

# DALK: Dynamic Co-Augmentation of LLMs and KG to answer Alzheimer’s Disease Questions with Scientific Literature

Anonymous ACL submission

## Abstract

Recent advancements in large language models (LLMs) have achieved promising performances across various applications. Nonetheless, the ongoing challenge of integrating long-tail knowledge continues to impede the seamless adoption of LLMs in specialized domains. In this work, we introduce DALK, a.k.a. Dynamic Co-Augmentation of LLMs and KG, to address this limitation and demonstrate its ability on studying Alzheimer’s Disease (AD), a specialized sub-field in biomedicine and a global health priority. With a synergized framework of LLM and KG mutually enhancing each other, we first leverage LLM to construct an evolving AD-specific knowledge graph (KG) sourced from AD-related scientific literature, and then we utilize a coarse-to-fine sampling method with a novel self-aware knowledge retrieval approach to select appropriate knowledge from the KG to augment LLM inference capabilities. The experimental results, conducted on our constructed AD question answering (ADQA) benchmark, underscore the efficacy of DALK. Additionally, we perform a series of detailed analyses that can offer valuable insights and guidelines for the emerging topic of mutually enhancing KG and LLM.

## 1 Introduction

Alzheimer’s Disease (AD) is a neurodegenerative disorder characterized by progressive declines in cognitive and functional status over a span of decades (Report, 2023). However, current AD therapy developments are facing critical challenges due to the lack of knowledge and understanding of the underlying etiological mechanisms of the disease. Although scientific literature and dedicated biomedical databases could supply rich sources of AD knowledge, manual review of relevant information is impossible due to the large volume.

As large language models (LLMs) (Brown et al., 2020; Zhang et al., 2022; Anil et al., 2023; Tou-

vron et al., 2023) demonstrate strong language capabilities across various tasks, there have been attempts to leverage LLMs-based systems in general biomedical and AD-related applications (Mao et al., 2023; Li et al., 2023b; Yan et al., 2024; Feng et al., 2023). However, while the LLMs have shown promising performances in many general tasks, recent studies revealed LLMs’ limitations in long-tail (Kandpal et al., 2023) and domain-specific (Li et al., 2023a, 2024) knowledge, thereby significantly impeding their adaptations in vertical fields such as AD. To deal with this issue, the most common strategies are retrieval augmented generation (RAG) and domain-specific LLMs training.

Nevertheless, directly applying these strategies in the context like AD would still suffer from several issues. First, **Data Quality**: As in many biomedical fields, scientific literature composes the largest publicly available corpus source in AD. Yet, the dense and information-overloaded nature of scientific literature, when combined with automatic retrieval methods, can lead to the retrieval of irrelevant and noisy information. Previous research has shown that noisy and irrelevant corpora can significantly undermine the performance of LLMs (Yu et al., 2023; Chen et al., 2024; Wu et al., 2024). Second, **Efficiency & Scale Issues**: Being an critical field of research, the knowledge of AD is rapidly evolving with scientific advancements at a remarkable pace and scale. However, retraining a domain-specific LLM or updating certain knowledge in it demands substantial computational resources (Hu et al., 2021; Ovadia et al., 2023). This efficiency issue would also limit the sizes of domain-specific LLMs, consequently affecting their performances.

To tackle these limitations, here we propose a Dynamic Co-Augmentation of LLMs and KG (DALK) framework that facilitates mutual benefits between LLMs and knowledge graphs (KG) for the AD domain. Initially, our framework addresses the data quality challenge by extracting more struc-

042  
043  
044  
045  
046  
047  
048  
049  
050  
051  
052  
053  
054  
055  
056  
057  
058  
059  
060  
061  
062  
063  
064  
065  
066  
067  
068  
069  
070  
071  
072  
073  
074  
075  
076  
077  
078  
079  
080  
081  
082

083 tural and accurate knowledge from unstructured  
084 and dense scientific literature and constructing a  
085 domain-specific knowledge graph tailored to AD.  
086 We employ two widely utilized knowledge graph  
087 construction methods, namely pair-wise construc-  
088 tion (Carta et al., 2023; Wadhwa et al., 2023) and  
089 generative construction (Han et al., 2023; Bi et al.,  
090 2024), to comprehensively assess their impact on  
091 knowledge graph quality. Then, we adopt a coarse-  
092 to-fine sampling method with a novel self-aware  
093 knowledge retrieval approach to select appropri-  
094 ate knowledge from the knowledge graph and thus  
095 further address the data quality problem. Notably,  
096 the tuning-free nature of our framework signifi-  
097 cantly enhances efficiency and facilitates its ap-  
098 plication in large-scale and API-based language  
099 models (OpenAI, 2022). In the evaluation sec-  
100 tion, we derive an Alzheimer’s Disease question  
101 answering (ADQA) benchmark from existing gen-  
102 eral medical QA datasets with millions of sam-  
103 ples filtered by a curated keyword list and self-  
104 sampling of LLMs. Our extensive experiment on  
105 ADQA demonstrates the effectiveness of our frame-  
106 work in domain-specific applications compared  
107 with general biomedical LLMs and retrieval aug-  
108 mented models. Further evaluation and analysis  
109 provide valuable insights into constructing high-  
110 quality knowledge graphs and sampling accurate  
111 knowledge from them.

112 In summary, our contribution in this work can  
113 be summarized as follows:

- 114 • We identify the constraints of the current methods  
115 for LLMs in domain-specific areas like AD and  
116 introduce DALK, a co-augmentation framework  
117 of the LLM and KG to address these issues.
- 118 • We build AD-specific KG and QA benchmark.  
119 Through extensive comparisons with other meth-  
120 ods, we showcase the effectiveness of DALK.
- 121 • We delve into a comprehensive analysis of our  
122 proposed method and provide valuable insights  
123 and guidance on how to construct a high-quality  
124 KG and sample accurate knowledge from it.

## 125 2 Related Work

126 **The interplay between LLMs and KGs**  
127 KGs (Miller, 1995; Speer et al., 2017; Vrandečić  
128 and Krötzsch, 2014) serve as structured represen-  
129 tations of factual knowledge, typically expressed  
130 as (head, relation, tail) triples. Their structured,  
131 factual, and interpretable nature renders them ex-  
132 cellent complements to parametric language mod-

133 els (Pan et al., 2024). Recently, with the rise of  
134 large language models (LLMs), numerous studies  
135 have delved into exploring the synergy between  
136 LLMs and KGs for various purposes. There are a  
137 lot of efforts in conducting knowledge graph con-  
138 struction (Carta et al., 2023; Wadhwa et al., 2023;  
139 Han et al., 2023; Bi et al., 2024; Datta et al., 2024),  
140 completion (Wei et al., 2023; Zhang et al., 2023b;  
141 Li et al., 2024) with the aid of LLMs. Conversely,  
142 other works aim to enhance LLMs by integrating  
143 knowledge sampled from KGs during both train-  
144 ing (Tang et al., 2023; Luo et al., 2024; Dernbach  
145 et al., 2024; Rangel et al., 2024) and inference (Kim  
146 et al., 2023; Wen et al., 2023; Jiang et al., 2023;  
147 Sun et al., 2023a) times. Our work distinguishes  
148 itself by proposing a co-augmentation framework  
149 for LLMs and KGs, facilitating their mutual en-  
150 hancement, and applying it to the domain of AD.

151 **LLMs and KGs for AD research** LLMs and  
152 KGs have both been applied to Alzheimer’s Dis-  
153 ease research in previous studies. Pre-trained lan-  
154 guage models are utilized to work on AD de-  
155 tection and many other related tasks based on  
156 speech recordings and transcripts (Balagopalan  
157 et al., 2020; Agbavor and Liang, 2022), electronic  
158 health records (EHRs) (Mao et al., 2023; Li et al.,  
159 2023b; Yan et al., 2024), and tabular data (Feng  
160 et al., 2023). KGs have been widely used in  
161 biomedical research, yet only a few are specifically  
162 for AD research (Romano et al., 2023; Pu et al.,  
163 2023; Hsieh et al., 2023; Nian et al., 2022; Daluwa-  
164 tumulle et al., 2023). These KGs were generally  
165 constructed from a variety of information derived  
166 from heterogeneous biomedical databases (e.g. for  
167 genes, drugs, pathways, etc.) or scientific literature  
168 related to AD. Despite the aforementioned efforts  
169 for LLMs and KGs in AD research, no prior study  
170 has explored using LLM to augment AD-KG, or  
171 vice versa, let alone the potential for mutual en-  
172 hancement between the two as we propose here.

## 173 3 Our Methodology

174 This section elaborates on our dynamic co-  
175 augmentation framework of LLMs and KG. Sec-  
176 tion 3.1 presents the details of augmenting an AD-  
177 specific evolving KG with LLMs and literature  
178 corpus in a time-slicing fashion (i.e. year by year).  
179 Following it, Section 3.2 describes the process of  
180 sampling appropriate knowledge from the evolving  
181 KG to enhance LLMs’ reasoning. Figure 1 illus-  
182 trates an overall pipeline of our method DALK.

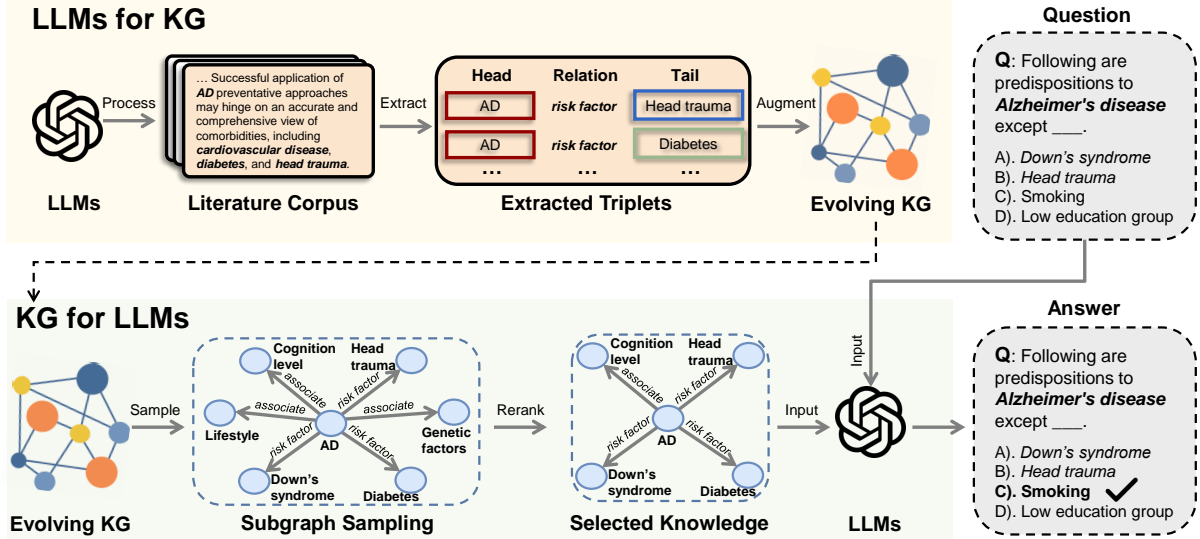


Figure 1: The overview pipeline of DALK. We first extract structural knowledge from unstructured corpora and construct a domain-specific knowledge graph tailored to AD (Section 3.1). Then, we utilize a coarse-to-fine sampling method with a novel self-aware knowledge retrieval approach to select appropriate knowledge from the knowledge graph (Section 3.2).

	$KG_{pair}$	$KG_{gen}$
#Corpus	9,764	9,764
#Nodes	13,509	20,545
#Relations	3,952	3,651
#Triples	171,431	53,585

Table 1: Detailed statistics about our augmented knowledge graph.

### 3.1 LLMs for KG

**Corpus Collection** To create an AD-specific knowledge graph, we follow (Pu et al., 2023) and use the AD corpus collected by a domain expert Professor Colin Masters at the University of Melbourne who discovered amyloid proteins being the potential cause of AD (Masters et al., 1985). The corpus is based on his extensive bibliography of representative AD-related papers and consists of more than 16K PMID (PubMed ID)-indexed articles from 1977 to 2021. For our study, we focus on the papers since 2011 which reflect the most recent knowledge in the field and get 9,764 articles.

**Entity Recognition** In order to identify knowledge at the proper granularity level for AD, we extract relevant entities from the corpus by utilizing the PubTator Central (PTC) (Wei et al., 2013) developed and continuously maintained by NCBI. PTC is a widely-used tool to provide state-of-the-art annotations of biomedical concepts for PubMed

abstracts and full-text articles, and it supports six bioconcept types including genes, diseases, chemicals, mutations, species and cell lines. We apply PTC to the abstracts of all our AD papers and obtain the relevant named entities which will serve as nodes in the knowledge graph.

**Relation Extraction** To build an accurate and high-quality knowledge graph on AD, we aim to assign a specific relation type between the two related entities. Through a comprehensive survey of relation extraction methods for knowledge graph construction, we categorize current approaches with LLMs into two main groups: (a). **Pair-wised Relation Extraction** (Carta et al., 2023; Wadhwa et al., 2023) aims to prompt the LLMs to describe the relationship between any two entities in a segment of text. (b). **Generative Relation Extraction** (Han et al., 2023; Bi et al., 2024; Datta et al., 2024), where LLMs directly output all related entity pairs and their corresponding relationships. As shown in Figure 2, we incorporate both of these relation extraction methods into our knowledge graph augmentation process to provide a comprehensive comparison between them. We denote the resulting knowledge graphs from these approaches as  $KG_{pair}$  and  $KG_{gen}$  respectively.

Table 1 presents the detailed statistics about our augmented knowledge graph, including the number of corpora we used, and the number of nodes,

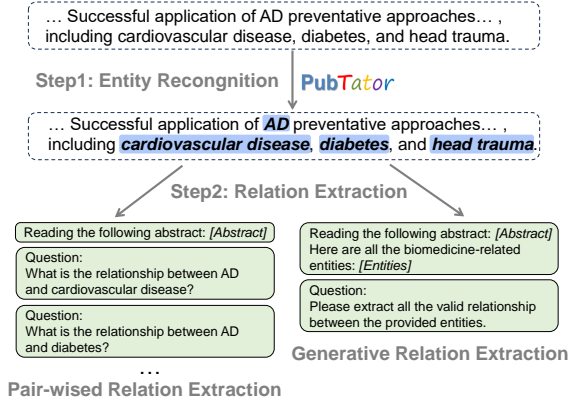


Figure 2: The detailed process of AD-specific KG construction.

relations and triples in  $KG_{pair}$  and  $KG_{gen}$ .

## 3.2 KG for LLMs

In this section, we begin by outlining our process for sampling coarse-grained augmented knowledge from our evolving knowledge graph (Section 3.2.1). Subsequently, we delve into detail regarding our self-aware knowledge retrieval method, which aims to filter out noise and retrieve the most pertinent knowledge to provide to the LLM (Section 3.2.2).

### 3.2.1 Coarse-grained Knowledge Sample

Given a question query  $Q$ , we first construct a prompt and ask LLMs to extract all the domain-specific entities  $E = \{e_1, e_2, \dots\}$  from it. Afterward, we adhere to the methodology proposed by Wen et al. (2023) and execute a similarity-based entity linking process to connect all entities within  $E$  to the entity structure in our knowledge graph  $G$ . Specifically, we employ a semantic similarity model (Reimers and Gurevych, 2019) to encode all entities in  $G$  and  $E$  into dense embeddings, denoted as  $H_G$  and  $H_E$ , respectively. Subsequently, utilizing cosine similarity, we establish links between each entity in  $E$  and its nearest neighbor entity in  $G$ . This procedure yields an initial entity set  $E_G$  for the subsequent knowledge sampling step.

To build an evidence sub-graph to boost LLMs’ reasoning process, we follow the previous study (Wen et al., 2023) and consider the following two kinds of explorations in our AD-KG:

**Path-based Exploration** entails the extraction of a sub-graph from  $G$  to encompass all entities within  $E_G$ . The process unfolds as follows: (a) Begin by selecting one node from  $e_Q^0$  as the initial node, denoted as  $e_1$ , and place the remaining nodes into a

candidate node set,  $E_{cand}$ . Explore at most  $k$  hops from  $e_1$  to identify the subsequent node,  $e_2$ , where  $e_1 \in E_{cand}$ . If  $e_2$  is successfully reached within  $k$  hops, update the start node to  $e_2$  and remove  $e_2$  from  $E_{cand}$ . In the event  $e_2$  cannot be found within  $k$  hops, concatenate the segment paths acquired thus far and store them in  $G_Q^{path}$ . Subsequently, choose another node  $e'_1$  from  $V_{cand}$  as the new start node, and eliminate both the original start node  $e_1$  and the current node  $e_2$  from  $E_{cand}$ . (b) Verify if  $E_{cand}$  is empty. If not, repeat step (a) to identify the next segment of the path. If  $E_{cand}$  is empty, combine all segments to construct a set of sub-graphs and place them into  $G_Q^{path}$ .

**Neighbor-based Exploration** endeavors to augment the evidence relevant to the query within  $G_Q$ . This process consists of two steps: (a) Initially, expand each node  $e$  within  $E_G$  by 1-hop to incorporate their neighbors  $e'$ , thus appending triples  $(e, r, e')$  to  $G_Q^{nei}$ . (b) Then assess whether each  $e'$  exhibits semantic relevance to the query. If affirmative, further expand the 1-hop neighbors of  $e'$ , consequently adding triples  $(e_{nei}, r', e')$  to  $G_Q^{nei}$ .

After obtaining the two sub-graphs  $G_Q^{path}$  and  $G_Q^{nei}$ , we perform post-processing to further prune redundant information in sub-graphs and prompt LLMs to describe the structure of each sub-graph.

### 3.2.2 Self-aware Knowledge Retrieval

In our initial experiment, we noticed the coarse-grained knowledge sampled with the above-mentioned approaches still contained redundant and irrelevant information. This issue of noise is a common challenge encountered in automatically-constructed knowledge graphs (Fang et al., 2021; Zhang et al., 2020; Li et al., 2022; Bi et al., 2024). Moreover, many recent works (Yu et al., 2023; Li et al., 2023c; Chen et al., 2024; Wu et al., 2024) have demonstrated LLMs can indeed be influenced by such noisy information. To address this challenge, we borrow insights from the recent self-powered LLMs (Wang et al., 2022; Pan et al., 2023; Yuan et al., 2024; Tong et al., 2024) and propose a self-aware knowledge retrieval method to leverage LLMs’ ranking capability (Sun et al., 2023b; Ma et al., 2023) to filter out noisy information.

In particular, we directly prompt the LLM to rerank the sampled knowledge and only retrieve top  $k$  triples to provide for itself in the final-round inference. Given the question  $Q$  and either the path-based or neighbor-based sub-graph  $G_Q$ , we create

prompt  $p_{self}$  by filling the pre-defined template:

$$p_{self} = \text{Template}_{self}(Q, G_Q, k). \quad (1)$$

Then, we use  $p_{self}$  as the input to prompt the LLM to obtain the self-retrieved knowledge:

$$G_Q^{self} = \text{LLM}(p_{self}), \quad (2)$$

Finally, we provide the question  $Q$  and fine-grained knowledge  $G_Q^{self}$  to the LLM for reasoning and get the predicted answer  $a$  in two steps:

$$p_{inference} = \text{Template}_{inference}(Q, G_Q^{self}), \quad (3)$$

$$a = \text{LLM}(p_{inference}). \quad (4)$$

We provide detailed examples in Appendix A and B to demonstrate the input and output in our DALK.

## 4 Main Experiment

### 4.1 ADQA Benchmark

For performance evaluation, we consider four widely-used medical QA datasets spanning diverse biomedical topics (Jin et al., 2021; Pal et al., 2022; Hendrycks et al., 2021; Peñas et al., 2013) and derive an AD-specific QA dataset from them. The four medical QA datasets are all multiple-choice based and include: 1) MedQA (Jin et al., 2021) consisting of US Medical Licensing Examination (USMLE)-style questions, 2) MedMCQA (Pal et al., 2022) containing medical school entrance exam questions from India, 3) MMLU (Hendrycks et al., 2021) consisting of diverse biomedical and clinical questions from various sources, 4) QA4MRE (Peñas et al., 2013) containing a subset of questions for AD derived from PubMed and Medline. In order to extract from the medical QA datasets a subset of samples related to AD for our evaluation, we referred to NIH’s Common Alzheimer’s and Related Dementias Research Ontology (CADRO)<sup>1</sup>. Jointly developed by the National Institute on Aging and the Alzheimer’s Association, CADRO is a three-tiered classification system with eight main categories and a dozen sub-categories for AD and related dementia, and it contains common terminologies or keywords used in the field. We derived from the CADRO a list of AD-related keywords most relevant to the medical QA datasets: <Aging, Alzheimer, Amyloid beta, APOE, Dementia, Lipoprotein, Microglia>. Then, we searched against each medical QA dataset for

<sup>1</sup><https://iadrp.nia.nih.gov/about/cadro>

matches with these keywords to find putative QA samples, then further asked GPT-3.5-turbo to judge for each putative sample whether the question is indeed related to AD or not. Finally, we filtered out a subset of such samples that are considered highly relevant to AD to conduct our evaluation (number of samples in each dataset is shown in Table 2). More details about ADQA can be found in Appendix C.

### 4.2 Experiment Settings

We apply our framework with OpenAI GPT-3.5-turbo models (OpenAI, 2022). We also include the following baseline methods for comparison:

**Biomedical LLMs** Both ChatDoctor (Yunxiang et al., 2023) and Med-Alpaca (Shu et al., 2023) are fine-tuned versions of LLaMA (Touvron et al.) on biomedical corpora. Compared with them, Meditron (Chen et al., 2023) is built on LLaMA-2 (Touvron et al., 2023) and extends its pretraining on a comprehensively curated medical corpus. BiomedGPT (Zhang et al., 2023a) is also based on LLaMA-2 and pioneer as the first open-source and generalist visual language AI for diverse biomedical tasks. Biomistral (Labrak et al., 2024) is an open-source LLM crafted specifically for the biomedical domain, optimized for efficiency through quantization and model merging techniques.

**Retrieval-Augmented LLMs** Furthermore, we also compare our method with several representative retrieval-augmented LLMs in the biomedical domain. Almanac (Zakka et al., 2024) is a novel approach utilizing OpenAI’s GPT model integrated with a Qdrant vector database to hold external sources of knowledge retrieved from local corpus, web search, and calculators, designed to answer open-domain clinical questions. Like Almanac, Lozano et al. (2023) introduced Clinfo.ai, which is an open-source, end-to-end retrieval-augmented LLM (GPT) to answer medical queries using scientific literature summarizations derived from PubMed search engine. We adopt both Almanac and Clinfo.ai with the same prompt as ours to answer multiple-choice questions to suit the ADQA benchmark. Lastly, we implement a simple retrieval-augmented GPT baseline with CoT prompting similar to our proposed DALK. All the GPT models used are set to GPT-3.5-turbo as detailed in the next paragraph, to be consistent.

	MedQA	MedMCQA	MMLU	QA4MRE	AVG
<b>Biomedical LLMs</b>					
ChatDoctor (Yunxiang et al., 2023)	25.7	36.4	46.9	51.4	40.1
Med-Alpaca (Shu et al., 2023)	41.4	42.8	44.9	57.1	46.5
BiomedGPT (Zhang et al., 2023a)	38.8	41.9	48.9	42.6	43.1
Meditron (Chen et al., 2023)	27.6	31.4	36.7	25.7	30.4
Biomistral (Labrak et al., 2024)	44.7	49.5	53.1	68.6	54.0
<b>Retrieval-augmented LLMs</b>					
GPT-3.5-turbo w/ Ada (OpenAI, 2024)	<u>57.2</u>	65.7	<u>83.7</u>	62.9	67.4
Almanac (Zakka et al., 2024)	48.0	69.5	71.4	60.0	62.2
Clinfo.ai (Lozano et al., 2023)	54.3	<b>77.0</b>	81.3	67.7	<u>70.1</u>
Clinfo.ai w/o PubMed API	49.3	68.6	79.6	<b>74.3</b>	67.9
GPT-3.5-turbo	50.0	71.9	83.6	62.9	67.1
<b>DALK</b>	<b>57.9</b>	<u>75.2</u>	<b>85.4</b>	<u>71.4</u>	<b>72.6</b>

Table 2: Experiment results on our constructed ADQA benchmark. The sample size is indicated after each dataset name. The best results of each metric are in bold and the second-best results are underlined. The ‘‘AVG’’ column represents the average accuracy score on the four sub-dataset.

**Implementation Details** We use the knowledge graph constructed with the generative approach ( $KG_{gen}$ ) in our main experiment and conduct an ablation study on the knowledge graph with RE method ( $KG_{pair}$ ) in Section 4.5. We use GPT-3.5-turbo with the version ‘‘gpt-3.5-turbo-0301’’ and set the sampling temperature to 0.7. We utilize 7B versions of all the biomedical LLMs baselines. For RAG methods, we split each document with a max length of 128 and retrieve the top 3 most relevant documents as the support evidence for LLMs to do inference. We set the parameter  $k$  in our self-aware knowledge retrieval to 5 and conduct further analysis on it in Section 5.2.

### 4.3 Main Result

Table 2 shows the experimental results on our ADQA benchmark. We note that upon applying our dynamic co-augmentation framework, DALK’s performance surpasses that of other biomedical LLMs and RAG methods overall. It consistently achieves either the best or the second-best accuracy score across all sub-datasets and attains the highest AVG score. Furthermore, the substantial improvement over vanilla GPT-3.5-turbo underscores the efficacy of our approach in domain-specific ADQA.

Furthermore, we observe that the performance of biomedical-specific LLMs generally lags behind that of GPT-3.5-turbo. We attribute this discrepancy to the smaller size of these biomedical LLMs. While they may perform adequately in general medical contexts, they fall short in the AD scenario, which demands more domain-specific knowledge. In the case of GPT-3.5-turbo combined with various RAG methods, it is evident that most RAG methods enhance the models’ performance.

Among them, GPT-3.5-turbo with Clinfo.ai yields the most significant improvement, boosting the accuracy score from 67.1 to 70.1 compared to vanilla GPT-3.5-turbo. However, it is important to note that the original Clinfo.ai necessitates access to the PubMed API, constituting an external resource. When we disable this access and solely utilize the same corpora as in DALK within the Clinfo.ai retrieval system, the improvement it brings becomes marginal and incomparable to our method. Due to the space limitation, we put more RAG results with different hyper-parameters in Appendix D.

### 4.4 Ablation Study on Self-aware Knowledge Retrieval

	MedQA	MedMCQA	MMLU	QA4MRE	AVG
AVG Length	107.4	23.8	342.9	17.6	122.9
GPT-3.5-turbo	50.0	71.9	83.6	62.9	67.1
<b>DALK</b>	<b>57.9</b>	<b>75.2</b>	<b>85.4</b>	71.4	<b>72.6</b>
DALK w/o self-aware knowledge retrieval	56.5	71.0	77.6	<b>77.1</b>	70.6

Table 3: Ablation study results with and without our proposed self-aware knowledge retrieval.

In this section, we evaluate the efficacy of our proposed self-aware knowledge retrieval method through an ablation study. As depicted in Table 3, we observe that while the dynamic co-augmentation framework without the self-aware knowledge retrieval module still enhances the model’s performance, the overall improvement is less pronounced. Furthermore, we observe that the efficacy of self-aware knowledge retrieval correlates with the length of queries within a given context. For instance, a notable enhancement in performance is evident within the MMLU sub-dataset upon the implementation of self-aware knowledge retrieval. We attribute this to the fact that questions

in the MMLU dataset typically contain longer contexts compared to other medical QA datasets integrated into ADQA. Consequently, irrelevant knowledge sourced from the context may exacerbate the issue of information noise thus underscoring the necessity for self-aware retrieval. Conversely, within QA4MRE, characterized by shorter query lengths, the application of self-aware knowledge retrieval can even lead to a decline in performance.

#### 4.5 Ablation Study on KG Construction

	AVG	#Triples
GPT-3.5-turbo	67.1	-
<b>DALK</b> w/ Generative KG	<b>72.6</b>	53,585
<b>DALK</b> w/ RE KG	66.3	171,431

Table 4: Ablation study results with generative construction and RE construction.

Table 4 illustrates the results of the ablation study conducted using generatively constructed KG and RE-constructed KG. Surprisingly, despite the RE method yielding a KG with a larger scale and more triples, knowledge sampled from it has unexpectedly resulted in a non-trivial drop in performance within ADQA. After a manual examination of the two constructed knowledge graphs, we find LLMs with the RE construction method have a strong inclination to wrongly assign a relationship to two unrelated entities, which has been exposed by the previous studies (Wan et al., 2023). In contrast, the generative construction approach exclusively outputs triples that LLMs confidently endorse, yielding a smaller yet more precise knowledge graph. This trade-off between coverage and accuracy underscores the critical importance of denoising in the construction of KGs by LLMs.

### 5 Further Analysis

#### 5.1 Co-augmentation Analysis

To comprehensively understand how the performance of LLMs evolves in response to increasing KG sizes, we undertake a detailed co-augmentation analysis. Illustrated in Figure 3, our experiments aim to discern the changing performance trends of LLMs as the knowledge triples accumulate annually. Our findings reveal that our framework effectively fosters the co-evolution of LLMs and KG, with the performance of KG-augmented LLMs exhibiting a generally upward trajectory as the KG

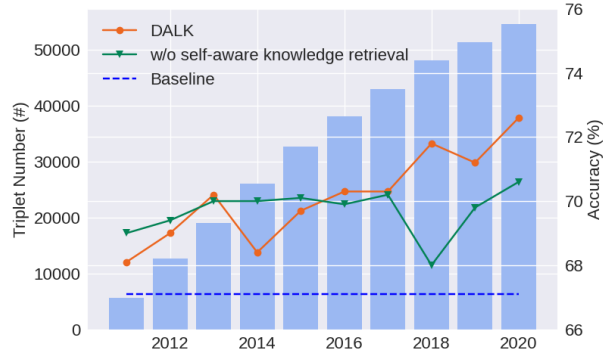


Figure 3: The size of the knowledge graph (triplet number) and the KG-augmented GPT-3.5-turbo’s performance (accuracy) over time.

expands. Notably, when we remove the self-aware knowledge retrieval module, this upward trend becomes less significant. This further implies the importance of sampling and selecting appropriate knowledge for LLMs when the KG’s size increases.

#### 5.2 Hyper-parameter Analysis

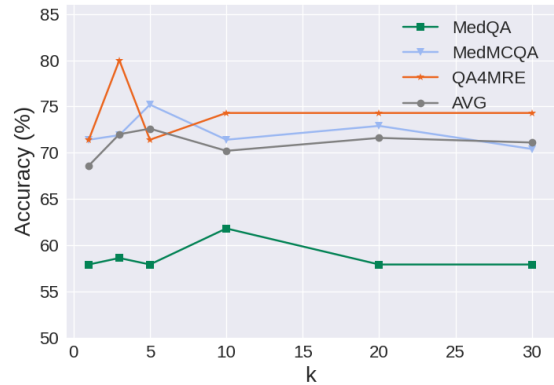


Figure 4: Different self-aware knowledge retrieval settings in MedQA, MedMCQA, QA4MRE and the average score (AVG).

In this section, we do a hyper-parameter analysis on the retrieval number  $k$  of our self-aware retrieval module. We select a group of value for  $k$  ([1,3,5,10,20,30]) and present the experiment results in Figure 4. We show the accuracy score on MedQA, MedMCQA, QA4MRE and AVG with different  $k$ . We find when  $k$  is small, an increment to it can lead to a performance enhancement. After the best performance shows up, continually increasing the value of  $k$  will cause a smooth decrease in the model accuracy score. This result

	Path-based Sub-graph	Answer
Baseline	-	C ✗
DALK -w/o self-aware knowledge retrieval	neurofibrillary tangles->FORM BY->microtubule-associated protein tau... ... entorhinal cortex->is a part of->brain->ASSOCIATES->mouse with Alzheimer’s disease->brain region->temporal lobe	C ✗
<b>DALK</b>	Reranked Triples1: entorhinal cortex ->is a part of ->brain Reranked Triples2: entorhinal cortex ->associates ->mouse with Alzheimer’s disease Reranked Triples3: temporal lobe ->affected by ->Alzheimer’s disease	D ✓

Table 5: A case to show the effectiveness of DALK. The question is: “The area of the brain resistant to Neurofibrillary tangles of Alzheimer’s disease is: A. Visual association areas B. Entorhinal coex C. Temporal lobe D.Lateral geniculate body”

indicates the knowledge ranked in the top positions is more helpful while the knowledge ranked behind is something not very useful, thus successfully validating the capability of LLMs to do a fine-grained knowledge reranking.

Moreover, we find the best  $k$  value is correlated with the length of queries in each single dataset. For example, the best performance in MedQA (average query length is 107.4) shows up when  $k = 10$  while the best performance in MedMCQA and QA4MRE shows up when  $k = 5$  and 3 respectively. This is consistent with our findings in Section 4.4 that a longer query corresponds to a larger and noisier sub-knowledge graph.

### 5.3 Sensitivity Analysis on ADQA Benchmark

Benchmark	DALK	DALK w/o self-aware knowledge retrieval
ADQA	72.6	70.6
w/o “Alzheimer”	72.1	70.4
w/o “Dementia”	72.4	71.3
w/o “APOE”	73.2	71.2
w/o “Amyloid beta”	73.5	70.7
w/o “Aging”	72.9	71.4
w/o “Lipoprotein”	73.1	71.0
w/o “Microglia”	72.8	70.9

Table 6: Sensitivity analysis for ADQA benchmark with a leave-one-out evaluation on AD-related keywords.

In this section, we conduct a sensitivity analysis for our constructed ADQA by conducting a leave-one-out evaluation on AD-related keywords. We do it by removing the samples with each keyword in our keyword list and calculating the AVG score of the remaining samples. As the result shown in Table 6, we find not all of the keywords are incorporated in our ADQA benchmark. Notably, the keywords “CSF Biomarkers”, “Neurogenesis”, “PET Amyloid”, “PET Tau”, “Tau Phosphorylation” lack corresponding samples in ADQA. We believe one critical work in the future for benchmarking

AD-related knowledge is to collect QA samples to cover these missing keywords. Moreover, analyzing the performance variation upon removing samples linked to each keyword offers insight into determining the relevance of the keyword to AD.

### 5.4 Case Study

We put an example in Table 5 to showcase the efficacy of DALK. We notice while the path-based sub-graph contains the relevant knowledge to exclude option C, it still involves other irrelevant information and finally fails to prompt the LLMs to produce the correct answer. In contrast, our self-aware knowledge retrieval method successfully chooses the top 3 most relevant triples for the given problem and results in the correct answer D.

## 6 Conclusion

In this research, we begin by analyzing the main limitations of adopting the existing LLMs-based methods in AD-specific areas. To address these issues, we propose a novel approach in the merging of large language models and knowledge graphs in the context of Alzheimer’s Disease. Our team provides an innovative dynamic co-augmentation framework for the refinement of large language models and knowledge graphs. Initially, our approach extracts structural insights from the unstructured scientific literature, crafting a specialized knowledge graph for AD. Subsequently, we employ a coarse-to-fine sampling technique coupled with a unique self-aware knowledge retrieval strategy to pinpoint relevant information from the knowledge graph. The extensive evaluation conducted in our constructed ADQA benchmark showcases the effectiveness of our method and provides further hints into the synergy of LLMs and knowledge graph in the context of AD. In the future, we will do more exploration in adopting and benchmarking LLMs in the AD areas.



## 7 Limitations

In the development of our AD-KG, our primary focus lies in the exploration of two distinct methods for extracting relationships between associated entities. For entity recognition, we employ a strong PubTator annotator directly, without delving into the utilization of LLMs in this context. However, we have observed that LLMs also exhibit promising entity extraction capabilities in Section 3.2.1. We defer the refinement of methods for extracting entities for KG construction with LLMs to future works. Furthermore, a significant contribution of our work is the establishment of the ADQA benchmark. Nonetheless, the datasets utilized in constructing ADQA primarily consist of medical school exam questions, potentially exhibiting a domain gap from the scientific literature informing AD-KG. One potential remedy is leveraging PubmedQA (Jin et al., 2019); however, it is hindered by limited data amount. In the future, we will keep gathering AD-related QA samples and expanding the size of our ADQA benchmark.

## 8 Ethics Statement

We have familiarized ourselves with and honour the ethical code set out in the ACL Code of Ethics<sup>2</sup>. The knowledge graphs constructed in the paper are based on published scientific literature from PubMed. The ADQA dataset used in the study is also derived from publicly available medical QA datasets that are properly cited. We strive to ensure our study upholds ethical principles and not cause any kind of safety or privacy concerns. Although not observed in our multiple-choice QA analysis, we recognize the possibility of factual errors and hallucinations when using pre-trained LLMs for medical QA tasks in general, and we do not recommend these models be applied in a practical setting at present.

## References

Felix Agbavor and Hualou Liang. 2022. [Predicting dementia from spontaneous speech using large language models](#). *PLoS Digital Health*, 1:1–14.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Aparna Balagopalan, Benjamin Eyre, Frank Rudzicz, and Jekaterina Novikova. 2020. [To BERT or not to BERT: Comparing Speech and Language-Based Approaches for Alzheimer’s Disease Detection](#). In *Proc. Interspeech 2020*, pages 2167–2171.

Zhen Bi, Jing Chen, Yinuo Jiang, Feiyu Xiong, Wei Guo, Huajun Chen, and Ningyu Zhang. 2024. Codekgc: Code language model for generative knowledge graph construction. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(3):1–16.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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*.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17754–17762.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. 2023. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

Geesa Daluwatumulle, Rupika Wijesinghe, and Ruvan Weerasinghe. 2023. [In silico drug repurposing using knowledge graph embeddings for alzheimer’s disease](#). In *Proceedings of the 9th International Conference on Bioinformatics Research and Applications, ICBRA ’22*, page 61–66, New York, NY, USA. Association for Computing Machinery.

Preetha Datta, Fedor Vitiugin, Anastasiia Chizhikova, and Nitin Sawhney. 2024. Construction of hyper-relational knowledge graphs using pre-trained large language models. *arXiv preprint arXiv:2403.11786*.

Stefan Dernbach, Khushbu Agarwal, Alejandro Zuniga, Michael Henry, and Sutanay Choudhury. 2024. Glam: Fine-tuning large language models for domain knowledge graph alignment via neighborhood partitioning and generative subgraph encoding. *arXiv preprint arXiv:2402.06764*.

Tianqing Fang, Hongming Zhang, Weiqi Wang, Yangqiu Song, and Bin He. 2021. Discos: bridging the gap between discourse knowledge and commonsense knowledge. In *Proceedings of the Web Conference 2021*, pages 2648–2659.

<sup>2</sup><https://www.aclweb.org/portal/content/acl-code-ethics>

697	Y. Feng, X. Xu, Y. Zhuang, and M. Zhang. 2023. <a href="#">Large language models improve alzheimerx27;s disease diagnosis using multi-modality data</a> . In <i>2023 IEEE International Conference on Medical Artificial Intelligence (MedAI)</i> , pages 61–66, Los Alamitos, CA, USA. IEEE Computer Society.	753
698		754
699		755
700		756
701		757
702		
703	Jiuzhou Han, Nigel Collier, Wray Buntine, and Ehsan Shareghi. 2023. Pive: Prompting with iterative verification improving graph-based generative capability of llms. <i>arXiv preprint arXiv:2305.12392</i> .	758
704		759
705		760
706		761
707		762
707	Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. <a href="#">Measuring massive multitask language understanding</a> . In <i>International Conference on Learning Representations</i> .	763
708		764
709		765
710		766
711		
712	Daniel Scott Himmelstein, Antoine Lizee, Christine Hessler, Leo Brueggeman, Sabrina L Chen, Dexter Hadley, Ari Green, Pouya Khankhanian, and Sergio E Baranzini. 2017. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. <i>Elife</i> , 6:e26726.	767
713		768
714		769
715		770
716		771
717		
718	Kang-Lin Hsieh, German Plascencia-Villa, Ko-Hong Lin, George Perry, Xiaoqian Jiang, and Yejin Kim. 2023. <a href="#">Synthesize heterogeneous biological knowledge via representation learning for alzheimer’s disease drug repurposing</a> . <i>iScience</i> , 26(1):105678.	772
719		773
720		774
721		775
722		776
723		777
723	Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In <i>International Conference on Learning Representations</i> .	778
724		779
725		780
726		781
727		782
728		
728	Jinhao Jiang, Kun Zhou, Zican Dong, Keming Ye, Wayne Xin Zhao, and Ji-Rong Wen. 2023. <a href="#">Structgpt: A general framework for large language model to reason over structured data</a> . In <i>Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing</i> , pages 9237–9251.	783
729		784
730		785
731		786
732		787
733		788
734		
734	Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. <i>Applied Sciences</i> , 11(14):6421.	789
735		790
736		791
737		792
738		
739		
739	Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. <i>arXiv preprint arXiv:1909.06146</i> .	793
740		794
741		795
742		796
743		797
743	Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In <i>International Conference on Machine Learning</i> , pages 15696–15707. PMLR.	798
744		799
745		800
746		801
747		802
748		803
748	Jiho Kim, Yeonsu Kwon, Yohan Jo, and Edward Choi. 2023. Kg-gpt: A general framework for reasoning on knowledge graphs using large language models. In <i>Findings of the Association for Computational Linguistics: EMNLP 2023</i> , pages 9410–9421.	804
749		805
750		
751		
752		
	Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. <i>arXiv preprint arXiv:2402.10373</i> .	806
		807
		808
	Dawei Li, Yanran Li, Jiayi Zhang, Ke Li, Chen Wei, Jianwei Cui, and Bin Wang. 2022. C3kg: A chinese commonsense conversation knowledge graph. In <i>Findings of the Association for Computational Linguistics: ACL 2022</i> , pages 1369–1383.	
	Dawei Li, Zhen Tan, Tianlong Chen, and Huan Liu. 2024. Contextualization distillation from large language model for knowledge graph completion. <i>arXiv preprint arXiv:2402.01729</i> .	
	Dawei Li, Hengyuan Zhang, Yanran Li, and Shiping Yang. 2023a. Multi-level contrastive learning for script-based character understanding. In <i>Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing</i> , pages 5995–6013.	
	Rumeng Li, Xun Wang, and Hong Yu. 2023b. Two directions for clinical data generation with large language models: Data-to-label and label-to-data. In <i>Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing</i> , pages 7129–7143.	
	Yucheng Li, Bo Dong, Frank Guerin, and Chenghua Lin. 2023c. Compressing context to enhance inference efficiency of large language models. In <i>Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing</i> , pages 6342–6353.	
	Alejandro Lozano, Scott L Fleming, Chia-Chun Chiang, and Nigam Shah. 2023. Clinfo. ai: An open-source retrieval-augmented large language model system for answering medical questions using scientific literature. In <i>PACIFIC SYMPOSIUM ON BIOCOMPUTING 2024</i> , pages 8–23. World Scientific.	
	Xindi Luo, Zequn Sun, Jing Zhao, Zhe Zhao, and Wei Hu. 2024. Knowla: Enhancing parameter-efficient finetuning with knowledgeable adaptation. <i>arXiv preprint arXiv:2403.14950</i> .	
	Yubo Ma, Yixin Cao, Yong Hong, and Aixun Sun. 2023. Large language model is not a good few-shot information extractor, but a good reranker for hard samples! In <i>Findings of the Association for Computational Linguistics: EMNLP 2023</i> , pages 10572–10601.	
	Chengsheng Mao, Jie Xu, Luke Rasmussen, Yikuan Li, Prakash Adekanattu, Jennifer Pacheco, Borna Bonakdarpour, Robert Vassar, Li Shen, Guoqian Jiang, Fei Wang, Jyotishman Pathak, and Yuan Luo. 2023. <a href="#">Ad-bert: Using pre-trained language model to predict the progression from mild cognitive impairment to alzheimer’s disease</a> . <i>Journal of Biomedical Informatics</i> , 144:104442.	
	C L Masters, G Simms, N A Weinman, G Multhaup, B L McDonald, and K Beyreuther. 1985. <a href="#">Amyloid plaque core protein in alzheimer disease and down</a>	

809	<a href="#">syndrome</a> . <i>Proceedings of the National Academy of Sciences</i> , 82(12):4245–4249.	Alzheimer’s Association Report. 2023. <a href="#">2023 alzheimer’s disease facts and figures</a> . <i>Alzheimer’s &amp; Dementia</i> , 19(4):1598–1695.	864
810			865
811	George A Miller. 1995. Wordnet: a lexical database for english. <i>Communications of the ACM</i> , 38(11):39–41.		866
812		Joseph Romano, Van Truong, Rachit Kumar, Mythreya Venkatesan, Britney Graham, Yun Hao, Nick Matsumoto, Xi Li, Zhiping Wang, Marylyn Ritchie, Li Shen, and Jason Moore. 2023. <a href="#">The alzheimer’s knowledge base – a knowledge graph for therapeutic discovery in alzheimer’s disease research (preprint)</a> . <i>Journal of Medical Internet Research</i> .	867
813	Yi Nian, Xinyue Hu, Rui Zhang, Jingna Feng, Jingcheng Du, Fang Li, Yong Chen, and Cui Tao. 2022. Mining on alzheimer’s diseases related knowledge graph to identity potential ad-related semantic triples for drug repurposing. <i>BMC Bioinformatics</i> , 23.		868
814			869
815			870
816			871
817			872
818	OpenAI. 2022. <a href="#">Introducing chatgpt</a> .		873
819	OpenAI. 2024. <a href="#">New embedding models and api updates</a> .	Chang Shu, Baian Chen, Fangyu Liu, Zihao Fu, Ehsan Shareghi, and Nigel Collier. 2023. Visual med-alpaca: A parameter-efficient biomedical llm with visual capabilities.	874
820			875
821	Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. 2023. Fine-tuning or retrieval? comparing knowledge injection in llms. <i>arXiv preprint arXiv:2312.05934</i> .		876
822			877
823		Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In <i>Proceedings of the AAAI conference on artificial intelligence</i> , volume 31.	878
824			879
825	Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In <i>Proceedings of the Conference on Health, Inference, and Learning</i> , volume 174 of <i>Proceedings of Machine Learning Research</i> , pages 248–260. PMLR.		880
826			881
827		Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Heung-Yeung Shum, and Jian Guo. 2023a. Think-on-graph: Deep and responsible reasoning of large language model with knowledge graph. <i>arXiv preprint arXiv:2307.07697</i> .	882
828			883
829			884
830			885
831			886
832	Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. 2023. Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies. <i>arXiv preprint arXiv:2308.03188</i> .	Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023b. Is chatgpt good at search? investigating large language models as re-ranking agents. In <i>Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing</i> , pages 14918–14937.	887
833			888
834			889
835			890
836			891
837	Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jipu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: A roadmap. <i>IEEE Transactions on Knowledge and Data Engineering</i> .	Jiabin Tang, Yuhao Yang, Wei Wei, Lei Shi, Lixin Su, Suqi Cheng, Dawei Yin, and Chao Huang. 2023. Graphgpt: Graph instruction tuning for large language models. <i>arXiv preprint arXiv:2310.13023</i> .	892
838			893
839			894
840			895
841			896
842	Anselmo Peñas, Eduard H. Hovy, Pamela Forner, Álvaro Rodrigo, Richard F. E. Sutcliffe, and Roser Morante. 2013. Qa4mre 2011-2013: Overview of question answering for machine reading evaluation. In <i>International Conference of the Cross-Language Evaluation Forum for European Languages</i> , pages 303–320. Springer.	Yongqi Tong, Dawei Li, Sizhe Wang, Yujia Wang, Fei Teng, and Jingbo Shang. 2024. Can llms learn from previous mistakes? investigating llms’ errors to boost for reasoning. <i>arXiv preprint arXiv:2403.20046</i> .	897
843			898
844			899
845			900
846			901
847			902
848			903
849	Yiyuan Pu, Daniel Beck, and Karin Verspoor. 2023. Graph embedding-based link prediction for literature-based discovery in alzheimer’s disease. <i>Journal of Biomedical Informatics</i> , 145:104464.	Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothee Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models.	904
850			905
851			906
852			907
853	Julio C Rangel, Tarcisio Mendes de Farias, Ana Claudia Sima, and Norio Kobayashi. 2024. Sparql generation: an analysis on fine-tuning openllama for question answering over a life science knowledge graph. <i>arXiv preprint arXiv:2402.04627</i> .	Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. <i>arXiv preprint arXiv:2307.09288</i> .	908
854			909
855			910
856			911
857			912
858	Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In <i>Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)</i> , pages 3982–3992.	Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. <i>Communications of the ACM</i> , 57(10):78–85.	913
859			914
860			915
861			916
862			917
863		Somin Wadhwa, Silvio Amir, and Byron C Wallace. 2023. Revisiting relation extraction in the era of large language models. In <i>Proceedings of the 61st</i>	918

919	<i>Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 15566–15589.	et al. 2024. Almanac—retrieval-augmented language models for clinical medicine. <i>NEJM AI</i> , 1(2):A10a2300068.	973 974 975
922	Zhen Wan, Fei Cheng, Zhuoyuan Mao, Qianying Liu, Haiyue Song, Jiwei Li, and Sadao Kurohashi. 2023. Gpt-re: In-context learning for relation extraction using large language models. In <i>Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing</i> , pages 3534–3547.	Hongming Zhang, Xin Liu, Haojie Pan, Yangqiu Song, and Cane Wing-Ki Leung. 2020. Aser: A large-scale eventuality knowledge graph. In <i>Proceedings of the web conference 2020</i> , pages 201–211.	976 977 978 979
928	Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. In <i>The Eleventh International Conference on Learning Representations</i> .	Kai Zhang, Jun Yu, Zhiling Yan, Yixin Liu, Eashan Adhikarla, Sunyang Fu, Xun Chen, Chen Chen, Yuyin Zhou, Xiang Li, et al. 2023a. Biomedgpt: A unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. <i>arXiv preprint arXiv:2305.17100</i> .	980 981 982 983 984 985
934	Chih-Hsuan Wei, Hung-Yu Kao, and Zhiyong Lu. 2013. Pubtator: a web-based text mining tool for assisting biocuration. <i>Nucleic acids research</i> , 41(W1):W518–W522.	Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. <i>arXiv preprint arXiv:2205.01068</i> .	986 987 988 989 990
938	Yanbin Wei, Qiushi Huang, Yu Zhang, and James Kwok. 2023. Kicgpt: Large language model with knowledge in context for knowledge graph completion. In <i>Findings of the Association for Computational Linguistics: EMNLP 2023</i> , pages 8667–8683.	Yichi Zhang, Zhuo Chen, Wen Zhang, and Huajun Chen. 2023b. Making large language models perform better in knowledge graph completion. <i>arXiv preprint arXiv:2310.06671</i> .	991 992 993 994
943	Yilin Wen, Zifeng Wang, and Jimeng Sun. 2023. Mindmap: Knowledge graph prompting sparks graph of thoughts in large language models. <i>arXiv preprint arXiv:2308.09729</i> .		
947	Siye Wu, Jian Xie, Jiangjie Chen, Tinghui Zhu, Kai Zhang, and Yanghua Xiao. 2024. How easily do irrelevant inputs skew the responses of large language models? <i>arXiv preprint arXiv:2404.03302</i> .		
951	Chao Yan, Monika Grabowska, Alyson Dickson, Bingshan Li, Zhexing Wen, Dan Roden, C. Stein, Peter Embí, Josh Peterson, Qiping Feng, Bradley Malin, and Wei-Qi Wei. 2024. Leveraging generative ai to prioritize drug repurposing candidates for alzheimer’s disease with real-world clinical validation. <i>npj Digital Medicine</i> , 7.		
958	Wenhao Yu, Hongming Zhang, Xiaoman Pan, Kaixin Ma, Hongwei Wang, and Dong Yu. 2023. Chain-of-note: Enhancing robustness in retrieval-augmented language models. <i>arXiv preprint arXiv:2311.09210</i> .		
962	Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-rewarding language models. <i>arXiv preprint arXiv:2401.10020</i> .		
966	Li Yunxiang, Li Zihan, Zhang Kai, Dan Ruilong, and Zhang You. 2023. Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge. <i>arXiv preprint arXiv:2303.14070</i> .		
970	Cyril Zakka, Rohan Shad, Akash Chaurasia, Alex R Dalal, Jennifer L Kim, Michael Moor, Robyn Fong, Curran Phillips, Kevin Alexander, Euan Ashley,		

## A Details of LLMs for KG

995  
996  
997  
998  
999  
1000  
1001

Table 7 and 8 present examples of our two KG construction methods respectively. For both methods, we adopt a select-or-generate prompt to instruct the LLM whether to choose a relation from hetionet (Himmelstein et al., 2017), a well-built general medical KG, or generate a new one to describe the relationship between two entities. In the RE construction method, we also conduct a type matching (Table 9) for each entity from type name of PubTator to that of Hetionet and ask the LLM to choose from the relation set that corresponds to the two entities' types (Table 10).

Input	<p>Read the following abstract, extract the relationships between each entity. You can choose the relation from: (covaries, interacts, regulates, resembles, downregulates, upregulates, associates, binds, treats, palliates), or generate a new predicate to describe the relationship between the two entities. Output all the extract triples in the format of "head   relation   tail". For example: "Alzheimer's disease   associates   memory deficits" Abstract: Thiamine pyrophosphate (TPP) and the activities of thiamine-dependent enzymes are reduced in Alzheimer's disease (AD) patients. In this study, we analyzed the relationship between thiamine deficiency (TD) and amyloid precursor protein (APP) processing in both cellular and animal models of TD. In SH-SY5Y neuroblastoma cells overexpressing APP, TD promoted maturation of beta-site APP cleaving enzyme 1 (BACE1) and increased beta-secretase activity which resulted in elevated levels of beta-amyloid (Abeta) as well as beta-secretase cleaved C-terminal fragment (beta-CTF). An inhibitor of beta-secretase efficiently reduced TD-induced up-regulation of Abeta and beta-CTF. Importantly, thiamine supplementation reversed the TD-induced alterations. Furthermore, TD treatment caused a significant accumulation of reactive oxygen species (ROS); antioxidants suppressed ROS production and maturation of BACE1, as well as TD-induced Abeta accumulation. On the other hand, exogenous Abeta(1-40) enhanced TD-induced production of ROS. A study on mice indicated that TD also caused Abeta accumulation in the brain, which was reversed by thiamine supplementation. Taken together, our study suggests that TD could enhance Abeta generation by promoting beta-secretase activity, and the accumulation of Abeta subsequently exacerbated TD-induced oxidative stress. Entity: Thiamine (thiamine), deficiency increases beta-secretase, TPP (Thiamine pyrophosphate), Alzheimer's disease (AD), patients, amyloid precursor protein (beta-amyloid (Abeta)), neuroblastoma, BACE1 (beta-site APP cleaving enzyme 1), Abeta, ROS (reactive oxygen species), Abeta accumulation, mice. Output:</p>
Output	<p>Thiamine deficiency   downregulates   TPP, Thiamine deficiency   regulates   Alzheimer's disease, Alzheimer's disease   involves   beta-secretase, Amyloid precursor protein   produces   beta-amyloid, Amyloid precursor protein   in   neuroblastoma, BACE1   is   beta-secretase, BACE1   regulates   beta-amyloid, Abeta   is   beta-amyloid, Abeta   in   mice, Abeta   causes   reactive oxygen species, Abeta   regulates   ROS</p>

Table 7: An example to extract knowledge triples using generative construction method.

Input	<p>Read the following abstract, answer the following question. Abstract: Thiamine pyrophosphate (TPP) and the activities of thiamine-dependent enzymes are reduced in Alzheimer’s disease (AD) patients. In this study, we analyzed the relationship between thiamine deficiency (TD) and amyloid precursor protein (APP) processing in both cellular and animal models of TD. In SH-SY5Y neuroblastoma cells overexpressing APP, TD promoted maturation of beta-site APP cleaving enzyme 1 (BACE1) and increased beta-secretase activity which resulted in elevated levels of beta-amyloid (Abeta) as well as beta-secretase cleaved C-terminal fragment (beta-CTF). An inhibitor of beta-secretase efficiently reduced TD-induced up-regulation of Abeta and beta-CTF. Importantly, thiamine supplementation reversed the TD-induced alterations. Furthermore, TD treatment caused a significant accumulation of reactive oxygen species (ROS); antioxidants suppressed ROS production and maturation of BACE1, as well as TD-induced Abeta accumulation. On the other hand, exogenous Abeta(1-40) enhanced TD-induced production of ROS. A study on mice indicated that TD also caused Abeta accumulation in the brain, which was reversed by thiamine supplementation. Taken together, our study suggests that TD could enhance Abeta generation by promoting beta-secretase activity, and the accumulation of Abeta subsequently exacerbated TD-induced oxidative stress. Entity: Thiamine (thiamine), deficiency increases beta-secretase, TPP (Thiamine pyrophosphate), Alzheimer’s disease (AD), patients, amyloid precursor protein (beta-amyloid (Abeta)), neuroblastoma, BACE1 (beta-site APP cleaving enzyme 1), Abeta, ROS (reactive oxygen species), Abeta accumulation, mice. Question: predict the relationship between Disease entity "sclerosis" and Disease entity "multiple sclerosis", first choose from the following options: A. resembles B. no-relation C. others, please specify by generating a short predicate in 5 words. Answer: Let’s think step by step:</p>
Output	<p>1. Sclerosis is a disease of the central nervous system. 2. Multiple sclerosis is the most common form of sclerosis. 3. So the relationship between sclerosis and multiple sclerosis should be "sclerosis is a kind of multiple sclerosis". So the answer is: A. resembles. So the answer is: A. resembles</p>

Table 8: An example to extract the relationship between “sclerosis” and “multiple sclerosis” using RE construction method.

PubTator Type	Hetionet Type
Gene	genes
Chemical	compounds
Disease	diseases

Table 9: Entity type match from PubTator to Hetionet.

Type-Type	Relations
genes-genes	["covaries", "interacts", "regulates"]
disease-disease	["resembles"]
compounds-compounds	["resembles"]
genes-diseases	["downregulates", "associates", "upregulates"]
genes-compounds	["binds", "upregulates", "downregulates"]
compounds-diseases	["treats", "palliates"]

Table 10: Type-type to relation match in Hetionet.

## B Details of KG for LLMs

1002

In this section, we provide detailed input and output for adopting KG to augment LLMs, including path-based and neighbor-based sub-graph sampling results (Table 11), self-aware knowledge retrieval (Table 12), describing sub-graphs with LLMs (Table 13) and inference with sampled knowledge (Table 14). The question we showcase here is “The area of the brain resistant to Neurofibrillary tangles of Alzheimer’s disease is: A. Visual association areas B. Entorhinal coex C. Temporal lobe D.Lateral geniculate body”, which same as the one we use in Section 5.4.

1003

1004

1005

1006

1007

1008

Path-based Sub-graph	neurofibrillary tangles->FORM BY->microtubule-associated protein tau->BINDS->(18)F-THK-5117->ADMINISTERED TO->rats->has->Alzheimer’s disease ->Alzheimer’s disease ->affects->human->has->AD->DISEASE OF->Brain entorhinal cortex->is a part of->brain->ASSOCIATES->mouse with Alzheimer’s disease->brain region->temporal lobe
Neighbor-based Sub-graph	neurofibrillary tangles->FORM BY->microtubule-associated protein tau Alzheimer’s disease ->causes->neuronal death Alzheimer’s disease ->associates->cognitive decline Alzheimer’s disease ->affects->human Alzheimer’s disease ->has subtype->neurodegenerative diseases

Table 11: An example of path-based and neighbor-based sub-graph for the question.

Input	<p>There is a question and some knowledge graph. The knowledge graphs follow entity-&gt;relationship -&gt;entity list format.</p> <p>Graph:  neurofibrillary tangles-&gt;FORM BY-&gt;microtubule-associated protein tau-&gt;BINDS-&gt;(18)F-THK-5117-&gt;ADMINISTERED TO-&gt;rats-&gt;has-&gt;Alzheimer’s disease -&gt;Alzheimer’s disease -&gt;affects-&gt;human-&gt;has-&gt;AD-&gt;DISEASE OF-&gt;Brain entorhinal cortex-&gt;is a part of-&gt;brain-&gt;ASSOCIATES-&gt;mouse with Alzheimer’s disease-&gt;brain region-&gt;temporal lobe }</p> <p>Question:</p> <p>The area of the brain resistant to Neurofibrillary tangles of Alzheimer’s disease is:  A. Visual association areas  B. Entorhinal coex  C. Temporal lobe  D.Lateral geniculate body</p>
Output	<p>Please rerank the knowledge graph and output at most 5 important and relevant triples for solving the given question. Output the reranked knowledge in the following format:  Reranked Triple1: xxx ——&gt;xxx  Reranked Triple2: xxx ——&gt;xxx  Reranked Triple3: xxx ——&gt;xxx  Reranked Triple4: xxx ——&gt;xxx  Reranked Triple5: xxx ——&gt;xxx</p> <p>Answer:</p> <p>Reranked Triples1: entorhinal cortex -&gt;is a part of -&gt;brain  Reranked Triples2: entorhinal cortex -&gt;associates -&gt;mouse with Alzheimer’s disease  Reranked Triples3: temporal lobe -&gt;affected by -&gt;Alzheimer’s disease</p>

Table 12: An example of our self-aware knowledge retrieval in path-based sub-graph.

	There are some knowledge graph paths. They follow entity->relationship->entity format.
	Reranked Triples1: entorhinal cortex ->is a part of ->brain Reranked Triples2: entorhinal cortex ->associates ->mouse with Alzheimer's disease Reranked Triples3: temporal lobe ->affected by ->Alzheimer's disease
Input	Use the knowledge graph information. Try to convert them to natural language, respectively. Use single quotation marks for entity name and relation name. And name them as Path-based Evidence 1, Path-based Evidence 2,...
	Output:
Output	Path-based Evidence 1: 'Entorhinal cortex' is a part of 'brain'. Path-based Evidence 2: 'Entorhinal cortex' associates 'mouse' with 'Alzheimer's disease'. Path-based Evidence 3: 'Temporal lobe' is affected by 'Alzheimer's disease'.

Table 13: An example to describe the self-retrieved path-based sub-graph with the LLM.

	Question: The area of the brain resistant to Neurofibrillary tangles of Alzheimer's disease is: A. Visual association areas B. Entorhinal cortex C. Temporal lobe D. Lateral geniculate body
Input	You have some medical knowledge information in the following: ###Path-based Evidence 1: 'Entorhinal cortex' is a part of 'brain'. Path-based Evidence 2: 'Entorhinal cortex' associates 'mouse' with 'Alzheimer's disease'. Path-based Evidence 3: 'Temporal lobe' is affected by 'Alzheimer's disease'. ###Neighbor-based Evidence 1: 'neurofibrillary tangles' are 'FORMED BY' 'microtubule-associated protein tau'. Neighbor-based Evidence 2: 'Alzheimer's disease' 'CAUSES' 'neuronal death'. Neighbor-based Evidence 3: 'Alzheimer's disease' 'ASSOCIATES' with 'cognitive decline'. Neighbor-based Evidence 4: 'Alzheimer's disease' 'AFFECTS' 'human'. Neighbor-based Evidence 5: 'Alzheimer's disease' 'HAS SUBTYPE' 'neurodegenerative diseases'."
	Answer: Let's think step by step:
Output	In Alzheimer's disease, neurofibrillary tangles are composed of abnormal aggregations of tau protein within neurons, leading to neuronal dysfunction and eventual cell death. The lateral geniculate body is a relay center in the thalamus that processes visual information from the retina to the visual cortex. Since the lateral geniculate body is not directly involved in memory and cognitive function, it is less likely to be affected by neurofibrillary tangles in Alzheimer's disease compared to other brain regions such as the entorhinal cortex and temporal lobe which play crucial roles in memory and cognitive processing. Therefore, the area of the brain resistant to neurofibrillary tangles of Alzheimer's disease is the lateral geniculate body (option D)

Table 14: An example of inference with selected knowledge

## C ADQA Benchmark

Table 15 shows the detailed prompt of our self-sample method to construct ADQA benchmark. Table 16 gives the statistics of our ADQA benchmark.

Input	Judge whether the question below is related to Alzheimer's Disease. Please answer yes or no. Question: Treatable causes of dementia are ____. a).AD b).Hypothyroidism c).Multi-infarct dementia d).SDH e).Hydrocephalus Is the question related to Alzheimer's Disease? Answer:
Output	Yes

Table 15: An example from MedMCQA to self-sample AD-related QA sample with LLMs.



Dataset	MedQA	MedMCQA	MMLU	QA4MRE	Total
Number	152	210	49	35	446

Table 16: Statistics of our ADQA benchmark.

## D Further Experiment for RAG

1012

	MedQA	MedMCQA	NMMLU	QA4MRE	AVG
Almanac w/ 256 chunk size	50.0	69.0	67.3	62.9	62.3
Almanac w/ top 10 docuemnt	48.7	68.6	65.3	62.9	61.4
Almanac w/ CoT	50.0	65.7	77.6	65.7	64.7
Clinfo.ai w/ 256 chunk size	48.6	66.7	81.6	65.7	65.7
Clinfo.ai w/ top 5 docuemnt	43.4	68.1	77.6	68.6	64.4
Clinfo.ai w/ CoT	48.7	68.6	79.6	68.6	65.0

Table 17: Further experiment in RAG methods with different hyper-parameter settings.