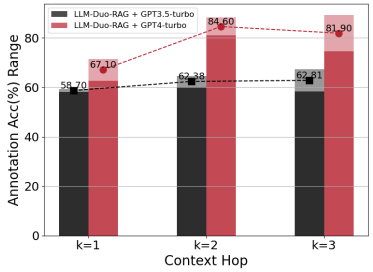
In the POP algorithm, the context size k determines the diversity and volume of information included in the annotation prompt. To assess the impact of context size on annotation quality, we conducted experiments using various k values to generate prompts for LLM-Duo-RAG annotations. As shown in Figure 5a, we annotate the 'participant' concept for the experiment, which was based on a random selection of 8 speech-language pathology literature.

The annotation accuracy is shown in Figure 5b. The results indicate that as the context size k increases, annotation accuracy improves significantly, suggesting that a larger context provides more informative prompts, thereby enhancing annotation quality. Moreover, GPT-4-turbo consistently outperforms GPT-3.5-turbo across all k values, demonstrating that more advanced language models can further improve annotation accuracy. Besides, we inspect the text chunks retrieved back by different annotation queries based on various k values.

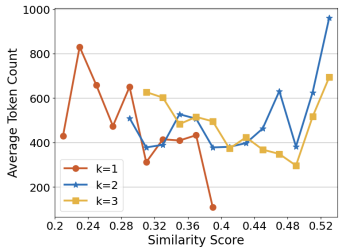
[6] https://docs.llamaindex.ai/en/stable/examples/evaluation/faithfulness_eval



(a) Illustration of 'participant' annotation under different k values.



(b) Range of 'participant' annotation accuracy using LLM-Duo-RAG with GPTs at k=1,2,3.



(c) Token count distribution of retrieved-back chunks across varying similarity scores using 'participant' annotation queries of k=1,2,3.

Figure 5: Evaluation of 'participant' annotation with POP of different context sizes.

We report the range of similarity scores and token count distribution of retrieved-back chunks in Figure 5c. The similarity score represents the semantically relevancy between retrieved texts to annotation queries. The results show that for $k = 1$, the retrieved text chunks generally have low similarity to the query, and the token count decreases as the similarity score increases, leading to lower annotation quality. In contrast, higher k values, especially $k = 2$, yield more relevant retrievals. For $k = 2$, the token count increases with higher similarity scores, indicating that richer and more relevant content is captured, resulting in improved annotation quality.

6.1.2 Prioritized BFS Analysis

LLM-Duo-RAG	GPT3.5-turbo	GPT4-turbo
POP✗	68.18	72.73
POP✓ Prioritized-BFS✗	77.28	83.20
POP✓ Prioritized-BFS✓	**81.82**	**86.37**

Table 1: Comparison of annotation results with and without the POP and prioritized BFS.

In the POP algorithm, we employ the out-to-in ratio to prioritize neighboring nodes during BFS-based prompt creation and scheduling. This strategy ensures that nodes with more outgoing edges are visited first, allowing them to provide more context for annotating other nodes. For example, one annotation sequence following prioritized POP over intervention ontology in Appendix B is 'TherapyActivity'→'TherapyGoal'→'Disorder'. In this section, we compare the 'Disorder' annotation results using the POP algorithm with and without prioritization, as well as one-shot annotation without using POP, where the entire KG schema is included on a single annotation prompt to extract all triplets. The results are presented in Table 1. Applying both the POP and prioritized BFS notably enhances annotation performance.

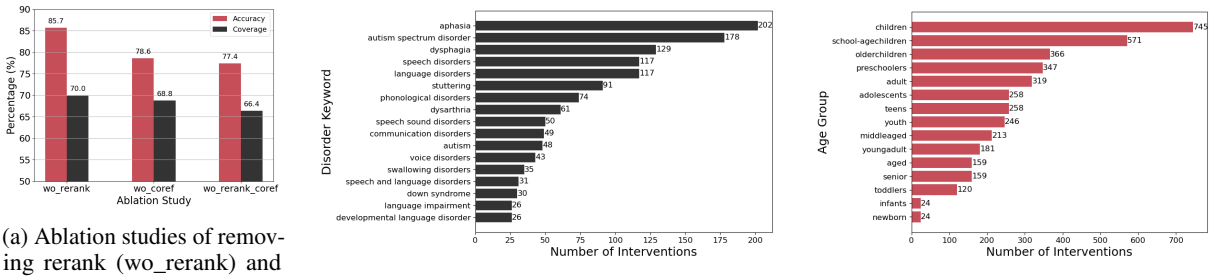
6.2 LLM-Duo with Baselines

In this section, we compare LLM-Duo with several advanced baselines using annotations of 8 randomly selected papers from our speech-language literature corpus. The evaluation focuses on three key dimensions: 1) Intervention Recognition (IR), identifying intervention entities within the literature; 2) Intervention Aspect Summary (IAS), annotating the key aspects (e.g., procedure, therapy activity, therapy goals) of the intervention by capturing and summarizing relevant information from the paper; and 3) Intervention Knowledge Completion (IKC), linking interventions to theme classes (e.g., speech awareness, speech articulation, comprehension, foundation skills, etc.) and setting concept nodes (e.g., home, healthcare facilities, schools, teletherapy, etc.). We use human annotators for the IR and IKC tasks to generate gold-standard results for comparison. In the IAS task, we only ask human annotators to tag relevant text fragments related to specific intervention aspects due to potential individual bias in human interpretation.

The experimental results are reported in Table 2. It should be noted that we implemented 'Short-Context' using Llama3-instruct-70b (FP16) and Mistral-instruct-8x22b models (INT8). However, directly prompting these models with full paper text fails to produce annotations in a zero-shot QA setting. Their generations do not align with the annotation questions. The results in Table 2 show that LLM-Duo-RAG outperforms all baselines. Although GPT4-turbo has a 128k context window length and is capable of generating annotations, its annotation coverage remains inadequate. Integrating it with the LLM-Duo framework can significantly improve both the accuracy and the comprehensiveness of the annotations. Addi-

Methods	LLM	IR			ICA				IKC	
		CR	ACC	Cover	CR	VI	EQ	Faith	CR	ACC
ShortContext	*GPT3.5-turbo*	-	36.9%	50%	-	0.0249	5.46	0.9667	-	48.2%
OpenAI Assistant	*GPT4-turbo*	-	76.1%	69.0%	-	0.0631	4.17	0.7857	-	53.3%
LongContext	*GPT4-turbo*	-	76.3%	57.1%	-	0.0919	8.64	**1.0**	-	61.2%
LLM-Duo-LongContext	*GPT4-turbo*	2.17	81.0%	68.7%	2.5	0.0926	8.68	0.8571	1.31	69.6%
RAG	*GPT3.5-turbo*	-	47.6%	50%	-	0.0319	7.96	0.8550	-	48.7%
CoT-RAG	*GPT3.5-turbo*	1.04	78.6%	81%	3.18	0.0771	10.37	0.7250	1.07	73.2%
Self-Refine-RAG	*GPT3.5-turbo*	1.19	78.5%	54.4%	2.85	0.0694	7.17	0.8125	1.12	54.8%
LLM-Duo-RAG	*GPT3.5-turbo*	1.84	**100%**	**86.4%**	2.58	**0.1159**	**13.71**	0.9285	1.46	**85.6%**
	Llama3-instruct-70b	2.71	78.6%	55.6%	2.59	0.0748	9.79	0.8648	1.52	61.0%
	Mistral-instruct-8x22b	2.30	81.9%	67.5%	2.16	0.0763	9.87	0.8875	1.46	67.2%

Table 2: Comparison of annotation results with baselines using different LLMs.



(a) Ablation studies of removing rerank (wo_rerank) and corefenence (wo_coref) modules in LLM-Duo-RAG.

(b) Number of discovered interventions for top 20 disorders.

(c) Number of discovered interventions for different age groups.

Figure 6: Ablation study and annotated speech-language intervention statistics.

tionally, compared with simple RAG, self-correct prompting methods such as CoT and Self-Refine can significantly enhance annotations, but their performance is still worse than LLM-Duo-RAG. Instead of utilizing costly GPT models, LLM-Duo-RAG, which employs open-source models including Llama3-instruct-70b and Mistral-instruct-8x22b, can achieve comparable annotation quality.

6.3 Ablation Study

In our implementation, the RAG technique serves as the backbone of *explorer*. We employ 'Fast-Coref' for coreference resolution and rerank retrieved chunks by similarity score using the 'cross-encoder/ms-marco-MiniLM-L-2-v2 model'. This section presents ablation studies for both components. We report the accuracy of intervention recognition in this section. As shown in Figure 6a, the results demonstrate that removing these components significantly decreases annotation accuracy, showing the necessity of each module.

6.4 Speech-Language Intervention Discovery

Through our automated KGC framework, we identified 2,421 interventions supported by case studies from 64,177 literature in the speech-language pathology domain. One complete intervention annotation example is provided in Appendix D. The statistics of discovered interventions are presented in Figure 6b and Figure 6c. More intervention examples are provided in Appendix E.1. 19 clinicians and students reviewed our annotations through online Google forms. In checking the intervention annotation and reading the corresponding literature, they reported that 89% of our annotations are valid and complete. We have constructed the first intervention knowledge graph in the speech-language pathology domain, which will be made publicly accessible upon acceptance. The knowledge graph overview is shown in Appendix E.2. This knowledge graph is anticipated to be a valuable resource for domain experts, facilitating evidence-based clinical decision-making, question-answering, and recommendation systems, ultimately contributing to improved healthcare outcomes.

7 Conclusion

In this paper, we developed a novel LLM-based framework for automatic knowledge graph construction, featuring a progressive ontology prompting algorithm and a dual-agent system. The proposed method achieves superior performance compared with advanced baselines, enabling more accurate knowledge graph construction automatically. Our approach is adaptable to various ontologies across different domains and offers practical value for automating knowledge extraction.

8 Limitations

Considering the significant challenges that domain experts and practitioners face in extracting knowledge from the vast volume of scientific literature, we propose an ontology-driven framework to automate domain knowledge graph construction. Even though there are document-level datasets available for tasks like NER and RE in the biomedical domain (Yao et al., 2019; Wei and Li, 2023; Doğan et al., 2014; Consortium, 2004), our annotation task is additionally compounded with information summarization at the document level, such as annotating 'intervention procedure', 'case study', and classification based on document-level insights, such as annotating 'theme' and 'setting' of intervention. Therefore, we only demonstrate the efficacy of our automated KGC framework through a real-world application of speech-language intervention discovery, even though the framework is applicable to general ontologies. Our future research will investigate the framework's adaptability and effectiveness across different domains.

References

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

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

Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, and Jian-Guang Lou. 2024. Make your llm fully utilize the context. *arXiv preprint arXiv:2404.16811*.

Salvatore Carta, Alessandro Giuliani, Leonardo Piano, Alessandro Sebastian Podda, Livio Pompianu, and Sandro Gabriele Tiddia. 2023. Iterative zero-shot llm prompting for knowledge graph construction. *arXiv preprint arXiv:2307.01128*.

Zeno Cauter and Nikolay Yakovets. 2024. Ontology-guided knowledge graph construction from maintenance short texts. In *Proceedings of the 1st Workshop on Knowledge Graphs and Large Language Models (KaLLM 2024)*, pages 75–84.

Weize Chen, Jiarui Yuan, Chen Qian, Cheng Yang, Zhiyuan Liu, and Maosong Sun. 2024. Optima: Optimizing effectiveness and efficiency for llm-based multi-agent system. *arXiv preprint arXiv:2410.08115*.

Yuetian Chen and Mei Si. 2024. Reflections & resonance: Two-agent partnership for advancing LLM-based story annotation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13813–13818, Torino, Italia. ELRA and ICCL.

Gene Ontology Consortium. 2004. The gene ontology (go) database and informatics resource. *Nucleic acids research*, 32(suppl_1):D258–D261.

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. 2024. Longrope: Extending llm context window beyond 2 million tokens. *arXiv preprint arXiv:2402.13753*.

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

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.

Lisa Ehrlinger and Wolfram Wöß. 2016. Towards a definition of knowledge graphs. *SEMANTiCS (Posters, Demos, SuCCESS)*, 48(1-4):2.

Yujian Gan, Massimo Poesio, and Juntao Yu. 2024. Assessing the capabilities of large language models in coreference: An evaluation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1645–1665.

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

Akshay Goel, Almog Gueta, Omry Gilon, Chang Liu, Sofia Erell, Lan Huong Nguyen, Xiaohong Hao, Bolous Jaber, Shashir Reddy, Rupesh Kartha, et al. 2023. Llms accelerate annotation for medical information extraction. In *Machine Learning for Health (ML4H)*, pages 82–100. PMLR.

Michael Gruninger. 1995. Methodology for the design and evaluation of ontologies. In *Proc. IJCAI'95, Workshop on Basic Ontological Issues in Knowledge Sharing*.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*.

Li He, Hayilang Zhang, Jie Liu, Kang Sun, and Qing Zhang. 2024. Zero-shot relation triplet extraction via knowledge-driven llm synthetic data generation. In

International Conference on Intelligent Computing, pages 329–340. Springer.

Sarthak Jain, Madeleine Van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy. 2020. Scirex: A challenge dataset for document-level information extraction. *arXiv preprint arXiv:2005.00512*.

Ziyan Jiang, Xueguang Ma, and Wenhu Chen. 2024. Longrag: Enhancing retrieval-augmented generation with long-context llms. *arXiv preprint arXiv:2406.15319*.

Hannah Kim, Kushan Mitra, Rafael Li Chen, Sajjadur Rahman, and Dan Zhang. 2024a. Meganno+: A human-llm collaborative annotation system. *arXiv preprint arXiv:2402.18050*.

Kyungha Kim, Sangyun Lee, Kung-Hsiang Huang, Hou Pong Chan, Manling Li, and Heng Ji. 2024b. Can llms produce faithful explanations for fact-checking? towards faithful explainable fact-checking via multi-agent debate. *arXiv preprint arXiv:2402.07401*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Vamsi Krishna Kommineni, Birgitta König-Ries, and Sheeba Samuel. 2024. From human experts to machines: An llm supported approach to ontology and knowledge graph construction. *arXiv preprint arXiv:2403.08345*.

Emanuele Laurenzi, Adrian Mathys, and Andreas Martin. 2024. An llm-aided enterprise knowledge graph (ekg) engineering process. In *Proceedings of the AAAI Symposium Series*, volume 3, pages 148–156.

James Law, Zoe Garrett, Chad Nye, Psychosocial Cochrane Developmental, and Learning Problems Group. 1996. Speech and language therapy interventions for children with primary speech and language delay or disorder. *Cochrane Database of Systematic Reviews*, 2015(5).

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. *arXiv preprint arXiv:1707.07045*.

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

Harry Li, Gabriel Appleby, and Ashley Suh. 2024a. A preliminary roadmap for llms as assistants in exploring, analyzing, and visualizing knowledge graphs. *arXiv preprint arXiv:2404.01425*.

Jiyi Li. 2024. A comparative study on annotation quality of crowdsourcing and llm via label aggregation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6525–6529. IEEE.

Loka Li, Zhenhao Chen, Guangyi Chen, Yixuan Zhang, Yusheng Su, Eric Xing, and Kun Zhang. 2024b. Confidence matters: Revisiting intrinsic self-correction capabilities of large language models. *Preprint*, arXiv:2402.12563.

Shuangjie Li, Wei He, Yabing Shi, Wenbin Jiang, Haijin Liang, Ye Jiang, Yang Zhang, Yajuan Lyu, and Yong Zhu. 2019. Duie: A large-scale chinese dataset for information extraction. In *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8*, pages 791–800. Springer.

Xingzuo Li, Kehai Chen, Yunfei Long, and Min Zhang. 2024c. Llm with relation classifier for document-level relation extraction. *arXiv preprint arXiv:2408.13889*.

Dancheng Liu, Amir Nassereldine, Ziming Yang, Chenhui Xu, Yuting Hu, Jiajie Li, Utkarsh Kumar, Changjae Lee, and Jinjun Xiong. 2024a. Large language models have intrinsic self-correction ability. *Preprint*, arXiv:2406.15673.

Yanming Liu, Xinyue Peng, Jiannan Cao, Shi Bo, Yanxin Shen, Xuhong Zhang, Sheng Cheng, Xun Wang, Jianwei Yin, and Tianyu Du. 2024b. Bridging context gaps: Leveraging coreference resolution for long contextual understanding. *arXiv preprint arXiv:2410.01671*.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. *arXiv preprint arXiv:1808.09602*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2024. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36.

Leslie A Mahler and Lorraine O Ramig. 2012. Intensive treatment of dysarthria secondary to stroke. *Clinical linguistics & phonetics*, 26(8):681–694.

Pedro Henrique Martins, Zita Marinho, and André FT Martins. 2019. Joint learning of named entity recognition and entity linking. *arXiv preprint arXiv:1907.08243*.

Bernadette Mazurek Melnyk and Ellen Fineout-Overholt. 2022. *Evidence-based practice in nursing & healthcare: A guide to best practice*. Lippincott Williams & Wilkins.

Nandana Mihindukulasooriya, Sanju Tiwari, Carlos F Enguix, and Kusum Lata. 2023. Text2kgbench: A benchmark for ontology-driven knowledge graph generation from text. In *International Semantic Web Conference*, pages 247–265. Springer.

Shon Otmazgin, Arie Cattan, and Yoav Goldberg. 2022. F-coref: Fast, accurate and easy to use coreference resolution. *arXiv preprint arXiv:2209.04280*.

Ciyuan Peng, Feng Xia, Mehdi Naseriparsa, and Francesco Osborne. 2023. Knowledge graphs: Opportunities and challenges. *Artificial Intelligence Review*, 56(11):13071–13102.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.

Lila J Finney Rutten, Xuan Zhu, Aaron L Leppin, Jennifer L Ridgeway, Melanie D Swift, Joan M Griffin, Jennifer L St Sauver, Abinash Virk, and Robert M Jacobson. 2021. Evidence-based strategies for clinical organizations to address covid-19 vaccine hesitancy. In *Mayo clinic proceedings*, volume 96, pages 699–707. Elsevier.

David L Sackett. 1997. Evidence-based medicine. In *Seminars in perinatology*, volume 21, pages 3–5. Elsevier.

Qi Sun, Kun Huang, Xiaocui Yang, Rong Tong, Kun Zhang, and Soujanya Poria. 2024. Consistency guided knowledge retrieval and denoising in llms for zero-shot document-level relation triplet extraction. In *Proceedings of the ACM on Web Conference 2024*, pages 4407–4416.

Hengzhu Tang, Yanan Cao, Zhenyu Zhang, Jiangxia Cao, Fang Fang, Shi Wang, and Pengfei Yin. 2020. Hin: Hierarchical inference network for document-level relation extraction. In *Advances in Knowledge Discovery and Data Mining: 24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 11–14, 2020, Proceedings, Part I 24*, pages 197–209. Springer.

Yi Tang, Chia-Ming Chang, and Xi Yang. 2024. Pdfchatannotator: A human-llm collaborative multimodal data annotation tool for pdf-format catalogs. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, pages 419–430.

Antonio Usai, Marco Pironti, Monika Mital, and Chiraz Aouina Mejri. 2018. Knowledge discovery out of text data: a systematic review via text mining. *Journal of knowledge management*, 22(7):1471–1488.

Krishna Kommineni Vamsi, Vamsi Krishna Kommineni, and Sheeba Samuel. 2024. From human experts to machines: An llm supported approach to ontology and knowledge graph construction.

Somin Wadhwa, Silvio Amir, and Byron C Wallace. 2023. Revisiting relation extraction in the era of large language models. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2023, page 15566. NIH Public Access.

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, et al. 2023a. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023b. Gpt-ner: Named entity recognition via large language models. *arXiv preprint arXiv:2304.10428*.

Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024. Human-llm collaborative annotation through effective verification of llm labels. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–21.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, et al. 2023. Zero-shot information extraction via chatting with chatgpt. *arXiv preprint arXiv:2302.10205*.

Ying Wei and Qi Li. 2023. Scdner: Span-based consistency-aware document-level named entity recognition. Association for Computational Linguistics.

Zhepei Wei, Jianlin Su, Yue Wang, Yuan Tian, and Yi Chang. 2019. A novel cascade binary tagging framework for relational triple extraction. *arXiv preprint arXiv:1909.03227*.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. Docred: A large-scale document-level relation extraction dataset. *arXiv preprint arXiv:1906.06127*.

Hongbin Ye, Ningyu Zhang, Hui Chen, and Huajun Chen. 2022. Generative knowledge graph construction: A review. *arXiv preprint arXiv:2210.12714*.

Bowen Zhang and Harold Soh. 2024. Extract, define, canonicalize: An llm-based framework for knowledge graph construction. *arXiv preprint arXiv:2404.03868*.

Hongxin Zhang, Weihua Du, Jiaming Shan, Qinhong Zhou, Yilun Du, Joshua B Tenenbaum, Tianmin Shu, and Chuang Gan. 2023. Building cooperative embodied agents modularly with large language models. *arXiv preprint arXiv:2307.02485*.

Weizhong Zhao, Jinyong Zhang, Jincai Yang, Tingting He, Huifang Ma, and Zhixin Li. 2021. A novel joint biomedical event extraction framework via two-level modeling of documents. *Information Sciences*, 550:27–40.

Wei Zheng, Hongfei Lin, Zhiheng Li, Xiaoxia Liu, Zhengguang Li, Bo Xu, Yijia Zhang, Zhihao Yang, and Jian Wang. 2018. An effective neural model extracting document level chemical-induced disease relations from biomedical literature. *Journal of biomedical informatics*, 83:1–9.

Lingfeng Zhong, Jia Wu, Qian Li, Hao Peng, and Xindong Wu. 2023a. A comprehensive survey on automatic knowledge graph construction. *ACM Computing Surveys*, 56(4):1–62.

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2023b. Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert. *arXiv preprint arXiv:2302.10198*.

Yijie Zhou, Xianhui Cheng, Qiming Zhang, Lei Wang, Wenchao Ding, Xiangyang Xue, Chunbo Luo, and Jian Pu. 2024. Algpt: Multi-agent cooperative framework for open-vocabulary multi-modal auto-annotating in autonomous driving. *IEEE Transactions on Intelligent Vehicles*.

Yuqi Zhu, Xiaohan Wang, Jing Chen, Shuofei Qiao, Yixin Ou, Yunzhi Yao, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2024. Llms for knowledge graph construction and reasoning: Recent capabilities and future opportunities. *World Wide Web*, 27(5):58.

A POP Algorithm

A.1 Algorithm Pseudocode

Algorithm 1 Progressive Ontology Prompting

```
function ONTOLOGYTRAVERSAL(𝒢, k)
    Initialize: queue, Order ← [], visited, Khops ← {}
    Identify the source node s
    enqueue(queue, (s, 0)), visited[s] ← true, Khops[s] ← []
    while not queue.empty() do
        (current, hop) ← dequeue(queue)
        Order.append(current)
        Gather unvisited neighbors of current into neighbors
        Sort neighbors by out-to-in ratio in descending order
        for neighbor in neighbors do
            if not visited[neighbor] then
                enqueue(queue, (neighbor, hop + 1))
                visited[neighbor] ← true
                Khops[neighbor] ← []
            end if
            if hop + 1 ≤ k then
                Khops[neighbor].append(current)
            end if
        end for
    end while
    return Order, Khops
end function
function MAIN(𝒢, k, T)
    Initialize annotations ← {}
    Order, Khops ← ONTOLOGYTRAVERSAL(𝒢, k)
    while not Order.empty() do
        current ← Order.dequeue(), context ← Khops[current]
        if not context.empty() then
            discoveries ← {annotations[t] for t in context}
            prompt ← T(current, G_sub(context), discoveries)
        else
            prompt ← T(current)
        end if
        annotations[current] ← LLM(prompt)
    end while
end function
```

A.2 Annotation Prompt Template Generation

In the POP algorithm, the annotation prompt for a specific concept node is generated from a template T, which is constructed using the node's context (the k-hop visited neighborhood) and the completed annotations within that context. We carefully design the prompt to task LLM to generate T as follows:

```
The following triplets outline an annotation ontology:{(Intervention, StudiedIn, Case Study), (Intervention, Include, Participant), (Case Study, UsedWith, Frequency)}. All concept nodes have been annotated except from {#Frequency}. Your task is to create all possible annotation prompt templates for the {#Frequency}, leveraging the ontology structure.
Example:
Ontology: [(Intervention, StudiedIn, Case Study), (Intervention, TargetAt, Disorder), (Case Study, Include, Participant), (Participant, Has, Disorder)].
Annotate: Participant
Prompts:
T1: Who is the participant in {#case study}?
T2: {#Intervention} is studied in {#case study}. Who is the participant in {#case study}?
T3: {#Intervention}, targeting {#Disorder}, is studied in {#case study}. Who is the participant in {#case study}?
```

Some example output of GPT3.5 for the above prompt is:

```
T1: What is the frequency used in {#case study}?
T2: In the context of {#intervention} studied in {#case study}, what is the frequency of the intervention?
T3: {#Intervention} is studied in {#case study} with {#Participant}. What is the frequency of the intervention in {#case study}?
```

Since template T is composed of concept nodes visited before the current node, the final annotation prompt for the current node is derived by incorporating the annotations associated with these concept nodes into T.

B Speech-Language Intervention Annotation

B.1 Literature Corpus

We cultivate a literature corpus of 64,177 research articles within the domain of speech-language pathology to facilitate intervention discovery. To conduct our literature search, we use a collection of carefully selected keywords drawn from a glossary of commonly used terms in speech-language pathology. These keywords include:

"*speech language therapy, speech language disorder, speech sound disorder, articulation disorder, speech intervention, language intervention, auditory discrimination, auditory processing disorder, phonological awareness, phonological processes, auditory perception, babbling, motor speech disorder, morpheme, phonology, prosody, stuttering, language impairment, speech language pathologist, speech and language therapist, babbling, expressive language delay, cleft speech disorder, autism spectrum disorder, developmental phonological disorder, developmental stuttering, phonological impairment, developmental dysarthria, down syndrome, swallowing disorder, communication*

Theme	Definition
speech awareness	work that involves recognizing and understanding speech sounds. It includes phonological awareness (identifying and manipulating sounds), auditory discrimination (distinguishing sounds), and sound identification (recognizing the meaning of sounds).
speech articulation	work that refers to the physical production of speech sounds using the mouth, lips, tongue, and respiratory system. It focuses on the clarity and accuracy of pronouncing phonemes and forming them into words.
comprehension	work that aims to improve the understanding of (receptive) language.
expressive language	work that aims to improve the children's expressive language, in quantity, vocabulary or structure.
self-monitoring	work designed to help the patient's awareness of their speech and language difficulties and how they might be able to overcome them.
generalisation	work to help make speech and language or therapy gains transferable to other situations and environments.
foundation skills	work to practise and improve a range of early skills, many of which might be considered foundations for speech and language development.
functional communication	work focusing on those aspects of communication that help the child's involvement and participation in life situations; this might be functional language, signing or the use of symbols.
adult understanding and empowerment	work that helps parents to understand the nature of their child's speech and language difficulty, what helps to improve it and why.
adult-child interaction	work on the interaction between the parent/adult and the child. All of the changes to adult/parent-child interactions were emphasised in terms of those that encourage speech and language development.

Table 3: Definition of intervention themes.

impairment, articulation impairment, dyslexia, apraxia, dysarthria, dysphagia, communication disorder, expressive language disorder, dyspraxia, aphasia, augmentative and alternative communication, central auditory processing disorder, cleft lip and palate, down syndrome, fluency disorders, hearing loss, orofacial myofunctional disorders, spoken language disorders, written language disorders, acquired brain injury, apraxia of speech, auditory comprehension, literacy impairments, voice difficulties, language-based learning disabilities."

B.2 Annotation Ontology

Speech-language pathology provides interventions for individuals with speech-language deficits, improving their quality of life at various stages. When choosing an intervention, evidence-based practice (EBP) is attractive as it integrates research evidence from literature into the decision-making process to ensure high-quality patient care (Law et al., 1996). Research on interventions, especially those presenting clear frameworks and comprehensive case studies, serves as valuable guidance for designing EBPs. In this paper, we apply our automated framework to speech-language intervention discovery, with the intervention ontology shown in Figure 4. A detailed explanation of the concepts within the ontology is provided below:

- **Intervention** represents a targeted treatment practice designed to enhance an individual's communication skills.
- **Disorder** represents the type of disorder that causes difficulties in an individual's voice, speech, language, or swallowing functions.
- **Setting** represents a specific environment where interventions are implemented. We identify six key settings: home, healthcare facilities (such as hospitals or rehabilitation centers), early childhood centers (like nurseries or daycare), schools, clinics and private practices, and teletherapy.
- **Theme** represents the theme of the intervention. As shown in Table 3, we categorize interventions into 10 themes based on their characteristics and therapy goals.
- **Therapy Activity** represents a task designed to address a particular speech or language challenge in an individual, such as using a minimal pairs activity to enhance phonological awareness.
- **Therapy Goal** represents a specific area that the intervention is designed to enhance.
- **Procedure** represents a comprehensive description of how the intervention is carried out.
- **Efficacy** represents the conclusion about the effectiveness of the intervention.
- **Frequency/Dosage/Duration** represents the frequency/dosage/ duration of the intervention practiced in the case study that demonstrates its efficacy.
- **Case Study** represents a detailed examination of the intervention on a particular individual or group with communication disorders. The purpose of a case study is to provide a deep understanding of the patient's unique needs and assess the intervention's effectiveness.
- **Participant** represents the individuals or populations that are involved in the case study of intervention.
- **Age** represent the age of experiment participant or claimed target population of the intervention. The age is quantified with a granularity of half a year. We additionally convert

age to age groups including *"newborn, infants, toddlers, children, preschoolers, school-age children, older children, youth, teens, adolescents, adult, young adult, middle aged, aged, senior"*. Each specific age may be associated with multiple age groups. For instance, an individual aged 13 years could be categorized into the 'teens,' 'adolescents,' and 'children' age groups.

- **Language** represents the speaking language of the experiment participant in the case study of the intervention.

C CoT & Self-Refine Baselines

Instead of using an external LLM to provide evaluation feedback to the explorer, we use CoT and SelfRefine prompting techniques as baselines to task explorer refine annotations independently. The prompts of CoT and SelfRefine are as follows:

```
Background: Your last answer to my
    question {#init annotation
    question} is: {#last annotation}.
---
*CoT:
    Prompt_CoT1: "{#Background} Make
        a plan to correctly answer my
         question again."
    Prompt_CoT2: "{#Background} Make
        a plan to answer my question
        again with more comprehensive
         results."
---
*SelfRefine:
    Prompt_SelfRefine_Feedback1: "{#
        Background} Reflect your
        answer. Analyze the
        correctness of the
        information provided. Provide
         critque to help improve the
        answer. Your feedback:"
    Prompt_SelfRefine_Refine1: "{#
        Background} Critics: {#
        feedback}. Based on your last
         answer and its critics,
        revise your answer to my
        question. Your answer:"
    Prompt_SelfRefine_Feedback2: "{#
        Background} Reflect your
        answer. Analyze the included
        aspects in your answer.
        Provide critque to help make
        the response more
        comprehensive. Your feedback:
        "
    Prompt_SelfRefine_Refine2: "{#
        Background} Critics: {#
        feedback}. Based on your last
         answer and its critics,
        revise your answer to my
        question. Your answer:"
```

D Annotation Example

Below is an example of an intervention annotation from the paper "Intensive Treatment of Dysarthria Secondary to Stroke" (Mahler and Ramig, 2012).

```
"Intervention": "Lee Silverman Voice
    Treatment (LSVT LOUD)",
"Disorder": [
        "stroke",
        "ataxia",
        "multiple sclerosis",
        "traumatic brain injury (TBI)
            "
],
"Procedure": "The therapy process of
    conducting the Lee Silverman
    Voice Treatment (LSVT LOUD)
    intervention involves intensive
    high-effort exercises aimed at
    increasing vocal loudness to a
    level within normal limits using
    healthy and efficient voice
    techniques. The treatment
    protocol includes sessions four
    times a week for 4 weeks,
    totaling 16 individual one-hour
    sessions. Each session consists
    of tasks such as maximal
    sustained vowel phonation, pitch
    range exercises, and reading
    functional phrases at individual
    target loudness levels. The
    second half of each session
    progresses to functional speech
    tasks, moving from words and
    phrases to conversation over the
    course of the 16 sessions.
    Additionally, participants are
    assigned daily homework to
    practice using normal loudness
    and facilitate generalization of
    normal loudness outside the
    treatment room.",
"Frequency": "four times a week",
"Dosage": "one-hour session.",
"Duration": "4 weeks, totaling 16
    individual one-hour sessions.",
"Efficacy": "The outcome of the Lee
    Silverman Voice Treatment (LSVT
    LOUD) intervention in this study
    showed positive changes in
    acoustic variables of speech for
    all participants with dysarthria
    secondary to stroke. There were
    statistically significant
    increases in vocal dB SPL for
    sustained vowel phonation and
    speech tasks, indicating
    improvements in loudness levels
    and phonatory stability.
    Additionally, post-treatment
    speech samples were rated as
    having better voice quality and
    sounding more natural, suggesting
     an amelioration of dysarthria
    characteristics. Participants
    also reported increased
    confidence in their speech during
```

Disorder Name	Number of Interventions	Intervention Examples
Aphasia	202	Phonological therapy, Semantic therapy, Syntax Stimulation Program, Melodic Intonation Therapy (MIT), Multimodal Speech Therapy with tDCS, Cross-Language Generalization Therapy (CLGT), Word Learning Paradigm
Autism Spectrum Disorder	178	Personalized Idiom Intervention (PII), Classroom-wide peer tutoring, Idiom Isolation Intervention, Hanen More Than Words, Parent-Mediated Communication-Focused Treatment, Picture Exchange Communication System (PECS)
Dysphagia	129	Swallowing Maneuver Therapy, Focal Vibration Therapy (FVT), Oral Neuromuscular Training and Vibrational Therapy, Prophylactic Swallowing Intervention, High-speed jaw-opening exercise, Palatal Augmentation Prostheses (PAP)
Stuttering	91	Lidcombe Program, Syllable-timed speech, Electronic devices for stuttering, Computer software for stuttering, Bone Conduction Delayed Feedback Therapy, Fluency Techniques and Fear Reduction, Cognitive Behavioral Therapy (CBT)
Phonological Disorder	74	Nonlinear Phonological Intervention Program, Metronome-paced Speech Therapy, Phonological Awareness and Articulatory Training (PAAT), Phonological Meaning Therapy, Motor-based Intervention Approach

Table 4: Intervention-disorder examples in our discoveries.

AgeGroup	Number of Interventions	Intervention Examples
Children	745	Pharyngeal flap procedure, National Health Service (NHS) 1-week intensive course, Ultrasound Visual Biofeedback (U-VBF), Intensive Speech Therapy, Community-Based Speech Therapy Models, Early Vocal Intervention, Auditory-Verbal Therapy (AVT)
School-age Children	571	Intensive Speech Therapy, Early Vocal Intervention, APD intervention, Auditory-Verbal Therapy (AVT), Multisensory Stimulation Therapy, Oral Functional Training (OFT), Rhythmic Reading Training (RRT), Rapid Syllable Transition Treatment (ReST)
Older children	366	Semantic Categorization Therapy, Early Vocal Intervention, Rhythmic Reading Training (RRT), Speech Bulb Reduction Program, Intensive speech and language therapy, Peer-Mediated Intervention, Lidcombe Program, Oral Functional Training (OFT)
Preschoolers	347	Early Vocal Intervention, Treatment-as-usual, The Lidcombe Program, Oral Functional Training (OFT), Speech Production Therapy with Reward System, Phonological Interventions and Contrast Therapy, Cycles Phonological Remediation Approach
Adult	319	Pharyngeal flap procedure, Linguistic Retrieval Therapy (LRT), Oral Hydration Intervention, Physiologic Swallowing Therapy, Myofunctional Intervention (OMT), Orthognathic speech therapy, Eye-Tongue Movement Training, Behavioral Voice Therapy

Table 5: Intervention-AgeGroup examples in our discoveries.


```
    post-treatment interviews.",
"Therapy Goal": [
        "Increase vocal loudness to a
            level within normal
            limits",
        "Use healthy and efficient
            voice techniques",
        "Progress from words and
            phrases to conversation
            over 16 sessions",
        "Facilitate generalization of
            normal loudness outside
            the treatment room"
],
"Participant": "Four participants (P0
    1 to P04).",
"Age": "Participants in the study
    ranged in age from 50 to 74 years
    .",
"Language": "assume to be English.",
"Case Study": "Case studies and
    experiments regarding the Lee
    Silverman Voice Treatment (LSVT
    LOUD) intervention in the paper
    include studies on people
    diagnosed with stroke, ataxia,
    multiple sclerosis, and traumatic
     brain injury (TBI). These
    studies have shown improvement in
     articulatory features as well as
     loudness.",
"Therapy Activity": [
        "Maximal sustained vowel
            phonation",
        "Pitch range exercises",
        "Reading 10 functional
            phrases at individual
            target loudness levels",
        "Functional speech tasks
            progressing from words
            and phrases to
            conversation"
],
"Setting": "home",
"Theme": "speech articulation"
```


E Speech-Language Intervention Knowledge Base

E.1 Intervention Examples

We separately provide some intervention-disorder and intervention-age group examples in Table 4 and Table 5.

E.2 Knowledge Graph Overview

Through our automated KGC framework, we constructed the first intervention knowledge graph in the speech-language pathology domain. Our intervention KG contains 33,148 nodes with 324,707 relations. We present some views of our intervention KG in Figure 7 and Figure 8.

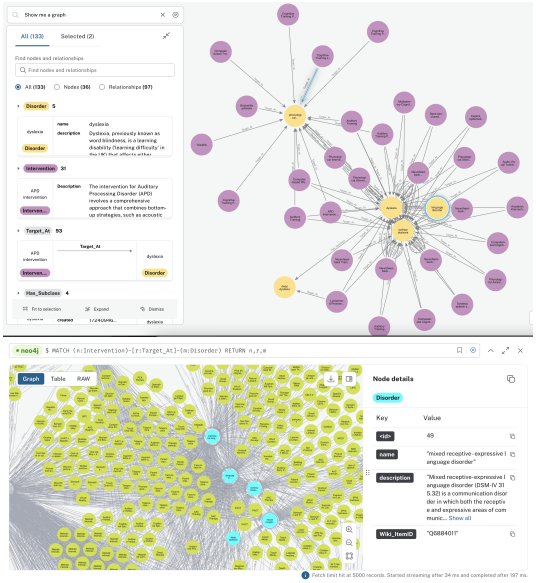


Figure 7: Partial views of Intervention-Disorder in the intervention KG.

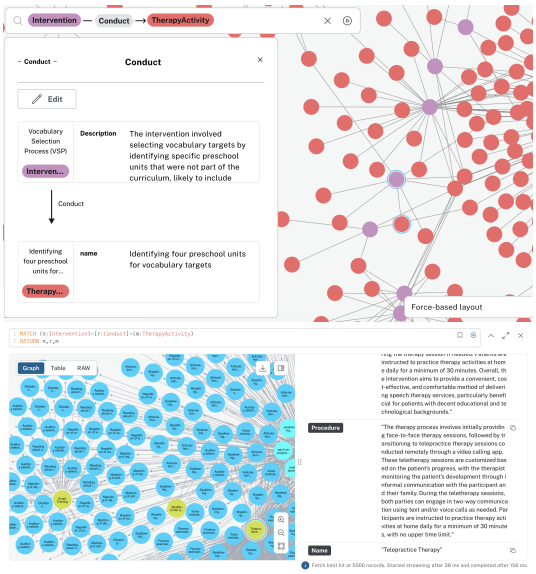


Figure 8: Partial views of Intervention-TherapyActivity in intervention KG.