CUET_SR34 at CQs-Gen 2025: Critical Question Generation via Few-Shot LLMs – Integrating NER and Argument Schemes

Sajib Bhattacharjee, Tabassum Basher Rashfi, Samia Rahman, Hasan Murad
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology, Bangladesh
{u2004003, u2004004, u1904022}@student.cuet.ac.bd, hasanmurad@cuet.ac.bd

Abstract

Critical Question Generation (CQs-Gen) improves reasoning and critical thinking skills through Critical Questions (CQs), which identify reasoning gaps and address misinformation in NLP, especially as LLM-based chat systems are widely used for learning and may encourage superficial learning habits. The Shared Task on Critical Question Generation, hosted at the 12th Workshop on Argument Mining and co-located in ACL 2025, has aimed to address these challenges. This study proposes a CQs-Gen pipeline using Llama-3-8B-Instruct-GGUF-Q8_0 with few-shot learning, integrating text simplification, NER, and argument schemes to enhance question quality. Through an extensive experiment testing without training, fine-tuning with PEFT using LoRA on 10% of the dataset, and few-shot fine-tuning (using five examples) with an 8-bit quantized model, we demonstrate that the few-shot approach outperforms others. On the validation set, 397 out of 558 generated CQs were classified as Useful, representing 71.1% of the total. In contrast, on the test set, 49 out of 102 generated CQs, accounting for 48% of the total, were classified as Useful following evaluation through semantic similarity and manual assessments.

1 Introduction

Critical Question Generation (CQs-Gen) is the automated process of generating questions to assess the strength, validity, and assumptions of arguments in a text. Instead of simple factual questions, critical questions promote deeper inference and reasoning, essential for critical thinking (Ennis, 2011). With the rise of LLM-based chats, there is concern that students may develop superficial learning habits, weakening their crucial critical thinking abilities. Critical Questions (CQs) sharpen one's mind by exposing weakness in arguments and forging stronger arguments (Walton, 2006).

The CQs-Gen shared task (Calvo Figueras et al., 2025), held as part of the 12th Workshop on Argument Mining (ACL 2025), focused on generating critical questions from argumentative texts. Unlike earlier QG models (Du et al., 2017; Heilman and Smith, 2010) that focused on surface-level question generation, this task emphasizes deeper reasoning. Previous models often missed logical structure and implicit assumptions, lacking the use of tools like NER, text simplification, or argument schemes that could improve understanding.

In this work, we propose a pipeline to generate high-quality CQs by combining diverse strategies to enhance question development. We initiated our experiment by testing some LLMs without any training. Thereafter, we fine-tuned these models on the given dataset, incorporating experiments with and without text simplification and NER. Ultimately, the best result was achieved with few-shot fine-tuning integrating text simplification, NER, and argument schemes, even with just five training examples. Our proposed pipeline generates contextually relevant and logically targeted CQs. We evaluated multiple systems, with our best model generating 397 out of 558 useful CQs (71.1%) on the validation set and 49 out of 102 CQs (48%) on the test set.

Our key contributions include:

- We integrated text simplification, NER, and argument schemes to improve the quality of generated questions.
- Our experiment demonstrated that few-shot fine-tuning with an 8-bit model outperforms traditional fine-tuning approaches.
- We provided a reproducible implementation available at: https://github.com/Sojib001/Critical-Question-Generation

Proceedings of the 12th Argument Mining Workshop, pages 295–300
July 31, 2025 ©2025 Association for Computational Linguistics

2 Related Work

Critical Question Generation (CQs-Gen) is an improvement over standard question generation by generating questions that probe the logical structure and weakness of argumentative texts, which was first introduced by Calvo Figueras and Agerri, 2024.

Transformer-based models have advanced QG by generating grammatically correct questions. Kriangchaivech and Wangperawong (2019) found that these models made a lot of mistakes on the SQuAD dataset. Later, it was pointed out that the models copied parts of the text directly or didn't even form proper questions, mainly because they were too influenced by patterns in their training data (Lopez et al., 2020).

LLMs like GPT-3 and T5 often struggle with understanding deeper context or specialized topics (Cuskley et al., 2024). Cuskley et al. (2024) highlighted LLMs reliance on unimodal text, leading to generic outputs. Pérez-Gállego et al. (2024) demonstrated that LLMs generate questions misaligned with educational goals. Recent multimodal approaches improved distractor generation but still lack focus on argumentative reasoning (Luo et al., 2024). Li and Zhang, 2024 uses LLM in a zero-shot setting to generate questions in a controlled setting. Various QG methods (Duan et al., 2017; Subramanian et al., 2018; Yao et al., 2022) used named entities to guide models in generating contextually relevant questions.

Prior work has not combined text simplification, NER, and argumentation schemes for CQs-Gen, nor explored few-shot learning with quantized LLMs in this context. Our system leverages these techniques with few-shot learning using an 8-bit quantized LLM.

3 Data

We have used the dataset (Figueras and Agerri, 2025) provided under the Shared Task on Critical Question Generation hosted at the 12th Workshop on Argument Mining and co-located in ACL 2025 (Calvo Figueras et al., 2025) which is segmented into a sample set containing 5 interventions with 133 critical questions (CQs) and a validation set comprising 186 interventions with 4,136 CQs. It consists of argumentative texts like political debates, economic policies, social issues, security, foreign policy, and social justice.

4 System

Our task was to generate exactly three critical questions using the given intervention with an LLM. The input is an argumentative text in English and the schemes of the argument.

4.1 Simplifying Text

As Van et al., 2021 suggested, text simplification improves downstream NLP tasks, so we preprocessed the data by simplifying the intervention text, employing the Llama-3-8B-Instruct-GGUF-Q8_0.[1] We fine-tuned the model using a few-shot technique with five examples to illustrate the expected input-output mapping to the model.

4.2 Named Entity Recognition Feature

Following Harrison and Walker, 2018, we used named entity recognition (NER) to boost question relevance. Using the flair/ner-english-large[2] model, we labeled entities as Person, Location, Organization, or MISC and appended them to each input to guide question generation.

4.3 Argument Schemes Feature

Baumtrog, 2021 and Yu and Zenker, 2020 advocate that argument schemes can enhance the generation of CQs by offering a structured framework. So, we integrated this feature in the input intervention across various configurations. For each data instance, argument schemes from the dataset were provided.

An illustration of how we integrated the simplification process, NER, and argument schemes with our input text is seen in Figure 1.

4.4 Initial Experimentation

We initially conducted experiments to fine-tune LLMs with a small portion of our dataset. The models evaluated were Llama-3.2-3B-Instruct,[3] Mistral-7B-Instruct-v0.3,[4] and Llama-3-8B-Instruct.[5] We used 4-bit quantization to reduce the size of the models when importing them from Hugging Face. Additionally, we applied Parameter Efficient Fine Tuning (PEFT) with Low-Rank Adaptation (LoRA) to reduce the number of trainable parameters. Finally, we have used the GGUF version of Llama-3-8B-Instruct, employing a few-shot training approach, which

[1] huggingface.co/bartowski/Meta-Llama-3-8B-Instruct-GGUF

[2] https://huggingface.co/flair/ner-english-large

[3] huggingface.co/meta-llama/Llama-3.2-3B-Instruct

[4] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

[5] huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

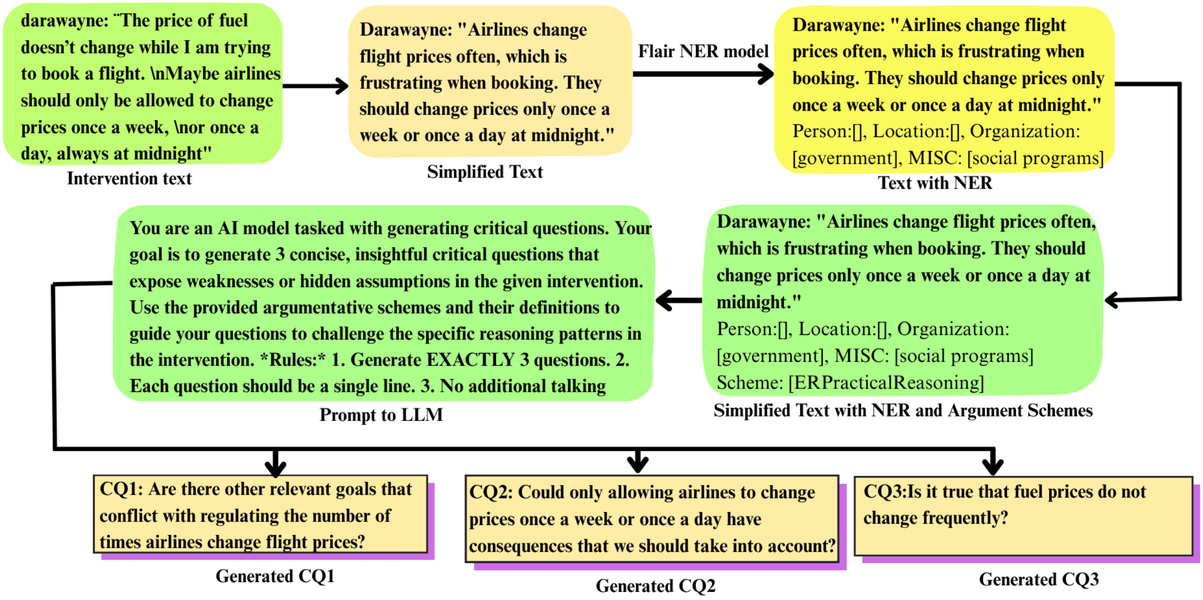
darawayne: ¨The price of fuel doesn't change while I am trying to book a flight. \nMaybe airlines should only be allowed to change prices once a week, \nor once a day, always at midnight"
Intervention text
Darawayne: "Airlines change flight prices often, which is frustrating when booking. They should change prices only once a week or once a day at midnight."
Simplified Text
Flair NER model
Darawayne: "Airlines change flight prices often, which is frustrating when booking. They should change prices only once a week or once a day at midnight."
Person:[], Location:[], Organization: [government], MISC: [social programs]
Text with NER
Darawayne: "Airlines change flight prices often, which is frustrating when booking. They should change prices only once a week or once a day at midnight."
Person:[], Location:[], Organization: [government], MISC: [social programs]
Scheme: [ERPracticalReasoning]
Simplified Text with NER and Argument Schemes
You are an AI model tasked with generating critical questions. Your goal is to generate 3 concise, insightful critical questions that expose weaknesses or hidden assumptions in the given intervention. Use the provided argumentative schemes and their definitions to guide your questions to challenge the specific reasoning patterns in the intervention. *Rules:* 1. Generate EXACTLY 3 questions. 2. Each question should be a single line. 3. No additional talking
Prompt to LLM
CQ1: Are there other relevant goals that conflict with regulating the number of times airlines change flight prices?
Generated CQ1
CQ2: Could only allowing airlines to change prices once a week or once a day have consequences that we should take into account?
Generated CQ2
CQ3:Is it true that fuel prices do not change frequently?
Generated CQ3

Figure 1: Overview of Our Proposed Critical Question Generation System.

outperformed other models due to its 8-bit quantization as opposed to the 4-bit quantization used in other models.

4.5 Overview of the Adopted Model

We used 8-bit quantized Llama-3-8B-Instruct-GGUF-Q8_0 due to its enhanced quantization compared to previously used models. We have used a few-shot technique to fine-tune it by using five interventions from the validation set and mapping expected input-output. For every example, we have provided the simplified intervention text, a categorized list of named entities, and the argument schemes of that intervention text. Figure 1 illustrates the workflow of our system.

5 Parameter Setting

For traditionally fine-tuned models, we have utilized 10% of our dataset, applying LoRA with a rank of 32 and an alpha value of 64. For our best-performing model, Llama-3-8B-Instruct-GGUF-Q8_0, we have employed a few-shot fine-tuning technique with five examples to map expected input-output. The same prompt was consistently applied throughout the experiment, as illustrated in Figure 1.

6 Evaluation Metric

The evaluation metric is based on semantic similarity with a reference question set. A Sentence-Transformer model (stsb-mpnet-base-v2)[6] was used to check semantic similarity, or the BLEURT score, between the generated CQs and reference CQs. If similarity crosses the predefined threshold of 0.65, the highest-scoring reference CQ is selected, and the corresponding label is assigned to the generated CQ. If the highest semantic similarity falls below the threshold, the question is labeled as "not_able_to_evaluate" and subjected to manual evaluation later.

7 Result and Analysis

The produced CQs were categorized into four classes based on semantic similarity: Useful (USE), Unhelpful (UN), Invalid (IN), and Not Able to Evaluate (NAE). Questions in the last class require manual evaluation. Based on Table 1, five-example few-shot learning outperformed no training and traditional training, producing the most useful CQs.

The number of successful CQs increased with larger model sizes in both no-training and fine-tuned conditions. Fine-tuning on 10% of the data, along with features like NER and text simplification, also improved performance. Llama-3-8B-Instruct showed consistent performance in all conditions, with its quantized GGUF version performing best under few-shot learning. Including the argument schemes as an input feature made the model optimal.

Table 2 presents the test results of our three submitted models, and the optimal performance is from the model with simplified text, NER, and argument schemes. Human assessment of uneval-

[6] https://huggingface.co/sentence-transformers/stsb-mpnet-base-v2

Type	Model	Count			
		USE	UN	IN	NAE
No Training	Llama-3.2-3B-Instruct	100	20	10	428
	Mistral-7B-Instruct-v0.3	248	42	16	252
	Llama-3-8B-Instruct	253	56	15	234
Fine-Tuned	Llama-3.2-3B-Instruct	150	30	11	367
	Mistral-7B-Instruct-v0.3	260	42	16	240
	Llama-3-8B-Instruct	260	50	13	235
	Llama-3-8B-Instruct (original text + NER)	266	47	14	231
	Llama-3-8B-Instruct (simplified text + NER)	267	54	13	224
Few-Shot	Llama-3-8B-Instruct-GGUF-Q8_0 (simplified text + No NER)	386	73	25	74
	Llama-3-8B-Instruct-GGUF-Q8_0 (simplified text + NER)	392	74	19	73
	Llama-3-8B-Instruct-GGUF-Q8_0 (simplified text + NER + schemes)	**397**	**79**	**18**	**64**

Table 1: Count metrics of models on the validation dataset.

uated CQs generated from the model is given in Table 3.

7.1 Error Analysis

While there was consistent overall performance, our model also made some mistakes by generating unhelpful and invalid CQs.

Unhelpful CQs were generated when the model did not fully understand the argument schemes. These questions were well-formed but did not challenge the speaker's reasoning or assumptions. As an example, in the "CLINTON_199_2" intervention, the model asked an abductive question regarding cooperation, missing the chance to challenge the logic or feasibility of Clinton's argument.

Invalid CQs resulted from misunderstanding the argument. For instance, in "CLINTON_25," the model generated a question about inflation and environmental destruction when Clinton actually spoke about clean energy and the economy, which did not match the actual topic.

Examples of both types of errors are shown in Table 4.

Run	Model	Count			
		USE	UN	IN	CAE
Test Run 1	**Llama-3-8B-Instruct-GGUF-Q8_0 (simplified text + NER + schemes)**	**44**	**29**	**17**	**12**
Test Run 2	Llama-3-8B-Instruct-GGUF-Q8_0 (simplified text + No NER)	33	28	17	24
Test Run 3	Llama-3-8B-Instruct-GGUF-Q8_0 (simplified text + NER)	43	24	14	21

Table 2: Count metrics of models on the test dataset.

Run	Model	Count		
		USE	UN	IN
Test Run 1	Llama-3-8B-Instruct-GGUF-Q8_0 (simplified text + NER + schemes)	49	34	19

Table 3: Final metrics of models on the test dataset after manual evaluation.

8 Conclusion

Though the task of critical question generation using LLMs is a new task, we have generated a huge number of useful critical questions using Llama-3-8B-Instruct-GGUF-Q8_0 using features like text simplification, NER, and argumentation schemes. We have contributed by exploring multiple models and enhancing the input by adopting text simplification, NER, and argumentation schemes as a feature. Among all the models, Llama-3-8B-Instruct-GGUF-Q8_0 outperformed others by generating 397 useful CQs (71.1%) in the validation set and 49 useful CQs (48%) in the test dataset after manual evaluation. However, it has struggled to understand the argument schemes and the right interpretation of a given intervention.

9 Limitations

While this study presents useful insights into CQs-Gen with LLMs, it is not without limitations. We were only able to utilize two models, Mistral and Llama, due to time and resource limitations. The fine-tuning was performed using 4-bit quantized models with only 10% of the data. Additionally, we were not able to evaluate the CQs labeled "Not Able to Evaluate (NAE)" because it was done by the task organizers. In the future, we plan to use larger models, train on more data, improve NER, and explore other languages to make our approach more generalizable.

References

Michael D Baumtrog. 2021. Designing critical questions for argumentation schemes. *Argumentation*.

Blanca Calvo Figueras and Rodrigo Agerri. 2024. Critical questions generation: Motivation and challenges. *arXiv preprint arXiv:2410.14335*.

Blanca Calvo Figueras, Jaione Bengoetxea, Maite Heredia, Ekaterina Sviridova, Elena Cabrio, Serena Villata, and Rodrigo Agerri. 2025. Overview of the critical questions generation shared task 2025. In *Proceedings of the 12th Workshop on Argument Mining (ArgMining 2025)*, Vienna, Austria. Association for Computational Linguistics.

Christine Cuskley, Rebecca Woods, and Molly Flaherty. 2024. The limitations of large language models for understanding human language and cognition. *Open Mind*.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. *arXiv preprint arXiv:1705.00106*.

Nan Duan, Duyu Tang, Peng Chen, and Ming Zhou. 2017. Question generation for question answering. Association for Computational Linguistics.

Robert H. Ennis. 2011. Critical thinking: Reflection and perspective. *Inquiry: Critical thinking across the Disciplines*.

Banca Calvo Figueras and Rodrigo Agerri. 2025. Benchmarking critical questions generation: A challenging reasoning task for large language models. *arXiv preprint arXiv:2505.11341*.

Vrindavan Harrison and Marilyn Walker. 2018. Neural generation of diverse questions using answer focus, contextual and linguistic features. *arXiv preprint arXiv:1809.02637*.

Michael Heilman and Noah A. Smith. 2010. Good question! statistical ranking for question generation. Association for Computational Linguistics.

Kettip Kriangchaivech and Artit Wangperawong. 2019. Question generation by transformers. *arXiv preprint arXiv:1909.05017*.

Kunze Li and Yu Zhang. 2024. Planning first, question second: An LLM-guided method for controllable question generation. Association for Computational Linguistics.

Luis Enrico Lopez, Diane Kathryn Cruz, Jan Christian Blaise Cruz, and Charibeth Cheng. 2020. Transformer-based end-to-end question generation. *arXiv preprint arXiv:2005.01107*.

Haohao Luo, Yang Deng, Ying Shen, See-Kiong Ng, and Tat-Seng Chua. 2024. Chain-of-exemplar: Enhancing distractor generation for multimodal educational question generation. Association for Computational Linguistics.

Pablo Pérez-Gállego, Marc Franco-Salvador, Luis García-Sardiña, Marco del Tredici, and Roberto Navigli. 2024. Analysis of large language models for educational question classification and generation. *Computers and Education: Artificial Intelligence*.

Sandeep Subramanian, Tong Wang, Xingdi Yuan, Saizheng Zhang, Adam Trischler, and Yoshua Bengio. 2018. Neural models for key phrase extraction and question generation. *arXiv preprint arXiv:1706.04560*.

Hoang Van, Zheng Tang, and Mihai Surdeanu. 2021. How may i help you? using neural text simplification to improve downstream nlp tasks. *arXiv preprint arXiv:2109.04604*.

Douglas Walton. 2006. *Fundamentals of Critical Argumentation*. Cambridge University Press.

Bingsheng Yao, Dakuo Wang, Tongshuang Wu, Zheng Zhang, Toby Jia-Jun Li, Mo Yu, and Ying Xu. 2022. It is AI's turn to ask humans a question: Question-answer pair generation for children's story books. *arXiv preprint arXiv:2109.03423*.

Shiyang Yu and Frank Zenker. 2020. Schemes, critical questions, and complete argument evaluation. *Argumentation*.

A Appendix: Examples of Invalid and Unhelpful CQ Errors

Label	Intervention_ID, CQ_ID	Critical Question (CQ)
Unhelpful	CLINTON_199_2 (CQ_ID 1)	How does Clinton's proposal to work closely with Muslim-majority nations in the Middle East and Europe address the concerns of Muslim communities, and what specific measures does she plan to take to ensure their cooperation?
	CLINTON_223_2 (CQ_ID 2)	What are the consequences of Clinton's approach to diplomacy, and how does she weigh the benefits of building coalitions against the potential risks of relying on other nations' cooperation?
Invalid	CLINTON_25 (CQ_ID 2)	What specific measures will be taken to ensure that the creation of jobs and stimulation of the economy will not lead to unintended consequences, such as increased inflation or environmental degradation?
	Elmattador_92 (CQ_ID 0)	What specific arguments or points made during the debate did Elmattador find unconvincing or problematic, rather than simply attacking the tone or demeanor of the debaters?

Table 4: Representative Examples of Invalid and Unhelpful CQ Errors