Ask to Understand: Question Generation for Multi-hop Question Answering

Jiawei Li, Mucheng Ren, Yang Gao, Yizhe Yang
School of Computer Science and Technology,
Beijing Institute of Technology, Beijing, China
Beijing Engineering Research Center of High Volume Language Information
Processing and Cloud Computing Applications, Beijing, China
{jwli,renm,gyang,yizheyang}@bit.edu.cn

Abstract

Multi-hop Question Answering (QA) requires the machine to answer complex questions by finding scattering clues and reasoning from multiple documents. Graph Network (GN) and Question Decomposition (QD) are two common approaches at present. The former uses the "black-box" reasoning process to capture the potential relationship between entities and sentences, thus achieving good performance. At the same time, the latter provides a clear reasoning logical route by decomposing multi-hop questions into simple single-hop sub-questions. In this paper, we propose a novel method to complete multi-hop QA from the perspective of Question Generation (QG). Specifically, we carefully design an end-to-end QG module on the basis of a classical QA module, which could help the model understand the context by asking inherently logical sub-questions, thus inheriting interpretability from the QD-based method and showing superior performance. Experiments on the HotpotQA dataset demonstrate that the effectiveness of our proposed QG module, human evaluation further clarifies its interpretability quantitatively, and thorough analysis shows that the QG module could generate better sub-questions than QD methods in terms of fluency, consistency, and diversity.

1 Intorduction

Unlike single-hop QA (Rajpurkar et al., 2016; Trischler et al., 2017; Lai et al., 2017) where the answers could usually be derived from a single paragraph or sentence, multi-hop QA (Welbl et al., 2018; Yang et al., 2018) is a challenging task that requires soliciting hidden information from scattered documents on different granularity levels and reasoning over it in an explainable way.

The HotpotQA (Yang et al., 2018) was published to leverage the research attentions on reasoning processing and explainable predictions. Figure 1 shows an example from HotpotQA, where the question requires first finding the name of the company (*Tata Consultancy Services*), and then the address of the company (*Mumbai*). While, a popular stream of Graph Network-based (GN) approaches (De Cao et al., 2019; Tu et al., 2019; Ding et al., 2019; Fang et al., 2020) was proposed due to the structures of scattered evidence could be captured by the graphs and reflected in the representing vectors. However, the reasoning process of the GN-based method is entirely different from human thoughts. Specifically, GN tries to figure out the underlying relations between the key entities or sentences from the context. However, the process is a "black-box"; we do not know which nodes in the network are involved in reasoning for the final answer, thus showing relatively poor interpretability.

Inspired by that human solves such questions by following a transparent and explainable logical route, another popular stream of Question Decomposition-based (QD) approaches became favored in recent years (Fu et al., 2021; Nishida et al., 2019; Min et al., 2019; Jiang and Bansal, 2019b). The method mimics human reasoning to decompose complex questions into simpler, single-hop sub-questions; thus, the interpretability is greatly improved by exposing intermediate evidence generated by each sub-question. Nevertheless, the general performance is usually much worse than GN-based ones due to error accumulation that arose by aggregating answers from each single-hop reasoning process. Furthermore, the sub-questions are generated mainly by extracting text spans from the original question to fill the template.

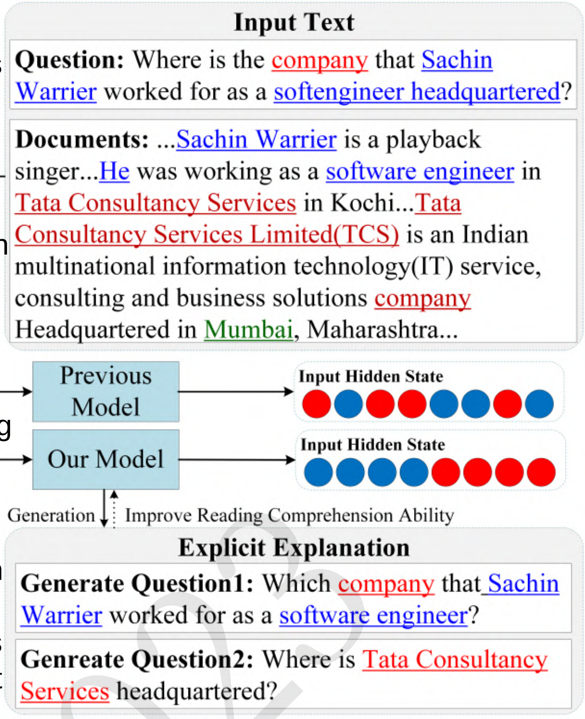
Corresponding author.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 569-582, Harbin, China, August 3 - 5, 2023.

Hence the sub-questions are challenging to guarantee in terms of quality, such as fluency, diversity, and consistency with the original question intention, especially when the original questions are linguistically complex.

In this work, we believe that asking the question is an effective way to elicit intrinsic information in the text and is an inherent step towards understanding it (Pyatkin et al., 2021). Thus, we propose resolving these difficulties by introducing an additional QG task to teach the model to ask questions. Specifically, we carefully design and add one end-to-end QG module based on the classical GN-based module. Unlike the traditional QD-based methods that only rely on information brought by the question, our proposed QG module could generate fluent and inherently logical sub-questions based on the understanding of the original context and the question simultaneously.

Our method enjoys three advantages: First, it achieves better performance. Our approach preserves the GN module, which could collect information scattered throughout the documents and allows the model to understand the context in depth by asking questions. Moreover, the end-to-end training avoids the error accumulation issue; Second, it brings better interpretability because explainable evidence for its decision making could be provided in the form of sub-questions; Thirdly, the proposed QG module has better generalization capability. Theoretically, it can be plugged and played on most traditional QA models.



Question: Where is the company that Sachin Warrier worked for as a softengineer headquartered?

Documents: ...Sachin Warrier is a playback singer...He was working as a software engineer in Tata Consultancy Services in Kochi...Tata Consultancy Services Limited(TCS) is an Indian multinational information technology(IT) service, consulting and business solutions company Headquartered in Mumbai, Maharashtra...

Generate Question1: Which company that Sachin Warrier worked for as a software engineer?

Genreate Question2: Where is Tata Consultancy Services headquartered?

Figure 1: An example from HotpotQA dataset. Text in blue is the first-hop information and text in red is the second-hop information. The mixed encoding of the first-hop information () and the second-hop information () will confuse models with weaker reading comprehension.

Experimental results on the HotpotQA dataset demonstrate the effectiveness of our proposed approach. It surpasses the GN-based model and QD-based model by a large margin. Furthermore, robust performance on the noisy version of HotpotQA proves that the QG module could alleviate the shortcut issue, and visualization on sentence-level attention indicates a clear improvement in natural language understanding capability. Moreover, a human evaluation is innovatively introduced to quantify improvements in interpretability. Finally, exploration on generated sub-questions clarifies diversity, fluency, and consistency.

2 Related Work

Multi-hop QA In multi-hop QA, the evidence for reasoning answers is scattered across multiple sentences. Initially, researchers still adopted the ideas of single-hop QA to solve multi-hop QA (Dhingra et al., 2018; Zhong et al., 2019). Then the graph neural network that builds graphs based on entities was introduced to multi-hop QA tasks and achieved astonishing performance (De Cao et al., 2019; Tu et al., 2019; Ding et al., 2019). While, some researchers paid much attention to the interpretability of the coreference reasoning chains (Fu et al., 2021; Nishida et al., 2019; Min et al., 2019; Jiang and Bansal, 2019b). By providing decomposed single-hop sub-questions, the QD-based method makes the model decisions explainable.

Interpretability Analysis in NLP An increasing body of work has been devoted to interpreting neural network models in NLP in recent years. These efforts could be roughly divided into structural analyses,

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 569-582, Harbin, China, August 3 - 5, 2023.

behavioral studies, and interactive visualization (Belinkov and Glass, 2019).

Firstly, the typical way of structural analysis is to design probe classifiers to analyze model characteristics, such as syntactic structural features (Elazar et al., 2021) and semantic features (Wu et al., 2021). Secondly, the main idea of behavioral studies is that design experiments that allow researchers to make inferences about computed representations based on the model's behavior, such as proposing various challenge sets that aim to cover specific, diverse phenomena, like systematicity exhaustivity (Gardner et al., 2020; Ravichander et al., 2021). Thirdly, for interactive visualization, neuron activation (Durrani et al., 2020), attention mechanisms (Hao et al., 2020) and saliency measures (Janizek et al., 2021) are three main standard visualization methods.

Question Generation QG is the task of generating a series of questions related to the given contextual information. Previous works on QG focus on rule-based approaches. Fabbri et al. (2020) used a template-based approach to complete sentence extraction and QG in an unsupervised manner. Dhole and Manning (2021) developed Syn-QG using a rule-based approach. The system consists of serialized rule modules that transform input documents into QA pairs and use reverse translation counting, resulting in highly fluent and relevant results. One of the essential applications of QG is to construct pseudo-datasets for QA tasks, thereby assisting in improving their performance (Zhang and Bansal, 2019; Alberti et al., 2019; Lee et al., 2020).

Our work is most related to Pyatkin et al. (2021), which produces a set of questions asking about all possible semantic roles to bring the benefits of QA-based representations to traditional SRL and information extraction tasks. However, we innovatively leverage QG into complicated multi-hop QA tasks and enrich representations by asking questions at each reasoning step.

3 Methods

Multi-hop QA is challenging because it requires a model to aggregate scattered evidence across multiple documents to predict the correct answer. Probably, the final answer is obtained conditioned on the first sub-question is correctly answered. Inspired by humans who always decompose complex questions into single-hop questions, our task is to automatically produce naturally-phrased sub-questions asking about every reasoning step given the original question and a passage. Following the reasoning processing, the generated sub-questions further explain why the answer is predicted. For instance, in Figure 1, the answer 'Mumbai' is predicted to answer Question 2 which is conditioned on Question 1's answer. More importantly, we believe that the better questions the model asks, the better it understands the reading passage and boosts the performance of the QA model in return.

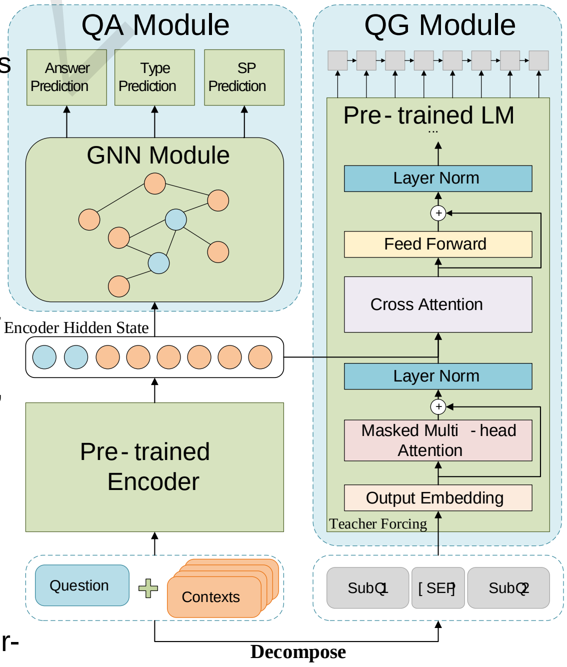


Figure 2: Overall model architecture.

Figure 2 illustrates the overall framework of our proposed model. It consists of two modules: QA module (Section §3.1) and QG module (Section §3.2). The QA module could help model to solve multi-hop QA in a traditional way, and the QG module allows the model to solve the question in an interpretable manner by asking questions. These two modules share the same encoder and are trained end-to-end with multi-task strategy.

3.1 Question Answering Module

Encoder A key point of the GN-based approach to solving QA problems is the initial encoding of entity nodes. Prior studies have shown that pre-trained models are beneficial for increasing the comprehension of the model (Yang et al., 2019; Qiu et al., 2020), which enables better encoding of the input text. In Section 3.2 we will mention that encoder will be shared to the QG module to further increase the model's reading comprehension of the input text through the QG task. Here we chose BERT as the encoder considering its strong performance and simplicity.

GNN Encode Module The representation ability of the model will directly affect the performance of QA. Recent works leverage graphs to represent the relationship between entities or sentences, which have strong representation ability (Xiao et al., 2019; Tu et al., 2020; Fang et al., 2020). We believe that the advantage of graph neural networks is essential for solving multi-hop questions. Thus, we adopt the GN-based model DFGN[1] (Xiao et al., 2019) that has been proven to be effective in HotpotQA.

Xiao et al. (2019) build graph edges between two entities if they co-exist in one single sentence. After encoding the question Q and context C by the pre-trained encoder, DFGN extracts the entities' representation from the encoder output by their location information. Both mean-pooling and max-pooling are used to represent the entities' embeddings. Then, a graph neural network propagates node information to its neighbors. A soft mask mechanism is used to calculate the relevance score between each entity and the question in this process. The soft mask score is used as the weight value of each entity to indicate its importance in the graph neural network computation. At each step, the query embedding should be updated by the entities embedding of the current step by a bi-attention network (Seo et al., 2018). The entities embeddings in the t-th reasoning step:

$$E^t = \mathrm{GAT}([m_1^{t\ \ 1} e_1^{t\ \ 1}; m_2^{t\ \ 1} e_2^{t\ \ 1}; :::; m_n^{t\ \ 1} e_n^{t\ \ 1}]); \tag{1}$$

where $e_i^{t\ \ 1}$ is the i-th entity's embedding at the $(t\ \ 1)$-th step and e_i^0 is the i-th entity's embedding produced both mean-pooling and max-pooling results from encoder output according to its position. $m_i^{t\ \ 1}$ is the relevance score, which is also called soft mask score in previous, between i-th entity and the question at the $(t\ \ 1)$-th step calculated by an attention network. GAT is graph attention networks proposed by Veličković et al. (2017).

In each reasoning step, every entity node gains some information from its neighbors. An LSTM layer is then used to produce the context representation:

$$C^t = \mathrm{LSTM}([C^{t\ \ 1}; ME^{t>}]); \tag{2}$$

where M is the adjacency matrix which records the location information of the entities.

The updated context representations are used for different sub-tasks: (i) answer type prediction; (ii) answer start position and answer end position; (iii) extract support facts prediction. All three tasks are jointly performed through multitasking learning.

$$L_{qa} = \ _1L_{start} + \ _2L_{end} + \ _3L_{type} + \ _4L_{para}; \tag{3}$$

where $_1$, $_2$, $_3$, $_4$ are hyper-parameters[2].

3.2 Question Generation Module

Question Generation Training Dataset A key challenge of training the QG module is that it is challenging to obtain the annotated sub-questions dataset. To achieve this, we take the following steps to generate sub-question dataset automatically:

First of all, according to the annotations provided by the HotpotQA dataset, the questions in the training set could be classified into the following two types Bridge (70%) and Comparison (30%), where

[1] QA module is not the main focus of this work, and DFGN is one of the representative off-the-shelf QA models. In fact, any QA model could be adopted to replace it.

[2] In our experiments, we set $_1 = \ _2 = \ _3 = 1$, $_4 = 5$

the former one requires finding evidence from first hop reasoning then use it to find second-hop evidence, while the latter requires comparing the property of two different entities mentioned in the question.

Then we leverage the methods proposed by Min et al. (2019) to process these two types respectively. Specifically, we adopt an off-the-shelf span predictor Pointer to map the question into several points, which could be for segmenting the question into various text spans.

Finally, we generated sub-questions by considering the type of questions and index points provided by Pointer. Concretely, for Bridge questions like Kristine Moore Gebbie is a professor at a university founded in what year?, Pointer could divided the question into two parts: Kristin Moore Gebbie be a professor at a university and founded in what year?. Then some question words are inserted into the first part as the first-hop evidence like Kristin Moore Gebbie be a professor at which university, denoted as S^A. Afterward, an off-the-shelf single QA model is used to find the answer for the first sub-question, and the answer would be used to form the second sub-question like Flinders University founded in what year?, denoted as S^B. On the other hand, for Comparison questions like Do The Importance of Being Icelandic and The Five Obstructions belong to different film genres?, Pointer would divide it into three parts: first entity (The Importance of Being Icelandic), second entity (The Five Obstructions), and target property (film genre). Then two sub-questions could be further generated by inserting question words to these parts like S^A :Do The Importance of Being Icelandic belong to which film genres? and S^B : Do The Five Obstructions belong to which film genres?

Pre-trained Language Model (LM) as Generator After automatically creating the sub-question dataset, the next step is to train the QG module from scratch. Specifically, the structure of whole QG module is designed as seq2seq, where it shares the encoder with QA module and adopts GPT-2 (Radford et al., 2019) as the decoder. During training stage, the input of decoder is formed as: $[bos; y_1^A; y_2^A; ...; y_n^A; [SEP]; y_1^B; y_2^B; ...; y_n^B; eos]$, where [SEP] is the separator token, bos is the start token and eos is the end token. y_i^A and y_i^B are the i-th token in constructed sub-questions S^A and S^B respectively.

Then the training objective of the QG module is to maximize the conditional probability of the target sub-questions sequence as follows:

$$L_{qg} = \sum_{i=1}^{n} \log P(y_t | y_{<t}; h); \qquad (4)$$

where h is encoder hidden state. Finally, QG module and QA module are trained together in end-to-end multi-task manner, and the overall loss is defined as:

$$L_{multitask} = L_{qa} + L_{qg}. \qquad (5)$$

4 Experiments

4.1 Dataset

We evaluate our approach on HotpotQA (Yang et al., 2018) under the distraction setting, a popular multi-hop QA dataset taking the explanation ability of models into accounts. Expressly, for each question, two gold paragraphs with ground-truth answers and supporting facts are provided, along with 8 `distractor' paragraphs that were collected via bi-gram TF-IDF retriever (i.e., 10 paragraphs in total). Furthermore, HotpotQA contains two types of subtasks: a) Answer prediction; and b) Supporting facts prediction; both subtasks adopt the same evaluation metrics: Exact Match (EM) and Partial Match (F1).

4.2 Implementation Details

We implement the model via HuggingFace library (Wolf et al., 2020). In detail, DFGN is selected as a QA module by following the details provided by (Xiao et al., 2019). While, for the QG module, the pre-trained decoder language model is initialized with GPT2 (Radford et al., 2019). The number of shared encoder layers is set as 12, the number of decoder layers is 6, the maximum sequence length is 512. We train the model on four TITAN RTX GPUs for 30 epochs at a batch size of 8, where each epoch tasks for

Model	Answer		Sup Fact		Joint	
	EM	F1	EM	F1	EM	F1
Baseline Model	44.44	58.28	21.95	66.66	11.56	40.86
DecompRC	55.20	69.63	-	-	-	-
DFGN* (Bridge)	53.38	69.14	47.72	84.44	29.79	58.67
DFGN* (Comparison)	63.75	69.48	70.68	89.98	46.74	63.56
DFGN* (Total)	55.46	69.21	52.33	82.12	33.19	59.66
DFGN (Total)	55.66	69.34	53.10	82.24	33.68	59.86
Ours (Bridge)	56.24	71.67	51.06	81.16	33.61	61.75
Ours (Comparison)	63.08	69.59	73.03	90.36	49.23	64.45
Ours (Total)	57.79	71.36	55.77	83.33	36.99	62.52

Table 1: Performance comparison on the development set of HotpotQA in the distractor setting. * indicates the results implemented by us.

around 2 hours. We select Adam (Kingma and Ba, 2017) as our optimizer with a learning rate of 5e-5 and a warm-up ratio of 10%. In general, we determine the hyperparameters by comparing the final EM and F1 scores.

4.3 Comparison Models

Baseline Model A neural paragraph-level QA model introduced in Yang et al. (2018) and original proposed by Clark and Gardner (2018).

DFGN The classic GN-based model (Xiao et al., 2019) , which is trained in an end-to-end fashion for multi-hop QA task. We select this as the primary QA module in our approach, and reproduce the DFGN model by using the BERT-base pre-trained model under the hyperparameter settings released by Yang et al. (2018).

DecompRC The classic QD-based model that decomposes each question into several sub-questions (Min et al., 2019). We reproduce the DecompRC model by following the same QD instruction illustrated in Min et al. (2019).

5 Analysis

Table 1 shows the performance of various models on the development set of HotpotQA. In general, our method attains substantial improvement across all tasks when compared to either the GN-based method or the QD-based approach. This demonstrates that the integration of the QG task can effectively augment the model's textual understanding capabilities. Additionally, our method exhibits consistent enhancement in performance for both types of questions. Notably, the performance on bridge-type questions, which necessitate linear reasoning chains, experiences a marked improvement, underscoring the efficacy of posing questions at each reasoning stage. In subsequent sections, we will further explore the functionality, interpretability, and quality of the sub-questions generated by the QG module, providing a comprehensive analysis of our proposed method's strengths and potential applications.

5.1 Does it alleviate shortcut problem by adding question generation module?

In order to validate the capacity of the QG module to concentrate on uncovering the authentic reasoning process, as opposed to exploiting shortcuts for predicting answers, we further undertake QA tasks using baselines and our model on Adversarial MultiHopQA. This dataset was initially introduced by Jiang and Bansal (2019a) and is designed to challenge the model's comprehension capabilities. Specifically, multiple noisy facts, constructed by substituting entities within the reasoning chain, are incorporated into the original HotpotQA dataset with the intent to confound the model. For instance, in the example provided in Figure 3, the noisy facts are formulated by replacing key entities present in Support Fact2.

Question: 2014 S/S is the debut album of a South Korean boy group that was formed by who?
Support Fact1: 2014 S/S is the debut album of South Korean group WINNER.
Support Fact2: Winner, often stylized as WINNER, is a South Korean boy group formed in 2013 by YG Entertainment and debuted in 2014.
reasoning chain: 2014 S/S WINNER YG Entertainment

Noisy Fact1: Juarez, often stylized as Juarez, is a South Korean boy group formed in 2013 by YG Arthur and debuted in 2014.
Noisy Fact2: Epic, often stylized as Epic, is a South Korean boy group formed in 2013 by YG Republic and debuted in 2014.
Noisy Fact3: ...
No reasoning chain with Support Fact1!

Right Answer: YG Entertainment (from ours)
Disturbances: YG Arthur; YG Republic (from baselines)

Figure 3: An example of the noisy dataset. The red text indicates a reasoning path with complete reasoning logic. The blue text indicates some other entities which have a similar structure with the red texts, but they can be inferred from the logical relationships.

These noisy facts retain the same sentence structure as the support facts but convey disparate meanings, thereby compelling the model to thoroughly comprehend the context. This additional layer of complexity serves to rigorously test our proposed QG module, ensuring it remains focused on elucidating the genuine reasoning process.

Table 2 shows the performance between the DFGN model and our model on the Adversarial-MultiHopQA dataset. In general, DFGN experiences a significant decline in performance, indicating that the existing QA model has poor robustness and is vulnerable to adversarial attacks. This further indicates that the model solves questions by mostly remembering patterns. On the other hand, by adding a QG module, the performance degradation of our method is significantly reduced. We think this is mainly because asking questions is an important strategy for guiding the model to understand the text.

Model	Answer	
	EM	F1
DFGN	55.66	69.34
DFGN*	48.08(-13.62%)	61.28(-11.62%)
Ours	57.79	71.36
Ours*	52.34(-9.43%)	65.12(-8.74%)

Table 2: Performance of DFGN model and ours on HotpotQA dataset and its noisy version Adversarial-MultiHopQA (marked with *).

We further prove this point through a case study shown in Figure 3. To answer the original question, the correct reasoning chain is 2014 S/S ! WINNER ! YG Entertainment. However, when there exists an overlap in the context between facts (i.e. South Korean boy group) the current main-stream method, which strengthens representation by solely capturing internal relationships over entities or documents, usually regards the incorrect entity (i.e. YG Arthur or YG Republic) as a key node of reasoning chain, where so-called shortcut issue. It does not understand the reasoning process but remembers certain context patterns. However, our method mitigates such issues by reinforcing representations by asking a question at each reasoning step. As such, it could remain robust despite these disturbances.

5.2 Does generated sub-question provide better interpretability?

Past works have proved that interpretability can be improved by exposing evidence from decomposed sub-questions. However, few quantitative analyses have been carried out on interpretability due to its subjective nature. In this paper, we use human evaluation to quantify the improvement of interpretability brought by our QG module.

Specifically, we design human evaluation by following steps: First, we assemble 16 well-educated vol-

Indicators	Methods	Win	Tie	Loss
Diversity	QG vs. QD	57.64%	26.70%	15.66%
LM Score	QG vs. QD	60.22%	-	39.78%
Attention weight	QG vs. w/o QG	79.51%	-	20.49%

Table 4: Comparison between sub-questions generated by QG and template on diversity, LM score and Attention weights.

unteers and divide them into two groups, A and B; Second, we randomly sample 8 Bridge type questions from the dev set and manually write out the correct two-hop reasoning chain for solving each question. Afterward, we replace the entity that appeared in each correct reasoning chain with other confusing entities selected from context to generate three more wrong reasoning chains (i.e., each question has 4 reasoning chains.). Then shuffle them and combine them with the original question to form a four-way multi-choice QA; Third, for each group, we ask them to figure out the correct reasoning chain and record the time elapsed for finishing all questions. To be noticed, besides original questions and reasoning chains, we provide different additional information for each group to facilitate them, all supporting facts for Group A, and all sub-questions generated by our QG for Group B. For more details, please refer to Appendix.

Table 3 presents the results of the two groups. Remarkably, Group B has higher accuracy and takes less time. Therefore, we could argue that sub-questions generated by our QG contain more concise and precise explanations for problem-solving and further proves that the QG module can indeed improve interpretability.

Group	Accuracy	Time(s)
A (Support Facts)	65.63%	981
B (Sub-questions)	85.94%	543

Table 3: Average results for accuracy and time elapsed of human evaluation.

5.3 Does asking questions enhance the natural language understanding capability?

In this work, we believe that the ability to exhaustively generate a set of logical questions according to a complex scenario allows for a comprehensive, interpretable, and flexible way of excavating the information hidden in natural language text, thereby enhancing the natural language understanding ability.

The self-attention mechanism in the pre-trained model is crucial for the model to understand the input information. Generally, the more critical a sentence is in its context, the greater attention weights it deserves. Thus, to verify whether the QG module could edify the model to carry out deep understanding intrinsically, we compare the sentence-level attention weight of our model with and without the QG module. In particular, we account for the number of increases in attention weight of support facts after adding the QG module. As shown in the last row of Table 4, the attention weight of around 80% of support facts is increased, which proves that the model is more prone to focus on meaningful information with the aid of the QG task.

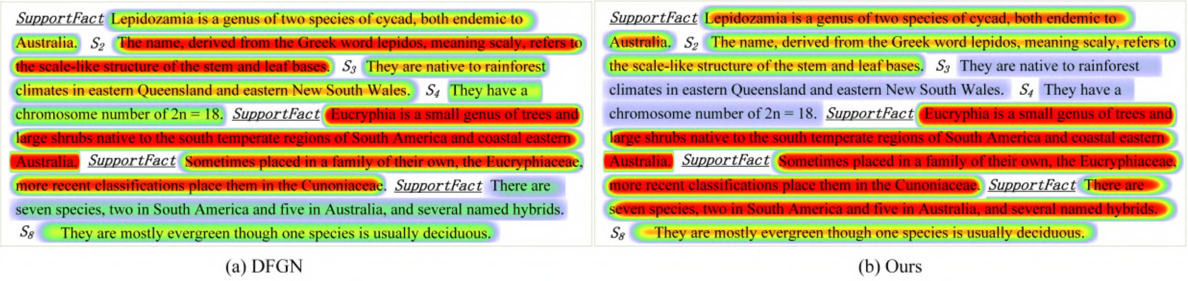
Furthermore, Figure 4 visualizes the changes in attention weights over supporting facts between DFGN and our method. In this case, sentences $S_{1,5,6,7}$ are considered as supporting facts. DFGN fails to predict all supporting facts and focuses on the wrong ones while our method works properly.

5.4 Characteristics of Generated Questions

QG can indeed promote an in-depth understanding of the model, but What are the characteristics of the generated questions that contribute to this? Specifically, what are the distinctive features of the sub-questions we generate using the QG module compared to the previous QD-based methods, which generate sub-questions using templates. Through case and statistical analysis, we find that the sub-questions generated by the QG module exhibit the following characteristics:

Consistency As mentioned in Section 3.2, prior QD-based methods necessitate the implementation of a span predictor to dissect questions into constituent text spans. During the segmentation process, errors are predisposed to accumulate, rendering the generated sub-questions susceptible to inconsisten-

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 569-582, Harbin, China, August 3 - 5, 2023.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China



SupportFact Lepidozamia is a genus of two species of cycad, both endemic to Australia. S_2 The name, derived from the Greek word lepidos, meaning scaly, refers to the scale-like structure of the stem and leaf bases. S_3 They are native to rainforest climates in eastern Queensland and eastern New South Wales. S_4 They have a chromosome number of 2n = 18. SupportFact Eucryphia is a small genus of trees and large shrubs native to the south temperate regions of South America and coastal eastern Australia. SupportFact Sometimes placed in a family of their own, the Eucryphiaceae, more recent classifications place them in the Cunoniaceae. SupportFact There are seven species, two in South America and five in Australia, and several named hybrids. S_8 They are mostly evergreen though one species is usually deciduous.

(a) DFGN

SupportFact Lepidozamia is a genus of two species of cycad, both endemic to Australia. S_2 The name, derived from the Greek word lepidos, meaning scaly, refers to the scale-like structure of the stem and leaf bases. S_3 They are native to rainforest climates in eastern Queensland and eastern New South Wales. S_4 They have a chromosome number of 2n = 18. SupportFact Eucryphia is a small genus of trees and large shrubs native to the south temperate regions of South America and coastal eastern Australia. SupportFact Sometimes placed in a family of their own, the Eucryphiaceae, more recent classifications place them in the Cunoniaceae. SupportFact There are seven species, two in South America and five in Australia, and several named hybrids. S_8 They are mostly evergreen though one species is usually deciduous.

(b) Ours

Figure 4: Visualization of attention weights at sentence-level between DFGN and our method. The depth of the color corresponds to the higher attention weights of the sentence.

ID			Question / Sub-question	Fluency	Diversity
1	Question		In 1991 Euromarche was bought by a chain that operated how any hypermarkets at the end of 2016?		
	QD	Q1	Which chain that operated how any hypermarkets?	×	×
		Q2	In 1991 Euromarche was bought by Euromarche at the end of 2016?		
	QG	Q1	In 1991 Euromarche was bought by which chain?	X	X
		Q2	Carrefour's oprated how many hypermarkets at the end of 2016?		
2	Question		Do The Importance of Being Icelandic and The Five Obstructions belong to different film genres?		
	QD	Q1	Do the Importance of Being Icelandic and The Five Obstructions belong to different film genres?	×	×
		Q2	Do the importance of?		
	QG	Q1	Does the Importance of Being Icelandic and The Five Obstructions belong to which film genres?	X	X
		Q2	Does The Five Obstructions belong to which film genres?		
			......		
7404	Question		Who was known by his stage name Aladin and helped organizations improve their performance as a consultant?		
	QD	Q1	Who was known by his stage name Aladin?	X	×
		Q2	Who helped organizations improve their performance as a consultant?		
	QG	Q1	His stage name Aladdin?	×	X
		Q2	Who was known by his stage name Aladdin and helped organizations improve their performance as a consultant?		
7405	Question		Which American film actor and dancer starred in the 1945 film Johnny Angel?		
	QD	Q1	Which 1945 file Johnny Angel?	×	-
		Q2	Which American film actor and dancer starred in noir?		
	QG	Q1	Which American file actor and dancer?	X	-
		Q2	Which starred in the 1945 film Johnny Angel?		

Table 5: Results on linguistic fluency and diversity of sub-questions generated by QG compared to those generated by template. X indicate the method performs better, × indicate performs worse, and - indicate performs competitively.

cies with the original question. This issue becomes increasingly prevalent when the original question exhibits linguistic complexity. As illustrated by the second example in Table 5, the pair of sub-questions generated by template-based approaches erroneously deconstruct the original question, culminating in a question intention that deviates significantly from the original intent. Consequently, such sub-questions characterized by incongruent intent can mislead the model. In contrast, our proposed QG module is designed to facilitate a comprehensive understanding of the original question, utilizing abundant contextual information to generate logically ordered sub-questions. Ultimately, this approach ensures that the intentions of the combined sub-questions remain consistent with the original question, mitigating the risk of misinterpretation by the model.

Fluency The fluidity and grammatical integrity of a sentence play a crucial role in accurately conveying meaning, particularly in the case of questions. When a question is plagued by grammatical inaccuracies or incoherence, it becomes challenging for individuals or computational models to comprehend, potentially leading to misinterpretation of the intended inquiry. This issue is widespread and inescapable in numerous datasets, primarily due to the manual construction of questions, as exemplified by the first instance in Table 5. In the original question, a typographical error (how many ! how any) causes a shift in the intended meaning. Nonetheless, it remains feasible to discern the correct response from the additional information offered by the original question and general knowledge. Regrettably, the sub-question produced by the QD-based technique incorporates the typographical error, and the model fails to ascertain the accurate intention due to the limited information available within the sub-question. Moreover, syntactic errors are prone to accrue since determining the boundaries and attributes of text spans proves to be a challenging task, leading to subpar readability.

Contrastingly, our QG module is capable of leveraging contextual information and the embedded knowledge within the language model to rectify typographical errors. Simultaneously, it can employ the capabilities of the pre-trained language model to generate coherent sentences, thus alleviating the impact of syntactic errors. To assess fluency, we utilize the Language Model Score[3] (LMS) as a metric. As demonstrated in Table 4, over 60% of the questions generated by QG modules exhibit higher scores compared to those produced by the QD method.

Diversity Sultan et al. (2020) highlight that the diversity of generated questions can directly impact QA performance. However, sub-questions produced by QD methods tend to be monotonous and laborious due to constraints on vocabulary and templates. In contrast, our proposed QG module can gently mitigate these challenges and enhance question diversity. Relying on the pretrained LM, the QG module is capable of incorporating contextually appropriate words into sub-questions, adapting to various situations. This is exemplified by the inclusion of Carrefour in the first example provided in Table 5, which results in more diverse and rational sub-questions. In our analysis, we consider the number of words in sub-questions that did not appear in the original question as a measure of diversity. As demonstrated in Table 4, approximately 57% of sub-questions generated by our method exhibit greater diversity, underlining the advantages of our proposed QG module.

6 Conclusion

In this paper, drawing inspiration from human cognitive behavior, we posit that the act of asking questions serves as a crucial indicator for determining whether a model genuinely comprehends the input text. Consequently, we introduce a QG module designed to tackle multi-hop QA tasks in an interpretable manner. Building upon traditional QA modules, the incorporation of the QG module effectively enhances natural language understanding capabilities, delivering superior and robust performance through the process of asking questions. Furthermore, we conduct a quantitative analysis of interpretability, as provided by sub-questions, utilizing human evaluation and elucidating interpretability through attention visualization. Ultimately, we substantiate that the sub-questions derived from the QG method surpass those obtained via the QD method in terms of linguistic fluency, consistency, and diversity, underscoring the benefits of our proposed approach.

7 Limitations

Although our research presents numerous advantages, certain limitations persist. The lack of comparison with extant SOTA methods and validation on alternative datasets constitute two principal shortcomings. Despite these issues, we maintain our advocacy for the "ask to understand" concept, positing that the integration of a QG task can bolster a model's interpretability and comprehension capabilities.

Primarily, the rationale behind our decision not to utilize top SOTA models as baselines is that these approaches often entail the application of meticulously designed, task-specific, and labor-intensive GNN to the encoder segment. Conversely, we posit that our method operates in a plug-and-play manner; validating its efficacy on two rudimentary baselines suggests that it may also be applicable to other models. Consequently, outperforming SOTA methods in terms of performance is not the central contribution of this paper.Additionally, question decomposition serves as a vital component of our work, and we employ DecompRC to parse multi-hop questions into single-hop queries. Since DecompRC is tailored specifically for HotpotQA, adapting it to other multi-hop QA datasets may not yield the anticipated results; thus, we solely verify our methods on HotpotQA.

Finally, grounded in our core concept of "asking to understand," the applicability and reliability of the QA model in industrial contexts are significantly enhanced. Our model delivers answers accompanied by comprehensive multi-hop questions, enabling agents to evaluate the accuracy of the response. Furthermore, our model aids agents in "understanding by asking," delineating the steps involved in obtaining the answer and facilitating a more profound comprehension of the information's origin.

[3] https://github.com/simonepri/lm-scorer

References

Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. 2019. Synthetic qa corpora generation with roundtrip consistency. *arXiv preprint arXiv:1906.05416*.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Christopher Clark and Matt Gardner. 2018. Simple and effective multi-paragraph reading comprehension. In *Proceedings of the 56th ACL (Volume 1: Long Papers)*, pages 845–855.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2019. Question answering by reasoning across documents with graph convolutional networks. In *ACL*, pages 2306–2317.

Bhuwan Dhingra, Qiao Jin, Zhilin Yang, William Cohen, and Ruslan Salakhutdinov. 2018. Neural models for reasoning over multiple mentions using coreference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 42–48.

Kaustubh D. Dhole and Christopher D. Manning. 2021. Syn-qg: Syntactic and shallow semantic rules for question generation.

Ming Ding, Chang Zhou, Qibin Chen, Hongxia Yang, and Jie Tang. 2019. Cognitive graph for multi-hop reading comprehension at scale. In *ACL*, pages 2694–2703.

Nadir Durrani, Hassan Sajjad, Fahim Dalvi, and Yonatan Belinkov. 2020. Analyzing individual neurons in pre-trained language models. *arXiv preprint arXiv:2010.02695*.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175.

Alexander R. Fabbri, Patrick Ng, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2020. Template-based question generation from retrieved sentences for improved unsupervised question answering.

Yuwei Fang, Siqi Sun, Zhe Gan, Rohit Pillai, Shuohang Wang, and Jingjing Liu. 2020. Hierarchical graph network for multi-hop question answering.

Ruiliu Fu, Han Wang, Xuejun Zhang, Jun Zhou, and Yonghong Yan. 2021. Decomposing complex questions makes multi-hop qa easier and more interpretable. In *EMNLP*, pages 169–180.

Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, et al. 2020. Evaluating models' local decision boundaries via contrast sets. *arXiv preprint arXiv:2004.02709*.

Yaru Hao, Li Dong, Furu Wei, and Ke Xu. 2020. Self-attention attribution: Interpreting information interactions inside transformer. *arXiv preprint arXiv:2004.11207*.

Joseph D Janizek, Pascal Sturmfels, and Su-In Lee. 2021. Explaining explanations: Axiomatic feature interactions for deep networks. *Journal of Machine Learning Research*, 22(104):1–54.

Yichen Jiang and Mohit Bansal. 2019a. Avoiding reasoning shortcuts: Adversarial evaluation, training, and model development for multi-hop qa. In *ACL*.

Yichen Jiang and Mohit Bansal. 2019b. Self-assembling modular networks for interpretable multi-hop reasoning. In *EMNLP-IJCNLP*, pages 4474–4484.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *ACL*, pages 785–794.

Dong Bok Lee, Seanie Lee, Woo Tae Jeong, Donghwan Kim, and Sung Ju Hwang. 2020. Generating diverse and consistent qa pairs from contexts with information-maximizing hierarchical conditional vaes. *arXiv preprint arXiv:2005.13837*.

Sewon Min, Victor Zhong, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2019. Multi-hop reading comprehension through question decomposition and rescoring. In *ACL*.

Kosuke Nishida, Kyosuke Nishida, Masaaki Nagata, Atsushi Otsuka, Itsumi Saito, Hisako Asano, and Junji Tomita. 2019. Multi-task learning for multi-hop qa with evidence extraction.

Valentina Pyatkin, Paul Roit, Julian Michael, Yoav Goldberg, Reut Tsarfaty, and Ido Dagan. 2021. Asking it all: Generating contextualized questions for any semantic role. In *Proceedings of the 2021 Conference on EMNLP*, pages 1429–1441.

XiPeng Qiu, TianXiang Sun, YiGe Xu, YunFan Shao, Ning Dai, and XuanJing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63:1872–1897, Sep.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *EMNLP*, pages 2383–2392.

Abhilasha Ravichander, Siddharth Dalmia, Maria Ryskina, Florian Metze, Eduard Hovy, and Alan W Black. 2021. Noiseqa: Challenge set evaluation for user-centric question answering. *arXiv preprint arXiv:2102.08345*.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2018. Bidirectional attention flow for machine comprehension.

Md Arafat Sultan, Shubham Chandel, Ramón Fernandez Astudillo, and Vittorio Castelli. 2020. On the importance of diversity in question generation for QA. In *ACL*, July.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, August.

Ming Tu, Guangtao Wang, Jing Huang, Yun Tang, Xiaodong He, and Bowen Zhou. 2019. Multi-hop reading comprehension across multiple documents by reasoning over heterogeneous graphs. In *ACL*, pages 2704–2713.

Ming Tu, Kevin Huang, Guangtao Wang, Jing Huang, Xiaodong He, and Bowen Zhou. 2020. Interpretable multi-hop reading comprehension over multiple documents.

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

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. *TACL*, 6:287–302.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Zhaofeng Wu, Hao Peng, and Noah A Smith. 2021. Infusing finetuning with semantic dependencies. *Transactions of the Association for Computational Linguistics*, 9:226–242.

Yunxuan Xiao, Yanru Qu, Lin Qiu, Hao Zhou, Lei Li, Weinan Zhang, and Yong Yu. 2019. Dynamically fused graph network for multi-hop reasoning.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *EMNLP*, pages 2369–2380.

An Yang, Quan Wang, Jing Liu, Kai Liu, Yajuan Lyu, Hua Wu, Qiaoqiao She, and Sujian Li. 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*, pages 2346–2357, Florence, Italy, July. Association for Computational Linguistics.

Shiyue Zhang and Mohit Bansal. 2019. Addressing semantic drift in question generation for semi-supervised question answering. *arXiv preprint arXiv:1909.06356*.

Victor Zhong, Caiming Xiong, Nitish Shirish Keskar, and Richard Socher. 2019. Coarse-grain fine-grain coattention network for multi-evidence question answering. *arXiv preprint arXiv:1901.00603*.

A Appedix: Human Evaluation Instruction

Specifically, we design human evaluation by following steps:

1. We assemble 16 well-educated volunteers and randomly divide them into two groups, A and B. Each group contains 8 volunteers and evenly gender.

2. We randomly sample 8 Bridge type[4] questions from the dev set, and manually write out the correct two-hop reasoning chain for solving each question.

3. We replace the entity that appeared in each correct reasoning chain with other confusing entities selected from context to generate three more wrong reasoning chains (i.e., each question has 4 reasoning chains.), then shuffle them and combine them with the original question to form a four-way multi-choice QA.

4. For group A, except the original question, final answer and four reasoning chains, we also provide supporting facts. Then volunteers are asked to find the correct reasoning chain.

5. For group B, except the original question, final answer and four reasoning chains, we also provide the sub-questions generated by our QG module. Then volunteers are asked to find the correct reasoning chain.

6. We count the accuracy and time elapsed for solving problem.

Beyond that, some details are worth noting:

- The volunteers participated in the human evaluation test are all well-educated graduate students with skilled English.

- We use the online questionnaire platform to design the electronic questionnaire.

- The questionnaire system can automatically score according to the pre-set reference answers, and count the time spent on answering the questions.

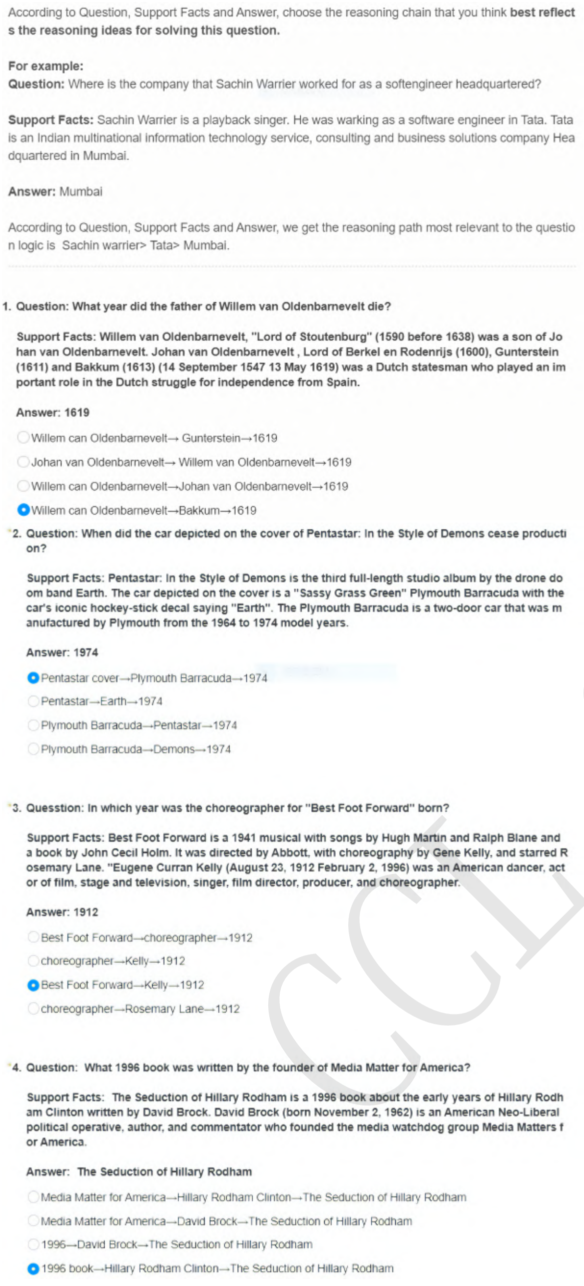
- The timer starts when the volunteer clicks “accept" button on the questionnaire, and ends when the volunteer clicks “submit" button.

- Volunteers are required to answer the questionnaire without any interruption, ensuring that all time spent is for answering questions.

- Before starting filling the questionnaire, we provide a sample example as instruction to teach the volunteers how to find the answer.

The interface of human evaluation for each group could be found in Figure 5 and Figure 6.

[4] Because Bride type questions always has deterministic linear reasoning chains.



According to Question, Support Facts and Answer, choose the reasoning chain that you think **best reflects the reasoning ideas for solving this question.**

For example:
Question: Where is the company that Sachin Warrier worked for as a softengineer headquartered?

Support Facts: Sachin Warrier is a playback singer. He was warking as a software engineer in Tata. Tata is an Indian multinational information technology service, consulting and business solutions company Headquartered in Mumbai.

Answer: Mumbai

According to Question, Support Facts and Answer, we get the reasoning path most relevant to the question logic is Sachin warrier> Tata> Mumbai.

1. **Question: What year did the father of Willem van Oldenbarnevelt die?**

   Support Facts: Willem van Oldenbarnevelt, "Lord of Stoutenburg" (1590 before 1638) was a son of Johan van Oldenbarnevelt. Johan van Oldenbarnevelt , Lord of Berkel en Rodenrijs (1600), Gunterstein (1611) and Bakkum (1613) (14 September 1547 13 May 1619) was a Dutch statesman who played an important role in the Dutch struggle for independence from Spain.

   Answer: 1619

   ○ Willem can Oldenbarnevelt→ Gunterstein→1619
   ○ Johan van Oldenbarnevelt→ Willem van Oldenbarnevelt→1619
   ○ Willem can Oldenbarnevelt→Johan van Oldenbarnevelt→1619
   ● Willem can Oldenbarnevelt→Bakkum→1619

2. **Question: When did the car depicted on the cover of Pentastar: In the Style of Demons cease production?**

   Support Facts: Pentastar: In the Style of Demons is the third full-length studio album by the drone doom band Earth. The car depicted on the cover is a "Sassy Grass Green" Plymouth Barracuda with the car's iconic hockey-stick decal saying "Earth". The Plymouth Barracuda is a two-door car that was manufactured by Plymouth from the 1964 to 1974 model years.

   Answer: 1974

   ● Pentastar cover→Plymouth Barracuda→1974
   ○ Pentastar→Earth→1974
   ○ Plymouth Barracuda→Pentastar→1974
   ○ Plymouth Barracuda→Demons→1974

3. **Quesstion: In which year was the choreographer for "Best Foot Forward" born?**

   Support Facts: Best Foot Forward is a 1941 musical with songs by Hugh Martin and Ralph Blane and a book by John Cecil Holm. It was directed by Abbott, with choreography by Gene Kelly, and starred Rosemary Lane. "Eugene Curran Kelly (August 23, 1912 February 2, 1996) was an American dancer, actor of film, stage and television, singer, film director, producer, and choreographer.

   Answer: 1912

   ○ Best Foot Forward→choreographer→1912
   ○ choreographer→Kelly→1912
   ● Best Foot Forward→Kelly→1912
   ○ choreographer→Rosemary Lane→1912

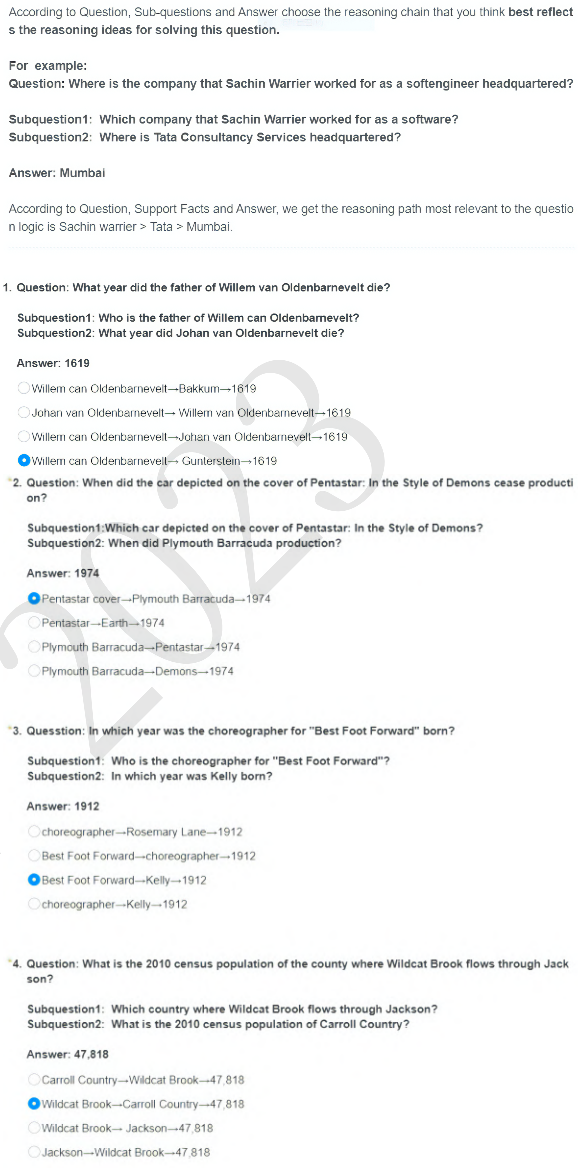
4. **Question: What 1996 book was written by the founder of Media Matter for America?**

   Support Facts: The Seduction of Hillary Rodham is a 1996 book about the early years of Hillary Rodham Clinton written by David Brock. David Brock (born November 2, 1962) is an American Neo-Liberal political operative, author, and commentator who founded the media watchdog group Media Matters for America.

   Answer: The Seduction of Hillary Rodham

   ○ Media Matter for America→Hillary Rodham Clinton→The Seduction of Hillary Rodham
   ○ Media Matter for America→David Brock→The Seduction of Hillary Rodham
   ○ 1996→David Brock→The Seduction of Hillary Rodham
   ● 1996 book→Hillary Rodham Clinton→The Seduction of Hillary Rodham

Figure 5: Interface for human evaluation of choosing reasoning chain based on support facts.



According to Question, Sub-questions and Answer choose the reasoning chain that you think **best reflects the reasoning ideas for solving this question.**

For example:
Question: Where is the company that Sachin Warrier worked for as a softengineer headquartered?

Subquestion1: Which company that Sachin Warrier worked for as a software?
Subquestion2: Where is Tata Consultancy Services headquartered?

Answer: Mumbai

According to Question, Support Facts and Answer, we get the reasoning path most relevant to the question logic is Sachin warrier > Tata > Mumbai.

1. **Question: What year did the father of Willem van Oldenbarnevelt die?**

   Subquestion1: Who is the father of Willem can Oldenbarnevelt?
   Subquestion2: What year did Johan van Oldenbarnevelt die?

   Answer: 1619

   ○ Willem can Oldenbarnevelt→Bakkum→1619
   ○ Johan van Oldenbarnevelt→ Willem van Oldenbarnevelt→1619
   ○ Willem can Oldenbarnevelt→Johan van Oldenbarnevelt→1619
   ● Willem can Oldenbarnevelt→ Gunterstein→1619

2. **Question: When did the car depicted on the cover of Pentastar: In the Style of Demons cease production?**

   Subquestion1:Which car depicted on the cover of Pentastar: In the Style of Demons?
   Subquestion2: When did Plymouth Barracuda production?

   Answer: 1974

   ● Pentastar cover→Plymouth Barracuda→1974
   ○ Pentastar→Earth→1974
   ○ Plymouth Barracuda→Pentastar→1974
   ○ Plymouth Barracuda→Demons→1974

3. **Quesstion: In which year was the choreographer for "Best Foot Forward" born?**

   Subquestion1: Who is the choreographer for "Best Foot Forward"?
   Subquestion2: In which year was Kelly born?

   Answer: 1912

   ○ choreographer→Rosemary Lane→1912
   ○ Best Foot Forward→choreographer→1912
   ● Best Foot Forward→Kelly→1912
   ○ choreographer→Kelly→1912

4. **Question: What is the 2010 census population of the county where Wildcat Brook flows through Jackson?**

   Subquestion1: Which country where Wildcat Brook flows through Jackson?
   Subquestion2: What is the 2010 census population of Carroll Country?

   Answer: 47,818

   ○ Carroll Country→Wildcat Brook→47,818
   ● Wildcat Brook→Carroll Country→47,818
   ○ Wildcat Brook→ Jackson→47,818
   ○ Jackson→Wildcat Brook→47,818

Figure 6: Interface for human evaluation of choosing reasoning chain based on sub-questions.s