Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25)
Special Track on AI and Social Good

Automating Intervention Discovery from Scientific Literature: A Progressive Ontology Prompting and Dual-LLM Framework

Yuting Hu[1], **Dancheng Liu**[1], **Qingyun Wang**[2], **Charles Yu**[2], **Chenhui Xu**[1], **Qingxiao Zheng**[1], **Heng Ji**[2] and **Jinjun Xiong**[1*]
[1]University at Buffalo
[2]University of Illinois at Urbana-Champaign
{yhu54,dliu37,cxu26,qingxiao,jinjun}@buffalo.edu, {qingyun4,ctyu2,hengji}@illinois.edu

Abstract

Identifying effective interventions from the scientific literature is challenging due to the high volume of publications, specialized terminology, and inconsistent reporting formats, making manual curation laborious and prone to oversight. To address this challenge, this paper proposes 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

Evidence-based interventions refer 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]. Intervention discovery from scientific literature enables researchers to keep abreast of the latest advancements and facilitate valuable insights that can significantly enhance the healthcare quality [Usai *et al.*, 2018][Wang *et al.*, 2023a]. However, due to the labor-intensive nature of human review, only a small fraction of intervention knowledge is systematically collected and curated. In healthcare, 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, large language models (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 [Achiam *et al.*, 2023][Li *et al.*, 2024a]. Many studies have utilized LLMs to streamline various subtasks involved in knowledge graph construction, such as named entity recognition (NER), relation extraction (RE), event extraction (EE), and entity linking (EL) [Wang *et al.*, 2023b][Zhu *et al.*, 2024]. 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, extracting intervention knowledge from long-range, domain-specific literature remains a significant challenge. On the one hand, developing human-annotated datasets for training deep learning models in NER and RE tasks requires specialized domain expertise to accurately interpret the literature. On the other hand, mining 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]. Particularly in healthcare contexts, these challenges are further compounded by the need for specialized therapeutic expertise, labor-intensive manual annotation, and difficulties in maintaining consistency and scalability [Zhao *et al.*, 2021]. In this context, LLMs offer a promising alternative through in-context learning, enabling scalable information extraction without the need for extensive human-labeled data. Despite these advancements, fully automated knowledge graph construction remains a challenge, particularly when dealing with long-range documents. Most cur-

*Corresponding Author.
Project website: slp-kg.xlabub.com

rent knowledge graph construction approaches focus on short texts, leaving significant potential for further development in handling more complex, lengthy content.

In this paper, we address the challenge of automating intervention discovery via LLMs by formulating it as a prompt design and annotation scheduling problem with a predefined intervention ontology graph structure and designing a framework leveraging two LLM agents to iteratively enhance the annotation quality. We introduce a progressive ontology prompting (POP) algorithm that employs an outdegree-prioritized breadth-first search (BFS) across the intervention 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, it integrates two LLM agents working both collaboratively and adversarially to refine annotation generation.

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 intervention discovery framework with several advanced 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 intervention discovery 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.

Related Work. Traditional approaches to automated knowledge discovery 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.*, 2023][Laurenzi *et al.*, 2024]. Recent advancements leverage LLMs to generate relational triplets in zero/few-shot settings for knowledge graph construction, 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 knowledge graph construction 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. Moreover, domain-specific knowledge often exhibits complex patterns that cannot be captured solely through sentence-level syntactic structures. As a result, most existing approaches [Du *et al.*, 2020][Rossanez *et al.*, 2020][Alam *et al.*, 2023] are limited to handling abstracts and fail to extract and summarize knowledge across long-range contexts.

2 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 [Peng *et al.*, 2023]. For intervention knowledge discovery 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 populating the intervention ontology to create the intervention knowledge graph.

In our methodology, the intervention 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 extract intervention knowledge anchored to $\mathcal{G}$ automatically, the design of annotation prompts and the query sequences plays a crucial role. We thereby frame the problem of automated intervention knowledge discovery 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 intervention 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 whole KG schema within the prompt such as the annotation methods used in [Mihindukulasooriya *et al.*, 2023][Kommineni *et al.*, 2024]. 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.

3 Methodology

In this section, we first introduce the POP algorithm that converts an intervention 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.

3.1 Progressive Ontology Prompting

We develop a progressive ontology prompting (POP) algorithm that employs a prioritized BFS on the intervention 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:

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

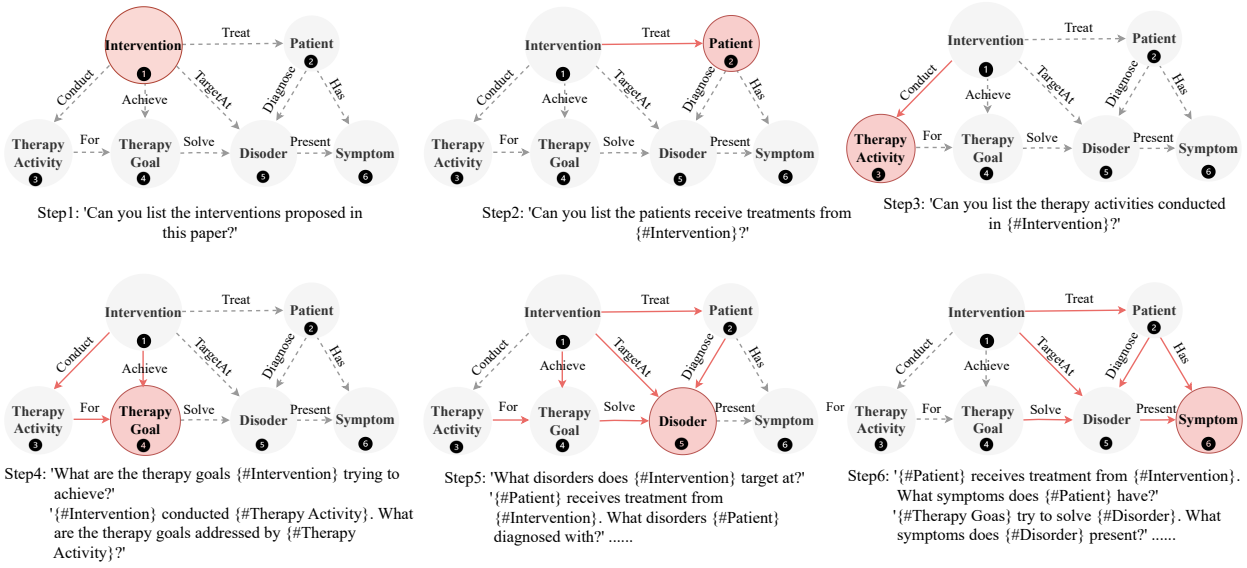


Figure 1: Illustration of prompt design and scheduling based on the progressive ontology prompting algorithm.

Our algorithm selects the neighboring node with the maximum R value to visit in the next step. For instance, in the example of Figure 1, visiting the 'Patient' node before the 'Disorder' node provides 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 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.

Our algorithm first follows prioritized BFS traversal to capture the local k hop context and visit order of concept node v, then composes the annotation prompts $Prompt_v$ based on its ontology substructure $N_k(v)$ and completed annotations within its context, which 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.

3.2 LLM-Duo Annotation Framework

To guarantee the integrity and reliability of LLM annotations, we propose LLM-Duo, a dual-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]. Drawing inspiration from the multi-agent debate idea in [Kim *et al.*, 2024b],

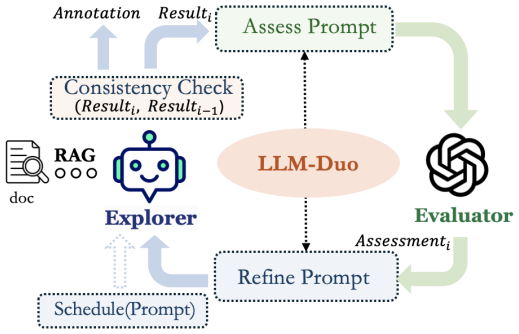


Figure 2: Iterative annotation with two LLM agents under the LLM-Duo framework.

we design a framework where agents work both collaboratively and adversarially to enhance the quality of annotations.

The architecture of LLM-Duo is shown in Figure 2, featuring two LLM agents: the *explorer* and the *evaluator*. The *explorer* is a chatbot performing annotation tasks using zero-shot question answering (QA). To break the context window limit of LLMs and ensure the generated annotations are faithful to the provided literature content, RAG is employed in *explorer* to reference relevant sources, reducing LLM hallucinations. To improve the accuracy and reliability of the annotations, the *evaluator* LLM is set up to review and validate the *explorer's* responses, ensuring higher-quality results.

LLM-Duo will be tasked with annotation prompts following the sequential order generated by the POP algorithm. During each annotation cycle for a specific concept node, when focusing on concepts that emphasize rationality (e.g., disorder, intervention efficacy), the *explorer* provides an answer and an explanatory rationale to the *evaluator*. The *evaluator* then reviews the reasoning and offers feedback. Based

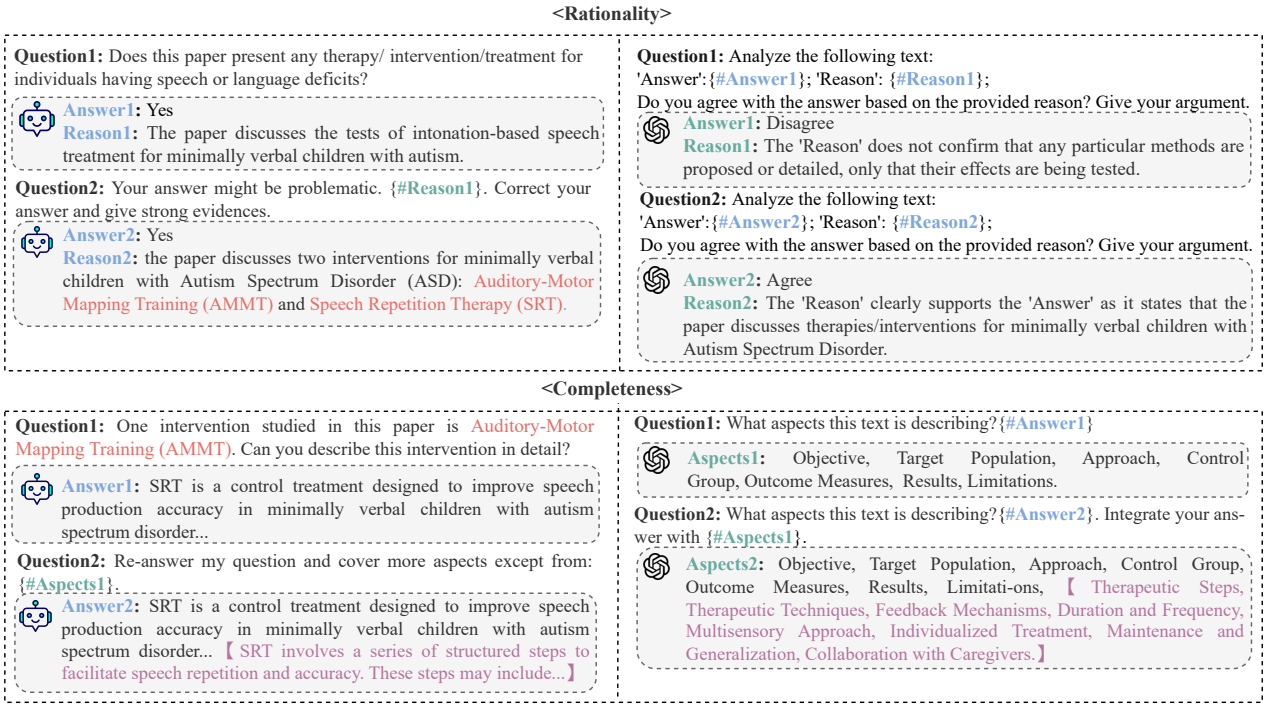


Figure 3: Annotation examples of speech-language intervention discovery using the LLM-Duo framework.

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 (e.g., intervention procedure, therapy activity), 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.

4 Experiments

4.1 Implementation

For LLM-Duo, 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 SentenceTransformerRerank[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*.

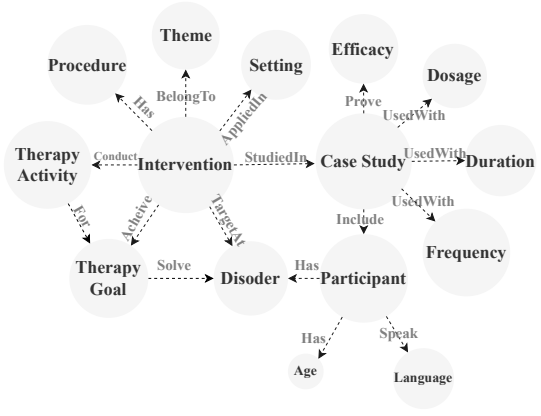


Figure 4: Ontology of speech-language intervention.

[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

4.2 Case Study: Speech-language Intervention Discovery

Speech-language therapy provides interventions for individuals with speech-language deficits, enhancing their quality of life across various life 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]. Intervention research, especially studies that offer clear intervention frameworks and comprehensive case studies, are valuable references to guide EBP designs. Intervention discovery aims to extensively gather speech-language interventions from the literature corpus as references to facilitate EBP design. It involves identifying relevant studies and extracting essential features of interventions including target disorder, procedure, efficacy, case study, therapy activity, etc., which is extremely labor-intensive for human reviewers, highlighting the efficiency of automating knowledge discovery based on LLMs.

To verify the effectiveness of our method in a realistic scenario, we employ our framework in a speech-language intervention discovery setting. The ontology is shown in Figure 4. To enable a large-scale discovery, we cultivate a literature base including 64,177 papers within the domain of speech-language therapy.

4.3 Annotation Baselines

The core idea of our automated intervention framework is to leverage the POP algorithm for guiding the annotation process while utilizing LLM-Duo to refine initial annotations by incorporating 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.*, 2024][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 explorer chatbot based on RAG with these two prompting methods for annotation and denote them as CoT-RAG and Self-Refine-RAG. 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.

4.4 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 the annotation generation achieving 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.); 4) Faithfulness (Faith): the extent of the annotation faithful to the provided literature literature, which is measured by FaithfulnessEvaluator[6] of Llamaindex. 5) Accuracy (ACC): the percentage of correct annotations in all LLM-generated annotations. 6) Cover: the percentage of correct LLM-generated annotations to the total mentioned concept entities in the provided literature.

5 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 intervention discovery framework.

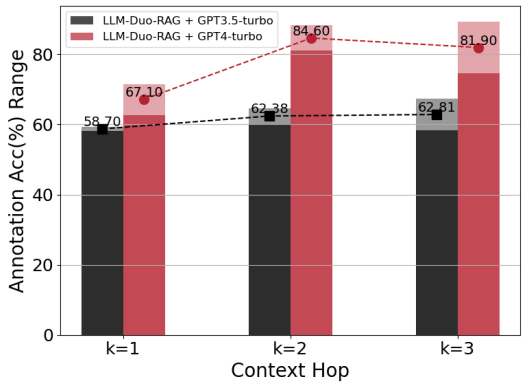
5.1 POP Algorithm Study

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

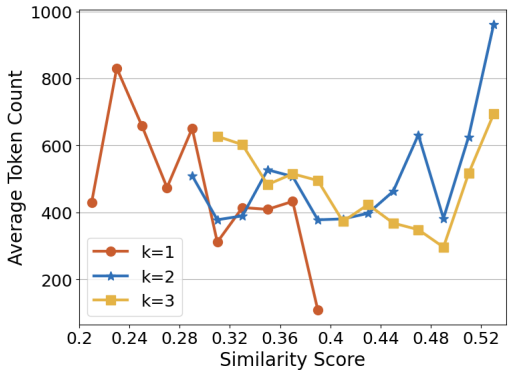
The annotation accuracy is shown in Figure 5a. The results indicate that as the context size k increases, annotation accuracy improves significantly, suggesting that a larger context provides more informative prompts, thereby enhancing annotation quality. Moreover, GPT-4-turbo consistently outperforms 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 informative annotation queries based on various k values. We report the range of similarity scores and token count distribution of retrieved-back chunks in Figure 5b. 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.

Prioritized BFS Analysis. 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 the speech-language intervention ontology is 'TherapyActivity'→'TherapyGoal'→'Disorder'. In this section, we compare the 'Disorder' annotation results using the POP algorithm with and without prioritization, as

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



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



(b) 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.

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 1: Comparison of annotation results with baselines using different LLMs.

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 2. We can observe that applying both the POP and prioritized BFS notably enhances annotation performance.

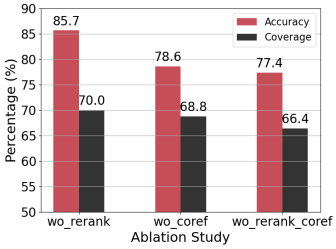
LLM-Duo-RAG	**GPT3.5-turbo**	**GPT4-turbo**
POP✗	68.18	72.73
POP✓ Prioritized-BFS✗	77.28	83.20
POP✓ Prioritized-BFS✓	**81.82**	**86.37**

Table 2: Comparison of annotation results with and without the POP and prioritized BFS.

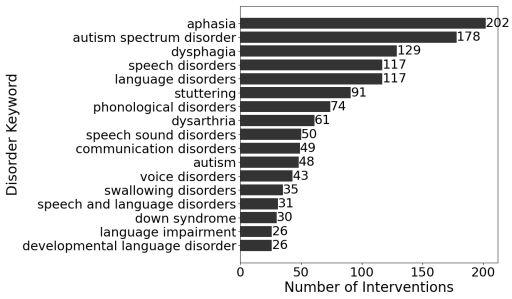
5.2 LLM-Duo with Baselines

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.

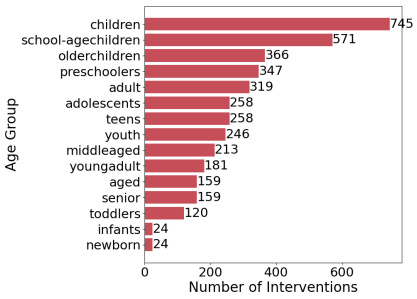
The experimental results are reported in Table 1. It should be noted that we implemented 'ShortContext' using Llama3-instruct-70b (FP16) and Mistral-instruct-8x22b models (INT8). However, directly prompting these models with full paper text fails to produce annotations in a zero-shot QA setting. Their generations do not align with the annotation questions. The results in Table 1 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 intervention annotations. Additionally, compared with simple RAG, self-correct prompting methods such as CoT and Self-



(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.

Disorder	**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
AgeGroup	**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 3: Intervention-disorder examples in our discoveries.

Refine can significantly enhance intervention 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.

5.3 Ablation Study

In our implementation, the RAG technique serves as the backbone of *explorer*. We employ 'FastCoref' 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.

5.4 Speech-Language Intervention Discovery

Through our automated intervention discovery framework, we identified 2,421 interventions supported by case studies from 64,177 literature in the speech-language pathology domain. The statistics of discovered interventions are presented in Figure 6b and Figure 6c. More intervention examples are provided in Table 3. 19 clinicians and students reviewed our annotations through online Google forms. We have constructed the first intervention knowledge graph in the speech-language pathology domain, which will be made publicly accessible upon acceptance. 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.

6 Conclusion

In this paper, we developed a novel LLM-based framework for automatic intervention discovery from literature, featuring a progressive ontology prompting algorithm and a dual-agent system. The proposed method achieves superior performance compared with advanced baselines, enabling more accurate intervention discoveries. Our approach is adaptable to various intervention ontologies in healthcare and offers practical value to improve healthcare quality.

Acknowledgments

The research was supported, in part, by the National AI Institute for Exceptional Education (NSF Award #2229873), Center for Early Literacy and Responsible AI (IES Award #R305C240046), FuSe-TG (NSF Award #2235364) and SaTC (NSF Award #2329704). The opinions expressed are those of the authors and do not represent the views of any sponsors.

References

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

[Alam *et al.*, 2023] Fakhare Alam, Hamed Babaei Giglou, and Khalid Mahmood Malik. Automated clinical knowledge graph generation framework for evidence based medicine. *Expert Systems with Applications*, 233:120964, 2023.

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

[Du *et al.*, 2020] Jian Du, Xiaoying Li, et al. A knowledge graph of combined drug therapies using semantic predications from biomedical literature: Algorithm development. *JMIR medical informatics*, 8(4):e18323, 2020.

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

[He *et al.*, 2024] Li He, Hayilang Zhang, Jie Liu, Kang Sun, and Qing Zhang. Zero-shot relation triplet extraction via knowledge-driven llm synthetic data generation. In *International Conference on Intelligent Computing*, pages 329–340. Springer, 2024.

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

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

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

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

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

[Li *et al.*, 2019] Shuangjie Li, Wei He, Yabing Shi, Wenbin Jiang, Haijin Liang, Ye Jiang, Yang Zhang, Yajuan Lyu, and Yong Zhu. 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, 2019.

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

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

[Liu *et al.*, 2024] Dancheng Liu, Amir Nassereldine, Ziming Yang, Chenhui Xu, Yuting Hu, Jiajie Li, Utkarsh Kumar, Changjae Lee, and Jinjun Xiong. Large language models have intrinsic self-correction ability, 2024.

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

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

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

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

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

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

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

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

[Rossanez *et al.*, 2020] Anderson Rossanez, Julio Cesar Dos Reis, Ricardo da Silva Torres, and Hélène de Ribaupierre. Kgen: a knowledge graph generator from biomedical scientific literature. *BMC medical informatics and decision making*, 20:1–24, 2020.

[Rutten *et al.*, 2021] 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. Evidence-based strategies for clinical organizations to address covid-19 vaccine hesitancy. In *Mayo clinic proceedings*, volume 96, pages 699–707. Elsevier, 2021.

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

[Sun *et al.*, 2024] Qi Sun, Kun Huang, Xiaocui Yang, Rong Tong, Kun Zhang, and Soujanya Poria. 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, 2024.

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

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

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

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

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

[Wang *et al.*, 2024] Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 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, 2024.

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

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

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

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

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

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

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

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

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

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