FiTs: Fine-grained Two-stage Training for Knowledge-aware Question Answering

Bowen Cao[1]*, **Qichen Ye**[1]*, **Nuo Chen**[3,4], **Weiyuan Xu**[1], **Yuexian Zou**[1,2]†

[1]ADSPLAB, School of ECE, Peking University, Shenzhen, China, [2]Peng Cheng Laboratory, Shenzhen, China
[3]Hong Kong University of Science and Technology (Guangzhou), [4]Hong Kong University of Science and Technology
{yeeeqichen, zouyx}@pku.edu.cn, {cbw2021, xuwy}@stu.pku.edu.cn, chennuo26@gmail.com

Abstract

Knowledge-aware question answering (KAQA) requires the model to answer questions over a knowledge base, which is essential for both open-domain QA and domain-specific QA, especially when language models alone cannot provide all the knowledge needed. Despite the promising result of recent KAQA systems which tend to integrate linguistic knowledge from pre-trained language models (PLM) and factual knowledge from knowledge graphs (KG) to answer complex questions, a bottleneck exists in effectively fusing the representations from PLMs and KGs because of *(i)* the semantic and distributional gaps between them, and *(ii)* the difficulties in joint reasoning over the provided knowledge from both modalities. To address the above two problems, we propose a **Fi**ne-grained **T**wo-**s**tage training framework (FiTs) to boost the KAQA system performance: The first stage aims at aligning representations from the PLM and the KG, thus bridging the modality gaps between them, named knowledge adaptive post-training. The second stage, called knowledge-aware fine-tuning, aims to improve the model's joint reasoning ability based on the aligned representations. In detail, we fine-tune the post-trained model via two auxiliary self-supervised tasks in addition to the QA supervision. Extensive experiments demonstrate that our approach achieves state-of-the-art performance on three benchmarks in the commonsense reasoning (*i.e.*, CommonsenseQA, OpenbookQA) and medical question answering (*i.e.*, MedQA-USMILE) domains.

Introduction

Recent advances in large pre-trained language models (PLM) have demonstrated distinguishable applicability in various tasks. (Chen et al. 2021; Cheng et al. 2022; Jin et al. 2022; Cao et al. 2022; Li et al. 2023). However, empirical studies show that PLMs may struggle when dealing with examples that are distributionally different from the pre-training and fine-tuning corpora (Kassner and Schütze 2020). This shortcoming limits their performance in the open domain question answering (QA) task that requires a wide range of factual knowledge. More recently, some works (Mihaylov and Frank 2018; Feng et al. 2020) focus on the knowledge-aware question answering (KAQA) task

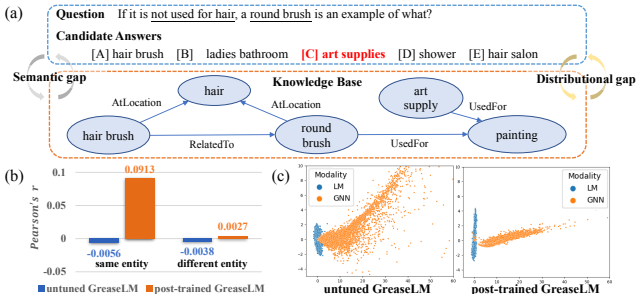

Figure 1: (a) An example of the knowledge-aware QA task from CommonsenseQA. (b) The Pearson correlation coefficient of representations for the same/different entity from LM and GNN, which represents the semantic consistency between the two modalities (+1: positive correlation; -1: negative correlation; 0: no linear dependency). (c) PCA visualization of the distributions of entity representations from the untuned GreaseLM (integrating RoBERTa-Large and GAT), and the corresponding post-trained GreaseLM. The latter serves as a better starting point for fine-tuning.

(cf. Figure 1(a)), which allows access to external knowledge bases, especially knowledge graphs (KG), because KGs capture a broad coverage of factual knowledge explicitly using triplets that encode the relationships between entities. Taking advantage of both PLMs and KGs, these systems achieve remarkable results for tasks requiring open domain knowledge and structured reasoning (Lin et al. 2019; Feng et al. 2020; Yasunaga et al. 2021).

The overwhelming majority of state-of-the-art KAQA methods can be classified into two categories: semantic parsing-based (*SP-based*) methods (Luo et al. 2018; Sun et al. 2020) and information retrieval-based (*IR-based*) methods (Chen et al. 2019; Zhang et al. 2022). This paper mainly focuses on the latter which follows a two-stage procedure: *(i)* retrieving relevant knowledge from KGs under the information conveyed in the question; and then *(ii)* fusing the retrieved knowledge and the contextualized representations captured by PLMs to perform joint reasoning.

However, IR-based KAQA models inevitably suffer from two problems: (1) **Modality Gaps.** There exist two intrinsic differences between two modalities, *i.e.*, the semantic

*These authors contributed equally.
†Corresponding author.

gap and distributional gap. On one hand, as shown in Figure 1(b), the weak dependency between representations of the same entity (*e.g.*, "`round brush`" in the QA context and in the knowledge base in Figure 1(a)) from the untuned GreaseLM (Zhang et al. 2022) suggests the **semantic gap** between the two modalities. On the other hand, we find that both the representations of the LM and the GNN are restricted to narrow cones with different shapes and apart from each other (which is consistent with the empirical findings of Liang et al. (2022)), indicating the **distributional gap**; (2) **Difficulties in Joint Reasoning**, which cause the problem in training the model to answer questions over the provided two sources of knowledge. We further discuss it in section *Knowledge-aware Fine-tuning*.

To address the above two problems, we propose a **Fi**ne-grained **T**wo-**s**tage training framework (**FiTs**), including post-training and fine-tuning stages (cf. Figure 2). **In the first stage**, we present a simple but effective post-training method with the knowledge adaptive (**KA**) objective that aligns the representations from PLMs and KGs. As shown in Figure 1(b), the semantic consistency between representations of the same entity is greatly improved after post-training; Figure 1(c) shows that the distributions of the two modalities are adapted to each other. They demonstrate that both the semantic and the distributional gaps are alleviated, leading to a better starting point for fine-tuning. **In the second stage**, our motivation is to train the model to efficiently and effectively reason with both sources of knowledge. To this end, we develop two auxiliary self-supervised learning objectives—*(i)* knowledge source distinction (**KSD**): we let the model distinguish whether a retrieved entity is related to the question, related to the answer, or an irrelevant entity to improve model's ability in knowledge understanding; and *(ii)* knowledge backbone regularization (**KBR**): we impose a regularization term on the retrieved KG knowledge triplets to enhance the joint reasoning ability of the model—in addition to the supervision signal to perform fine-tuning.

Experimental results on three benchmarks (*i.e.*, CommonsenseQA, OpenbookQA, and MedQA-USMILE) demonstrate that FiTs boost the model performance for multiple choice question answering, a typical task in KAQA. Specifically, our model achieves an absolute improvement in accuracy by 2.6%, 1.8%, and 0.6% on the above three benchmarks, respectively, with only **1%** additional trainable parameters, suggesting the effectiveness and the domain generality of the proposed method[1].

Related Work

Current KAQA methods can be categorized into two groups: IR-based approaches and SP-based approaches.

SP-based methods (Kapanipathi et al. 2020) reason over KGs with logic forms generated by conducting syntactic and semantic analysis on the question. Since the quality of the logic form is highly dependent on the parsing module which converts unstructured text into structured representations, complex questions with compositional semantics increase the difficulties in linguistic analysis, leading to a bottleneck in performance improvement. Meanwhile, manually annotating logic forms is costly and labor-intensive, which poses another limitation.

[1]The code can be found at https://github.com/yeeeqichen/FiTs

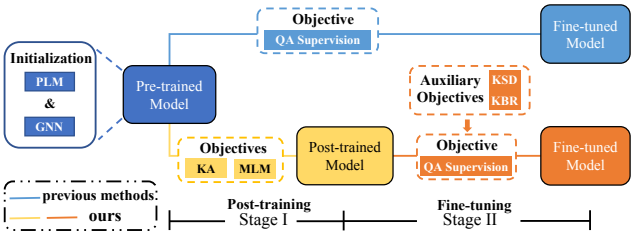

Figure 2: The pipelines of previous methods and our two-stage training framework for KAQA tasks.

IR-based KAQA systems typically consist of the modules of KG retrieval and joint reasoning. For KG retrieval, we follow the procedure from Yasunaga et al. (2021) to retrieve relevant KG attributes. To perform joint reasoning over PLMs and the retrieved KG sub-graphs, the most critical challenge is that models have to fuse the semantically and distributionally different knowledge encoded in PLMs and KGs. Some works separately capture language representations and graph representations with two-tower models (Wang et al. 2019). They suffer from the above issue due to a lack of interactions between the two modalities. Other works take one modality as auxiliary knowledge to ground the other, such as *(i)* exploiting the textual representation to augment a graph reasoning model (Lv et al. 2020), and *(ii)* augmenting the language representation of a QA example with the encoded graph knowledge (Lin et al. 2019; Yang et al. 2019). As for these methods, the information flows one way between the two modalities, which still limits the knowledge interaction.

More recently, GreaseLM mixes the multi-modal representations in the intermediate layers of the LM and the GNN to enable two-way interactions between both modalities. However, GreaseLM neglects to adapt the primary representations of PLMs and KGs to each other, thus poorly seeding the joint structure. In order to bridge the gaps between PLMs and KGs before joint learning, we propose a knowledge adaptive post-training objective, which brings a better starting point to the joint reasoning module. Simultaneously, we retain the structure of GreaseLM as our backbone model.

Additionally, due to the imperfectness of the KG retrieval module, models have to reason with inadequate and noisy knowledge retrieved from KGs. Although some works evaluate KAQA model robustness in severe settings, how to explicitly improve this ability remains an open question (Gu et al. 2021). To this end, in addition to the supervision signal, we design two auxiliary self-supervised fine-tuning objectives to improve model performance in utilizing useful knowledge and distinguishing irrelevant knowledge.

Proposed Method

In this section, we will explain the problem of KAQA we are investigating and introduce the objectives for our post-training and fine-tuning methods. The pipeline of our two-

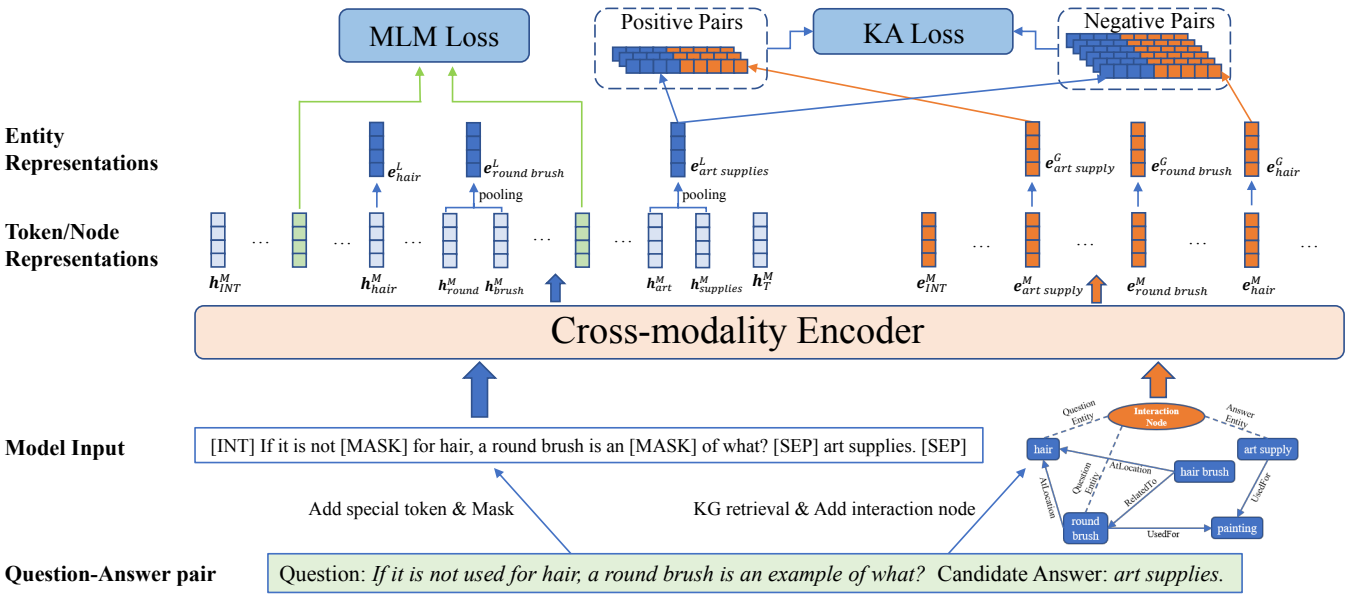

Figure 3: An overview of the post-training process. First, we transform the question-answer pair to model input (*i.e.*, masked context and KG sub-graph), and use the cross-modality encoder to get the fused token(node) representations $\mathbf{h}^M(\mathbf{e}^M)$. Then *(i)* the token representations corresponding to the [MASK] token are used to calculate the MLM loss, and *(ii)* the entity representations $\mathbf{e}^L$ for text entities and $\mathbf{e}^G$ for KG nodes are used to calculate the knowledge adaptive (KA) loss.

stage training framework is illustrated in Figure 2.

Problem Formalization

We mainly focus on the knowledge-aware multiple choice question answering (KAMCQA) task. Generally speaking, an MCQA-type dataset consists of examples with a context paragraph c, a question q, and a candidate answer set A, all in text format. Treating c and q as a whole, an MCQA example can be regarded as a Q-A pair. The KAMCQA task is an extension of MCQA, where an external knowledge graph G is accessible to provide auxiliary knowledge relevant to a given MCQA example.

In practice, due to computational factors, a KAQA system first retrieves a sub-graph G_{sub} from G for each MCQA example which consists of a certain number of entities that are most relevant to that example. We follow the procedure from Yasunaga et al. (2021) to compute the relevance score between each KG entity and the MCQA example. Given an example (c, q, A) and the retrieved G_{sub} as input, the task is to identify which answer $a \in A$ is correct. For simplicity, when computing the probability of a candidate answer being the correct answer, we refer to that answer as a.

Cross-modality Encoder

The cross-modality encoder extracts and fuses the information from text and KG. We use GreaseLM as the cross-modality encoder and give a brief introduction here (see more details in our supplementary material). GreaseLM employs a PLM to encode the textual input (c, q, a) and a GNN model to process the G_{sub}, and further use a two-layer MLP to mix these representations. Specifically, the textual context is appended with a special **interaction token** and passed through N LM-based unimodal encoding layers to get the pre-encoded language representations $\{\mathbf{h}_{int}, \mathbf{h}_1, ..., \mathbf{h}_T\}$. Simultaneously, an **interaction node**, whose representation will be used to interact with the representation of the interaction token, is also added to the G_{sub}. In each of the following M GreaseLM layers, the language representations are fed into transformer LM encoder blocks that continue to encode textual context, and graph representations are fed into a GNN layer to perform a round of information propagation between nodes in the graph:

$$\{\tilde{\mathbf{h}}^l_{int}, \mathbf{h}^l_1, .., \mathbf{h}^l_T\} = \text{LM-Enc}\left(\{\mathbf{h}^{l-1}_{int}, \mathbf{h}^{l-1}_1, .., \mathbf{h}^{l-1}_T\}\right) \quad (1)$$

$$\{\tilde{\mathbf{e}}^l_{int}, \mathbf{e}^l_1, ..., \mathbf{e}^l_J\} = \text{GNN}\left(\{\mathbf{e}^{l-1}_{int}, \mathbf{e}^{l-1}_1, ..., \mathbf{e}^{l-1}_J\}\right) \quad (2)$$

Then the representations of the interaction token and the interaction node are concatenated and passed through the MLP to get a modality-wise mixed representation:

$$\left[\mathbf{h}^l_{int}; \mathbf{e}^l_{int}\right] = \text{MLP}\left(\left[\tilde{\mathbf{h}}^l_{int}; \tilde{\mathbf{e}}^l_{int}\right]\right) \quad (3)$$

Consequently, the cross-modality encoder ensures that information propagates between both modalities.

Knowledge Adaptive Post-training

In fusing the representations from PLMs and KGs, two main challenges exist: *(i)* since PLMs and KGs are usually pre-trained on different corpora, the representations from the two modalities may be semantically contradictory, thus leading to confusion in joint reasoning; and *(ii)* the distribution of them may be discordant, which means that even similar semantics may be embedded in an opposite direction in the latent space. In order to close the gap between both modalities, *i.e.*, provide a better initialization for fine-tuning, we

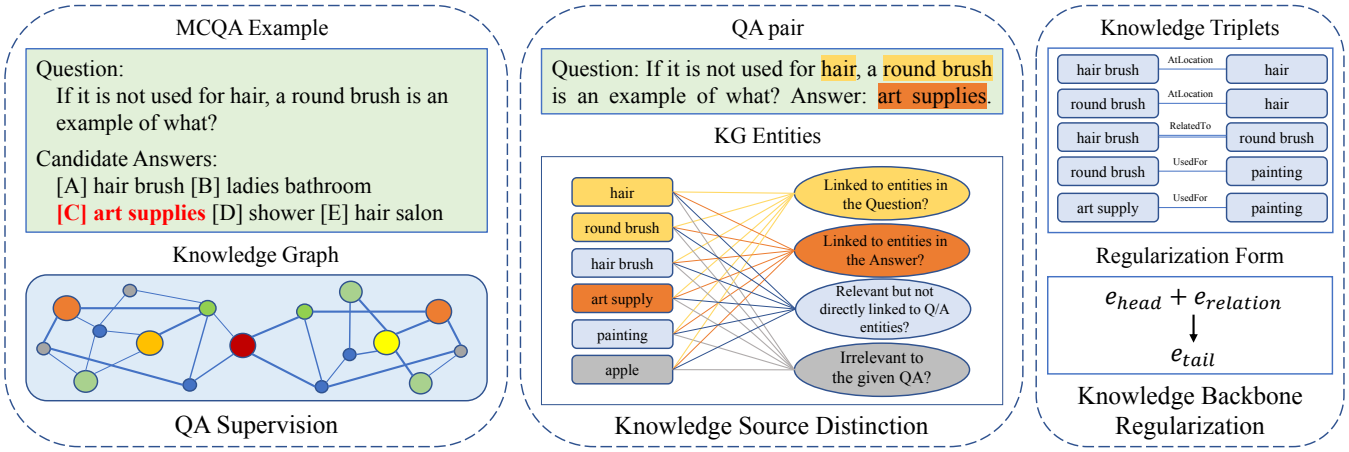

Figure 4: An overview of the knowledge-aware fine-tuning objectives.

propose a knowledge adaptive (KA) loss in addition to the MLM (Devlin et al. 2018) loss to perform post-training.

As shown in Figure 3, we get token representations $\mathbf{H}^M = \{\mathbf{h}^M_{int}, \mathbf{h}^M_1, ..., \mathbf{h}^M_T\}$ and KG node representations $\mathbf{E}^M = \{\mathbf{e}^M_{int}, \mathbf{e}^M_1, ..., \mathbf{e}^M_K\}$ through the cross-modality encoder. We then obtain text entity representations $\mathbf{E}^L = \{\mathbf{e}^L_{int}, \mathbf{e}^L_1, ..., \mathbf{e}^L_J\}$ by pooling over the corresponding token representations, where each $\mathbf{e}^L_i \in \mathbf{E}^L$ represents an entity in the input Q-A pair. Meanwhile, we regard KG node representations as KG entity representations $\mathbf{E}^G = \{\mathbf{e}^G_{int}, \mathbf{e}^G_1, ..., \mathbf{e}^G_J\}$, *i.e.*, $\mathbf{E}^G = \mathbf{E}^M$, because each node in the KG represents an entity. Then, we adopt a contrastive learning framework to align the inherent knowledge in $\mathbf{E}^L$ and $\mathbf{E}^G$, where each pair of $\mathbf{e}^L_i$ and $\mathbf{e}^G_j$ that represents the same entity constitutes a positive pair (*e.g.*, $\mathbf{e}^L_{art\ supplies}$ and $\mathbf{e}^G_{art\ supply}$ in Figure 3) and $\mathbf{e}^L_i$ is treated as a negative example for entity representations other than $\mathbf{e}^G_j$ in $\mathbf{E}^G$, vice versa. Given a positive or negative pair of entity representations, we concatenate them and pass the joint representation through an MLP to get the probability $\hat{y} \in R$ that indicates whether the two entities are matched:

$$\hat{y} = \mathbf{W}_1 \, \mathrm{ReLU}\left(\mathbf{W}_0 \left[\mathbf{e}^L_i; \mathbf{e}^G_j\right] + \mathbf{b}_0\right) \tag{4}$$

where $\mathbf{W}_1 \in R^d$, $\mathbf{W}_0 \in R^{d \times d}$ and $\mathbf{b}_0 \in R^d$ are trainable parameters, $d = d_l + d_g$, d_l and d_g are the dimension of $\mathbf{e}^L_i$ and $\mathbf{e}^G_j$, respectively.

For each Q-A pair, we choose k positive pairs and k corresponding negative pairs. We use labels $\mathbf{y} = [y_1, y_2, ..., y_{2k}]$ to distinguish positive and negative pairs, where $y_i = 1$ if and only if the i-th one is a positive pair, otherwise $y_i = 0$. The calculation of the KA loss is as follows:

$$\mathcal{L}_{KA} = -\frac{1}{2k} \sum_{i=1}^{2k} (1 - y_i) \log\left(1 - \hat{y}_i\right) + y_i \log\left(\hat{y}_i\right) \tag{5}$$

Similar to BERT (Devlin et al. 2018), we randomly choose 15% tokens in the MCQA example to perform MLM.

The overall post-training loss is the sum of the two losses:

$$\mathcal{L}_{post} = \mathcal{L}_{KA} + \mathcal{L}_{MLM} \tag{6}$$

Knowledge-aware Fine-tuning

To fully exploit the encoded knowledge in PLMs and KGs, we propose two auxiliary self-supervised learning objectives in addition to the QA supervision signal.

QA Supervision We follow the procedure from Zhang et al. (2022) to formulate the fully-supervised objective: for a n-way MCQA example (c, q, A), the probability $p\left(a_i|q, c\right)$ that $a_i \in A$ is the correct one is computed as:

$$p\left(a_i \mid q, c\right) \propto \exp\left(\mathrm{MLP}\left(\mathbf{h}^M_{int}, \mathbf{e}^M_{int}, \mathbf{g}\right)\right) \tag{7}$$

where $\mathbf{g}$ is computed by attentively pooling over the KG node embeddings $\{\mathbf{e}^M_1, ..., \mathbf{e}^M_J\}$ using $\mathbf{h}^M_{int}$ as query. Then we compute the cross-entropy loss:

$$\mathcal{L}_{Sup} = -\sum_{i=1}^{n} y_i \log\left(p\left(a_i \mid q, c\right)\right) \tag{8}$$

where $y_i = 1$ if a_i is the correct answer, otherwise $y_i = 0$.

In the inference time, the answer is predicted by:

$$a_p = \mathrm{argmax}_{a \in A} \; p\left(a \mid q, c\right) \tag{9}$$

Knowledge Source Distinction In the process of KG retrieval, entities related to the question or answer and edges connecting these entities are retrieved to form a sub-graph. Distinguishing whether a retrieved entity is related to the question or the answer is an essential aspect of knowledge understanding. For example,

A weasel has a thin body and short legs to easier burrow after prey in a what?
(A) tree (B) mulberry bush (C) chicken coop
(D) viking ship (E) rabbit warren

To pick out the correct answer *(E) rabbit warren*, knowledge about the predator-prey relationship and the narrowness of the prey's lair is required. Under real circumstances, preparing for this knowledge will introduce noise like weasels'

dens are narrow. So the model has to be clear that rabbits, in the candidate answer (E), have narrow warrens, while weasel appears in the question as the predator. Otherwise, guided by the ambiguous noisy knowledge, the wrong candidate answer (C) may be chosen.

Moreover, due to the imperfectness of the KG retrieval module, suspicious entities are often introduced to G_{sub}, *i.e.*, containing lots of irrelevant entities. Thus, it is important to let the model distinguish whether a retrieved entity is relevant to the Q-A pair. However, distinguishing whether a retrieved entity is relevant to the context is quite challenging due to lack of supervision, so we want to train the model in a heuristic way. To this end, we manually add k_{irr} irrelevant entities to G_{sub} to guide the model to learn distinguishable representations for significantly irrelevant entities and gradually differentiate existing irrelevant entities from the others.

To achieve the above two goals, given the representations $\mathbf{E}^G = (\mathbf{e}_1^G, \mathbf{e}_2^G, \ldots, \mathbf{e}_{m+k_{irr}}^G)$ of $m + k_{irr}$ entities in G_{sub}, we formulate the knowledge source distinction loss as:

$$\mathcal{L}_{KSD} = -\sum_{i=1}^{m+k_{irr}} \sum_{j=1}^{4} y_{ij} \log\left(\hat{y}_{ij}\right) \tag{10}$$

$$\hat{\mathbf{y}}_i = \text{softmax}\left(\mathbf{W}_3 \,\text{ReLU}\left(\mathbf{W}_2 \mathbf{e}_i^G + \mathbf{b}_1\right)\right) \tag{11}$$

where $\mathbf{W}_3 \in R^{4\times d_g}$, $\mathbf{W}_2 \in R^{d_g \times d_g}$ and $\mathbf{b}_1 \in R^{d_g}$ are trainable parameters, d_g is the dimension of $\mathbf{e}_i^G$, $\mathbf{y}_i = [y_{i1}, \ldots, y_{i4}]$ is the one-hot vector indicating whether the i-th KG entity is: (1) linked to entities in the question, (2) linked to entities in the candidate answer, (3) an entity in the retrieved multi-hop neighborhood but not directly linked to mentioned entities, or (4) an irrelevant entity.

Knowledge Backbone Regularization The knowledge backbone regularization objective is designed to guide the model to better understand the internal relation of KG knowledge triplets $< h, r, t >$, where h and t are head entity and tail entity, respectively, r is the relationship between them. Inspired by TransE (Bordes et al. 2013), where relationships are represented as translations in the embedding space, we assume that, in the latent representation space, the summation of the head entity representation $\mathbf{e}_h$ and the relationship representation $\mathbf{e}_r$ should be close to the tail entity representation $\mathbf{e}_t$ as much as possible, *i.e.*, $\mathbf{e}_h + \mathbf{e}_r \rightarrow \mathbf{e}_t$.

To this end, given the entity and relationship representations produced by the cross-modality encoder, we introduce a regularization for the k_{reg} knowledge triplets:

$$\mathcal{L}_{KBR} = \sum_{i=1}^{k_{reg}} \left(1 - \cos\left(\mathbf{e}_{h_i} + \mathbf{e}_{r_i}, \mathbf{e}_{t_i}\right)\right) \tag{12}$$

The overall fine-tuning loss is as follows:

$$\mathcal{L}_{finetune} = \mathcal{L}_{Sup} + \mathcal{L}_{KSD} + \mathcal{L}_{KBR} \tag{13}$$

Experimental Setups

See implementation details in our supplementary material.

Datasets

We evaluate our proposed post-training and fine-tuning methods on three MCQA datasets: CommonsenseQA (Talmor et al. 2019) and OpenbookQA (Mihaylov et al. 2018) for commonsense reasoning, and MedQA-USMILE (Jin et al. 2021) as a medical QA benchmark.

The **CommonsenseQA** dataset includes 12,102 5-way multiple-choice questions. Each question requires extra commonsense knowledge beyond the surface-level textual information given by the context. Due to the fact that the official test is hidden, our experiments are conducted using the in-house data split of Lin et al. (2019).

The **OpenbookQA** dataset consists of 5,957 4-way multiple-choice questions about elementary scientific knowledge, along with an open book of scientific facts. We perform experiments using the official data splits of Mihaylov and Frank (2018).

The **MedQA-USMILE** dataset includes 12,723 4-way multiple-choice questions that are originally collected from the National Medical Board Examination in the USA. These questions assess a model's ability to apply medical and clinical knowledge, concepts, and principles.

Baselines

For commonsense reasoning tasks, we include Roberta-Large (Liu et al. 2019) and several advanced knowledge graph enhanced question answering systems as baselines for comparison: (1) RGCN (Schlichtkrull et al. 2018), (2) KagNet (Lin et al. 2019), (3) MHGRN (Feng et al. 2020), (4) QA-GNN (Yasunaga et al. 2021), and (5) GreaseLM (Zhang et al. 2022). All of these methods, except for Roberta-Large, follow the paradigm of LM+KG. GreaseLM is the top-performing one that fuses representations from the LM and KG with modality interaction layers. For MedQA-USMILE, besides LM+KG methods, we compare with BioBERT-Large (Lee et al. 2020), a pre-trained biomedical language representation model based on BERT-Large (Devlin et al. 2018), and SapBERT (Liu et al. 2021), a state-of-the-art model for medical entity representation learning, which improves model's ability to capture entity relationships with the help of entity disambiguation objectives.

Results and Analysis

Our results in Tables 1 and 2 demonstrate consistent improvements on the CommonsenseQA (CSQA) and OpenBookQA (OBQA) datasets. On CSQA, our model's test performance improves by 7.5% over RoBERTa-Large (finetuned LM without KG), 2.8% over LM+KG methods before GreaseLM, and 2.6% over our implementation of the best prior LM+KG system, GreaseLM, on which we evaluate our proposed post-training and fine-tuning methods. As for the OBQA dataset, our model's test performance improves by 7.6% over AristoRoBERTa (finetuned LM without KG), 3.2% over existing LM+KG systems other than GreaseLM, and 1.8% over the GreaseLM that we have implemented. The boost over GreaseLM reveals the superiority of our proposed knowledge-adaptive post-training and

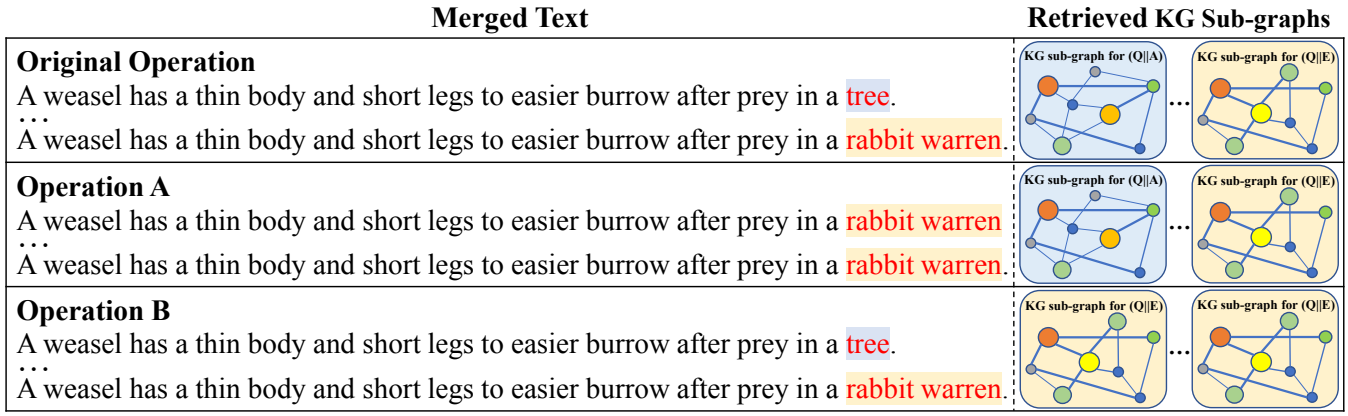

Figure 5: Operations A and B are used for quantitative analysis (section). The unprocessed MCQA example is shown in Figure 1. $Q||A$ denotes the text obtained by merging the question Q and the candidate answer A.

Model	IHtest-Acc (%)
RoBERTa-Large (w/o KG)♣	68.7
RGCN (Schlichtkrull et al. 2018)♣	68.4
KagNet (Lin et al. 2019)♣	69.0
MHGRN (Feng et al. 2020)♣	71.1
QA-GNN (Yasunaga et al. 2021)♣	73.4
GreaseLM (Zhang et al. 2022)♣	74.2
GreaseLM (Our implementation)	73.6
+ **FiTs (Ours)**	**76.2**

Table 1: Performance comparison on CommonsenseQA in-house split. We report the in-house Test (IHtest) accuracy using the data split of Lin et al. (2019), because the official test is hidden. ♣: results from Zhang et al. (2022); all other results are reproduced by ourselves.

knowledge-aware fine-tuning methods in making use of the inherent knowledge from PLMs and KGs.

Quantitative Analysis

Empirically, both the parametric knowledge and the joint reasoning ability of the model benefit KAQA tasks (Longpre et al. 2021). Provided the overall performance improvements, we investigated whether both knowledge sources make contributions or not by conducting new operations on the dataset. Originally, given an unprocessed MCQA example (where a question and candidate answers are separate), we separately merge the question and each candidate answer together and obtain KG sub-graphs based on the merged text to generate training/validation/testing data (the Original Operation in Figure 5). To evaluate the contribution of the model's joint reasoning ability, we apply **operation A**, where each candidate answer is replaced with the correct answer while their retrieved sub-graphs remain unchanged. This operation restricts the model to infer only based on the question and the KG sub-graph. As for the model's parametric knowledge, we apply **operation B**, where each sub-graph is replaced with the one obtained based on the correct answer while all candidate answers remain unchanged. The KG knowledge provided is equivalent and relevant to the correct answer, so the model is forced to make judgments based on an understanding of the relationship between a question and its candidate answers.

Model	Test-Acc (%)
AristoRoBERTa (no KG)♣	78.4
RGCN (Schlichtkrull et al. 2018)♣	74.6
MHGRN (Feng et al. 2020)♣	80.6
QA-GNN (Yasunaga et al. 2021)♣	82.8
GreaseLM (Zhang et al. 2022)♣	84.8
GreaseLM (Our implementation)	84.2
+ **FiTs (Ours)**	**86.0**

Table 2: Test accuracy comparison on OpenBookQA. ♣: results from Zhang et al. (2022); all other results are reproduced by ourselves.

Model	test-reason (A)	test-param (B)
GreaseLM (Zhang et al. 2022)	73.4	69.0
+ post-training	73.9	70.9
+ knowledge-aware fine-tuning	74.2	71.2
+ **FiTs (Ours)**	**74.5**	**71.6**

Table 3: The test-reason set evaluates models' joint reasoning ability, while the test-param set measures models' parametric knowledge.

We conduct experiments using the IHtest set of CommonsenseQA. The new test set obtained based on operation A is named test-reasoning, test-reason for short, and that based on operation B is named test-parametric, test-param for short. The results in Table 3 demonstrate that both the proposed post-training and fine-tuning methods can significantly increase the model's parametric knowledge. In comparison, improvements in the model's joint reasoning ability are marginal, which may be partly due to the imperfectness of the KG retrieval module, *i.e.*, a bottleneck for reasoning.

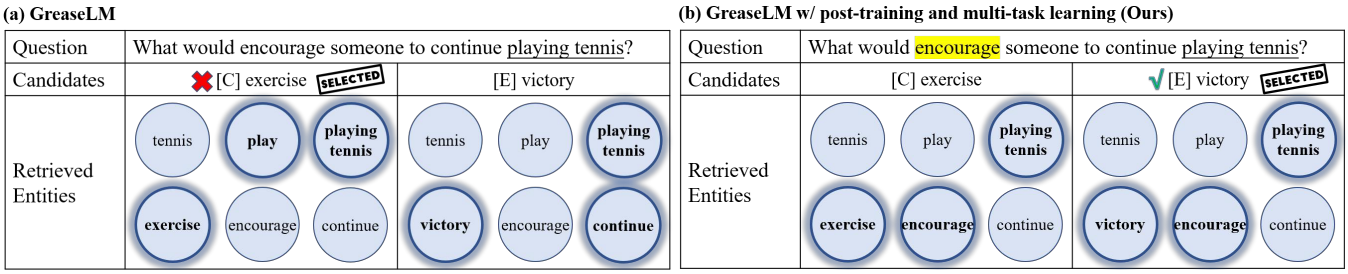

Figure 6: Attention analysis of GreaseLM w/ and w/o our methods. Entities with higher attention weights are highlighted. Our model demonstrates the expected pattern by consistently focusing on the "encourage" entity.

Model	Test-Acc (%)
BioBERT-Base (Lee et al. 2020)♣	34.1
BioBERT-Large (Lee et al. 2020)♣	36.7
QA-GNN (Yasunaga et al. 2021)♣	38.0
GreaseLM (Zhang et al. 2022)♣	38.5
GreaseLM (Our implementation)	38.6
+ **FiTs (Ours)**	**39.2**

Table 4: Test accuracy comparison on MedQA-USMLE. ♣: results from Zhang et al. (2022); all other results are reproduced by ourselves.

Qualitative Analysis

Analyzing the attention mechanism is a crucial way to examine a model's behavior. In this section, we analyze graph attention weights in the last layer of GAT (*i.e.*, the attention weight between each entity in G_{sub} and the context) to find out which entity the model focuses on and investigate whether our model demonstrates a sensible reasoning process. As the example in Figure 6 demonstrates, while GreaseLM concentrates on "play" (given the candidate answer [C] exercise) and "continue" (given the candidate answer [E] victory), our model shows consistent interest on "encourage", which is the core factor in clarifying the relationship between a candidate answer and "playing tennis". Intuitively, our model performs a more expected behavior of human reasoning, suggesting the success of our fine-grained two-stage training framework.

Domain Generality

So far, our model's performance demonstrates the effectiveness of our proposed post-training and fine-tuning methods in the commonsense reasoning domain. Here we further investigate whether they are applicable in the medical domain. Following Zhang et al. (2022), we evaluate our model's performance on the MedQA-USMLE dataset. The results in Table 4 illustrate the generality of our methods, achieving an improvement of 0.6% over the backbone model.

Ablation Studies

We investigated the impact of each part of the post-training and fine-tuning methods through a series of ablation experiments using the ConmensenseQA IHtest set.

Model	IHtest-Acc
GreaseLM (Only QA supervision)	73.6
+ MLM	74.1
+ KA	74.3
+ MLM + KA	**74.5**

Table 5: Ablation studies of the post-training objectives.

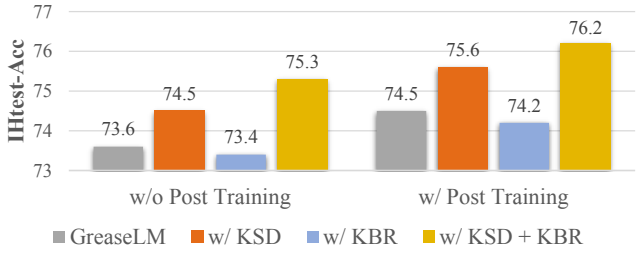

Figure 7: Ablation studies of the fine-tuning objectives.

Post-training: Results in Table 5 suggest that both MLM and KA are helpful for improving the model's test performance, and they complement each other, bringing an improvement by 0.9% over raw GreaseLM. Additionally, the comparison between the left and right half of Figure 7 demonstrates that our post-training method clearly benefits the subsequent fine-tuning process.

Fine-tuning: The usefulness of the two self-supervised objectives and the interdependence between them are demonstrated in Figure 7. Specifically, with or without post-training, *(i)* KSD alone can significantly improve model performance, but KBR alone degrades that because the model without KSD lacks discrimination on the relevance and importance of the external knowledge; *(ii)* KSD and KBR together bring the best result, suggesting that these two objectives complement each other—KSD improves model's discrimination on the relevance and importance of the external knowledge; KBR acts as a commonsense-knowledge-oriented regularization to avoid task-specific overfitting.

Conclusion

This paper introduces **FiTs**, a fine-grained two-stage training framework for KAQA tasks, including the knowledge adaptive post-training stage (with MLM and KA objectives)

and the knowledge-aware fine-tuning stage (with KBR and KSD objectives). Post-training alleviates the gaps between the representations from PLMs and KGs, leading to a better starting point for fine-tuning. Fine-tuned with the proposed objectives, the model is better at identifying how relevant and essential each entity in the retrieved KG sub-graph is to the given question. Experimental results on benchmarks in the commonsense reasoning and medical domains show the great improvements our method brings to the backbone model, which are further demonstrated to be reflected in both model's parametric knowledge and joint reasoning ability. Our work reveals a better way to integrate knowledge from PLMs and KGs and gives insights on how to design learning objectives for KAQA tasks.

Acknowledgement

This paper was partially supported by Shenzhen Science & Technology Research Program (No: GXWD202012311658-07007-20200814115301001) and NSFC (No: 62176008)

References

Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; and Yakhnenko, O. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26.

Cao, M.; Yang, T.; Weng, J.; Zhang, C.; Wang, J.; and Zou, Y. 2022. LocVTP: Video-Text Pre-training for Temporal Localization. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI*, 38–56.

Chen, N.; Liu, F.; You, C.; Zhou, P.; and Zou, Y. 2021. Adaptive bi-directional attention: Exploring multi-granularity representations for machine reading comprehension. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7833–7837. IEEE.

Chen, Z.-Y.; Chang, C.-H.; Chen, Y.-P.; Nayak, J.; and Ku, L.-W. 2019. UHop: An Unrestricted-Hop Relation Extraction Framework for Knowledge-Based Question Answering. In *Proceedings of NAACL-HLT*, 345–356.

Cheng, X.; Dong, Q.; Yue, F.; Ko, T.; Wang, M.; and Zou, Y. 2022. M3ST: Mix at Three Levels for Speech Translation. *CoRR*, abs/2212.03657.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Feng, Y.; Chen, X.; Lin, B. Y.; Wang, P.; Yan, J.; and Ren, X. 2020. Scalable Multi-Hop Relational Reasoning for Knowledge-Aware Question Answering. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, 1295–1309. Association for Computational Linguistics.

Gu, Y.; Kase, S.; Vanni, M.; Sadler, B.; Liang, P.; Yan, X.; and Su, Y. 2021. Beyond IID: three levels of generalization for question answering on knowledge bases. In *Proceedings of the Web Conference 2021*, 3477–3488.

Jin, D.; Pan, E.; Oufattole, N.; Weng, W.-H.; Fang, H.; and Szolovits, P. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14): 6421.

Jin, P.; Huang, J.; Liu, F.; Wu, X.; Ge, S.; Song, G.; Clifton, D. A.; and Chen, J. 2022. Expectation-Maximization Contrastive Learning for Compact Video-and-Language Representations. In *Thirty-Sixth Conference on Neural Information Processing Systems*.

Kapanipathi, P.; Abdelaziz, I.; Ravishankar, S.; Roukos, S.; Gray, A.; Astudillo, R.; Chang, M.; Cornelio, C.; Dana, S.; Fokoue, A.; et al. 2020. Question answering over knowledge bases by leveraging semantic parsing and neuro-symbolic reasoning. *arXiv preprint arXiv:2012.01707*.

Kassner, N.; and Schütze, H. 2020. Negated and Misprimed Probes for Pretrained Language Models: Birds Can Talk, But Cannot Fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7811–7818.

Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; and Kang, J. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4): 1234–1240.

Li, H.; Cao, M.; Cheng, X.; Zhu, Z.; Li, Y.; and Zou, Y. 2023. Generating Templated Caption for Video Grounding. *CoRR*, abs/2301.05997.

Liang, W.; Zhang, Y.; Kwon, Y.; Yeung, S.; and Zou, J. 2022. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *arXiv preprint arXiv:2203.02053*.

Lin, B. Y.; Chen, X.; Chen, J.; and Ren, X. 2019. KagNet: Knowledge-Aware Graph Networks for Commonsense Reasoning. In Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, 2829–2839. Association for Computational Linguistics.

Liu, F.; Shareghi, E.; Meng, Z.; Basaldella, M.; and Collier, N. 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*, 4228–4238.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.

Longpre, S.; Perisetla, K.; Chen, A.; Ramesh, N.; DuBois, C.; and Singh, S. 2021. Entity-Based Knowledge Conflicts in Question Answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 7052–7063.

Luo, K.; Lin, F.; Luo, X.; and Zhu, K. 2018. Knowledge base question answering via encoding of complex query graphs. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2185–2194.

Lv, S.; Guo, D.; Xu, J.; Tang, D.; Duan, N.; Gong, M.; Shou, L.; Jiang, D.; Cao, G.; and Hu, S. 2020. Graph-based reasoning over heterogeneous external knowledge for commonsense question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 8449–8456.

Mihaylov, T.; Clark, P.; Khot, T.; and Sabharwal, A. 2018. Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering. In Riloff, E.; Chiang, D.; Hockenmaier, J.; and Tsujii, J., eds., *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, 2381–2391. Association for Computational Linguistics.

Mihaylov, T.; and Frank, A. 2018. Knowledgeable Reader: Enhancing Cloze-Style Reading Comprehension with External Commonsense Knowledge. In Gurevych, I.; and Miyao, Y., eds., *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, 821–832. Association for Computational Linguistics.

Schlichtkrull, M. S.; Kipf, T. N.; Bloem, P.; van den Berg, R.; Titov, I.; and Welling, M. 2018. Modeling Relational Data with Graph Convolutional Networks. In Gangemi, A.; Navigli, R.; Vidal, M.; Hitzler, P.; Troncy, R.; Hollink, L.; Tordai, A.; and Alam, M., eds., *The Semantic Web - 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3-7, 2018, Proceedings*, volume 10843 of *Lecture Notes in Computer Science*, 593–607. Springer.

Sun, Y.; Zhang, L.; Cheng, G.; and Qu, Y. 2020. SPARQA: skeleton-based semantic parsing for complex questions over knowledge bases. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 8952–8959.

Talmor, A.; Herzig, J.; Lourie, N.; and Berant, J. 2019. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 4149–4158. Association for Computational Linguistics.

Wang, X.; Kapanipathi, P.; Musa, R.; Yu, M.; Talamadupula, K.; Abdelaziz, I.; Chang, M.; Fokoue, A.; Makni, B.; Mattei, N.; et al. 2019. Improving natural language inference using external knowledge in the science questions domain. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 7208–7215.

Yang, A.; Wang, Q.; Liu, J.; Liu, K.; Lyu, Y.; Wu, H.; She, Q.; and Li, S. 2019. Enhancing pre-trained language representations with rich knowledge for machine reading comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2346–2357.

Yasunaga, M.; Ren, H.; Bosselut, A.; Liang, P.; and Leskovec, J. 2021. QA-GNN: Reasoning with Language Models and Knowledge Graphs for Question Answering. In Toutanova, K.; Rumshisky, A.; Zettlemoyer, L.; Hakkani-Tür, D.; Beltagy, I.; Bethard, S.; Cotterell, R.; Chakraborty, T.; and Zhou, Y., eds., *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, 535–546. Association for Computational Linguistics.

Zhang, X.; Bosselut, A.; Yasunaga, M.; Ren, H.; Liang, P.; Manning, C. D.; and Leskovec, J. 2022. GreaseLM: Graph REASoning Enhanced Language Models for Question Answering. *CoRR*, abs/2201.08860.