Can Large Language Models Be a Good Evaluator for Review-based Product Question Answering?

Tony D. Huang Yongli Ren Xiuzhen Zhang

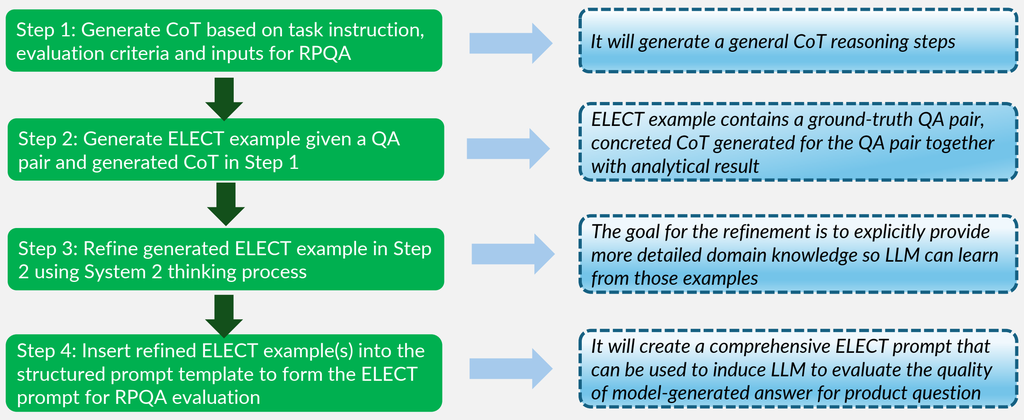
RMIT University, Australia

Introduction

- Review-based Product Question Answering (RPQA) evaluation, as a domain knowledge-intensive QA evaluation, still relies on lexicon-based and frozen embedding-based metrics. Those metrics fail if the reference answers absent.
- While LLMs have demonstrated impressive performance on evaluating NLG tasks, little research has been done on RPQA evaluation.
- In-context learning offers built-in knowledge through one or few demonstration examples, allowing LLMs to learn domain specific knowledge without updating their parameters.
- System 2 thinking process (one of the thinking processes of dual-process theories) is involving hypothetical thinking and reasoning, and requiring explicit knowledge to solve complex problems, such as Chain-of-Thought (CoT).
- Our goal is to investigate the ability of LLMs serving as evaluators for domain knowledge intensive RPQA. We are aiming to ask the following two research questions:
  - **RQ1:** Can LLM-based metrics outperform traditional metrics in RPQA setting?
  - **RQ2:** How does ELECT prompting help with LLM-based RPQA evaluation?

Methodology

- **Structured prompt template**
  - *Task instruction*: we define the role of the evaluation task followed by a task description together with the desired output format.
  - *Evaluation criteria*: we design an evaluation criteria to assess the quality of the model-generated answer given a product question.
  - *Inputs*: inputs include a given question and model-generated answer pair.
- **Our proposed method** – **E**xamp**L**e-**E**nhanced **C**o**T** prompting (**ELECT**)
  - A structured prompt template with one or few-shot in-context learning, incorporating the CoT Mechanism and the System 2 thinking process for RPQA evaluation.
  - **ELECT** prompting



Real examples of structured prompt template, general CoT, concreted CoT and refined CoT in ELECT prompting

Structured prompt template – direct prompt

You are a product question answering task evaluator.
You will be given a product question, and a model generated answer. Your task is to rate how well the model generated answer can answer the question based on the evaluation criteria. Please do not generate any opening, closing, and explanations. The response should be one word with a number between 0 and 3 only.
Evaluation Criteria: Rating Score (0-3)
0: the answer was irrelevant to the question;
1: the answer was related but didn't answer the question;
2: the answer somehow answered or partially answered the question;
3: the answer directly answer the question.
Product Question:
I know that every review says they're great for close-up viewing, but how about far distance?
Model Generated Answer:
for distance viewing, you can do much better.

General CoT

1. Understanding the Product Question
Clearly define the product-related question being asked. Identify any specific details or requirements the question is asking for (e.g., features, specifications, price, availability).
2. Analyzing the model-generated Answer
Check whether the model-generated answer addresses the key aspects of the question.
Determine if the model-generated answer provides the necessary information (or if something is missing).
Verify if the model-generated answer is directly related to the product in question.
Evaluate if the model-generated answer includes irrelevant or extraneous information.
3. Rating the model-generated Answer Based on the Evaluation Criteria
***Score 0**: If the model-generated answer is entirely irrelevant or doesn't address any part of the question.*
***Score 1**: If the model-generated answer is somewhat related to the question, but it doesn't answer it.*
***Score 2**: If the model-generated answer partially answers the question but is incomplete or lacks clarity.*
***Score 3**: If the model-generated answer fully and directly answers the question with clear and relevant information.*
4. Conclusion
Based on the analysis, assign a score (0–3) to the model's answer.

General demonstration example

Product Question: Is this difficult to clean?
Model Generated Answer: it can be a little tedious to clean.
Rating Score: 3

ELECT demonstration example without refinement

***Product Question**: Is this difficult to clean?*
***model-generated Answer**: it can be a little tedious to clean.*
***Chain of Thought**:*
1. Understanding the Product Question
The question asks, "Is this difficult to clean?" The user is seeking information about the ease or difficulty of cleaning the product. The expected answer should clearly address whether the product is easy or difficult to clean.
2. Analyzing the model-generated Answer
The model-generated answer says, "it can be a little tedious to clean."
The phrase "a little tedious" suggests that cleaning is not extremely difficult but requires some effort.
3. Rating the model-generated Answer Based on the Evaluation Criteria
Score 3: The model-generated answer directly answers the question by explaining that cleaning the product is somewhat difficult or tedious. It addresses the key aspect of the user's question.
4. Conclusion
The model-generated answer is relevant, clear, and directly addresses the question about difficulty in cleaning.
Rating: 3 (Directly answers the question).

Refined ELECT demonstration example

***Product Question**: Is this difficult to clean?*
***model-generated Answer**: it can be a little tedious to clean.*
***Chain of Thought**:*
1. Understanding the Product Question
The question asks, "Is this difficult to clean?" The user is seeking information about the ease or difficulty of cleaning the product. The expected answer should clearly address whether the product is easy or difficult to clean.
2. Analyzing the model-generated Answer
The model-generated answer says, "it can be a little tedious to clean."
The model-generated answer indicates that cleaning the product may not be easy and can take some effort, which aligns with the user's inquiry about difficulty.
The phrase 'a little tedious' implies some difficulty, though not extreme, and provides a direct answer to the question.
3. Rating the model-generated Answer Based on the Evaluation Criteria
Score 3: The model-generated answer directly answers the question by explaining that cleaning the product is somewhat difficult or tedious. It addresses the key aspect of the user's question.
4. Conclusion
The model-generated answer is relevant, clear, and directly addresses the question about difficulty in cleaning.
Rating: 3 (Directly answers the question).

Experiments and Results

Table 1: Results of correlation scores between human judgments and various evaluation metrics including LLM-based metrics and traditional metrics.

Product Category	GPT-4o		Prometheus		Llama		DGML		BERTScore	
	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r
Automotive	**0.698**	0.654	0.633	0.600	0.531	0.544	0.596	0.500	0.509	0.529
Cell phones	**0.747**	0.711	0.636	0.632	0.537	0.531	0.617	0.418	0.496	0.509
Electronics	**0.716**	0.729	0.609	0.617	0.497	0.529	0.540	0.471	0.527	0.578
Patio lawn and garden	**0.719**	0. 674	0.667	0.654	0.591	0.570	0.599	0.522	0.403	0.391
Office products	**0.837**	0.786	0.736	0.729	0.631	0.616	0.668	0.459	0.541	0.567
Pet supplies	0.649	0.621	0.653	0.638	0.612	0.606	**0.661**	0.530	0.429	0.479
Average	**0.728**	0.696	0.656	0.645	0.567	0.566	0.614	0.483	0.484	0.509

Table 2: Results of correlation scores between human judgments and various prompting strategies for Llama. 'G-Eval': direct prompt with general CoT; 'DP': direct prompt without CoT and demonstration; 'One-shot DP': direct prompt with one demonstration without example-enhanced CoT.

Product Category	ELECT		G-Eval		DP		One-shot DP	
	ρ	r	ρ	r	ρ	r	ρ	r
Automotive	**0.607**	0.589	0.549	0.556	0.531	0.544	0.574	0.552
Cell phones	0.539	0.548	0.527	0.512	0.537	0.531	**0.543**	0.521
Electronics	**0.614**	0.603	0.546	0.534	0.501	0.537	0.521	0.537
Patio lawn and garden	**0.682**	0.665	0.574	0.559	0.591	0.570	0.583	0.537
Home and kitchen	**0.484**	0.445	0.475	0.445	0.456	0.413	0.407	0.325
Office products	**0.661**	0.640	0.633	0.611	0.631	0.616	0.613	0.598
Pet supplies	**0.648**	0.643	0.604	0.592	0.612	0.606	0.635	0.615
Sports and outdoors	**0.676**	0.649	0.632	0.582	0.594	0.577	0.632	0.606
Tools and home impr.	**0.642**	0.636	0.599	0.581	0.592	0.595	0.619	0.614
Toys and games	**0.661**	0.630	0.575	0.545	0.581	0.563	0.624	0.595
Average	**0.621**	0.605	0.571	0.552	0.563	0.555	0.575	0.550

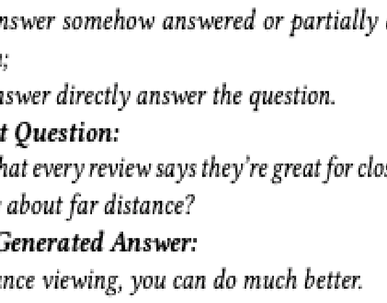


Figure 2: scores distribution of 'Automotive' product category with one-shot, two-shot, three-shot and four- shot ELECT demonstration for Llama.

Datasets: Amazon CQA dataset and reviews collected by Julian McAuley et al.
RPQA model: UnifiedQA (Daniel Khashabi et al.)
Human annotation: We employed master workers from Amazon Mechanical Turk platform. 1526 questions are collected after filtered by the quality control process.
Baselines: Llama, Prometheus, GPT-4o, DGML and BERTScore
Correlation measures: Spearman and Pearson

Discussions

Findings in RQ1:

- Table 1 overall results show that LLM-based metrics have strong alignments with human judgements, which can be good evaluators in RPQA setting. This indicates that the capability of LLM-based evaluator is not only distinguishing between good and bad model-generated answers, but also measuring the quality level of answers and reflecting the absolute differences in the scores.

Findings in RQ2:

- Table 2 shows that ELECT prompting is statistically higher than other three prompting strategies (G-Eval, DP and One-shot DP) in terms of Spearman and Pearson correlation scores. This indicates that ELECT prompting method can help Llama understand more on the demonstration context and latent thinking process, and it further guides Llama to generate more reliable metric scores aligned with human judgements.
- Figure 2 shows the impact of the number of examples provided with ELECT in the demonstration. We found that the few-shot demonstration does not result in a more significant improvement than the one-shot demonstration.

Conclusion

In this work, we experimentally study the capability of LLMs as evaluators for RPQA. Specially, we design ELECT prompting method that can be used with one or few-shot in-context learning, where examples of demonstration are adapted with CoT reasoning using System 2 thinking process. Experiment results show that LLMs can be effective evaluators for RPQA. In particular, ELECT prompting method can significantly improve the performance comparing with other LLMs evaluation methods.