Knowledge-based Question Answering with Large Language Models

Yunshi Lan

2023 Nov

East China Normal University
Overview

Knowledge-based Question Answering with LLMs

- Multi-modal Questions
- Solving Reasoning Problems under Noisy Context
- Generating Complex Questions

Future Directions

- The Applications of LLMs on more NLP tasks
- Instruction-tuning of LLMs for KGQA
Overview

Knowledge-based Question Answering with LLMs

• **Multi-modal Questions**
  • Solving Reasoning Problems under Noisy Context
  • Generating Complex Questions

Future Directions

• The Applications of LLMs on more NLP tasks
• Instruction-tuning of LLMs for KGQA
What is the appliance the woman is holding used for?

Step 1: external knowledge retrieval

ConceptNet
An open, multilingual knowledge graph

Step 2: knowledge-image-question reasoning

Knowledge
Question
Image

A New Paradigm

This is a blow dryer in a bathroom

What is the appliance the woman is holding used for?

C: A red bird; Q: What is the breed? A: Parrot

Exemplar
Selection/Generation
Strategy

PICa[1]/Img2Prompt[2]

Exemplar prompts

LLMs

A


Our Motivation

1. The current methods entirely rely on the understanding capability of LLMs to resolve the ambiguity and infer the intent of the questions, which might involve unexpected bias.
2. LLMs are brittle to ill-posed questions, especially under the zero-shot setting.
Our Motivation

Reasoning Question Prompts: converting original questions into self-contained questions by editing the segments of the question.

Yunshi Lan, Alex Xiang Li, Xin Liu, Yang Li, Wei Qin, Weining Qian. Improving Zero-shot Visual Question Answering via Large Language Models with Reasoning Question Prompts. ACM MM. 2023
Our prompting method generates reasoning question prompts and enables LLMs to perform VQA tasks with two-step reasoning.
Prompting for Answer Generation

Question Reasoning Prompt: What unpleasant emotional does clouds and lightning often cause?
Unsupervised Question Edition: Our scoring function for evaluating the quality of the candidates:

- **LM Score.** \( f_{LM}(\tilde{Q}) = \ln \prod_{i=1}^{T} P(w_i|w_{i-1}, \ldots, w_1) \)
- **Semantic Integrity.** \( f_{semantic}(\tilde{Q}) = \cos(\tilde{Q}, Q) \)
- **Syntactic Invariance.** \( f_{syntactic}(\tilde{Q}) = \mathbb{I}(\text{Tag}_Q[i] = \text{Tag}_{\tilde{Q}}[j]) \)

The overall scoring function:

\[
 f(\tilde{Q}) = f_{LM}(\tilde{Q})^\alpha f_{semantic}(\tilde{Q})^\beta f_{syntactic}(\tilde{Q})
\]

Prompt Design:

- **Instruction:** Please answer the question according to the contexts.
- **Context:** [caption].
- **Question:** [reasoning question prompt].
- **Answer:**
Prompting for Answer Choosing

Answer Heuristics Construction:

\[ P(A) = \sum_{LLM(\tilde{Q}) \rightarrow A} P(\tilde{Q}) P_{LLM}(A|\tilde{Q}) \]

Prompt Design:

**Instruction:** Please answer the question according to the contexts and candidates.

**Context:** [caption].

**Question:** [original question].

**Candidates:** \([A_1 \ P(A_1)]; [A_2 \ P(A_2)]; \ldots; [A_m \ P(A_m)]\)

**Answer:**

Datasets:
• OK-VQA: 5,046 test questions.
• A-OKVQA: 1,100 and 6,700 questions for validation and testing, respectively.
• VQAv2: 214,354 validation questions.

Comparable Methods:
• LLM-based methods: PICa, Img2Prompt
• Pre-trained zero-shot VQA methods: Flamingo, Frozen VL-T5, FewVLM and VLKD.
### Experimental Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Model size</th>
<th>Shot number</th>
<th>Exemplar number</th>
<th>OK-VQA test</th>
<th>VQAval2 val</th>
<th>A-OKVQA val</th>
<th>test</th>
</tr>
</thead>
<tbody>
<tr>
<td>PICa {GPT-3}</td>
<td>175B</td>
<td>0</td>
<td>0</td>
<td>17.7</td>
<td>—</td>
<td>23.8°</td>
<td>—</td>
</tr>
<tr>
<td>Img2Prompt {OPT}</td>
<td>6.7B</td>
<td>0</td>
<td>30</td>
<td>38.2</td>
<td>52.2°</td>
<td>33.3</td>
<td>32.2</td>
</tr>
<tr>
<td>Img2Prompt {OPT}</td>
<td>30B</td>
<td>0</td>
<td>30</td>
<td>41.8</td>
<td>54.2°</td>
<td>36.9</td>
<td>33.0</td>
</tr>
<tr>
<td>Img2Prompt {GPT-3}</td>
<td>175B</td>
<td>0</td>
<td>30</td>
<td>42.8</td>
<td>—</td>
<td>38.9°</td>
<td>43.4°</td>
</tr>
<tr>
<td>Img2Prompt {OPT}</td>
<td>175B</td>
<td>0</td>
<td>30</td>
<td>45.6</td>
<td>60.6</td>
<td>42.9</td>
<td>40.7</td>
</tr>
<tr>
<td>PICa+RQ prompt {GPT-3} {Ours}</td>
<td>175B</td>
<td>0</td>
<td>0</td>
<td>20.3(↑ 2.6)</td>
<td>—</td>
<td>29.0(↑ 5.2)</td>
<td>—</td>
</tr>
<tr>
<td>Img2Prompt+RQ prompt {OPT} {Ours}</td>
<td>6.7B</td>
<td>0</td>
<td>30</td>
<td>38.5(↑ 0.3)</td>
<td>52.9(↑ 0.7)</td>
<td>36.3(↑ 3.0)</td>
<td>31.5</td>
</tr>
<tr>
<td>Img2Prompt+RQ prompt {OPT} {Ours}</td>
<td>30B</td>
<td>0</td>
<td>30</td>
<td>42.1(↑ 0.3)</td>
<td>54.5(↑ 0.3)</td>
<td>38.1(↑ 1.2)</td>
<td>35.2(↑ 3.0)</td>
</tr>
<tr>
<td>Img2Prompt+RQ prompt {GPT-3} {Ours}</td>
<td>175B</td>
<td>0</td>
<td>30</td>
<td>46.4(↑ 3.6)</td>
<td>—</td>
<td>43.2(↑ 4.3)</td>
<td>43.9(↑ 0.5)</td>
</tr>
</tbody>
</table>

### Zero-shot Evaluation with Frozen LLMs

<table>
<thead>
<tr>
<th>Method</th>
<th>Model size</th>
<th>Shot number</th>
<th>Exemplar number</th>
<th>OK-VQA test</th>
<th>VQAval2 val</th>
<th>A-OKVQA val</th>
<th>test</th>
</tr>
</thead>
<tbody>
<tr>
<td>VL-T5 {no-vqa}</td>
<td>224M</td>
<td>0</td>
<td>0</td>
<td>5.8</td>
<td>13.5</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>FewVLM {large}</td>
<td>740M</td>
<td>0</td>
<td>0</td>
<td>16.5</td>
<td>47.7</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>VLKD {ViT-L/14}</td>
<td>408M</td>
<td>0</td>
<td>0</td>
<td>13.3</td>
<td>44.5</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Frozen</td>
<td>7B</td>
<td>0</td>
<td>0</td>
<td>5.9</td>
<td>29.5</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Flamingo</td>
<td>80B</td>
<td>0</td>
<td>0</td>
<td>50.6</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

### Few-shot Evaluation with Frozen LLMs

<table>
<thead>
<tr>
<th>Method</th>
<th>Model size</th>
<th>Shot number</th>
<th>Exemplar number</th>
<th>OK-VQA test</th>
<th>VQAval2 val</th>
<th>A-OKVQA val</th>
<th>test</th>
</tr>
</thead>
<tbody>
<tr>
<td>PICa {GPT-3}</td>
<td>175B</td>
<td>16</td>
<td>16</td>
<td>46.5</td>
<td>54.3</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Prophet {GPT-3}</td>
<td>175B</td>
<td>20</td>
<td>20</td>
<td>61.1</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

RQ prompts can generally improve VQA tasks under zero-shot setting.
RQ Prompts with Different LLMs

Table 3: Zero-shot performance A-OKVQA validation set having Img2Prompt as baselines but with different LLMs. Δ denotes the performance gain brought by QR prompts.

<table>
<thead>
<tr>
<th>LLMs</th>
<th>Img2Prompt</th>
<th>+QR prompt</th>
<th>Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPT-3 175B</td>
<td>38.9</td>
<td>43.2</td>
<td>↑ 4.3</td>
</tr>
<tr>
<td>GPT-3.5 175B</td>
<td>37.1</td>
<td>40.3</td>
<td>↑ 3.2</td>
</tr>
<tr>
<td>GPT-Neo 2.7B</td>
<td>29.7</td>
<td>31.5</td>
<td>↑ 1.8</td>
</tr>
<tr>
<td>BLOOM 7.1B</td>
<td>29.8</td>
<td>32.1</td>
<td>↑ 2.3</td>
</tr>
<tr>
<td>GPT-J 6B</td>
<td>32.5</td>
<td>33.1</td>
<td>↑ 0.6</td>
</tr>
<tr>
<td>OPT 125M</td>
<td>10.8</td>
<td>13.3</td>
<td>↑ 2.5</td>
</tr>
</tbody>
</table>

RQ prompts can generally collaborate with LLMs
More Analysis of RQ Prompts

Larger improvement gain brought by RQ prompts can be shown when: (1) shot number is low; (2) the size of LLM is relatively large
Case Study

Caption: This is a blow dryer in a bathroom.

Question: What is the appliance the woman is holding used for?
GT Answer: drying hair
Original Answer: cutting hair

Prompting for Answer Generation:
Q1: What is the appliance a blow dryer used for?  
A1: drying hair  \( P(A|Q) = 0.15 \)
A2: drying hair  \( P_{LLM}(A|Q) = 0.10 \)

Prompting for Answer Choosing:
Question: What is the appliance the woman is holding used for?
Candidates: drying hair (1.00)  
Predicted Answer: drying hair
(a)

Caption: A little girl holding a cup with rice in dishes in front of her

Question: What is the child eating?
GT Answer: rice
Original Answer: spaghetti

Prompting for Answer Generation:
Q1: what is dishes in front of her?  \( P(Q) = 0.15 \)
A1: rice  \( P_{LLM}(A|Q) = 0.20 \)
Q2: What is the child eating?  \( P(Q) = 0.7 \)
A2: spaghetti  \( P_{LLM}(A|Q) = 0.10 \)
Q3: what is a cup with food in dishes in front of her?  \( P(Q) = 0.15 \)
A3: rice  \( P_{LLM}(A|Q) = 0.30 \)

Prompting for Answer Choosing:
Question: What is the child eating?
Candidates: spaghetti (0.48); rice (0.51)
Predicted Answer: rice
(b)

The questions become self-contained with RQ prompts.
Conclusions

• RQ prompts are helpful to bridge the gap between questions and captions, such that it can boost performance of leveraging LLMs to VQA tasks.

• RQ prompts show general improvement on different LLMs. It could achieve SOTA results on three of four VQA tasks on zero-shot setting.

• RQ prompts show more effect on zero-shot setting and large LLMs.
Overview

Knowledge-based Question Answering with LLMs
  • Multi-modal Questions
  • **Solving Reasoning Problems under Noisy Context**
    • Generating Complex Questions

Future Directions
  • The Applications of LLMs on more NLP tasks
  • Instruction-tuning of LLMs for KGQA
Figure 1: Chain-of-thought prompting enables large language models to tackle complex arithmetic, commonsense, and symbolic reasoning tasks. Chain-of-thought reasoning processes are highlighted.
LLMs under Noisy Context

Original Problem
Jessica is six years older than Claire. In two years, Claire will be 20 years old. How old is Jessica now?

Modified Problem
Jessica is six years older than Claire. In two years, Claire will be 20 years old. Twenty years ago, the age of Claire’s father is 3 times of Jessica’s age. How old is Jessica now?

Standard Answer 24

Table 1. An example problem from GSM-IC. An irrelevant sentence (italic and underlined) that does not affect the standard answer is added immediately before the question.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Sch.rli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. ICML. 2023
Comparison of Existing CoTs

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schrli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. ICML. 2023


Our Method

Review Prompt

Q: Josh had 7 marbles in his collection. Josh's father works 8 hours a day. He lost 8 marbles and found 10 new ones. How many more marbles did he find than those he lost?

A: To answer this question, we need to notice:
1. He lost 8 marbles
2. He found 10 new ones

Key Sentences

Stage 2

Rephrase Prompt

Josh had 7 marbles in his collection. Josh's father works 8 hours a day. He lost 8 marbles and found 10 new ones. How many more marbles did he find than those he lost?

Q: Translate following sentences into equation:
1. He lost 8 marbles
2. He found 10 new ones

A: 1. Number of lost marbles = 8.
2. Number of found marbles = 10.

Variables

Stage 3

Resolve Prompt

Q: Josh had 7 marbles in his collection. Josh's father works 8 hours a day. He lost 8 marbles and found 10 new ones. How many more marbles did he find than those he lost?

A: With the Equation Hints: 1. Number of lost marbles = 8, 2. Number of found marbles = 10, we will answer the question.

The number of how many more marbles he found than he lost = Number of found marbles - Number of lost marbles = 10 - 8 = 2.
The answer is 2.

Figure 2: A running example of the inputs and outputs of R³ prompting in LLMs at each prompting stage. Green: In-topic noisy context. Red: Off-topic noisy context. Blue: Key sentences.
Experiments

Datasets:
- AddSub
- SVAMP
- GSM-IC
  - MultiArith-IC
  - SingEq-IC

Comparable Methods:
- One-turn interaction: Manual-CoT, Auto-CoT, Instructed-CoT
- Multi-turn interaction: Least-to-Most, PHP

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Sample</th>
<th>Ave. in-topic</th>
<th>Ave. off-topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSM-IC</td>
<td>1000</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>MultiArith-IC</td>
<td>600</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>SingEq-IC</td>
<td>508</td>
<td>0.48</td>
<td>0.52</td>
</tr>
</tbody>
</table>

Table 1: Details of constructed datasets. ‘Ave. in-topic’ and ‘Ave. off-topic’ denotes average number of in-topic sentences and off-topic sentences, respectively.
Experimental Results

<table>
<thead>
<tr>
<th>Methods</th>
<th>SVAMP</th>
<th>MultiArith-IC</th>
<th>SingleEq-IC</th>
<th>AddSub</th>
<th>GSM-IC</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-turn</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Manual-CoT</td>
<td>79.9</td>
<td>79.5</td>
<td>77.7</td>
<td>85.3</td>
<td>81.0</td>
<td>79.7</td>
</tr>
<tr>
<td>Auto-CoT</td>
<td>83.6</td>
<td>79.7</td>
<td>77.6</td>
<td>88.0</td>
<td>81.5</td>
<td>82.1</td>
</tr>
<tr>
<td>Instructed-CoT</td>
<td>81.3</td>
<td>80.1</td>
<td>78.2</td>
<td>87.3</td>
<td>82.0</td>
<td>81.8</td>
</tr>
<tr>
<td>Interactive</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Least-to-Most</td>
<td>80.8</td>
<td>77.4</td>
<td>76.2</td>
<td>85.3</td>
<td>81.1</td>
<td>80.2</td>
</tr>
<tr>
<td>PHP</td>
<td>83.1</td>
<td>80.0</td>
<td>79.0</td>
<td>85.3</td>
<td>85.1</td>
<td>82.5</td>
</tr>
<tr>
<td>R³ Prompting (Ours)</td>
<td>87.3</td>
<td>82.2</td>
<td>81.5</td>
<td>90.0</td>
<td>88.0</td>
<td>85.8</td>
</tr>
</tbody>
</table>

Table 2: Main result on five evaluated datasets. The best and second best results are boldfaced and underlined respectively.

(1) $R^3$ prompting performs well for CoT reasoning in LLMs under noisy context;
(2) The design of interactive prompts are important for denoising.
More Analysis

(1) Improvement of R^3 prompting is still significant with self-consistency;
(2) R^3 prompting exhibits robust performance under noisy context while Instructed-CoT and Manual-CoT are vulnerable when facing a large amount of noisy information.
Conclusions

- By comparison, one-turn CoTs are more robust than interactive CoTs when conducting reasoning under noisy context.

- R^3 prompting can effectively restrain the influence of noisy context. The three steps (i.e. review, rephrase and resolve) collaborative contribution to the good performance.
Knowledge-based Question Answering with LLMs
  • Multi-modal Questions
  • Solving Reasoning Problems under Noisy Context
  • Generating Complex Questions

Future Directions
  • The Applications of LLMs on more NLP tasks
  • Instruction-tuning of LLMs
Knowledge Base Question Generation

Figure 1: Overview of KQG-CoT framework.

Yuanyuan Liang, Jianing Wang, Hanlun Zhu, Lei Wang, Yunshi Lan, Weining Qian. Prompting Large Language Models with Chain-of-Thought for Few-Shot Knowledge Base Question Generation. EMNLP. 2023
Existing Methods

The existing methods solve few-shot KBQG tasks via designing prompter for the pair of sub-graph description and generated questions and conducting pre-training Language Models.

Motivations

- A **substantial amount of annotated data** is required, and acquiring it can be challenging.

- A logical form is made up of entities, relations, and query grammar. It's impossible to **encompass all the possible combinations of these fundamental components**.

- Certain **logical forms can become complex** when operations such as aggregation, superlatives, and comparisons are involved.
Our Method

- LLMs have the strong capability to accurately capture the semantics of relations between values in the data, enabling to transform the structured context to narrative text.
  - Structured logical forms to **natural language questions**

- LLMs have proven their strong generalizability on a wide range of few-shot and zero-shot tasks with Chain-of-Thought.
  - **Apply CoT** to solve few-shot KBQG

Our Method

Supportive Logical Forms Selection

Logical Forms

(AND medicine.routed_drug (JOIN medicine.routed_drug.marketed_formulations m.0hqs1x_))

... 

(AND music.producer (JOIN music.producer.tracks_produced (JOIN music.recording.producer m.01w8w6z)))

Structures

(A AND r (JOIN r e))

... 

(A AND r (JOIN r e)))

Clusters

1

k

Supportive Logical Forms

1 (AND sports.sport (JOIN sports.sport.team_coaches John Russo))

2 (COUNT (AND music.album (JOIN music.album.featured_artists Sierra Leone)))

... 

K (JOIN (R.people.person.place_of_birth) (JOIN music.producer.releases_produced Poguery In Motion))

Prompt Construction

Input: (AND sports.sport (JOIN sports.sport.team_coaches John Russo))

Subgraph1: (JSON (JOIN sports.sport.team_coaches John Russo))

Subgraph2: (AND sports.sport Subgraph1)

Subquestion1: sport team coach john russo

Subquestion2: Which sport does john russo coach?

Input: (JOIN (R.people.person.place_of_birth) (JOIN music.producer.releases_produced Poguery In Motion))

Subgraph1: (JOIN music.producer.releases_produced Poguery In Motion)

Subgraph2: (JOIN (R.people.person.place_of_birth) Subgraph1)

Subquestion1: music producer of released produce poguery in motion

Subquestion2: Where is the birth place of the music producer of poguery in motion?

Input: (COUNT (AND aviation.aircraft_manufacturer (JOIN organization. organization.legal_structure.S.A.)))

Prediction

Subgraph1: (JOIN organization.organization.legal_structure.S.A.)

Subgraph2: (AND aviation.aircraft_manufacturer Subgraph1)

Subgraph3: (COUNT Subgraph2)

Subquestion1: organization legal structure of s.a.

Subquestion2: aircraft manufacturer in the legal structure of s.a.

Subquestion3: What is the number of aircraft manufacturer in the legal structure of s.a.?
Supportive Logical Forms Selection

Step 1: Structure Encoding and Clustering
1. We extract the structure of logical form by converting the schema items into symbolic variables.

\[(\text{AND medicine.routed\_drug} (\text{JOIN medicine.routed\_drug.marketed\_formulations} m.0hqs1x))].\]

\[\downarrow\]

\[(\text{AND r} (\text{JOIN r e}))\]

2. We encode the contexts of the sequence with Sentence-Transformers.

Step 2: Logical Form Sampling
1. We utilize the **K-means clustering algorithm** to group the encoded structure into k-clusters based on their syntactic similarity.

2. We greedily pick up a candidate with **least semantic similarity** to the selected logical forms, where the similarity is measured by the encoding of the original logical forms.
• Generate a straightforward question that queries a **one-hop relation** from the topic entity.
• One-hop relation **subgraph1** leads to a simple **subquestion1**.
• Generate a question that inquires about a **two-hop relation chain** involving the aforementioned one-hop relation.
  The Step 2 includes the parsed logical form appended to the previous step as a component and generates **subquestion2** based on the **subgraph2** and **subquestion1**.
• Repeat until the entire logical forms have been traversed.
Experiments

Datasets:
• WebQuestions (WQ)
• PathQuestions (PQ)
• GrailQA (GQ)

Comparable Methods:
• LLMs+CoT methods: Standard Prompt, Random-CoT, Manual-CoT, Active-CoT, Auto-CoT
• Fine-trained methods: DSM, LFKQG, IGND, JointGT, T5-Large, etc.
• Few-shot methods: BiGraph2Seq, JointGT, AutoQGS
• Our methods: KQG-CoT, KQG-CoT+ (further display the examplers from short to long.)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Q</th>
<th>#R</th>
<th>#E</th>
<th>#T</th>
</tr>
</thead>
<tbody>
<tr>
<td>WQ</td>
<td>22,989</td>
<td>672</td>
<td>25,703</td>
<td>2/99/5.8</td>
</tr>
<tr>
<td>PQ</td>
<td>9,731</td>
<td>378</td>
<td>7,250</td>
<td>2/3/2.7</td>
</tr>
<tr>
<td>GQ</td>
<td>64,331</td>
<td>3,720</td>
<td>32,585</td>
<td>1/4/1.4</td>
</tr>
</tbody>
</table>

Table 1: Statistics of the evaluated datasets. #Q denotes the number of questions. #R and #E denote the total number of relations and entities, respectively. #T denotes the minimum/maximum/average number of triplets involved in each question.
Experimental Results

<table>
<thead>
<tr>
<th>Method</th>
<th>B</th>
<th>WQ M</th>
<th>R</th>
<th>B</th>
<th>PQ M</th>
<th>R</th>
<th>B</th>
<th>GQ M</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard Prompt</td>
<td>24.86</td>
<td>29.01</td>
<td>52.74</td>
<td>55.87</td>
<td>42.24</td>
<td>76.83</td>
<td>29.17</td>
<td>33.52</td>
<td>42.95</td>
</tr>
<tr>
<td>Random-CoT</td>
<td>25.02</td>
<td>29.37</td>
<td>53.16</td>
<td>56.42</td>
<td>42.61</td>
<td>77.03</td>
<td>29.81</td>
<td>33.75</td>
<td>43.31</td>
</tr>
<tr>
<td>Manual-CoT</td>
<td>28.44</td>
<td>30.24</td>
<td>54.30</td>
<td>60.37</td>
<td>42.88</td>
<td>77.48</td>
<td>30.18</td>
<td>33.61</td>
<td>44.89</td>
</tr>
<tr>
<td>Active-CoT</td>
<td>26.02</td>
<td>29.55</td>
<td>54.01</td>
<td>58.78</td>
<td>43.86</td>
<td>76.78</td>
<td>30.27</td>
<td>33.71</td>
<td>44.07</td>
</tr>
<tr>
<td>Auto-CoT</td>
<td>28.42</td>
<td>29.65</td>
<td>53.47</td>
<td>59.59</td>
<td>43.16</td>
<td>77.13</td>
<td>30.17</td>
<td>34.22</td>
<td>44.47</td>
</tr>
<tr>
<td>KQG-CoT (Ours)</td>
<td>28.89</td>
<td>30.41</td>
<td>54.38</td>
<td>60.81</td>
<td>43.54</td>
<td>77.35</td>
<td>30.51</td>
<td>34.26</td>
<td>44.91</td>
</tr>
<tr>
<td>KQG-CoT+ (Ours)</td>
<td>29.73</td>
<td>31.08</td>
<td>55.14</td>
<td>61.71</td>
<td>44.27</td>
<td>78.41</td>
<td>31.24</td>
<td>34.94</td>
<td>45.36</td>
</tr>
</tbody>
</table>

KQG-CoT outperforms the existing LLMs+CoT methods
Experimental Results

<table>
<thead>
<tr>
<th>Method</th>
<th>B</th>
<th>WQ</th>
<th>M</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full Training</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L2A (Du et al., 2017)</td>
<td>6.01</td>
<td>26.95</td>
<td>25.24</td>
<td></td>
</tr>
<tr>
<td>Transformer (Vaswani et al., 2017)</td>
<td>8.94</td>
<td>13.79</td>
<td>32.63</td>
<td></td>
</tr>
<tr>
<td>MHQG (Kumar et al., 2019)</td>
<td>11.57</td>
<td>29.69</td>
<td>35.53</td>
<td></td>
</tr>
<tr>
<td>BiGraph2Seq (Chen et al., 2023)</td>
<td>29.45</td>
<td>30.96</td>
<td>55.45</td>
<td></td>
</tr>
<tr>
<td>T5-Large (Raffel et al., 2020)</td>
<td>28.78</td>
<td>30.55</td>
<td>55.12</td>
<td></td>
</tr>
<tr>
<td>JointGT (Ke et al., 2021)</td>
<td>30.02</td>
<td>32.05</td>
<td>55.60</td>
<td></td>
</tr>
<tr>
<td>IGND (Fei et al., 2021)</td>
<td>30.62</td>
<td>31.41</td>
<td>55.82</td>
<td></td>
</tr>
<tr>
<td>LFKQG (Fei et al., 2022)</td>
<td>31.66</td>
<td>32.69</td>
<td>56.75</td>
<td></td>
</tr>
<tr>
<td>DSM (Guo et al., 2022)</td>
<td>28.62</td>
<td>-</td>
<td>64.25</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Few-shot Evaluation</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>KQG-CoT</td>
<td>28.89</td>
<td>30.41</td>
<td>54.87</td>
<td></td>
</tr>
<tr>
<td>KQG-CoT+</td>
<td>29.73</td>
<td>31.08</td>
<td>55.46</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>B</th>
<th>PQ</th>
<th>M</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full Training</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L2A (Du et al., 2017)</td>
<td>17.00</td>
<td>50.38</td>
<td>19.72</td>
<td></td>
</tr>
<tr>
<td>Transformer (Vaswani et al., 2017)</td>
<td>56.43</td>
<td>43.45</td>
<td>73.64</td>
<td></td>
</tr>
<tr>
<td>MHQG (Kumar et al., 2019)</td>
<td>25.99</td>
<td>33.16</td>
<td>58.94</td>
<td></td>
</tr>
<tr>
<td>BiGraph2Seq (Chen et al., 2023)</td>
<td>61.48</td>
<td>44.57</td>
<td>77.72</td>
<td></td>
</tr>
<tr>
<td>AutoQGS (Xiong et al., 2022)</td>
<td>65.13</td>
<td>47.50</td>
<td>76.80</td>
<td></td>
</tr>
<tr>
<td>T5-Large (Raffel et al., 2020)</td>
<td>58.95</td>
<td>44.72</td>
<td>76.58</td>
<td></td>
</tr>
<tr>
<td>IGND (Fei et al., 2021)</td>
<td>61.69</td>
<td>45.11</td>
<td>77.28</td>
<td></td>
</tr>
<tr>
<td>LFKQG (Fei et al., 2022)</td>
<td>63.92</td>
<td>46.91</td>
<td>78.40</td>
<td></td>
</tr>
<tr>
<td>JointGT (Ke et al., 2021)</td>
<td>65.89</td>
<td>48.25</td>
<td>78.87</td>
<td></td>
</tr>
<tr>
<td>DSM (Guo et al., 2022)</td>
<td>61.03</td>
<td>-</td>
<td>86.06</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Few-shot Evaluation</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>BiGraph2Seq (Chen et al., 2023)</td>
<td>1.01</td>
<td>4.99</td>
<td>12.07</td>
<td></td>
</tr>
<tr>
<td>JointGT (Ke et al., 2021)</td>
<td>43.15</td>
<td>35.91</td>
<td>69.57</td>
<td></td>
</tr>
<tr>
<td>AutoQGS (Xiong et al., 2022)</td>
<td>43.46</td>
<td>33.55</td>
<td>68.23</td>
<td></td>
</tr>
<tr>
<td>KQG-CoT</td>
<td>60.81</td>
<td>43.54</td>
<td>77.35</td>
<td></td>
</tr>
<tr>
<td>KQG-CoT+</td>
<td>61.71</td>
<td>44.27</td>
<td>78.41</td>
<td></td>
</tr>
</tbody>
</table>

(1) KQG-CoT outperforms existing few-shot methods with large margins;
(2) KQG-CoT achieves competitive results compared with full training methods.
More Analysis

(1) KQG-CoT outperforms existing LLM-CoT methods with various k;
(2) KQG-CoT results in supportive logical forms with larger diversity.
Conclusions

• When constructing prompts, the selection and arrangement of exemplars are paramount.

• KQG-CoT outperforms the existing CoT methods significantly and achieves performance levels comparable to those of fine-tuned methods.

• The utilization of LLMs in conjunction with CoT proves highly effective for handling generation tasks with structured inputs.