DRLK: Dynamic Hierarchical Reasoning with Language Model and Knowledge Graph for Question Answering

Miao Zhang[1,2,3], **Rufeng Dai**[2,3,4], **Ming Dong**[2,3,4*], **Tingting He**[2,3,4*]
[1]National Engineering Research Center for E-Learning,
[2]Hubei Provincial Key Laboratory of Artificial Intelligence and Smart Learning,
[3]National Language Resources Monitoring and Research Center for Network Media,
[4]School of Computer, Central China Normal University, Wuhan, China
{zmzhangmiao,dairufeng}@mails.ccnu.edu.cn
{dongming,tthe}@ccnu.edu.cn

Abstract

In recent years, Graph Neural Network (GNN) approaches with enhanced knowledge graphs (KG) perform well in question answering (QA) tasks. One critical challenge is how to effectively utilize interactions between the QA context and KG. However, existing work only adopts the identical QA context representation to interact with multiple layers of KG, which results in a restricted interaction. In this paper, we propose DRLK (Dynamic Hierarchical Reasoning with Language Model and Knowledge Graphs), a novel model that utilizes dynamic hierarchical interactions between the QA context and KG for reasoning. DRLK extracts dynamic hierarchical features in the QA context, and performs inter-layer and intra-layer interactions on each iteration, allowing the KG representation to be grounded with the hierarchical features of the QA context. We conduct extensive experiments on four benchmark datasets in medical QA and commonsense reasoning. The experimental results demonstrate that DRLK achieves state-of-the-art performances on two benchmark datasets and performs competitively on the others[1].

1 Introduction

Question answering (QA) system is a hot research area in natural language processing, requiring the robot to clearly understand the scenario described in the question and then reason with relevant domain knowledge (Jin et al., 2022). Recently, large-scale pre-trained language models (LMs) (Gu et al., 2022; Liu et al., 2021) have become a popular solution in several QA datasets (Mutabazi et al., 2021), achieving excellent performance. By training on an ultra large-scale corpus, LMs learn the latent domain knowledge and perform well in downstream tasks through fine-tuning (Lewis et al., 2020; Chakraborty et al., 2020). However, fine-tuned LMs perform poorly when downstream tasks involve complex reasoning or require explicit knowledge. The fine-tuned approach relies on similar task patterns and sample forms, while black-box models result in uninterpretable behavior (McCoy et al., 2019).

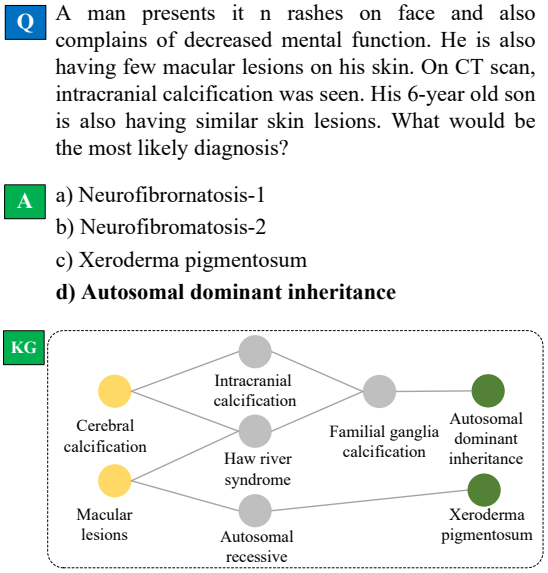

Figure 1: An example from the MedMCQA dataset with KG, where the correct answer is indicated in bold.

One research topic is to introduce external knowledge graphs (KGs) for effective and interpretable joint reasoning. Large-scale KGs, such as UMLS (Bodenreider, 2004) and DrugBank (Wishart et al., 2018), explicitly define structured knowledge through triples, entities, and relations. Existing work (Yasunaga et al., 2021; Zhang et al., 2022) demonstrates that KGs perform well in reasoning tasks involving knowledge. However, the question in QA task is always in natural language rather than a structured and logical query. The inevitable challenge is to constrain and integrate the structured KG according to the question, so as to

*Corresponding author.
[1]Our code is available at https://github.com/MZ-MiaoZhang/DRLK

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 5123–5133
December 7-11, 2022

extend the reasoning advantage to QA.

Furthermore, well-designed graph neural networks (GNNs) (Scarselli et al., 2009; Schlichtkrull et al., 2018; Yasunaga et al., 2021) are employed for structured knowledge processing. Related approaches follow a two-stage paradigm (Lin et al., 2019; Feng et al., 2020): retrieval and modeling. First, researchers construct the knowledge subgraph by retrieving triples relevant to question via character matching or entity recognition. Fig. 1 shows an example of knowledge subgraph, where "*cerebral calcification*" and "*macular lesions*" are core entities in the question. Then, designed GNN modules are utilized to constrain the knowledge subgraph and perform inference. Structured inference paths are contained in multi-hop relations in the subgraph. However, these methods only focus on isolated modeling of multi-hop relationships in KG. They only interact in a shallow manner, fusing the QA context and KG representations on the output layer or enhancing KG representations by the QA context statically (Zhang et al., 2022; Sun et al., 2022). Consequently, these methods demonstrate limited ability to exchange useful information. Effective interaction, especially in a non-shallow way between KG and QA context, is critical to breaking the bottleneck of correctly understanding the complex knowledge relationships in the question.

According to the above consideration, we propose DRLK, a novel model that utilizes hierarchical interactions between QA context and KG for reasoning (See in Fig. 2). DRLK extracts dynamic hierarchical features in the QA context, and performs inter-layer and intra-layer interactions on each iteration, allowing KG representations to be grounded with the hierarchical features of the QA context. Specifically, we design the hierarchical awareness module and the heterogeneous relationship module for dynamic hierarchical interactions. The former extracts hierarchical features of the QA context and KG, while the latter performs the message passing mechanism on the heterogeneous relational network to update the KG. DRLK employs dynamic hierarchical interactions between the QA context and KG via inter-layer and intra-layer interactions, accomplishing correct reasoning via an iterative execution of above interactions.

In summary, our contributions are three-fold:

- We propose DRLK, a novel approach that focuses on the hierarchical features of KG and the QA context, employing joint reason between LM and KG through inter-layer and intra-layer hierarchical interactions.
- We design a heterogeneous relationship graph to perform effective hierarchical interactions over the heterogeneous relationships, and ensure reasoning with correct knowledge relationships.
- We conduct experiments on four benchmark datasets in medical QA and commonsense reasoning. The results show that DRLK outperforms existing KG enhancement methods.

2 Related Work

Integrating KG has become a hot research topic for enhancing QA systems. Due to the formal heterogeneity between structured knowledge and natural language, some work (Lv et al., 2020; Bian et al., 2021) unifies two description forms during input, such as transforming structure knowledge into text via templates or grammar. These methods use PLMs as an encoder to perform end-to-end inference on KG and QA context. Such formal transformations inevitably lose the original formal characteristics. Other work (Bosselut et al., 2019, 2021) models structure knowledge with GNN and integrates them at the embedding representation level. Wang et al. (2019a) directly integrate the graphical representation and context of knowledge via the twin-tower model. Lin et al. (2019) enhance QA context through the KG representation. In contrast, Feng et al. (2020) augment the reasoning of KG by the QA context, which is usually static. Among these methods, enhancing KG with the QA context has the highest ceiling and is now the most popular method. However, in these methods, there is no interaction between two representations or only limited interaction with a static representation. Although such methods can model two representations separately, the shallow interaction limits the extraction of effective features. Our proposed approach DRLK improves mainly on this point.

Additionally, PLMs show excellent performance on QA tasks, such as fine-tuning (Su et al., 2019; Chakraborty et al., 2020) and prompt learning (Paranjape et al., 2021; Zhong et al., 2022). These methods do not require extra knowledge, but they are limited by the reasoning capability of PLM. Other researchers propose integrating the advantages of LM and KG for joint reasoning. QA-GNN (Yasunaga et al., 2021) proposes to consider the

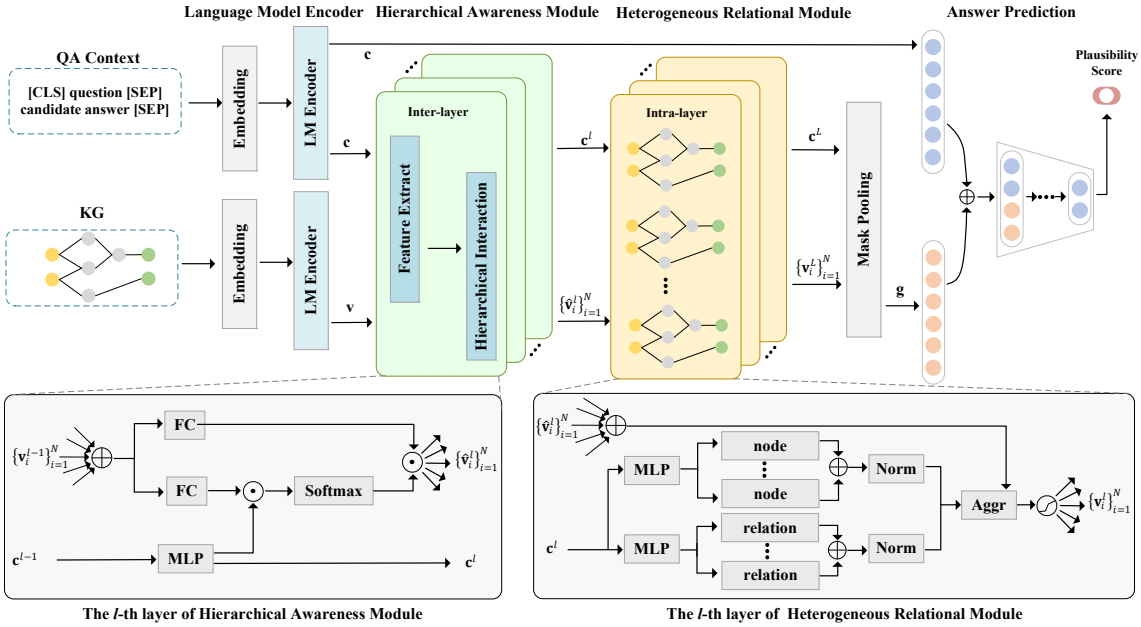

Figure 2: Overview of DRLK architecture.

QA context as a node directly connected to KG, and update the representation of both the context and the node through the message passing mechanism of the graph. However, the QA context focuses on one node, limiting the deep interaction between LM and GNN. GreaseLM (Zhang et al., 2022) and JointJK (Sun et al., 2022) are further extensions of QA-GNN to enhance the interaction while retaining the individual structure of both models. GreaseLM mixes LM and GNN node representations in the transformer module to achieve communication between the two modes. JointJK considers fine-grained interactions between tokens in the question and entities in KG via intensive bi-directional attention. In contrast to previous work, we focus on the cascading features of the QA context and KG, and update the QA context representation by a designed hierarchical feature extraction. For the interaction between them, we consider both before and inside the GNN layer for hierarchical features. In addition, we preserve the heterogeneous relational network and assess the interaction of heterogeneous relations with the QA context, making the structured inference paths interpretable.

3 Method

3.1 Task Definition

The task of multiple choice question answering (MCQA) in this paper can be formulated as $\mathbf{Y} = \{\mathbf{Q}, \mathbf{A}\}$, where $\mathbf{Q}$ denotes the question and $\mathbf{A}$ denotes the set of candidate answers. As the example in Fig. 1, each question has multiple candidate answers $\{a_1, a_2, \ldots, a_k\}$. The set of ground truth labels is $y = \{y_i\}$, where $y_i \in \{\mathbf{0}, \mathbf{1}\}^k$ is a one-hot vector. k is the number of candidate answers. The target of MCQA is to select one answer with highest plausibility from the candidate answer set, by learning a prediction function $f : \mathbf{Y} \rightarrow y$.

Additionally, for each QA sample, a domain KG is assumed to be accessible, providing the necessary background knowledge. We extract a subgraph from the external KG, guided by the question and candidate answers. We define the knowledge subgraph $G = (V, R)$, where V is the set of nodes from the external KG entities, R is the set of relationships. $E = V \times R \times V$ defines a set of edges that connect the nodes.

3.2 Language Model Encoder

In the encoding component, we use a pre-trained LM encoder as shown in Fig. 2, such as SapBERT-Base (Liu et al., 2021) and RoBERTa-Large (Liu et al., 2019), to encode the QA context and entities in KG separately.

Given a QA context $\{w_m\}_{m=1}^{M}$ (question and candidate answer), we first obtain its representation via the pre-trained LM encoder.

$$\mathbf{c} = \text{LM}_{\text{encoder}}(\{w_1, w_2, \ldots, w_M\}) \quad (1)$$

where $\mathbf{c}$ is the last hidden layer embedding of the

[CLS] token, which represents the embedding of the QA context.

For the set of entities $V = \{v_i\}_{i=1}^{N}$ in KG, each entity $v_i \in V$ is regarded as a sequence of tokens $\{v_{i,t}\}_{t=1}^{N'}$. We concatenate v_i and the QA context, and then encode them into a sequence of embeddings by sequence-to-sequence structure in LM encoder.

$$\begin{aligned}&\{\hat{\mathbf{w}}_1, \ldots, \hat{\mathbf{w}}_M, \hat{\mathbf{v}}_{i,1}, \ldots, \hat{\mathbf{v}}_{i,N'}\} \\ =&\text{LM}_{\text{encoder}}(\{w_1, \ldots, w_M, v_{i,1}, \ldots, v_{i,N'}\})\end{aligned} \tag{2}$$

Next, an average pooling operation is performed to get the initial representation for v_i, where we only operate on the entity tokens and drop the question tokens.

$$\mathbf{v}_i = \text{Pool}_{ave}(\{\hat{\mathbf{v}}_{i,1}, \ldots, \hat{\mathbf{v}}_{i,N'}\}) \tag{3}$$

We apply Eq. 2 and Eq. 3 on each node in $\{v_i\}_{i=1}^{N}$ to get the embedding set $\{\mathbf{v}_i\}_{i=1}^{N}$. Then the representation of QA context $\mathbf{c}$ and entities $\{\mathbf{v}_i\}_{i=1}^{N}$ will be provided to the hierarchical awareness module for further interactions.

3.3 Hierarchical Awareness Module

To achieve an effective inter-layer interaction, we capture the corresponding hierarchical features of the QA context and entities in the hierarchical awareness module, then integrate them via a multi-head attention mechanism.

The hierarchical awareness module accepts the embedding of the QA context and entities as input. As shown in Fig. 2, the input of the l-th layer are the QA context embedding $\mathbf{c}^{l-1}$ and entity embedding set $\{\mathbf{v}_i^{l-1}\}_{i=1}^{N}$.

We first employ a special MLP, with two layers of neural networks, to extract hierarchical features of the QA context.

$$\mathbf{c}^l = \text{MLP}(\mathbf{c}^{l-1}) \tag{4}$$

Then we update the representation of entities by a multi-headed attention interaction (Devlin et al., 2019), aiming to let it focus on the features in the current layer.

$$\hat{\mathbf{v}}_i^l = \text{Softmax}(\frac{\mathbf{c}^l \odot \mathbf{v}_i^{l-1}}{\sqrt{d}})\mathbf{v}_i^{l-1} \tag{5}$$

where d is the dimension of $\hat{\mathbf{v}}_i^{l-1}$. The updated embedding of entity set $\{\hat{\mathbf{v}}_i^l\}_{i=1}^{N}$ and the QA context $\mathbf{c}^l$ will be provided as input to the l layer of the heterogeneous relational module. The QA context representation will also be used as input to the l+1 layer of the hierarchical awareness module.

3.4 Heterogeneous Relational Module

The heterogeneous relations module is a graph neural network with heterogeneous relation and node types, achieving the intra-layer interaction and message passing. As shown in Fig. 2, the input of the l layer are the QA context embedding $\mathbf{c}^l$ and entity embedding set $\{\hat{\mathbf{v}}_i^l\}_{i=1}^{N}$.

We apply linear transformations (Feng et al., 2020) on relation and node types to make the model sensitive to heterogeneous networks.

$$\begin{aligned}\mathbf{t}(i) &= \mathbf{W}_{t(i)}\mathbf{c}_i^l + \mathbf{b}_{t(i)} \\ \mathbf{r}(i) &= \mathbf{W}_{r(i)}\mathbf{c}_i^l + \ \mathbf{b}_{r(i)}\end{aligned} \tag{6}$$

where $\mathbf{W}_{t(i)}$, $\mathbf{W}_{r(i)}$, $\mathbf{b}_{t(i)}$ and $\mathbf{b}_{r(i)}$ are learnable parameters for node v_i on relation and node type.

We apply message passing over heterogeneous graphs, which is built on the RGCN (Schlichtkrull et al., 2018). For brevity, we formulate the update and aggregation process of each entity in KG as:

$$\mathbf{v}_i^l = \text{GeLU}(\sum_{i,j=1}^{N} \mathbf{t}_j \mathbf{r}_j \hat{\mathbf{v}}_j^{l-1} \mathbf{W}^l) \tag{7}$$

where $\mathbf{W}_l$ is the learnable parameter. Node v_j is the neighbor of v_i and $\text{GeLU}(\cdot)$ is the activation function. We define $\{\mathbf{v}_i^l\}_{i=1}^{N}$ as the output of the heterogeneous relationship module, distinguishing from $\{\hat{\mathbf{v}}_i^l\}_{i=1}^{N}$ in Section 3.3.

After multiple layers of iterative message passing, we obtain the last layer output $\{\mathbf{v}_i^L\}_{i=1}^{N} = \{\mathbf{v}_1^L, \mathbf{v}_2^L, \ldots, \mathbf{v}_N^L\}$ as the final output of KG, which fuses incorporating contextual information. An average attention-based pooling and mask operation are applied to obtain the core KG representation $\mathbf{g}$:

$$\mathbf{g} = \text{AttPool}_{ave}(\{\mathbf{v}_1^L, \mathbf{v}_2^L, \ldots, \mathbf{v}_N^L\} \odot V_{mask}) \tag{8}$$

where V_{mask} is the masked matrix of the entity nodes, such as masking filled KG nodes.

3.5 Answer Prediction

By means of the iterations of the hierarchical awareness module and the heterogeneous relational module, we obtain the QA context representation $\mathbf{c}$ and

the KG representation **g**. We calculate the scores of candidate answers via:

$$p = (a|q, g) = \text{MLP}(\mathbf{c}; \mathbf{g}) \tag{9}$$

Finally, we utilize the softmax function to normalize all candidate answers, and obtain one choice with $argmax_{a \in A}\ p(a|q, g)$.

4 Experiments Setup

All experiments are conducted on one GPU (RTX-8000-48GB). The framework of our code relies on the PyTorch[2] and Transformer[3] packages.

4.1 Datasets and Metrics

Datasets. We evaluate DRLK on four benchmark datasets across two domains: MedMCQA (Pal et al., 2022) and MedQA-USMLE (Jin et al., 2021) are in medical QA; OpenBookQA (Mihaylov et al., 2018) and CommonsenseQA (Talmor et al., 2019) are in commonsense reasoning.

MedMCQA is a 4-choice question answering dataset of medical entrance exams, including more than 194k questions from AIIMS and NEET PG entrance exams. We conduct experiments on the original data splits from Pal et al. (2022).

MedQA-USMLE is a 4-choice question answering dataset about biomedical and clinical question based on the United States Medical License Exams. We conduct experiments on the official data splits in Jin et al. (2021).

OpenBookQA is a 5-choice question answering dataset about scientific knowledge. We conduct experiments on the official data splits from Mihaylov et al. (2018).

CommonsenseQA is a 4-choice question answering dataset about commonsense knowledge beyond real world. Since the test data is inaccessible, we conduct experiments on the in-house data split in Lin et al. (2019).

Dataset	Train	Dev	Test	Choices
MedQA-USMLE	10178	1272	1273	4
MedMCQA	182822	4183	6150	4
OpenBookQA	4957	500	500	4
CommonsenseQA	9741	1221	1140	5

Table 1: Overall statistics of Datasets.

Metrics. We follow baselines (Feng et al., 2020; Zhang et al., 2022) to utilize the accuracy score (Acc) as the metric protocol. We report the overall accuracy score on all benchmark datasets and subject accuracy score on MedMCQA.

4.2 Implementation Details

In pre-processing, we extract and construct knowledge subgraphs for each sample from DDB (Yasunaga et al., 2021) on Medical QA followed Zhang et al. (2022), while ConceptNet (Speer et al., 2017) on commonsense reasoning followed Feng et al. (2020). The maximum of nodes in the subgraph is set to 200 by truncating or completing. The QA context is concatenated as "`[CLS]` question `[SEP]` candidate answer `[SEP]`" and encoded via LM.

In training, we set the early stop mechanism with the guidance of the dev set (Lin et al., 2019). For hyperparameters, we set them empirically and make manual tuning. The batch size is set to 128 or a mini-size to be applied to computations on one single GPU. We use cross-entropy loss and RAdam optimizer. Separate learning rates is set in DRLK, {1e-5, 2e-5, 5e-5} for LM encoder and {1e-3, 2e-3, 3e-4} for other modules. We set the number of layers (L = 4) of GNN module, with dropout rate 0.1 applied to each layer. For MedQA-USMLE, we set the batch size to 128, maximum rounds to 50, and early stopping rounds to 10 as the best config. The overall training takes 8 hours on average, while the training and testing in one epoch take 18 minutes and 1.6 minutes on average.

4.3 Compared Methods

Due to the excellent performance in NLP, we use LM as a language encoder to obtain an initial representation of the input. For reasoning, DRLK focuses on enhancing the hierarchical interaction between the QA context and KG, so we use the strong model associated with LM and KG as the comparison model.

Since LM is KG-agnostic, a comparison with the fine-tuned model shows the improvement of KG on reasoning intuitively. We choose the corresponding pre-trained LM on different datasets, such as SapBERT-Base (Liu et al., 2021), PubmedBERT (Gu et al., 2022), BioBERT-Base and BioBERT-Large (Lee et al., 2020), BioRoBERTa-Base (Gururangan et al., 2020), SciBERT (Beltagy et al., 2019), ClinicalBERT (Alsentzer et al., 2019), BERT-Base (Devlin et al., 2019), and RoBERTa-Large (Liu et al., 2019).

[2]https://pytorch.org/
[3]https://huggingface.co/docs

Methods	Dev	Test
PMI (Clark et al., 2016)	29.8	31.1
MAX OUT (Mihaylov et al., 2018)	28.9	28.6
IR-ES (Chen et al., 2017)	34.0	35.5
IR-CUSTOM (Chen et al., 2017)	38.3	36.1
BERT-Base (Devlin et al., 2019)	33.9	34.3
ClinicalBERT-Base (Alsentzer et al., 2019)	33.7	32.4
BioBERT-Base (Lee et al., 2020)	34.3	34.1
BioRoBERTa-Base (Gururangan et al., 2020)	35.1	36.1
RoBERT-Large (Liu et al., 2019)	35.2	35.0
BioBERTa-Large (Lee et al., 2020)	36.1	36.7
SapBERT-Base (Liu et al., 2021)	-	37.2
+ QA-GNN (Yasunaga et al., 2021)	-	38.0
+ GreaseLM (Zhang et al., 2022)	38.3	38.5*
+ DRLK **(Ours)**	39.1	**40.4**

Table 2: Performance of baseline models on MedQA-USMLE. Here * indicates the improvement of DRLK is statistically significant ($p < 0.05$).

Methods	Dev	Test
BERT-Base (Devlin et al., 2019)	35.0	33.0
ClinicalBERT-Base (Alsentzer et al., 2019)	34.7	-
BioBERT-Base (Lee et al., 2020)	38.0	37.0
BioRoBERTa-Base (Gururangan et al., 2020)	34.6	-
SciBERT (Beltagy et al., 2019)	39.0	39.0
PubMedBERT (Gu et al., 2022)	40.0	41.0
SapBERT-Base (Liu et al., 2021)	40.3	40.0
+ QA-GNN (Yasunaga et al., 2021)	48.7	50.8
+ GreaseLM (Zhang et al., 2022)	49.3	51.0*
+ DRLK **(Ours)**	51.3	**52.5**

Table 3: Performance of baseline models on MedMCQA (without context).

For the evaluation of KG, we compare with similar approaches, which also use LM as a language encoder, with differences in the use of KG. RGCN (Schlichtkrull et al., 2018), RN (Santoro et al., 2017), GconAttn (Wang et al., 2019b), KageNet (Lin et al., 2019), MHGRN (Feng et al., 2020), QA-GNN, GreaseLM (Zhang et al., 2022), and JointJK (Sun et al., 2022) are QA paradigms with KG augmentation. GreaseLM and JointJK are the best-performing models as we know, which enhance reasoning by fusing LM and KG representations in GNNs. The main difference between DRLK and these approaches is that we not only perform a hierarchical feature extraction of the QA context but also perform a two-step interaction to enhance reasoning. We use LM to initialize these baselines for a fair comparison, consistent with DRLK.

5 Results and Analysis

5.1 Main Results

The results of MedQA-USMLE and MedMCQA are shown in Table 2-4, where Table 4 shows the

Subject Name	GreaseLM		DRLK **(Ours)**	
	Dev	Test	Dev	Test
Anaesthesia	44.1	45.8	47.1	45.8
Anatomy	54.3	49.4	61.1	**50.2**
Biochemistry	67.8	58.8	69.6	58.8
Dental	39.2	42.3	40.2	42.0
ENT	52.8	52.3	62.3	46.5
FM	37.3	56.8	38.8	**62.9**
O&G	58.5	50.2	64.3	46.1
Medicine	53.2	50.0	59.0	**57.5**
Microbiology	54.1	52.1	55.7	**57.5**
Ophthalmology	65.5	54.2	67.2	**62.1**
Orthopaedics	60.0	-	50.0	-
Pathology	54.9	55.7	55.5	**59.3**
Pediatrics	58.5	45.8	56.0	45.8
Pharmacology	61.7	56.1	60.9	**58.7**
Physiology	49.7	45.9	53.8	**47.9**
Psychiatry	56.2	50.0	68.8	50.0
Radiology	47.8	57.1	49.3	53.8
Skin	76.5	51.7	76.5	46.7
PSM	49.6	50.6	51.2	**51.9**
Surgery	42.0	51.7	44.2	**52.1**
Unknown	100.0	60.9	100.0	**66.0**
Average	49.3	51.0	51.3	**52.5**

Table 4: Subject performance on MedMCQA.

subject wise accuracies in MedMCQA. We observe that DRLK outperforms all LM models and KG enhanced models. In addition to DRLK, SapBERT-Base and GreaseLM on MedQA-USMLE are the best LM fine-tuning and KG augmentation models. DRLK has an absolute improvement of 3.2% relative to the SapBERT-Base fine-tuning model and 1.9% absolute improvement over the best model GreaseLM. On MedMCQA, DRLK also show the best performance, with absolute improvement of 12.5% relative to the SapBERT-Base fine-tuned model and 1.5% absolute improvement over the best model GreaseLM[4]. As shown in Table 4, DRLK outperforms more than half of the subjects in MedMCQA, equal on 5 subjects, and underperform only on 5 subjects. The state-of-the-art performance on MedQA-USMLE and MedMCQA demonstrates the effectiveness of hierarchical interactions and heterogeneous relational reasoning networks.

In addition to the medical domain, we also perform further robustness validation in the commonsense reasoning domain. Table 5 shows the comparison results on CommonsenseQA and OpenBookQA. As a result, DRLK outperforms the best

[4]We run it on MedMCQA following the open-source code of GreaseLM.

Dataset	OBQA	CSQA
RoBERTa-large (w/o KG)	64.8	68.7
+ RGCN (Schlichtkrull et al., 2018)	62.5	68.4
+ GconAttn (Wang et al., 2019b)	64.8	68.6
+ KagNet (Lin et al., 2019)	-	69.0
+ RN (Santoro et al., 2017)	65.2	69.1
+ MHGRN (Feng et al., 2020)	66.9	71.1
+ QA-GNN (Yasunaga et al., 2021)	67.8	73.4
+ GreaseLM (Zhang et al., 2022)	-	74.2
+ JointLK (Sun et al., 2022)	**70.3**	74.4*
+ DRLK **(Ours)**	70.2	**74.5**

Table 5: Performance of baseline models on OpenBookQA (OBQA) and CommonsenceQA (CSQA). Here * indicates the improvement of DRLK is statistically significant ($p < 0.05$).

Dataset	Dev-Acc. (%) (Overall)	Dev-Acc. (%) (Question w/ negation)	Dev-Acc. (%) (Question w/ $\leq$ 5 entities)	Dev-Acc. (%) (Question w/ $>$ 5 entities)
GreaseLM	49.3	48.0	49.4	50.0
DRLK **(Ours)**	51.3	50.1	51.4	55.0

Table 6: Performance of DRLK on MedMCQA dev set on questions with negative words and different number of entities.

Dataset	Test
w/o hierarchical feature (§3.3)	38.3
w/o hierarchical interaction (§3.3)	39.4
w/o feature and interaction (§3.3)	38.1
w/o heterogeneous relation type (§3.4)	40.0
w/o heterogeneous node type (§3.4)	39.8
w/o relation and node type (§3.4)	38.9
DRLK (layer = 4)	**40.4**

Table 7: Ablation study on different components.

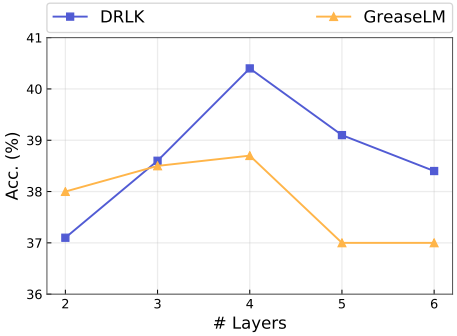

Figure 3: Impact on stacked of layers.

JoinkJK on CommonsenseQA and slightly underperforms on OpenBookQA. Overall, DRLK shows significant competitiveness with the best models in the commonsense reasoning domain. The experimental results demonstrate the robustness of DRLK in different domains.

5.2 Ablation Studies

Table 7 shows the further analysis of the different components of the model. We show the accuracy of the ablation experiments on MedMCQA.

Numbers of Layers. We investigate the effect of the layer number on KG reasoning. As shown in Fig. 3, the growth of layers is beneficial until N = 4. The performance starts to decrease when N > 4. Our analysis is that high layers make neighbors nodes be averaged too much, which leads to overfitting.

Hierarchical Features. We perform a separate evaluation of the layered feature extraction operation, which was the original motivation for our idea. As shown in Table 7, disabling the hierarchical feature extraction operation leads to 2.1% drop in performance, which indicates that the operation extracts features from the corresponding layers and impacts the subsequent reasoning.

Hierarchical Interactions. The purpose of the hierarchical interaction operation is to extract the hierarchical features of the nodes in the KG at each layer. We disable this operation to evaluate its impact on the overall effect. From the results in Table 7, disabling the operation results in 1% and 2.3% drop in performance. The ablation experiments indicate that the hierarchical interaction is crucial for subsequent reasoning.

Heterogeneous Relational Module. We evaluate the relation and node type in the heterogeneous awareness module. Disabling them results in 0.4%, 0.6%, 1.5% drop in performance. The ablation experiments show that interactions in the heterogeneous relational module are indispensable for the final reasoning.

5.3 Quantitative Analysis

To investigate whether the holistic performance improvement of DRLK is reflected in questions requiring complex inference, we analyze the inference complexity of the question prediction, such as the inclusion of negatives and multiple entities (Sun et al., 2022). Table 6 shows the comparison results of DRLK and GreaseLM (Zhang et al., 2022), the previous best-performing KG-enhanced models. First, the model faces enhanced noise interference and increased relational complexity when the question contains more entities. For example, the question in Fig. 1 involves multiple symptoms, e.g., "*rashes*", "*decreased mental function*", "*macular lesions*", and "*intracranial calcification*". Answering the question requires a comprehensive con-

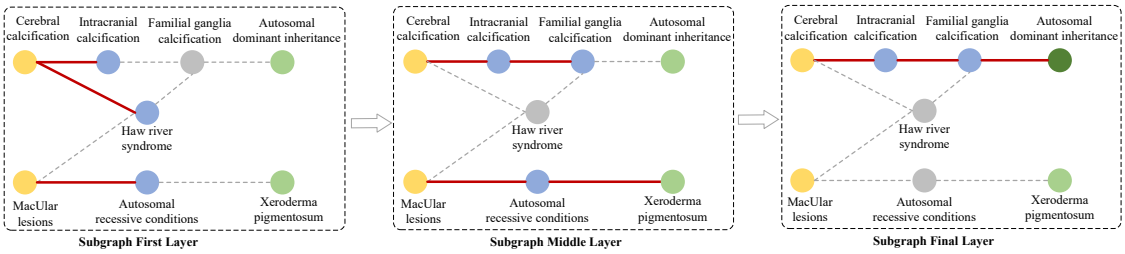

Figure 4: Case study of the variation in attention during DRLK reasoning, with red lines and blue nodes indicating high-weighted attention. The question corresponding to this case is that a man has a few macular lesions and intracranial calcification, while his 6-year-old son also has similar skin lesions.

sideration of the diseases corresponding to these symptoms. We classify the questions into two categories: entities less than 5 and entities more than 5. DRLK improves 2% in absolute on more entity questions and 5% in absolute on fewer entity questions, indicating that our model is more advantageous in dealing with more entity questions. Furthermore, the negation is a specific relationship, and the LM model is easily distracted by it. For example, the question "*What is not a major criteria for rheumatic heart disease* __?" contains the negative item "not". The model needs to focus on the negative relationship of "*criteria*", not only on "*rheumatic heart disease*". We retrieve questions with negatives and measure reasoning ability by the accuracy of negated questions. Compared to GreaseLM, DRLK improves 2.1% in absolute, indicating that it benefits in negative QA.

5.4 Case Study

We perform an interpretability analysis of DRLK's reasoning process, with the shift of attention in the knowledge subgraph. Fig. 4 shows a case where the bolded red lines and blue nodes indicate the high-weighted attention. Correspondingly, we use the dotted lines and gray nodes to represent the low-weighted parts. In this case, DRLK correctly answers the question and infers a reasonable inference path. The flow in Fig. 4 from left to right represents updating model attention in KG. The critical entities in the question are *Cerebral calcification* and *Macular lesions*. In the left subgraph, "*Cerebral calcification* → *Intracranial calcification*", "*Cerebral calcification* → *Haw River syndrome*", and "*Macular lesions* → *Autosomal recessive conditions*", are all reasonable speculations about critical entities. From left to middle, DRLK focuses on the critical evidence of KG's second layer over the previous layer. In terms of KG association, "*Autosomal dominant inheritance*" and "*Xeroderma pigmentosum*" are possible scenarios, but the former is related to the semantics of the question. Therefore, DRLK keeps only "*Autosomal dominant inheritance*" as the correct answer in the subgraphs from middle to right.

5.5 Error Analysis

To understand why DRLK fails in reason, we analyze the error cases of MedMCQA, which with a low correct and suitable sentence length between several datasets. The following is a categorization of 100 randomly selected error cases:

Incomprehensible Questions. Complex questions in medical scenarios usually involve diagnosing causes and proper treatments in specific medical situations. Such questions usually need to be developed over multiple symptoms, diseases, and treatments to choose the most appropriate option. This particular medical situation is difficult for humans too. The failure of the model to understand specific complex medical situations (individual patients) may cause erroneous predictions.

Indistinguishable Answers. The candidate answers may be similar. Distinguishing similar entities, such as {"*Anti C*", "*Anti D*", "*Anti E*", "*Anti Lewis*"}, may be self-evident to humans, but it is a huge challenge for the model. In addition, there is another situation in the dataset: the candidate answers may all be correct and the choice may depend on the person. Such a situation is also indistinguishable from the model.

Missing Evidence Entity. One critical aspect of reasoning with KG is obtaining as much background knowledge as possible. Although we use domain-related KGs, the retrieval of issue-related entities may miss some evidence entities due to the limitation of the KG coverage. In addition, the length of medical entities is usually long, making it difficult to achieve complete identification by character matching. The missing evidence entity

leads to an incomplete reasoning process, which may cause erroneous predictions.

Numerical Reasoning. Numerical questions are always a challenge in reasoning. Numerical reasoning shows some wrong predictions, where different doses of treatments need to be selected according to different symptoms in medical scenarios. For question "*Concentration of triple antibiotic paste (TAP) in treatment of revascularisation is?*", the candidate answers are "*1 mg*", "*0.1 mg*", "*100 mg*", "*10 mg*". We predict a wrong answer as "*0.1 mg*".

6 Conclusion and Future Work

In this paper, we propose a novel model that enables accurate reasoning through a hierarchical interaction between the QA context and KG. Compared to the fine-tuning based LM and KG-enhanced methods, DRLK achieves SOTA performances on two medical QA benchmark datasets. On commonsense reasoning benchmark datasets, DRLK performs competitively. Experimental results show that dynamic hierarchical interactions achieve superiority in dealing with complex knowledge relationships. In addition, the results on different domains show that DRLK possesses generalizations for the question answering task.

7 Limitations

To validate the effectiveness of DRLK, we conduct extensive experiments on four benchmark datasets, with different domains and scales. The results on four datasets show that DRLK achieves SOTA performance on two, just slightly inferior on the others. Nonetheless, DRLK relies on the domain KG, where the absence or low quality of the KG will directly affect the performance of DRLK.

Moreover, we follow Feng et al. (2020) to extract a knowledge subgraph with 200 nodes for each question. Due to the limitation of GPU resources, we do not test the effect of different knowledge subgraph scales, as it is not a major concern of DRLK. In response to the above two limitations, we will conduct further research in the future.

Ethical Considerations

We honor and support the ACL code of Ethics. This paper focuses on the question answering task, which aims to answer questions with an extra knowledge graph. Moreover, the datasets we use in this paper are from published open-source work and have no privacy or ethical implications. We neither introduce any social or ethical bias to the model nor amplify any bias in the data, so we do not foresee any direct social consequences or ethical issues.

Acknowledgements

This work is partially supported by the Key Research and Development Program of Hubei Province (2020BAB017), and Scientific Research Center Program of National Language Commission (ZDI135-135), and the Fundamental Research Funds for the Central Universities (KJ02502022-0155, CCNU22XJ037).

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Ning Bian, Xianpei Han, Bo Chen, and Le Sun. 2021. Benchmarking knowledge-enhanced commonsense question answering via knowledge-to-text transformation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12574–12582. AAAI Press.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32:267–270.

Antoine Bosselut, Ronan Le Bras, and Yejin Choi. 2021. Dynamic neuro-symbolic knowledge graph construction for zero-shot commonsense question answering. pages 4923–4931, Proceedings of the AAAI Conference on Artificial Intelligence. AAAI Press.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Souradip Chakraborty, Ekaba Bisong, Shweta Bhatt, Thomas Wagner, Riley Elliott, and Francesco

Mosconi. 2020. BioMedBERT: A pre-trained biomedical language model for QA and IR. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 669–679, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Peter Clark, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter D. Turney, and Daniel Khashabi. 2016. Combining retrieval, statistics, and inference to answer elementary science questions. pages 2580–2586, Proceedings of the AAAI Conference on Artificial Intelligence. AAAI Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yanlin Feng, Xinyue Chen, Bill Yuchen Lin, Peifeng Wang, Jun Yan, and Xiang Ren. 2020. Scalable multi-hop relational reasoning for knowledge-aware question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1295–1309, Online. Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3:2:1–2:23.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

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. *Applied Sciences*, 11:6421.

Qiao Jin, Zheng Yuan, Guangzhi Xiong, Qianlan Yu, Huaiyuan Ying, Chuanqi Tan, Mosha Chen, Songfang Huang, Xiaozhong Liu, and Sheng Yu. 2022. Biomedical question answering: A survey of approaches and challenges. *ACM Computing Surveys*, 55:35:1–35:36.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36:1234–1240.

Patrick Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov. 2020. Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 146–157, Online. Association for Computational Linguistics.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. KagNet: Knowledge-aware graph networks for commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2829–2839, Hong Kong, China. Association for Computational Linguistics.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4228–4238, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *arXiv:1907.11692*.

Shangwen Lv, Daya Guo, Jingjing Xu, Duyu Tang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, and Songlin Hu. 2020. Graph-based reasoning over heterogeneous external knowledge for commonsense question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8449–8456. AAAI Press.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Emmanuel Mutabazi, Jianjun Ni, Guangyi Tang, and Weidong Cao. 2021. A review on medical textual question answering systems based on deep learning approaches. *Applied Sciences*, 11:5456.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on Health, Inference, and Learning*, pages 248–260. PMLR.

Bhargavi Paranjape, Julian Michael, Marjan Ghazvininejad, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2021. Prompting contrastive explanations for commonsense reasoning tasks. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4179–4192, Online. Association for Computational Linguistics.

Adam Santoro, David Raposo, David G. T. Barrett, Mateusz Malinowski, Razvan Pascanu, Peter W. Battaglia, and Tim Lillicrap. 2017. A simple neural network module for relational reasoning. In *Annual Conference on Neural Information Processing Systems*, pages 4967–4976.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2009. The graph neural network model. *IEEE transactions on neural networks*, 20:61–80.

Michael Sejr Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *European semantic web conference*, pages 593–607, Cham. Springer.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 4444–4451.

Dan Su, Yan Xu, Genta Indra Winata, Peng Xu, Hyeondey Kim, Zihan Liu, and Pascale Fung. 2019. Generalizing question answering system with pre-trained language model fine-tuning. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 203–211, Hong Kong, China. Association for Computational Linguistics.

Yueqing Sun, Qi Shi, Le Qi, and Yu Zhang. 2022. JointLK: Joint reasoning with language models and knowledge graphs for commonsense question answering. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5049–5060, Seattle, United States. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Xiaoyan Wang, Pavan Kapanipathi, Ryan Musa, Mo Yu, Kartik Talamadupula, Ibrahim Abdelaziz, Maria Chang, Achille Fokoue, Bassem Makni, Nicholas Mattei, and Michael Witbrock. 2019a. Improving natural language inference using external knowledge in the science questions domain. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7208–7215. AAAI Press.

Xiaoyan Wang, Pavan Kapanipathi, Ryan Musa, Mo Yu, Kartik Talamadupula, Ibrahim Abdelaziz, Maria Chang, Achille Fokoue, Bassem Makni, Nicholas Mattei, and Michael Witbrock. 2019b. Improving natural language inference using external knowledge in the science questions domain. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7208–7215.

David S. Wishart, Yannick D. Feunang, An Chi Guo, Elvis J. Lo, Ana Marcu, Jason R. Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, Nazanin Assempour, Ithayavani Iynkkaran, Yifeng Liu, Adam Maciejewski, Nicola Gale, Alex Wilson, Lucy Chin, Ryan Cummings, Diana Le, Allison Pon, Craig Knox, and Michael Wilson. 2018. Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic acids research*, 46:D1074–D1082.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. Qa-gnn: Reasoning with language models and knowledge graphs for question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 535–546, Online. Association for Computational Linguistics.

Xikun Zhang, Antoine Bosselut, Michihiro Yasunaga, Hongyu Ren, Percy Liang, Christopher D. Manning, and Jure Leskovec. 2022. Greaselm: Graph reasoning enhanced language models for question answering. In *International Conference on Learning Representations*.

Wanjun Zhong, Yifan Gao, Ning Ding, Yujia Qin, Zhiyuan Liu, Ming Zhou, Jiahai Wang, Jian Yin, and Nan Duan. 2022. ProQA: Structural prompt-based pre-training for unified question answering. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4230–4243, Seattle, United States. Association for Computational Linguistics.