# Mind_Matrix at CQs-Gen 2025: Adaptive Generation of Critical Questions for Argumentative Interventions

**Sha Newaz Mahmud, Shahriar Hossain, Samia Rahman,**
**Momtazul Arefin Labib, Hasan Murad**
Department of Computer Science and Engineering,
Chittagong University of Engineering and Technology, Bangladesh
{u2004081,u2004069,u1904022,u1904111}@student.cuet.ac.bd,
hasanmurad@cuet.ac.bd

## Abstract

To encourage computational argumentation through critical question generation (CQs-Gen), we propose an ACL 2025 CQs-Gen shared task system to generate critical questions (CQs) with the best effort to counter argumentative text by discovering logical fallacies, unjustified assertions, and implicit assumptions. Our system integrates a quantized language model, semantic similarity analysis, and a meta-evaluation feedback mechanism including the key stages such as data preprocessing, rationale-augmented prompting to induce specificity, diversity filtering for redundancy elimination, enriched meta-evaluation for relevance, and a feedback-reflect-refine loop for iterative refinement. Multi-metric scoring guarantees high-quality CQs. With robust error handling, our pipeline ranked 7th among 15 teams, outperforming baseline fact-checking approaches by enabling critical engagement and successfully detecting argumentative fallacies. This study presents an adaptive, scalable method that advances argument mining and critical discourse analysis.

## 1 Introduction

Critical Questions (CQs) are designed specifically to challenge argumentative texts by uncovering logical fallacies, unsupported claims, and underlying assumptions (Walton et al., 2008). In accordance with the theory of argumentation, CQs promote rational discourse by stimulating a more detailed evaluation of claims; thus, they are critical to applications such as debate analysis, pedagogy, and policy critique (Lawrence and Reed, 2019). Investigating CQs-Gen is valuable because it adds to computational argumentation, enabling systems to enhance critical thinking and debunk false information without solely relying on fact-checking, which is often limited by consensus or data availability.

The ACL 2025 CQs-Gen shared task (Calvo Figueras et al., 2025) aims to advance computational argumentation by generating CQs that uncover these logical fallacies and assumptions. Previous CQs-Gen systems, which were commonly rule-based templates or early NLP-based, could not produce diverse, context-aware questions, instead yielding imprecise or redundant responses (Cao and Wang, 2021). These limitations necessitate adaptive and scalable solutions.

This paper describes our submission to the CQs-Gen Shared Task, which is designed to generate three high-quality and diverse CQs through a five-stage pipeline: (1) Data Preprocessing to normalize interventions, (2) CQs Generation using a quantized LLaMA-3 model, (3) Post-processing and validation to ensure well-formed questions, (4) semantic ranking to select the top three questions, and (5) an Adaptive Meta-Evaluation Loop to refine question quality, which finalizes and packages three CQs per intervention into a JSON file. This approach ensures contextually appropriate and useful CQs that enhance critical engagement with argumentative text. The implementation details have been provided in the following GitHub repositories[1] for reproducibility purposes.

## 2 Related Work

CQs are rooted in argumentation schemes that formalize reasoning patterns and associated questions to check assumptions, evidence, and logical consistency (Walton et al., 2008). Computational argumentation CQs assess argument quality and identify fallacies, allowing applications such as educational software (Pinkwart and McLaren, 2012). Corpora such as the Argument Reasoning Comprehension Task (Habernal et al., 2018) and Argument Annotated Essay Corpus (Stab and Gurevych, 2017) support argument mining but rarely include explicit CQs; thus, the CQs-Gen dataset is a new

[1] https://github.com/SM-Shaan/shared-task-critical-questions-generation.



*Proceedings of the 12th Argument Mining Workshop*, pages 332–339

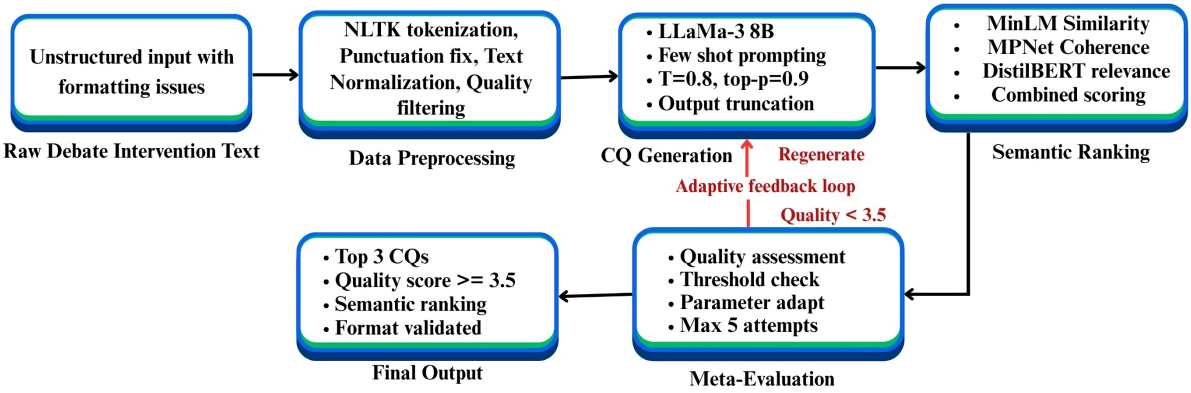


Figure 1: Workflow diagram of our proposed methodology.

contribution. Recent advances in question generation (QG) depend on transformer models to produce controllable questions, such as why-questions and counterfactuals (Cao and Wang, 2021). CQ generation is distinct and must aim at argumentative weaknesses, evaluated using fine-grained metrics such as utility (Scialom et al., 2021). Recent work by (Calvo Figueras and Agerri, 2024) underscores the motivation and challenges of computationally generating critical questions, highlighting the need for systems that produce context-aware, diverse, and argumentatively relevant questions to effectively challenge such claims. This study supplements these studies by employing rationale-augmented prompting and meta-evaluation to enhance the quality of CQs for the CQs-Gen shared task.

## 3 Dataset Description

The CQs-Gen dataset, as described in (Calvo Figueras and Agerri, 2025), includes debate interventions annotated with argumentation schemes and reference CQs labeled Useful, Unhelpful, or Invalid. Participants were provided with a small development sample and a larger validation set. An overview of the dataset is presented in Table 1. Combining the sample and validation datasets, all the schemes are listed in Figure 2 with their frequencies across the entire dataset.

| Set | # Int. | # CQs | % U | % UN | % IN |
|---|---|---|---|---|---|
| Sample | 6 | 122 | 48.36 | 29.51 | 22.13 |
| Validation | 186 | 4,136 | 67.46 | 21.59 | 10.95 |
| Test | 34 | 806 | 42.68 | 31.02 | 26.30 |

Table 1: Statistics of the CQs-Gen dataset.

## 4 Methodology

In this section, we describe the end-to-end pipeline of our CQs-Gen system, illustrated in Figure 1, organized into five-stage pipeline.

### 4.1 Data Preprocessing

We begin by normalizing each intervention to ensure well-formed sentence boundaries and punctuation. Raw debate texts often contain line breaks, missing periods, and irregular capitalization, which can confuse the language model. We apply NLTK's sent_tokenize [2] to split the text into sentences, then append a period to any sentence that does not end in one of '.', ';', '!', or '?'. Finally, we recombine the sentences into a single string. This "enhanced_normalize_text" step not only improves downstream tokenization but also maintains a minimum punctuation ratio (default 0.4) to prevent the occurrence of degenerate inputs.

### 4.2 CQ Generation

Our core generator is a quantized LLaMA-3 (8B, 4-bit GGUF) model. We employ two Appendix B prompts: the Few-Shot Prompt (Short) to support fast generation with few exemplars, and the Structured 3-Q Prompt in order to enforce a strict three-question structure. We sample with temperature $T = 0.8$ and top-$p = 0.9$, truncating at the fourth question indicator ("4.") to ensure three interrogative, numbered items without commentary.

### 4.3 Post-processing and Validation

The raw model output may contain irrelevant texts or malformed questions. We apply a regular expression (`((?m)^\s*(\d+)\.\s*(.+\?)\s*$`) in multi-

[2] https://www.nltk.org/api/nltk.tokenize.sent_tokenize.html

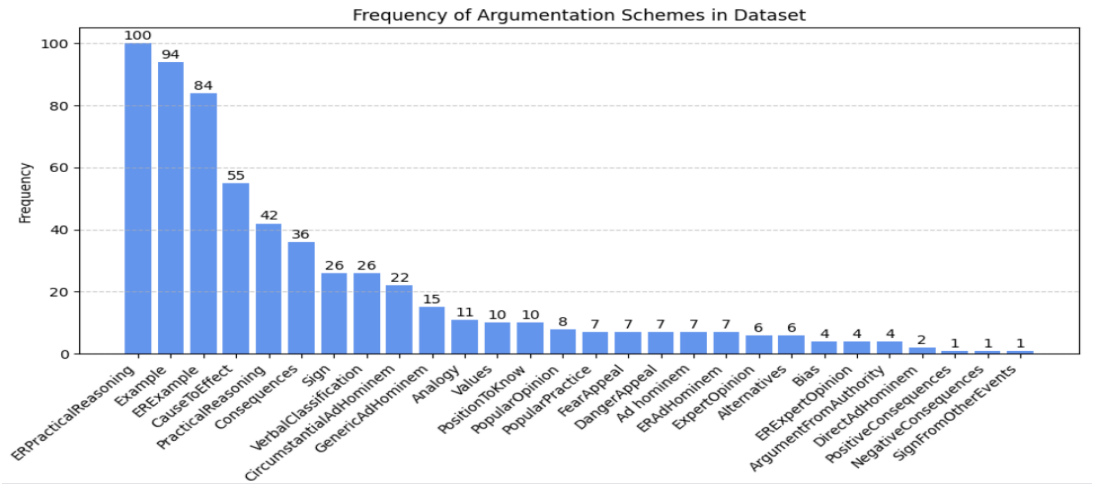


Figure 2: Frequency of argumentation schemes across the full dataset.

line mode to extract lines starting with an integer, followed by a period, and ending with a question mark. If fewer than three questions are found, we split on newlines, retaining only lines ending in a question mark. Each candidate question must have at least six words and be in interrogative form. A heuristic diversity check discards question pairs with a word-overlap ratio above 0.6, promoting varied content.

### 4.4 Semantic Ranking

To choose the top three questions when more than three pass validation, we embed the intervention and each CQ using three SentenceTransformer models:

- all-MiniLM-L6-v2[3] – measures semantic similarity, ensuring CQs align closely with the intervention's meaning.
- all-mpnet-base-v2[4] – evaluates coherence, assessing the logical clarity of CQs.
- msmarco-distilbert-base-v3[5] – determines relevance, focusing on CQs that target argumentative weaknesses.

We compute the cosine similarities between the intervention embedding and each question embedding and then calculate a weighted sum:

$$\text{score} = 0.4 \times \text{sim} + 0.3 \times \text{coh} + 0.3 \times \text{rel}.$$

Weights were empirically optimized via sensitivity analysis (Section 5.3, Table 3) to prioritize contextual alignment while ensuring clarity and argumentative focus. Then, Questions are sorted by this score, and the top three are retained for final evaluation.

### 4.5 Adaptive Meta-Evaluation Loop

To further ensure usefulness, we embed a feedback loop: the top three CQs are fed back into the LLaMA-3 (8B, 4-bit) model via a meta-evaluation prompt that asks for a 1–5 rating on how effectively the questions challenge the argument. If the average score is below 3.5 or the heuristic diversity checks (word-overlap ratio >0.6) fail, we adapt the generation parameters—either lowering the temperature by 0.1 (down to 0.5) or switching to the alternate prompt template—and retry up to five attempts. If no set meets the threshold, the highest-scoring set from prior iterations is retained. This loop enhances the relevance and diversity of CQ, addressing the limitations of semantic ranking alone.

Finally, we apply this adaptive pipeline to each intervention in the development or validation sets. The generated CQs (exactly three per intervention) are packaged alongside the intervention metadata into a JSON file conforming to the shared task submission format.

## 5 Experiments and Results

### 5.1 Experimental Setup

We evaluated our CQ-Gen pipeline using a quantized LLaMA-3 (8B, 4-bit GGUF) model, DeepHermes-3-Llama-3-8B, which was chosen for

[3] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[4] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

[5] https://huggingface.co/sentence-transformers/msmarco-distilbert-base-v3

its efficiency in few-shot prompting. For the CQs-Gen 2025 shared task, two systems were submitted for testing: DeepHermes-3-Llama-3-8B and TheBloke/Mistral-7B-OpenOrca-GPTQ. In the validation phase, three additional models were evaluated: meta-llama/Llama-2-7b-chat-hf, Zero-Shot LLaMA-3 (as a baseline), and google/flan-t5-large (as baselines). All models were hosted on a 16 GB VRAM GPU. We adopt the shared task's utility-based scoring: each Useful CQ receives 0.33 points, Unhelpful and Invalid receive 0. The per-intervention score is the sum of three questions (max = 1.0).

### 5.2 Overall Performance

Table 2 lists punctuation-scores for the CQs-Gen task, with test results in Table 4 (similarity metric). DeepHermes-3-Llama-3-8B topped with a validation score of 0.53 and test score of 0.42, surpassing TheBloke/Mistral-7B-OpenOrca-GPTQ (0.46 validation, 0.36 test), meta-llama/Llama-2-7b-chat-hf (0.50 validation), Zero-Shot LLaMA-3 (0.26 validation), and google/flan-t5-large (0.20 validation). In the test phase (similarity metric), it produced 43 Useful, 20 Unhelpful, 32 Not Able to Evaluate, and 7 Invalid CQs, versus Mistral-7B's 37 Useful, 14 Unhelpful, 43 Not Able to Evaluate, and 8 Invalid. Under the manual metric for ACL 2025 CQs-Gen, DeepHermes scored 0.559, with 57 Useful (55.88%), 27 Unhelpful (26.47%), and 18 Invalid (17.65%) CQs ($57 \times 0.33 \approx 0.559$).

| Model | Punctuation Score |
|---|---|
| **DeepHermes-3-Llama-3-8B** | **0.53** |
| Mistral-7B-OpenOrca-GPTQ | 0.46 |
| meta-llama/Llama-2-7b-chat-hf | 0.50 |
| Zero-Shot LLaMA-3 | 0.26 |
| google/flan-t5-large | 0.20 |

Table 2: Validation results

### 5.3 Sensitivity Analysis

To justify the semantic ranking weights (0.4 × sim + 0.3 × coh + 0.3 × rel), we tested alternative configurations and ablations on the dataset, as shown in Table 3.

### 5.4 Error Analysis

Despite strong overall performance, our system made errors in three key areas (Appendix A): (1) vague questions missing the intervention logic due to fallback or prompt drift, (2) redundant CQs by-passing word-level diversity filters, and (3) misaligned scoring from hallucinated outputs. Future work should explore embedding-based diversity re-ranking and CQs-Gen-aware external judges.

| Config | W(Sim, Coh, Rel) | Utility | 3 Useful |
|---|---|---|---|
| **Original** | **(0.3, 0.4, 0.3)** | **0.53** | **61.54%** |
| Equal | (0.5, 0.25, 0.25) | 0.47 | 53.85% |
| Sim-Heavy | (0.6, 0.2, 0.2) | 0.46 | 53.84% |
| No Sim | (0.75, 0.0, 0.25) | 0.26 | 26.67% |
| No Rel | (0.57, 0.43, 0) | 0.2 | 23.08% |

Table 3: Sensitivity analysis for ranking weights.

| Model | Test | U | UN | I |
|---|---|---|---|---|
| **DeepHermes-3-Llama-3-8B** | **0.42** | **43** | **20** | **7** |
| Mistral-7B-OpenOrca-GPTQ | 0.36 | 37 | 14 | 8 |

Table 4: Test run results based on similarity metric

## 6 Conclusion

We present an adaptive CQs-Gen system using few-shot prompting, semantic ranking, and meta-evaluation to enhance output diversity, relevance, and specificity. Achieving a punctuation score of 0.559 on the ACL 2025 shared task dataset, our system demonstrates the effectiveness of hybrid generation-evaluation loops for argument mining. Future studies will explore rationale-conditioned decoding, structured decoding, and human-in-the-loop refinement.

## 7 Limitations

Although our system is strong, it suffers from timely sensitivity, considering that the generated quality for important questions largely relies on short, few-shot, well-crafted prompts and thus limits new domain applicability. LLaMA-generated hallucinations during generation and meta-evaluation result in questionable question quality scores. Moreover, the use of iterative generation, multi-model encoding, and meta-evaluation introduces considerable inference time and resources.

## References

Blanca Calvo Figueras and Rodrigo Agerri. 2024. Critical questions generation: Motivation and challenges. In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 105–116, Miami, FL, USA. Association for Computational Linguistics.

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

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.

Shuyang Cao and Lu Wang. 2021. Controllable open-ended question generation with a new question type ontology. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6424–6439, Online. Association for Computational Linguistics.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. *Proceedings of NAACL-HLT*.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Niels Pinkwart and Bruce M. McLaren, editors. 2012. *Educational Technologies for Teaching Argumentation Skills*. Bentham Science Publishers.

Thomas Scialom, Paul-Antoine Dray, Sylvain Lamprier, and 1 others. 2021. Questeval: Summarization asks for fact-based evaluation. *EMNLP*.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3).

Douglas Walton, Chris Reed, and Fabrizio Macagno. 2008. *Argumentation schemes*. Cambridge University Press.

# A Appendix A: Examples of Error Types in Generated CQs

| Error Type | Example (Intervention_CQ ID) | Critical Question (CQ) |
|---|---|---|
| **Vague or Generic** | CLINTON_47 (id 1) | "Were there alternative ways to negotiate or resolve the issue without having to come to an agreement on the terms you disagreed with?" |
| | CLINTON_47 (id 0) | "Can you provide specific examples of where the terms of the negotiated agreement were not accurate?" |
| **Overlapping or Redundant** | Feedback-Commenter_183 (id 0) | "What specific evidence or data supports the claim that airlines treat passengers as a 'nuisance'?" |
| | Feedback-Commenter_183 (id 2) | "What specific examples or data support the claim that Southwest Airlines is a 'shining example' of how legacy airlines should treat customers?" |
| **Scoring Misalignment** | CLINTON_277 (id 1) | "Does Clinton address the potential for voter suppression or other issues that might prevent people from voting?" |
| | AB_68 (id 2) | "How does the argument define 'domination' in the context of family dynamics, and what criteria are used to determine when intervention is necessary?" |

Table 5: Representative Examples of Common CQ Error Types

# B Appendix B: Prompt Variations

| ID | Prompt Type | Description and Format |
|---|---|---|
| 1 | Basic Prompt | Generate three CQs to challenge the following argument: {text}. |
| 2 | Varied 3-Q Prompt | Multiple paraphrased instructions asking for three CQs to reveal fallacies, unsupported claims, and hidden assumptions. Each prompt ends with a numbered list starting from 1. |
| 3 | Varied 8-Q Prompt | Similar to Prompt 2, but requests exactly eight CQs. The format and objective remain the same: to highlight weaknesses in reasoning. |
| 4 | Few-Shot Prompt (Long) | Includes labeled examples of "Useful" questions and defines what makes CQ effective. Then asks the model to generate at least 5 diverse CQs for a new intervention. |
| 5 | Structured 3-Q Prompt | Direct instruction to write exactly three CQs focusing on fallacies, missing evidence, and hidden assumptions. The output must be a numbered list (no explanations). |
| 6 | Few-Shot Prompt (Short) | Includes a short example with three questions. Then prompts the model to generate exactly 3 CQs following similar logic with no added explanation. |
| 7 | Scheme-Guided Prompt (Walton) | Incorporates our own modified version of Walton's argumentation scheme to guide question generation, ensuring that questions map to specific schemes (see Appendix C). |
| 8 | Zero-Shot Prompting | Direct prompt without examples, instructing the model to generate CQs solely based on the instruction. |
| 9 | Chain-of-Thought Prompting | Prompts the model to articulate its reasoning process step-by-step before generating CQs, enhancing the depth and transparency. |
| 10 | Role-Based Prompting | Specifies a persona or expert role (e.g., "As a Critical thinker…"), steering tone and depth of the generated questions. |
| 11 | Iterative Refinement Prompting | Uses previous outputs as feedback to iteratively improve and refine CQs over multiple turns. |
| 12 | Dynamic Few-Shot Selection | Automatically selects and rotates few-shot examples based on similarity to the target argument for more tailored prompting. |

Table 6: Prompt variations used for CQs-Gen.

# C Appendix C: Walton-Style Argumentation Schemes

| Argumentation Scheme | Critical Questions |
|---|---|
| **Sign** | Is this sign always a reliable indicator of an underlying condition?<br>Could there be alternative explanations for this?<br>Is there evidence that contradicts the suggested interpretation of the sign? |
| **Practical Reasoning** | Are there other actions that could achieve the same goal more effectively?<br>What are the potential risks or downsides of taking these actions?<br>Is there strong evidence that this action will lead to the expected outcomes? |
| **Expert Opinion** | Are the experts truly qualified in this specific domain?<br>Do the experts have any biases or conflicts of interest?<br>Is the expert's opinion supported by strong evidence? |
| **Danger Appeal** | Is the danger real and supported by the evidence?<br>Are there alternative ways to mitigate this danger?<br>Is the warning of danger exaggerated for persuasive effect? |
| **Bias** | Does the alleged bias undermine the argument?<br>Can the claim be independently verified?<br>Is the same standard applied to all arguments or just this one? |
| **Popular Opinion** | Are people who believe this claim knowledgeable about the topic?<br>Can the claim be supported by objective evidence?<br>Has popular opinion been incorrect on similar issues in the past? |
| **Generic Ad Hominem** | Does this attack address the substance of the arguments?<br>Could the personal characteristics of the arguer be irrelevant to the claim itself?<br>Is there independent evidence to support or refute this argument? |
| **Example** | Are the examples provided representative of the general case?<br>Could there be counterexamples that weaken this argument?<br>Is there statistical or empirical evidence supporting this claim beyond these examples? |
| **Negative Consequences** | Are the predicted negative consequences likely to occur?<br>Is there evidence supporting this cause-and-effect relationship?<br>Could other factors influence the outcome? |
| **Fear Appeal** | Is the fear induced proportionate to the actual risk involved?<br>Could the threat be exaggerated to manipulate public opinion?<br>Are there alternative interpretations of the risk that are less alarming? |
| **Verbal Classification** | Is the classification accurate and relevant to the argument?<br>Could the labels be misleading or oversimplified?<br>Does the classification obscure the important nuances? |
| **Sign from Other Events** | Are the other events sufficiently similar to justify this inference?<br>Could these similarities be coincidental rather than causal?<br>Is there direct evidence linking these events to the condition? |
| **Popular Practice** | Does popular practice imply that the practice is correct or effective?<br>Are there cultural or contextual reasons for this practice that might not apply universally?<br>Is this practice supported by empirical evidence? |
| **Consequences** | Are the predicted consequences likely to occur?<br>What evidence supports the causal link between the action and its outcomes?<br>Could alternative actions lead to different consequences? |
| **Analogy** | Are the two cases truly comparable in terms of relevant aspects?<br>What are the key differences that might undermine this analogy?<br>Is the analogy oversimplifying complex issues? |
| **Circumstantial Ad Hominem** | Do the arguer's circumstances actually bias their arguments?<br>Is the argument being dismissed solely on personal circumstances?<br>Can the claim be evaluated independently of the arguer's situation? |
| **Argument from Authority** | Is the authority figure truly an expert on the subject?<br>Does the authority provide evidence beyond their status?<br>Can the claim be validated using independent evidence? |
| **Alternatives** | Are the alternatives plausible in the given context?<br>What evidence supports these proposed alternatives?<br>Could the original claim still hold despite these alternatives? |
| **Positive Consequences** | Are the predicted positive consequences likely to be realized?<br>What evidence supports the link between the action and positive outcomes?<br>Could there be unforeseen negative effects despite positive predictions? |
| **Position to Know** | Does the arguer's position guarantee an accurate insight?<br>Could their proximity to the issue bias their perspectives?<br>Is there independent evidence supporting the arguer's claims? |

Table 7: Templates of CQs for selected Walton-style argumentation schemes.