Hierarchy-Aware Multi-Hop Question Answering over Knowledge Graphs

Junnan Dong*
The Hong Kong Polytechnic University
Hung Hom, Hong Kong SAR
hanson.dong@connect.polyu.hk

Qinggang Zhang*
The Hong Kong Polytechnic University
Hung Hom, Hong Kong SAR
qinggangg.zhang@connect.polyu.hk

Xiao Huang
The Hong Kong Polytechnic University
Hung Hom, Hong Kong SAR
xiaohuang@comp.polyu.edu.hk

Keyu Duan
National University of Singapore
Singapore
k.duan@u.nus.edu

Qiaoyu Tan
Texas A&M University
College Station, TX, USA
qytan@tamu.edu

Zhimeng Jiang
Texas A&M University
College Station, TX, USA
zhimengj@tamu.edu

ABSTRACT

Knowledge graphs (KGs) have been widely used to enhance complex question answering (QA). To understand complex questions, existing studies employ language models (LMs) to encode contexts. Despite the simplicity, they neglect the latent relational information among question concepts and answers in KGs. While question concepts ubiquitously present hyponymy at the semantic level, e.g., *mammals* and *animals*, this feature is identically reflected in the hierarchical relations in KGs, e.g., *a_type_of*. Therefore, we are motivated to explore comprehensive reasoning by the hierarchical structures in KGs to help understand questions. However, it is non-trivial to reason over tree-like structures compared with chained paths. Moreover, identifying appropriate hierarchies relies on expertise. To this end, we propose **HamQA**, a novel Hierarchy-aware multi-hop Question Answering framework on knowledge graphs, to effectively align the mutual hierarchical information between question contexts and KGs. The entire learning is conducted in Hyperbolic space, inspired by its advantages of embedding hierarchical structures. Specifically, (i) we design a context-aware graph attentive network to capture context information. (ii) Hierarchical structures are continuously preserved in KGs by minimizing the Hyperbolic geodesic distances. The comprehensive reasoning is conducted to jointly train both components and provide a top-ranked candidate as an optimal answer. We achieve a higher ranking than the state-of-the-art multi-hop baselines on the official OpenBookQA leaderboard with an accuracy of 85%.

CCS CONCEPTS

• **Information systems** → **Question answering**; • **Computing methodologies** → **Semantic networks**.

*Both authors contributed equally to this research.

WWW '23, May 1–5, 2023, Austin, TX, USA

ACM ISBN 978-1-4503-9416-1/23/04…$15.00
https://doi.org/10.1145/3543507.3583376

KEYWORDS

multi-hop question answering, knowledge graphs, graph neural networks

ACM Reference Format:
Junnan Dong, Qinggang Zhang, Xiao Huang, Keyu Duan, Qiaoyu Tan, and Zhimeng Jiang. 2023. Hierarchy-Aware Multi-Hop Question Answering over Knowledge Graphs. In *Proceedings of the ACM Web Conference 2023 (WWW '23), May 1–5, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3543507.3583376

1 INTRODUCTION

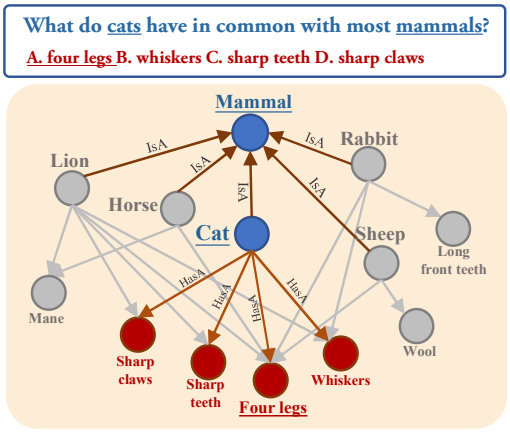

Figure 1: A running example of inference for a real question in CommonsenseQA over ConceptNet. It naturally depicts the tree-like properties among question concepts (blue) and answers(red).

Complex question answering (QA) has been widely studied, where models are trained to simulate how human beings make inferences [15]. Because of the powerful capability of knowledge graphs (KGs) in modeling structural real-world assertions in the form of *(head entity, relation, tail entity)* [8, 41], they manifest promising potentials to facilitate multi-hop QA. Holding the premise that answers could be found as entities that are K hops away from the question concepts in KGs [14], many methods have been proposed to infer over the neighbor entities and select the most relevant ones as answers. Among them, *question understanding* and *KG reasoning*

are two typical tasks. For KG reasoning, previous studies mainly dedicate to locating answers by modeling paths from questions to answers to extract the structural information [15, 26], or applying graph neural networks (GNNs) to the retrieved question-specific subgraph from KGs [9, 28, 40]. Meanwhile, regarding understanding questions, existing state-of-the-art (SOTA) methods solely rely on large language models (LMs), such as *BERT* [7] and *RoBERTa* [19]. They neglect the fruitful underlying structure information in KGs, which could significantly help the model to understand questions. This may result in the limitation of the reasoning capability in various real-world scenarios.

As shown in Figure 1, we provide an authentic running example from CommonsenseQA [29] over a real-world KG, ConcepNet [27]. In this question, *cats* and *mammals* are a semantically hierarchical pair that could be empirically recognized. Meanwhile, in ConceptNet, we find an identically hierarchical structure among their neighbors, i.e., *is_a*, as well as *has_a* among their 2-hop neighbors. The reasoning processes intuitively shed light on the hierarchical structure among question entities and answer entities. In conclusion, question concepts ubiquitously present hyponymy at the semantic level, while this feature is also correspondingly reflected in the hierarchical relations in KGs. It would be beneficial to perform hierarchy-aware reasoning for multi-hop KGQA.

To this end, we are motivated to explore comprehensive reasoning by leveraging the hierarchical structures in KGs to facilitate understanding the complex natural language questions. However, performing a hierarchy-aware multi-hop KGQA is challenging because of two major reasons. Firstly, it is non-trivial to reason over a hierarchical structure. Compared with chained paths, reasoning over tree-like structures is more complicated. Secondly, the hierarchical structure itself in multi-hop KGQA is complex. It relies on expertise to identify appropriate hierarchies from multifarious sub-trees that are highly related to the question.

To tackle these challenges, we propose a Hierarchy-aware multi-hop Question Answering framwork on knowledge graphs, namely **HamQA**, which effectively learns from the mutual hierarchical information between question contexts and KGs. As an essential research basis of this paper, we uniformly conduct the model learning in Hyperbolic space, which is inspired by its remarkable advantages of representing hierarchical structures [2, 5, 11]. For instance, [22] visualizes the embeddings of WordNet, which intuitively shows clear hierarchical structures between classes of mammals and animals with relation *is_a*. Therefore, in this paper, (*i*) we integrate the context semantics and guide a tailored context-aware graph attentive propagation with Hyperbolic manipulations. We derive the architecture on a Riemannian manifold. (*ii*) To preserve latent hierarchical structure in the subgraph, we continuously approximate the hierarchies in the Hyperbolic space, by minimizing the geodesic distances among relation-specific entities. (*iii*) A joint optimization scheme is designed to regularize both tasks with each other to achieve interdependently optimal hierarchies. As a final output, HamQA is expected to provide the correct answer with a top-ranked candidate with the highest probability score. Our main contributions are summarized as follows:

- We formally define the problem of hierarchy-aware multi-hop question answering over knowledge graphs.
- We capture the latent hyponymy from questions with a tailored context-aware graph attentive network, and preserve hierarchical structures in KGs with minimized Hyperbolic geodesic distances.
- We perform comprehensive reasoning to align both components and understand complex question with KGs. The whole framework is jointly optimized to provide top-ranked answers.
- Extensive experiments are conducted on two benchmark question answering datasets. Our model consistently outperforms the SOTA multi-hop KGQA baselines and achieve higher rankings on the official OpenBookQA leaderboard.

2 PROBLEM STATEMENT

To distinguish notations in this paper, we denote scalars as lowercase alphabets (e.g., c), vectors as boldface lowercase alphabets (e.g., $\boldsymbol{e}$), and matrices as boldface uppercase alphabets (e.g., $\mathbf{W}$).

Definition 1. *(Complex Question) Following previous study [14], we consider a natural language question as a complex question when its answer is an entity that could be located multi-hop away from the question concepts in a knowledge graph.*

Given a complex question with multiple choices, and a domain-related knowledge graph $\mathcal{G}$ that contains a number of triples (h, r, t). We first extract question entities $\mathcal{Q} = \{q_1, q_2, q_3...\}$ and answer entities $\mathcal{A} = \{a_1, a_2, a_3...\}$ from the KG. A subgraph is then retrieved containing Q, A, and their k-hop neighbors, denoted as $\mathcal{G}_{sub}$. A triple embedding is represented as $(\boldsymbol{e}_h, \boldsymbol{e}_r, \boldsymbol{e}_t)$. In this paper, we formally define the problem of hierarchy-aware multi-hop question answering over knowledge graphs as below.

> Given a domain-related $\mathcal{G}$ and a complex question, associated with $\mathcal{Q}$ and $\mathcal{A}$ as input, we aim to answer this question over a question-specific subgraph $\mathcal{G}_{sub}$. The model is trained to automatically capture the mutual hierarchical information between question contexts and KG topological structures, and iteratively optimized to provide a predicated answer $a_{pred} \in \mathcal{A}$ with the highest probability score. The overall performance is evaluated by the prediction accuracy comparing a_{pred} and ground truths.

3 APPROACH: HAMQA

This section elaborates on the rationales of how we tackle the aforementioned challenges of conducting a hierarchy-aware multi-hop QA on KGs. We first introduce the fundamental basis of this paper with a Riemannian Hyperbolic space. The overall framework, as illustrated in Figure 2, consists of three main components. First, we capture the latent semantics from question contexts with a context-aware Graph Neural Network in Hyperbolic space. Second, we approximate to preserve the hierarchical structures on KGs with a Hyperbolic KG embedding method in parallel. Finally, a joint optimization scheme is proposed to train both components and achieve the optimal comprehensive reasoning.

3.1 Curvature towards Hyperbolic Geometry

3.1.1 Riemannian Manifold. By approximating a small section of the curve as part of a circle with radius x in a 2D euclidean space, curvature $c = 1/x$ intuitively describes the bending degree. Extending from a constantly curved surface to higher dimensions, we obtain a d-dimensional manifold $\mathcal{M}_c^d$. To facilitate computation,

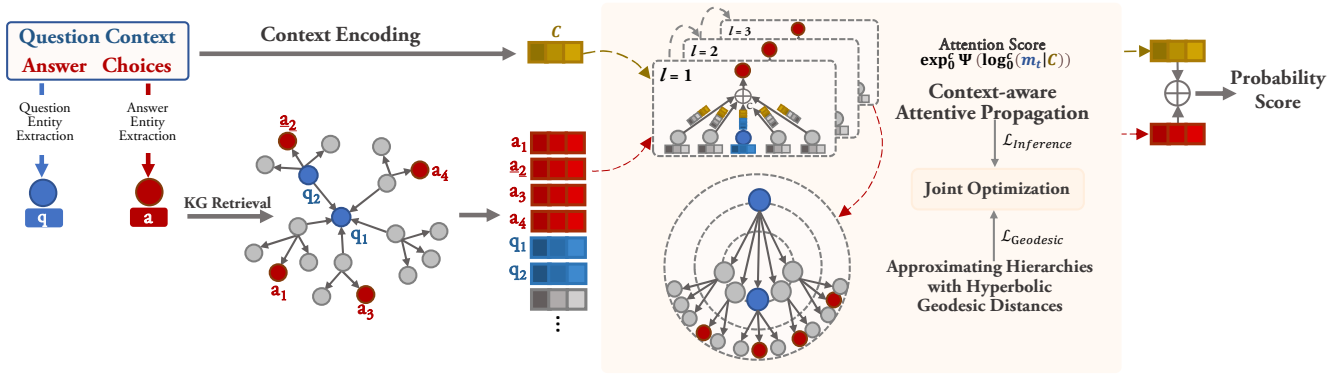

Figure 2: Our proposed HamQA, a comprehensive reasoning framework for multi-hop KGQA. We take a retrieved KG subgraph as input and match the hyponymy in question contexts with the reflected hierarchical structure in KGs. The final score is calculated to demonstrate the probability of an entity being a correct answer. The framework provides answers by prioritizing the top-ranked candidates.

Table 1: Major manipulations and formalism for mapping between the Euclidean Space and the Hyperbolic Space.

Manipulations	Formalism in Hyperbolic Space
Exponential Map	$\exp_0^c(v) = \tanh(\sqrt{c}\|v\|)\frac{v}{\sqrt{c}\|v\|}$
Logarithmic Map	$\log_0^c(v) = \tanh^{-1}(\sqrt{c}\|v\|)\frac{v}{\sqrt{\|c\|}\|v\|}$
Möbius Addition	$u \oplus_c v = \frac{(1+2c\langle u,v\rangle+c\|v\|^2)u+(1-c\|u\|^2)v}{1+2c\langle u,v\rangle+c^2\|u\|^2\|v\|^2}$
Multiplication	$u \otimes_c v = \exp_0^c\left(u \log_0^c(v)\right)$

each node u is associated with a *tangent* space $\mathcal{T}_u\mathcal{M}_c^d$, which is the d-dimensional first-order Euclidean approximation of u. A Riemannian metric g is applied to each u as an inner product [2], with which we obtain a Riemannian manifold defined as $(\mathcal{M}^d, g^{\mathcal{M}})$, and allow induction of the geometry with $\mathcal{T}_u\mathcal{M} \times \mathcal{T}_u\mathcal{M} \to \mathbb{R}$.

3.1.2 Hyperbolic space. In this paper, we define Hyperbolic space $\mathcal{H}_c^d$ with negative curvature $-c$ as a Poincaré ball of a radius $1/\sqrt{c}$, $\mathcal{H}_c^d = \{u \in \mathbb{R}^d : c\|u\|^2 < 1\}$, $c > 0$. To further simplify the calculation, a bi-directional mapping is carried out for projection between Hyperbolic $\mathcal{H}$ and Euclidean *tangent* spaces $\mathcal{T}_u\mathcal{H}$ via an exponential map $\exp_u^c(v)$ and a logarithmic map $\log_u^c(v)$ below, where $\|\cdot\|$ denotes the L_2 Euclidean norm and $\lambda_u^c = 2/(1-c\|u\|^2)$ is the conformal factor to Euclidean metric $g^{\mathcal{E}}$, which enables $g^{\mathcal{H}} = (\lambda_u^c)^2 g^{\mathcal{E}}$ [2, 11].

$$\exp_u^c(v) = u \oplus_c \left(\tanh(\sqrt{c}\frac{\lambda_u^c\|u\|}{2})\frac{v}{\sqrt{c}\|v\|}\right), \tag{1}$$

$$\log_u^c(v) = \frac{2}{\sqrt{c}\lambda_u^c}\tanh^{-1}(\sqrt{c}\,\|-u \oplus_c v\|)\frac{-u \oplus_c v}{\|-v \oplus_c v\|}. \tag{2}$$

3.2 Context-aware Graph Attentive Network

To capture the implicit hierarchies in questions, in this component, we design a context-aware Graph Attentive Network which is tailored in Hyperbolic space, **CGAT** in short. The message passing is built upon an attentive propagation subject to the context information, where important nodes are valued with higher attention scores. We employ pre-trained LMs to obtain the representation vector of contexts. To avoid confusion, we specially denote this vector as $\mathbf{C}$ instead of $\mathbf{c}$, which could be hardly distinguished from curvature c.

As the model learning is uniformly conducted to understand questions with the hierarchical information in KGs, we extend the traditional graph attention network [34], $\boldsymbol{e}_i^{(\ell+1)} = f\left(\sum_{j\in\mathcal{N}_i}\alpha_{ij}\boldsymbol{m}_{ij}\right) + \boldsymbol{e}_i^{(\ell)}$ to Hyperbolic space. In order to enable efficient computation on a Euclidean tangent space $\mathcal{T}_u\mathcal{H}$, we realize the projection back and forth between two spaces with $\exp_u^c(v)$ and $\log_u^c(v)$. As listed in Table 1, we summarize all the related Hyperbolic manipulations and formalism, which are corresponding to the ones in Euclidean spaces, e.g., Addition and Multiplication. Notably, Möbius multiplication is obtained via $\exp_0^c(v)$ and $\log_0^c(v)$, where the tangent space is mapped at node $\mathbf{0} \in \mathcal{H}_c^d$.

Different from existing models, e.g., [9] and [40], where message are modeled for capturing the entity types and relation types, we employ a triple-level message passing with all the tail entity embeddings e_t from $(\boldsymbol{e}_h, \boldsymbol{e}_r, \boldsymbol{e}_t) \in \mathbb{H}_c^{3\times d}$, which is headed by the target head entity embedding e_h. This recursively captures the higher-order relational information in the Hyperbolic space and does good in learning from the hierarchies. We denote this community of neighbors of h as $\mathcal{N}_h$. Accordingly, the message is defined as below,

$$\boldsymbol{m}_t = \boldsymbol{W} \otimes_c (\boldsymbol{e}_h, \boldsymbol{e}_r, \boldsymbol{e}_t), \tag{3}$$

where the message from a neighbor triple $\boldsymbol{m}_t (t \in \mathcal{N}_h)$ is computed by projecting the Hyperbolic triple embedding $(\boldsymbol{e}_h, \boldsymbol{e}_r, \boldsymbol{e}_t)$ with a trainable matrix $\boldsymbol{W}$ for the linear transformation, obtained by the Möbius multiplication $\otimes_c$.

3.2.1 Context-aware Attention. Given a mass of messages from $t \in \mathcal{N}_h$, identifying their importance respectively is what traditional graph attention networks aim to achieve. In this component, we are dedicated to only prioritizing the information that contributes to answering the question. Therefore, we propose context-aware

attention tailored to multi-hop KGQA. Specifically, we effectively concatenate the context representation $\mathbf{C}$ with each message $\boldsymbol{m}_t$ during the calculation of its attention score. This serves as a constraint to guide message propagation by focusing on more important neighbors. $\hat{a}(h,r,t)$ is calculated by

$$\hat{a}_{(h,r,t)} = \exp_0^c \Psi \left(\log_0^c(\boldsymbol{m}_t \| \mathbf{C})\right), \tag{4}$$

where Ψ is a LeakyReLU activation function, and $\|$ denotes the concatenation function. The concatenated vector $(\boldsymbol{m}_t \| \mathbf{C})$ is first projected to the tangent space $\mathcal{T}_0\mathcal{H}$ with $\log_0^c$ for computation by Ψ, and then mapped back to the Hyperbolic space with $\exp_0^c$.

To allocate proportions of all the messages from $\mathcal{N}_h$ for representing entity h, through normalizing all weights with a softmax function, we have,

$$\alpha_{(h,r,t)} = \frac{exp(\hat{a}_{(h,r,t)})}{\sum_{(h,r',t')\in\mathcal{N}_h} exp(\hat{a}_{(h,r',t')})}. \tag{5}$$

Once weighted, all $\alpha_{(h,r,t)}$ from $\mathcal{N}_h$ are used to derive a linear combination of all neighbor embeddings as the final output of the current target entity h during current propagation, formulated as follows,

$$\boldsymbol{e}^{\ell}_{\mathcal{N}_h} = \sum_{(h,r,t)\in\mathcal{N}_h} \alpha_{(h,r,t)} \otimes_c \boldsymbol{m}^{\ell}_t. \tag{6}$$

Under the overall architecture proposed for a context-aware attentive propagation, we effectively aggregate the information transmitted, and the Hyperbolic embedding $\boldsymbol{e}_h^{(\ell)} \in \mathbb{H}_c^d$ of entity h is updated in each layer as

$$\boldsymbol{e}_h^{(\ell+1)} = \exp_0^c f\left(\log_0^c\Big(\sum_{(h,r,t)\in\mathcal{N}_h} \alpha_{(h,r,t)} \otimes_c \boldsymbol{m}_t\Big)\right) \oplus_c \boldsymbol{e}_h^{(\ell)}. \tag{7}$$

where $f(\cdot)$ is a multi-layer perceptron. The previously learned $\boldsymbol{e}^{\ell}_{\mathcal{N}_h} \in \mathbb{H}_c^d$ is first similarly projected to the tangent space with $\log_0^c(v)$ to simplify the activation computation in a Euclidean space, and immediately projected back to the Hyperbolic space $\mathcal{H}_c^d$ with $\exp_0^c(v)$ for manipulations with $\boldsymbol{e}_h^{(\ell)}$. After reaching the plateau of model training, we could eventually obtain the final context-aware embedding for each entity from $\mathcal{G}_{sub}$.

3.3 Preservation of Hierarchical Structures

In parallel with capturing context information, we introduce the details of KG reasoning in this subsection, where the model continuously approximates hierarchical structures among question entities and answer entities in $\mathcal{G}_{sub}$.

We focus on representing entities in a relation-specific manner since relations primarily lay decisive influences on forming the hierarchies. To maximally preserve the geometrically hierarchical features, we leverage a Hyperbolic KG embedding referring to [2] as an auxiliary constraint. Based on traditionally translational models [3], $e_h + e_r = e_t$, where Euclidean distances are used to measure the similarity between e_h and e_t, the basic idea is extended from Euclidean to Hyperbolic by introducing the Hyperbolic distances, namely geodesic distances. This measures the relatedness between a pair of head entity h and tail entity t, formulated as below,

$$d_{\mathcal{M}}(h,t) = \frac{2}{\sqrt{c}} \tanh(\sqrt{c}\| -e_h \oplus_c e_t\|). \tag{8}$$

Additionally, two entity-specific scalar biases b_h and b_t are utilized to determine the decision boundaries of h and t in the Poincaré Ball subject to current relation r. The score function for this component is written as

$$\phi_{\mathbb{H}}(e_h, r, e_t) = -d^{(r)}_{\mathbb{H}^d}\left(\exp_0^c(\mathbf{R}\log_0^c(e_h)), e_t \oplus_c r_h\right)^2 + b_h + b_t, \tag{9}$$

where $\phi_{\mathbb{H}}$ is the logistic sigmoid indicating the probability of the link prediction in KG reasoning, $\mathbf{R} \in \mathbb{H}_c^{d\times d}$ is a diagonal relation matrix used to transform the head entity embedding $e_h \in \mathbb{H}_c^d$, and $r_h \in \mathbb{H}_c^d$ is a Hyperbolic translation vector of relation r. The Hyperbolic embedding e_h is first projected to the tangent space with $\log_0^c(v)$ for transformation with $\mathbf{R}$ likewise, and projected back to Hyperbolic with $\exp_0^c(v)$. After applying Möbius addition on e_t and r_h, the geodesic distances is obtained via $d^{(r)}_{\mathbb{H}^d}$. It would be considered to constitute a part of the hierarchical structure when the embeddings of a pair of h and t among question and answer entities satisfy a sphere with a radius of $\sqrt{b_h + b_t}$.

3.4 Joint Optimization Scheme

To perform comprehensive reasoning, we effectively integrate the training of two components above with a joint optimization scheme. In this subsection, we first introduce the training targets for each task respectively, then joint training is proposed with the overall combination of loss functions.

3.4.1 Discrimination-Encouraged Inference. The model is expected to identify the answers guided by the integrated context information $\mathbf{C}$. During the final inference stage, a probability score of an answer entity being correct is calculated by $MLP(\boldsymbol{e}_a \oplus \mathbf{C})$. We encourage the discrimination between correct answers and distractors. Therefore, we adopt a *cross-entropy loss* for training the context-aware attentive propagation and model inference, namely $\mathcal{L}_{Inference}$,

$$\mathcal{L}_{Inference} = -log\frac{exp(MLP(\boldsymbol{e}_a \oplus \mathbf{C})}{\sum_{a'\in A} exp(MLP(\boldsymbol{e}_{a'} \oplus \mathbf{C})}, \tag{10}$$

where e_a and $e_{a'}$ are Hyperbolic embeddings of correct and disturbing answers respectively, $a, a' \in A$. The loss is expected to be minimized for prioritizing the correct answers.

3.4.2 Margin-Based Hierarchy Approximation. As $\mathcal{G}_{sub}$ contains limited entities, to effectively train the approximation into hierarchies, we augment the negative sample set following [2] with synthetic triples by randomly replacing the tail entities as (h, r, t'), $t' \in \mathcal{G}_{sub}$. The KG embedding is trained to minimize the triple margin loss, denoted as $\mathcal{L}_{Geodesic}$,

$$\mathcal{L}_{Geodesic} = max\left((\phi_{\mathbb{H}}(e_h, r, e_t) + \gamma - \phi_{\mathbb{H}}(e_h, r, e_{t'})), 0\right), \tag{11}$$

where γ is a scalar margin value used to train the model to distinguish positive and negative triples. By setting a suitable γ, the Hyperbolic geodesic distance between positive pairs $\phi_{\mathbb{H}}(e_h, r, e_t)$ is expected to be smaller than $\phi_{\mathbb{H}}(e_h, r, e_{t'})$ between negative pairs.

3.4.3 Joint Learning. In this part, we jointly train both components in the framework by combining two loss functions above,

$$\mathcal{L}_{Joint} = \mathcal{L}_{Inference} + \omega\mathcal{L}_{Geodesic}. \tag{12}$$

Table 2: The statistical information of two QA datasets.

Dataset	Question	Choices	Train	Dev	Test
CSQA	12102	5	9741	1221	1140
CSQA(IH)*	12102	5	8500	1221	1241
OBQA	5957	4	4957	500	500

* As the official test set of CSQA is not publically available (predictions are evaluated bi-weekly via the leaderboard), we employ in-house (IH) data split used in [15].

To mitigate impacts from irrelevant neighbors, we employ ω as a weight to constrain $\mathcal{L}_{Geodesic}$.

4 EXPERIMENTS

To evaluate the effectiveness of HamQA, in this section, we aim to answer four research questions:

- **RQ1 (Effectiveness):** How effective is HamQA compared with the state-of-the-art multi-hop KGQA models?
- **RQ2 (Parameter analysis):** How do hyperparameters influence the performance of HamQA?
- **RQ3 (Ablation study):** How does each component of HamQA contribute to its performance?
- **RQ4 (Case study):** How does our proposed HamQA perform comprehensive reasoning in real-world QA scenarios?

4.1 Experimental Setup

4.1.1 QA Datasets. For fair comparison, we conduct experiments on two benchmark QA datasets following previous studies [9, 40], CommonsenseQA [29] and OpenBookQA [21].

CommonsenseQA, abbreviated as *CSQA*, is constructed as a multiple choice QA dataset with five choices for each question. To answer 12,102 questions in it, a background of commonsense knowledge is required. As the official test set is only used for a leaderboard, we first evaluate the model performance with the in-house (IH) data split used in [15]. **OpenBookQA**, *OBQA* for short, contains 5,957 multiple choice questions equipped with 4 choices. Answering *OBQA* questions requires a broad common knowledge of core science facts and applications. The statistical details are summarized in Table 2, where CSQA and CSQA(IH)* represent the official dataset and the IH split respectively.

Since both owners of two datasets maintain an authoritative leaderboard, we further provide our rankings and compare with related records in the following subsection.

4.1.2 Background Knowledge. We adopt **ConceptNet** [27], a large commonsense KG as the knowledge background. It consists of over 8 million commonsense entities linked by 34 condensed relations, describing abundant hierarchical relationships between real-world entities with *Part_Of*, *Has_Prerequisite*, *Has_A* and *Is_A*, *Causes*, etc. It serves as the background knowledge to facilitate our comprehensive reasoning on both datasets. Each time before a question is inputted to the model, we retrieve a question-specific subgraph $\mathcal{G}_{sub}$ from ConceptNet. $\mathcal{G}_{sub}$ consists of all 2-hop neighbors of both question entities and answer entities.

4.1.3 Baselines. For fair evaluation, we aim to compare HamQA with existing models from three important perspectives in general.

- **Four** fine-tuned language models, BERT-Base, BERT-Large [7], RoBERTa-Large [19], AristoRoBERTa [6]. Notably, AristoRoBERTa is only applicable for tasks on OpenBookQA. Methods equipped with AristoRoBERTa could integrate additional scientific facts as pieces of evidence for inference.
- HamQA is expected to outperform **three** embedding-based reasoning methods: RGCN [26], RN [24] and GconAttn [36]. We extend these relational graph embedding methods to predict answers following prevailing work.
- **Three** SOTA multi-hop KGQA algorithms, KagNet [15], MHGRN [9] and QA-GNN [40] are included to show the effectiveness of comprehensive reasoning with hierarchical information.

4.2 Main Results

To answer **RQ1**, we evaluate HamQA and conduct experiments on two benchmark datasets respectively. We use *Accuracy* (ACC.) as the main evaluation metric, which measures the proportion of questions that are predicted correctly among total questions. For CommonsenseQA in-house split, we apply the pre-trained *RoberTa-Large* to all baseline models and HamQA, main results are summarized in the first two columns of Table 3. 'w/o KG' in the first line means we directly use a fine-tuned LM to predict answers. HamQA achieves comparable improvements of 0.75% and 0.87% when compared with the best model. For OpenBookQA, we leverage the benchmark LM *AristoRoBERTa* to enhance the inferential ability of all models with additional facts, and list the results in the first two columns of Table 5. HamQA significantly outperforms the best model QA-GNN with improvements around 2.74% and 2.27%. Also, its superior performance over embedding-based methods also reflects the value of learning from $\mathcal{G}_{sub}$. We also make an interesting observation that additional information in LMs could significantly improve the model inference. In CommonsenseQA, the performance of *RoberTa-Large* without KG is worse than all baselines, however, *AristoRoBERTa* itself could surprisingly perform better than most of the baselines by integrating additional evidence.

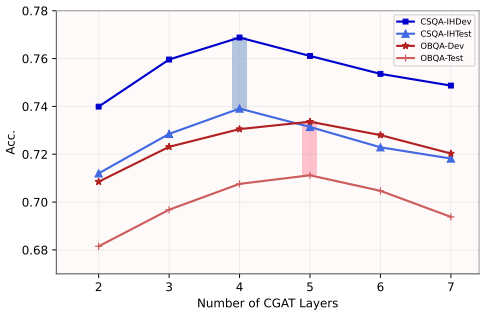

Figure 3: Changes in performance as the number of layers in CGAT increases.

4.2.1 Leaderboard Rankings. For authoritativeness, we submit our prediction results on the official test sets to the official leaderboards of CommonsenseQA and OpenBookQA respectively. Table 4 lists all the related records and methods, including both off-the-shelf methods and anonymous submissions. On the bi-weekly reviewed CommonsenseQA leaderboard, we achieve a comparable ranking to

Table 3: Performance comparison on CommensenseQA in-house split.

Methods	RoBERTa-Large		BERT-Large		BERT-Base	
	IHdev-Acc.	IHtest-Acc.	IHdev-Acc.	IHtest-Acc.	IHdev-Acc.	IHtest-Acc.
w/o KG	73.07%	68.69%	61.06%	55.39%	57.31%	53.47%
RN [24]	74.57%	69.08%	63.04%	58.46%	58.27%	56.20%
RGCN [26]	72.69%	68.41%	62.98%	57.13%	56.94%	54.50%
GconAttn [36]	72.61%	68.59%	63.17%	57.36%	57.27%	54.4%
KagNet [15]	73.48%	68.94%	62.35%	57.19%	55.67%	56.15%
MHGRN [9]	74.45%	71.34%	63.28%	60.59%	60.35%	57.19%
QA-GNN [40]	76.31%	73.27%	63.19%	59.68%	62.32%	58.30%
HamQA (Imp.%)	**76.88%**(+0.75)	**73.91%**(+0.87)	**63.85%**(+0.90)	**61.04%**(+0.74)	**62.79%**(+0.75)	**59.27%**(+1.66)

Table 4: Leaderboard records of related models for CommonsenseQA and OpenBookQA (sorted by rankings).

CommonsenseQA		OpenBookQA	
RoBERTa	**Records**	**AristoRoBERTa**	**Records**
+ KagNet [15]	0.589	+ w/o KG	0.778
+ CSPT	0.696	+ PG [35]	0.802
+ IR	0.721	+ MHGRN [9]	0.806
+ FreeLB [42]	0.731	+ HGN [39]	0.814
+ KE	0.733	+ AMR-SG [38]	0.816
+ KEDGN	0.744	+ CGR [37]	0.824
+ MHGRN [9]	0.754	+ QA-GNN [40]	0.828
+ QA-GNN [40]	**0.761**	UnifiedQA(T5-11B) [13]	0.872
+ **HamQA (Ours)**	0.759	+ **HamQA (Ours)**	**0.850**

QA-GNN with the accuracy at 75.9%. On OpenBookQA leaderboard, we are ranked higher than all the KGQA baselines with the performance of Acc.=85%, regardless of KagNet and MHGRN which are not ranked with no submissions. We obtain improvements of 9.25% over *AristoRoBERTa* and 2.66% over QA-GNN. Though UnifiedQA ranks higher than ours, both the LM they used, i.e., T5 [23] and the model itself own a much larger size of parameters than ours.

4.3 Parameter Analysis

We investigate the impacts of different hyperparameters in this paper. Since the hypothesis is carried out based on Hyperbolic space, curvature c lays decisive influences on the representation ability. As empirically studied by previous works, we set $c = 1$ as a hyperparameter without further tuning.

4.3.1 Graph Layer ℓ of CGAT. Empirically, the layer number in the architecture of CGAT allows informative aggregation from valuable neighbors within corresponding reaches, which makes choosing an appropriate layer number an important problem. Figure 3 shows the changes of model performance with *RoBERTa* when employing an increasing number of layers. It could be easily found that the performance reaches the summit at $l = 4$ for both IHDev and IHTest sets on CommonsenseQA, and it is $l = 5$ for both Dev and Test datasets on OpenBookQA. In conclusion, we consider the layers to be congenitally determined by the hierarchical structures in $\mathcal{G}_{sub}$.

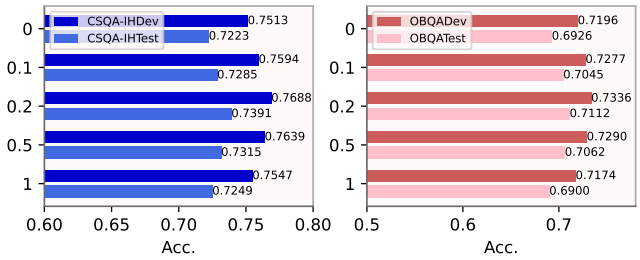

Figure 4: Impact Analysis of the margin value γ on CommonsenseQA and OpenBookQA.

4.3.2 Search of Margin γ. To facilitate the training in the second component, margin γ is required to be cautiously decided. We face a dilemma that on one hand, we expect a larger γ for a more sensitive perception of tough negative samples that are similar to the positive ones. However, this may lead to harder convergence of the model performance. On the other hand, easy training with a smaller γ would damage the model's capability to identify negative triples. Figure 4 compares the performance of HamQA with *RoBERTa* as γ increases. Specifically, we search for an appropriate value from $[0, 0.1, 0.2, 0.5, 1]$. As a result, we find that $\gamma = 0.2$ is consistently the best and uniformly apply it on both datasets.

4.4 Ablation Studies

In this subsection, we answer **RQ3** from three aspects, by elaborating on details of studies on different components in HamQA.

4.4.1 Analysis on Varying Language Models. Sufficient ablation studies on SOTA methods reveal the crucial role of an LM for multi-hop KGQA. The performance of the same architecture with different LMs may surprisingly vary by around 13.84% to 15.76% on CommonsenseQA IHdev set [9], which shows LMs' decisive influences. Thus, we vary different LMs to evaluate the dependency of HamQA on the current benchmark LM. Specifically, we replace *RoBERTa-Large* with *BERT-Base* and *BERT-Large* on CommonsenseQA following

Table 5: Performance comparison on OpenBookQA official split.

Methods	AristoRoBERTa*		RoBERTa-Large		BERT-Large	
	Dev-Acc.	Test-Acc.	Dev-Acc.	Test-Acc.	Dev-Acc.	Test-Acc.
w/o KG	79.91%	78.40%	66.76%	64.80%	62.55%	60.17%
RN [24]	78.40%	75.35%	67.00%	65.20%	63.01%	61.79
RGCN [26]	77.95%	74.60%	64.65%	62.45%	62.54	60.98
GconAttn [36]	74.02%	71.80%	64.30%	61.90%	62.04	59.22
MHGRN [9]	81.44%	80.60%	68.15%	66.87%	63.40%	61.79%
QA-GNN [40]	83.58%	82.71%	72.40%	70.39%	64.98%	63.26%
HamQA (Imp.%)	**85.87%**(+2.74)	**84.59%**(+2.27)	**73.36%**(+1.33)	**71.12%**(+1.04)	**65.85%**(+1.34)	**64.33%**(+1.69)

* Notably, AristoRoBERTa is only applicable for OpenBookQA. Methods equipped with AristoRoBERTa could integrate additional scientific facts as pieces of evidence for inference.

Table 6: Comparison between HamQA and HomQA.

Methods	CommonsenseQA		OpenBookQA	
	IHdev-Acc.	IHtest-Acc.	Dev-Acc.	Test-Acc.
HomQA	0.7512	0.7220	0.7170	0.6905
HamQA	**0.7688**	**0.7391**	**0.7336**	**0.7112**

prevailing settings. Results are elaborately listed in Table 3. HamQA outperforms all the baselines under two LMs, and achieves the most improvements of 1.66% with *BERT-Base* on IHtest set, and 0.9% with *BERT-Large* on IHdev set.

Similarly, we apply *RoBERTa-Large* and *BERT-Large* on OpenBookQA in addition, and show the results in Table 5, where HamQA consistently achieves the best performance with at least 1.04% improvements over the best baseline QA-GNN.

4.4.2 Investigation with hierarchy-only QA. In this part, we intuitively illustrate the importance of each component in comprehensive reasoning, where understanding question from the perspective of semantic hierarchies is indispensable. To be specific, we simply remove the context-aware part and come up with a hierarchy-only multi-hop KGQA, namely **HomQA**. Table 6 compares the performance between HamQA and HomQA on both datasets equipped with *RoBERTa*. This suggests the indispensable effects of being context-aware during the message propagation for multi-hop KGQA. Our intuition is that the attention score is calculated with regard to the importance to current context. HomQA neglects the context information, thus the training signal would merely rely on the prediction loss, which results in a damage to the holistic reasoning performance.

4.4.3 Exploration on the Importance of Hierarchy Approximation. After investigating the importance of the first component, we further ablate the second component in a decremental way. Specifically, to explore the impacts of continuous approximation of hierarchical structures, we adjust the weight ω of $\mathcal{L}_{Geodesic}$ within $[1, 0.1, 0.01, 0.001, 0.0001, 0]$. The variation tendency of the overall performance is shown as a heat map in Figure 5, colors changing from shallow to deep indicates the performance increasing, and vice versa. We make an interesting observation of two stages in the

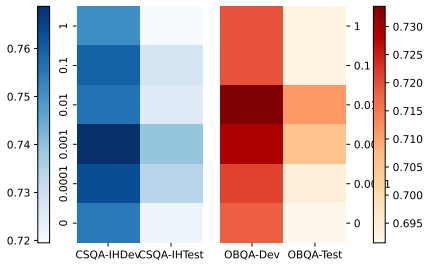

Figure 5: Ablation study on the weight of approximation of hierarchical structures in KGs.

chart. Before reaching the turning points (i.e., the deepest lumps), the prediction accuracy keeps increasing, while it promptly drops as ω continuously decreases. This is mainly due to the complex hierarchical structures in KGs. While information is attentively propagated and the entity embedding is accordingly updated, some irrelevant neighbors should no longer be kept inside current hierarchy anymore. Thus, we still face a dilemma that on the one hand, we expect a strict hierarchical structure with a larger weight of $\mathcal{L}_{Geodesic}$. Nevertheless, the noise would be inevitably introduced if a compelling approximation is applied to the final training loss; On the other hand, an inappropriate small weight may lose the constraints on tree structures. When $\omega = 0$, the model could be considered to be context-aware only. This demonstrates the importance of the second component. In this paper, we adopt $\omega = 0.001$ and $\omega = 0.01$ as final weights of $\mathcal{L}_{Geodesic}$ on two datasets respectively.

4.5 Case Study on Interpretability

To provide insights into the inference capability of HamQA, we conduct case studies with real examples in CommonsenseQA. In general, we aim to answer **RQ4** from two aspects.

First, we compare the prediction results of RoBERTa, QA-GNN and our HamQA with two examples in Table 7. (i) The ability to handle negation is very important for multi-hop QA [40]. To compare interpretability on *negation*, we adopt the first example, which requires consciousness of '**not** have a pen or pencil'. Although RoBERTa makes a wrong prediction, 'write down' is the closest

Table 7: Case studies on model interpretability, compared with RoBERTa, and QA-GNN.

CommonsenseQA Examples	RoBERTa	QA-GNN	HamQA
How would you express information if you do **not** have a pen or pencil? A. Disagree B. Close mouth C. Write down D. Talk E. Eyes	C. Write down (✗)	D. Talk (✓)	D. Talk (✓)
James's niece asked him about her grandfather. She was interested in what? A. Family tree B. Family reunion C. Brother's house D. Heirlooms	D. Heirlooms (✗)	B. Family reunion (✗)	A.Family tree (✓)

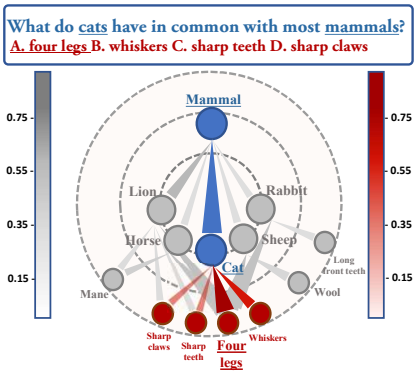

Figure 6: Case study on HamQA's comprehensive reasoning over commonsense questions containing hyponymy.

distractor regardless of the negation word. QA-GNN and HamQA could provide correct answers as they both integrate the contexts to understand questions. By using the second example, we study questions containing semantic hyponymy. Since only HamQA predicts correctly, this demonstrates the superiority of our model to identify the latent hierarchical information. Second, we further investigate the advantage of HamQA on answering questions containing hyponymy. Both RoBERTa and QA-GNN fail to predict correctly since it requires comprehensive reasoning over hierarchical structures. We interpret the reasoning process of HamQA when dealing with such questions in Figure 6. It intuitively demonstrates the attention weights of each entity in the aforementioned running example. Both deeper colors and thicker edges represent larger attention weights of one entity to another. The overall comparison generally shows HamQA's comparable performance of negated questions and superior comprehension of hyponymy-involved questions, as well as normal interrogative sentences.

5 RELATED WORK

5.1 Hyperbolic KG Embedding

The majority of existing efforts dedicate to learning KG representations through Euclidean geometry, which proves to be merely applicable for grids [10]. With the deluge of research interests on Hyperbolic space [1, 4, 5, 18], many models have been proposed to learn hierarchical structures in KGs with Hyperbolic geometry [16]. MURP [2] first introduces Hyperbolic to the community of KG embedding with the exploration on Poincaré Ball. It shows the superior performance on hierarchical relations, e.g., *hypernym* and *has_part* from the real-world KG, WN18RR. ATTH [4] leverages rotation and a tailored Hyperbolic attention. It proves that Hyperbolic space could achieve the same performance as Euclidean space with a much lower dimension. In this paper, we construct Hyperbolic space with Poincaré Ball, and leverage the Möbius operations to facilitate the learning of mutual hierarchical information in question contexts and KG structures.

5.2 Multi-hop KGQA

Recent efforts have been devoted to providing answers by reasoning over a knowledge graph [12, 14**?**], mainly including both semantic parsing and information retrieval methods[17, 25]. KagNet [15] constructs a schema graph to effectively encode paths between topic entities. [20] proposes to enhance the model reasoning ability by learning from multiple external knowledge sources, e.g., Wikipedia. However, both of them treat the question understanding and KG reasoning as separate tasks. This makes models vulnerable to the implicit information in question contexts, e.g., negation and constraints. Recently, several methods realize the shortcoming and consider joint reasoning. MHGRN [9] iteratively updates the question context embeddings during the reasoning with graph neural networks [30–33]. It provides both interpretability and scalability by combining path encoders and GNNs together. Followed by QA-GNN [40], authors construct a working graph that embeds the context as an entity, connected with topic entities by synthetic relations.

6 CONCLUSION

In this paper, we present hierarchy-aware reasoning for multi-hop KGQA at the first attempt. It sheds light on a novel and effective perspective of understanding questions comprehensively. Through three main components, (i) context-aware graph attention network, (ii) preservation of hierarchical structures on KGs, and (iii) a joint optimization scheme, we effectively align the hierarchical information between both question contexts semantically and KG structures topologically. Extensive experiments on two benchmark datasets demonstrate our outperforming performance over the SOTA methods. We also achieve a higher ranking than existing multi-hop KGQA baselines on the official OpenBookQA leaderboard. While it still remains a tough task to understand real-world complex questions like a human, as future work, we are motivated to further explore comprehensive reasoning with sufficient topological structures among question concepts and answers in KGs, e.g., chains, trees and circles, by adaptively modeling the inference over three embedding spaces for various questions.

ACKNOWLEDGMENTS

The work described in this paper was fully supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. PolyU 25208322).

REFERENCES

[1] Gregor Bachmann, Gary Bécigneul, and Octavian Ganea. 2020. Constant curvature graph convolutional networks. In *ICML*. PMLR, 486–496.

[2] Ivana Balazevic, Carl Allen, and Timothy Hospedales. 2019. Multi-relational poincaré graph embeddings. *NeurIPS* 32 (2019).

[3] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. *NeurIPS* 26 (2013).

[4] Ines Chami, Adva Wolf, Da-Cheng Juan, Frederic Sala, Sujith Ravi, and Christopher Ré. 2020. Low-Dimensional Hyperbolic Knowledge Graph Embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 6901–6914.

[5] Ines Chami, Zhitao Ying, Christopher Ré, and Jure Leskovec. 2019. Hyperbolic graph convolutional neural networks. *NeuIPS* 32 (2019).

[6] Peter Clark, Oren Etzioni, Tushar Khot, Daniel Khashabi, Bhavana Mishra, Kyle Richardson, Ashish Sabharwal, Carissa Schoenick, Oyvind Tafjord, Niket Tandon, et al. 2020. From 'F'to 'A'on the NY regents science exams: An overview of the aristo project. *AI Magazine* 41, 4 (2020), 39–53.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[8] Junnan Dong, Qinggang Zhang, Xiao Huang, Qiaoyu Tan, Daochen Zha, and Zhao Zihao. 2023. Active Ensemble Learning for Knowledge Graph Error Detection. In *WSDM*. 877–885.

[9] 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 *EMNLP*. 1295–1309.

[10] Xingcheng Fu, Jianxin Li, Jia Wu, Qingyun Sun, Cheng Ji, Senzhang Wang, Jiajun Tan, Hao Peng, and S Yu Philip. 2021. ACE-HGNN: Adaptive curvature exploration hyperbolic graph neural network. In *ICDM*. IEEE, 111–120.

[11] Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. 2018. Hyperbolic neural networks. *NeurIPS* 31 (2018).

[12] Xiao Huang, Jingyuan Zhang, Dingcheng Li, and Ping Li. 2019. Knowledge graph embedding based question answering. In *WSDM*. 105–113.

[13] Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UNIFIEDQA: Crossing Format Boundaries with a Single QA System. In *EMNLP 2020*. 1896–1907.

[14] Yunshi Lan, Gaole He, Jinhao Jiang, Jing Jiang, Wayne Xin Zhao, and Ji-Rong Wen. 2021. A Survey on Complex Knowledge Base Question Answering: Methods, Challenges and Solutions. In *IJCAI*.

[15] Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. KagNet: Knowledge-Aware Graph Networks for Commonsense Reasoning. In *EMNLP-IJCNLP*. 2829–2839.

[16] Mengqi Lin, Qing Liao, Yan Jia, and Ye Wang. 2021. Survey on Knowledge Graph Embedding Based on Hyperbolic Geometry. In *DSC*. IEEE, 152–158.

[17] Lihui Liu, Boxin Du, Jiejun Xu, Yinglong Xia, and Hanghang Tong. 2022. Joint Knowledge Graph Completion and Question Answering. In *KDD*. 1098–1108.

[18] Qi Liu, Maximilian Nickel, and Douwe Kiela. 2019. Hyperbolic graph neural networks. *NeurIPS* 32 (2019).

[19] 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 preprint arXiv:1907.11692* (2019).

[20] 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 *AAAI*.

[21] 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 *EMNLP*. 2381–2391.

[22] Maximillian Nickel and Douwe Kiela. 2017. Poincaré embeddings for learning hierarchical representations. *NeurIPS* 30 (2017).

[23] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR* 21, 140 (2020), 1–67.

[24] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. 2017. A simple neural network module for relational reasoning. *NeurIPS* 30 (2017).

[25] Apoorv Saxena, Aditay Tripathi, and Partha Talukdar. 2020. Improving multi-hop question answering over knowledge graphs using knowledge base embeddings. In *ACL*. 4498–4507.

[26] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *ESWC*. Springer, 593–607.

[27] Robyn Speer, Joshua Chin, and Catherine Havasi. 2016. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. (2016).

[28] Haitian Sun, Tania Bedrax-Weiss, and William Cohen. 2019. PullNet: Open Domain Question Answering with Iterative Retrieval on Knowledge Bases and Text. In *EMNLP-IJCNLP*. 2380–2390.

[29] 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)*. 4149–4158.

[30] Qiaoyu Tan, Ninghao Liu, and Xia Hu. 2019. Deep representation learning for social network analysis. *Frontiers in Big Data* 2 (2019), 2.

[31] Qiaoyu Tan, Ninghao Liu, Xiao Huang, Soo-Hyun Choi, Li Li, Rui Chen, and Xia Hu. 2023. S2GAE: Self-Supervised Graph Autoencoders are Generalizable Learners with Graph Masking. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*. 787–795.

[32] Qiaoyu Tan, Ninghao Liu, Xing Zhao, Hongxia Yang, Jingren Zhou, and Xia Hu. 2020. Learning to hash with graph neural networks for recommender systems. In *Proceedings of The Web Conference 2020*. 1988–1998.

[33] Qiaoyu Tan, Xin Zhang, Ninghao Liu, Daochen Zha, Li Li, Rui Chen, Soo-Hyun Choi, and Xia Hu. 2023. Bring Your Own View: Graph Neural Networks for Link Prediction with Personalized Subgraph Selection. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*. 625–633.

[34] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903* (2017).

[35] Peifeng Wang, Nanyun Peng, Filip Ilievski, Pedro A Szekely, and Xiang Ren. 2020. Connecting the Dots: A Knowledgeable Path Generator for Commonsense Question Answering. In *EMNLP (Findings)*.

[36] Xiaoyan Wang, Pavan Kapanipathi, Ryan Musa, Mo Yu, Kartik Talamadupula, Ibrahim Abdelaziz, Maria Chang, Achille Fokoue, Bassem Makni, Nicholas Mattei, et al. 2019. Improving natural language inference using external knowledge in the science questions domain. In *AAAI*, Vol. 33. 7208–7215.

[37] Weiwen Xu, Yang Deng, Huihui Zhang, Deng Cai, and Wai Lam. 2021. Exploiting Reasoning Chains for Multi-hop Science Question Answering. In *EMNLP 2021*. 1143–1156.

[38] Weiwen Xu, Huihui Zhang, Deng Cai, and Wai Lam. 2021. Dynamic Semantic Graph Construction and Reasoning for Explainable Multi-hop Science Question Answering. In *ACL-IJCNLP 2021*. 1044–1056.

[39] Jun Yan, Mrigank Raman, Aaron Chan, Tianyu Zhang, Ryan Rossi, Handong Zhao, Sungchul Kim, Nedim Lipka, and Xiang Ren. 2020. Learning contextualized knowledge structures for commonsense reasoning. In *ACL*.

[40] Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. QA-GNN: Reasoning with language models and knowledge graphs for question answering. *NAACL* (2021).

[41] Qinggang Zhang, Junnan Dong, Keyu Duan, Xiao Huang, Yezi Liu, and Linchuan Xu. 2022. Contrastive Knowledge Graph Error Detection. In *CIKM*. 2590–2599.

[42] Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. 2020. FreeLB: Enhanced Adversarial Training for Natural Language Understanding. In *ICLR*.