COGNAC at CQs-Gen 2025: Generating Critical Questions with LLM-Assisted Prompting and Multiple RAG Variants

Azwad Anjum Islam*, Tisa Islam Erana* and Mark A. Finlayson
Florida International University
Knight Foundation School of Computing and Information Sciences
11200 SW 8th Street, Miami, FL 33199, USA
{aisla028, tisla016, markaf}@fiu.edu

Abstract

We describe three approaches to solving the Critical Questions Generation Shared Task at ArgMining 2025. The task objective is to automatically generate critical questions that challenge the strength, validity, and credibility of a given argumentative text. The task dataset comprises debate statements ("interventions") annotated with a list of named argumentation schemes and associated with a set of critical questions (CQs). Our three Retrieval-Augmented Generation (RAG)-based approaches used in-context example selection based on (1) embedding the intervention, (2) embedding the intervention plus manually curated argumentation scheme descriptions as supplementary context, and (3) embedding the intervention plus a selection of associated CQs and argumentation scheme descriptions. We developed the prompt templates through GPT-4o-assisted analysis of patterns in validation data and the task-specific evaluation guideline. All three of our submitted systems outperformed the official baselines (0.44 and 0.53) with automatically computed accuracies of 0.62, 0.58, and 0.61, respectively, on the test data, with our first method securing the 2nd place in the competition (0.63 manual evaluation). Our results highlight the efficacy of LLM-assisted prompt development and RAG-enhanced generation in crafting contextually relevant critical questions for argument analysis.

1 Introduction

While LLM-based chat interfaces (e.g., ChatGPT, Gemini) provide convenient access to information, they can inadvertently promote superficial learning habits by delivering direct answers and hindering critical thinking. The *Critical Questions Generation (CQs-Gen) Shared Task* (Figueras et al., 2025) addresses this concern by challenging participants to build systems to generate insightful critical questions (CQs) about argumentative texts. Such questions serve to probe the underlying premises and implications of arguments, thereby encouraging deeper engagement and analytical reasoning. These questions are then evaluated based on their strength, relevance, and validity, and are categorized as *Useful*, *Unhelpful*, or *Invalid*.

Our approach to the task includes a series of prompting-based strategies using large language models (LLMs). First, we used a state-of-the-art model (GPT-4o) (OpenAI, 2024a) to analyze the validation data which we used to generate high-quality prompt templates. We then experimented with multiple variants of Retrieval Augmented Generation (RAG) (Lewis et al., 2021) with a smaller, less resource intensive model (GPT-4o-mini) (OpenAI, 2024b). Our RAG-based approaches include (1) few-shot prompting with in-context example selection based on embedding similarity of the interventions, (2) incorporation of manually curated argumentation scheme descriptions as supplementary context to the first approach, and (3) few-shot prompting with in-context example selection based on embedding similarity of the intervention plus a selection of associated CQs and argumentation scheme descriptions. Our experiments showed that these approaches significantly outperformed baseline prompting techniques. Our best-performing system (approach 1) achieved a maximum validation accuracy of 0.83—defined as the proportion of generated questions labeled as useful—and secured second place overall in the official evaluation with a test accuracy of 0.63.

The remainder of the paper is structured as follows. We first provide background on the task of critical question generation and related work on prompt-based and retrieval-augmented approaches (§2). We next describe the dataset and task formulation provided by the shared task organizers (§3).

*The first two authors shared equally in the ideation, implementation, and writing.

Proceedings of the 12th Argument Mining Workshop, pages 340–348
July 31, 2025 ©2025 Association for Computational Linguistics

We then detail our methodology and experimental setup (§4). Section §5 presents the results from our experiments as well as official submissions. Finally, we summarize our contributions and discuss key findings, limitations, and directions for future research (§6).

2 Related Work

The concept of Critical Questions (CQs) comes from argumentation theory, designed to expose the "blind spots" or missing information in an argument by questioning the validity of assumptions and inference (Calvo Figueras and Agerri, 2024). Walton et al.'s work provided a theoretical foundation with a comprehensive catalog of argumentation schemes each accompanied by a set of critical questions. Computational approaches to automatically generating CQs have only been explored in the past few years. Calvo Figueras and Agerri introduced *CQs-Gen* as a new NLP task. They leveraged LLMs to generate questions that dig into the hidden assumptions behind an argument. They built datasets in two ways: using templates from Walton's theory and prompting LLMs to produce critical questions. Their findings showed that only 27% of CQs generated by LLMs were valid in relation to the argumentative texts.

Beyond CQs-Gen, recent advances in LLMs have highlighted the importance of prompt engineering in guiding the model for complex reasoning tasks. Early methods such as Shin et al. (2020) showed that task-specific prompts could be optimized automatically, while more recent work like Zhou et al. (2023) demonstrated that LLMs themselves can iteratively propose and evaluate improved prompts that outperform manually written prompts. Building on this insight, we used a state-of-the-art model (GPT-4o) (OpenAI, 2024a) to analyze validation data and systematically derive better prompt structures for CQs-Gen.

Parallel to prompt development, Retrieval-Augmented Generation (RAG) (Lewis et al., 2021; Gao et al., 2024) has emerged as a powerful framework to enhance LLM performance by conditioning generation in retrieved external knowledge. A RAG framework consists of two steps: retrieval and generation. RAG provides richer context at inference time by retrieving semantically similar examples that are incorporated into the prompt, helping the model generate more relevant critical questions. This aligns with the findings by Zebaze et al. (2024); Liu et al. (2021), who showed that similarity-based in-context example selection can substantially improve LLM outputs in low-resource settings.

3 Shared Task Data

CQs are defined as inquiries that help determine whether an argument is acceptable or problematic by challenging inferences and exposing assumptions (Calvo Figueras and Agerri, 2024). The dataset consists of real debate interventions annotated with argumentation schemes and associated with sets of CQs. The validation set includes 186 interventions, each with 8 to 56 annotated CQs, while the test set comprises 34 interventions with no labeled CQ. Each annotated intervention includes the name of the speaker, annotated argumentation scheme(s), and a set of CQs labeled as:

- **Useful (USE):** The answer to this question can potentially challenge one of the arguments in the text.
- **Unhelpful (UN):** The question is valid, but unlikely to challenge the argument.
- **Invalid (IN):** The question is flawed—unrelated, overly general, or non-critical.

Participating systems were required to generate exactly three CQs per intervention, all intended to be *Useful*. Each CQ is evaluated independently: 0.33 for useful, and 0 for unhelpful or invalid CQs, with three *Useful* CQs achieving a score of 1.

The debate interventions in the validation set were also annotated with argumentation scheme labels such as *Bias*, *AdHominem*, *ArgumentFromAuthority*, etc.. While many of these tags correspond to well-known argumentative structures, no official documentation, list, or definitions were provided as part of the task. The full list of argument structures named in the data is found in Appendix A.

4 Our Approaches

As a baseline system for generating CQs with LLMs, we first developed a simple zero-shot prompt using the information provided on the task description website. The prompt is given in Appendix B. We then used a state-of-the-art LLM model, GPT-4o (OpenAI, 2024a), to analyze and identify the distinguishing characteristics of *Useful* CQs—both in terms of their semantics and syntactic patterns—by feeding it the validation data and the evaluation guidelines using the ChatGPT[1] interface. The prompt used for this step is shown in Appendix C. This step unearthed some key characteristics of *Useful*, *Unhelpful*, and *Invalid* questions, as shown in Table 4 in the Appendix F. We then manually incorporated these insights into a modified prompt template, given in Appendix D. While some of the findings were questionable—such as categorizing *"If...then...?"* style questions as indicative of *Invalid*, while this style of questions also appear as *Useful* in the data—Table 1 shows that including these findings into the prompt resulted in a significant boost in overall performance. This revised prompt template formed the foundation for all our subsequent experiments. Although we used GPT-4o for the purpose of a one-time analysis of the validation data, we conducted the rest of our experiments on a much smaller and less resource-intensive model, GPT-4o-mini (OpenAI, 2024b), due to compute limitations.

4.1 Approach 1: RAG on Interventions Alone

We experimented with few-shot prompting strategies to provide the model with contextual examples of high- and low-quality critical questions. Our baseline setup for a few-shot configuration includes augmenting the prompt with two randomly selected example interventions from the validation data. For each example, we included three random *Useful* CQs as well as one *Unhelpful* and one *Invalid* CQs.

In the first method, for each intervention, we identified the most similar interventions other than itself in the validation set using cosine similarity between intervention embeddings. We computed embeddings using the `stsb-mpnet-base-v2` sentence-transformer model (Reimers and Gurevych, 2019), which is the same model used in the official evaluation script. Note that in this method we only compared embeddings of the interventions, not the associated CQs. In a standard RAG the retrieval step fetches top-k similar documents using cosine similarity over the text embeddings. We experimented with the value of k and found that fetching the top-2 relevant documents performed best (the value of k=2 was optimal for all the methods discussed below as well). We then included these two similar interventions in the prompt as examples, along with three useful, one unhelpful, and one invalid CQs associated with each identified example, selected at random.

4.2 Approach 2: RAG on Interventions plus Argumentation Schemes

In our second method, we experimented with incorporating information about identified argumentation schemes to the selections of the first method. However, the lack of official definitions for the argumentation schemes identified in the validation data was a problem. Thus, we wrote brief descriptions for the argumentation schemes found in the validation dataset using external resources such as Walton (2013), and GPT-4o (OpenAI, 2024a). These descriptions explain the core reasoning behind each scheme and also highlight the types of concerns or weaknesses that a critical question should explore. For instance, we described *Argument from Authority* as "Argument that relies on the credibility of an expert or authoritative figure. Critical questions may examine if the authority cited is credible and relevant." For schemes without an obvious meaning—such as *ERPracticalReasoning*—we approximated their meaning by categorizing them under broader, more familiar scheme types[2]. In this case, *ERPracticalReasoning* was treated as a variant of *Practical Reasoning*. All the argumentation scheme descriptions are provided in Appendix A. We then included the scheme descriptions of the target interventions in the prompt as additional information with the goal of grounding the model in the underlying reasoning structure. However, Table 1 shows that inclusion of argumentation schemes in the prompt did not result in any noticeable improvement.

4.3 RAG on Annotated Examples Alone

Another approach we explored, but which we ultimately did not submit to the competition, was a standard RAG pipeline that retrieves semantically similar examples based on an embedding interventions along with their CQs. To generate embeddings of the documents, we used OpenAI's *text-embedding-3-large* (OpenAI, 2024c) model. Each document in the RAG vector store combines the original intervention with a set of labeled CQs: three *Useful*, one *Unhelpful*, and one *Invalid*, selected at random. We carried out the generation step using the GPT-4o-mini model using the prompt shown in Appendix E.

[1] chat.openai.com

[2] There were four schemes in this category: ERExpertOpinion, ERPracticalReasoning, ERAdHominem and SignFromOtherEvents.

Experiment Setup	**Useful**	**Unhelpful**	**Invalid**	**Unable to Evaluate**	**Score**
Baseline prompt	348	85	19	106	0.62
Baseline zero-shot prompting	424	54	43	37	0.76
2-shot prompting with random examples	435	50	32	41	0.78
RAG on interventions alone	**463**	38	23	34	**0.83**
RAG on interventions + argumentation schemes	**452**	55	29	22	**0.81**
RAG on annotated examples alone	440	55	31	32	0.79
RAG on annotated examples + argumentation schemes	**457**	21	29	51	**0.82**

Table 1: Detailed results of our different approaches on the validation dataset

4.4 Approach 3: RAG on Annotated Examples plus Argumentation Schemes

For our final approach, we enhanced the methods outlined in Section 4.3 by incorporating descriptions of the argumentation schemes associated with each target intervention as shown in the prompt template in Appendix E. These descriptions aimed to clarify the reasoning structure and guide the generation of more targeted questions. We formulated the scheme descriptions as detailed in 4.2. This method improved generation quality compared to using annotated examples alone.

5 Evaluation and Results

Automatic evaluation is conducted by comparing each generated question against the set of reference questions for that intervention using a sentence similarity model. If a generated question is sufficiently similar to a labeled reference question based on a predefined similarity threshold, it inherits the corresponding label. The scoring mechanism for different labels is described in Section 3. If no reference exceeds the similarity threshold, the generated question is flagged for manual evaluation.

All experiments described in Section 4 were conducted on the validation dataset, with results summarized in Table 1. Our three best approaches that we submit for official evaluation on the test data are highlighted in bold. These scores are conservative, treating all interventions flagged for manual evaluation as failures. The findings highlight that LLM-assisted prompt development yielded the greatest performance boost, with retrieval-augmented generation providing additional gains.

Table 2 shows the final score of the top-5 teams in the competition along with the distribution of *Useful* (USE), *Unhelpful* (UN), and *Invalid* (IN) CQs after manual evaluation by the task organizers. The results show that all three of our submissions—scoring 0.62, 0.61, and 0.58 with only automatic evaluation—would place in the top-5.

Team	**USE**	**UN**	**IN**	**Score**
ellisalicante	69	18	15	0.68
COGNAC*	64	24	14	0.**63**
CtCloud	61	25	16	0.60
DayDreamer	60	25	17	0.59
gottfried-wilhelm-leibniz	58	23	20	0.57

Table 2: Official final results on test data (top-5). Our submission is marked with an asterisk(*) symbol

6 Conclusion and Limitation

In this paper, we presented a set of RAG-based approaches for CQs-Gen using LLMs as part of the ArgMining 2025 Shared Task. Our methods focused on creating high-quality prompt using LLM-assisted data analysis and incorporating contextual supervision via retrieval-augmented generation (RAG). We submitted three RAG-based variant systems in the competition, all of which produced competitive performance against other participating systems. Our approach of in-context example selection using semantic similarity on the intervention alone produced the best score (0.63) on the test data and secured second place in the official evaluation.

While our approaches demonstrated strong performance, we acknowledge several limitations. First, our reliance on the validation set for example retrieval may have constrained generalization to novel argument types or schemes underrepresented in the data. This limitation is evident in the significant difference between the validation and test scores. Second, the lack of standardized definitions for argumentation schemes limited the effectiveness of scheme-based guidance. Our manually curated descriptions may not have captured the nuances of each scheme. Lastly, it was not qualitatively evaluated how effectively LLMs could identify the characteristics of different CQ labels. Complete reliance on LLMs at this stage risks overgeneralization.

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, 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*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3? *Preprint*, arXiv:2101.06804.

OpenAI. 2024a. Gpt-4o: An omnimodal ai model. Accessed: 2025-04-02.

OpenAI. 2024b. Gpt-4o mini: Advancing cost-efficient intelligence. Accessed: 2025-04-02.

OpenAI. 2024c. Openai text-embedding-3-large model. Accessed: 2025-04-01.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *Preprint*, arXiv:2010.15980.

Douglas Walton. 2013. *Argumentation schemes for presumptive reasoning*. Routledge, New York.

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

Armel Zebaze, Benoît Sagot, and Rachel Bawden. 2024. In-context example selection via similarity search improves low-resource machine translation. *Preprint*, arXiv:2408.00397.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023. Large language models are human-level prompt engineers. *Preprint*, arXiv:2211.01910.

A Argumentation Schemes Description

Argumentation Scheme	Description
Example	Relies on specific instances/examples as evidence. Critical questions may ask if examples are representative or sufficient.
CauseToEffect	Draws a causal connection between events. Critical questions may challenge the causal link or suggest alternatives.
PracticalReasoning	Appeals to practical considerations; weighing costs, benefits, or feasibility. Critical questions may ask for evidence that the action will achieve the outcome.
Consequences	Focuses on predicted outcomes. Critical questions may query likelihood, scope, or unintended side effects.
PopularOpinion	Appeals to what is widely believed or done. Critical questions may ask if popular sentiment justifies the claim.
Values	Based on ethical or normative standards. Critical questions may challenge how these values are defined or whether they are universally accepted.
Analogy	Makes a comparison between two cases. Critical questions may ask if the analogy holds and whether differences matter.
Sign	Uses observable indicators as evidence. Critical questions may ask if the sign reliably implies the conclusion.
FearAppeal	Employs threats or fear to persuade. Critical questions may evaluate whether the fear is justified or exaggerated.
DangerAppeal	Uses potential dangers to motivate. Critical questions may examine the realism and evidence behind the danger.
VerbalClassification	Categorizes an issue in a particular way. Critical questions may ask if the classification is appropriate or arbitrary.
ExpertOpinion	Cites expert testimony. Critical questions may scrutinize the credibility and potential bias of the expert.
Bias	Explores prejudices or preconceptions influencing the argument. Critical questions may assess their source and impact.
Alternatives	Suggests the existence of alternatives. Critical questions may ask if alternatives are viable or properly considered.
ERExpertOpinion	An elaborated form of expert opinion. Critical questions may probe the details and context of the expert evidence.
ERPracticalReasoning	An elaborated form of practical reasoning. Critical questions may examine thoroughness and realism.
AdHominem	Attacks the opponent's character instead of addressing the argument. Critical questions may challenge the relevance of the attack.
ERAdHominem	An extended ad hominem attack. Critical questions may examine whether the personal attack detracts from the actual argument.
CircumstantialAdHominem	Attacks based on opponent's circumstances. Critical questions may assess relevance to the argument.
GenericAdHominem	Makes a general personal attack. Critical questions may evaluate relevance to the argument's substance.
DirectAdHominem	A direct personal insult. Critical questions may examine if it distracts from the argument's merits.
NegativeConsequences	Highlights potential harmful outcomes. Critical questions may assess the likelihood and evidential support for these predictions.
PositiveConsequences	Highlights potential beneficial outcomes. Critical questions may examine whether these benefits are realistically attainable.
PositionToKnow	Assumes that holding a certain position grants special insight. Critical questions may assess whether the position truly provides reliable knowledge.
SignFromOtherEvents	Draws parallels between signs observed in different events. Critical questions may challenge whether the comparison is appropriate and meaningful.
ArgumentFromAuthority	Appeals to an authority's credibility to support a claim. Critical questions may evaluate the authority's reliability, expertise, and relevance.
PopularPractice	Bases claims on the commonality of a behavior or practice. Critical questions may examine whether popularity alone justifies the claim.

Table 3: Summary of argumentation schemes and associated critical questioning strategies.

B Baseline Prompt

```
You are a critical thinker. Your task is to generate three critical questions about a political or argumentative text. These questions are meant to help students evaluate the strength, validity, and credibility of the argument.
As an expert, you know that a critical question is a question that challenges the argument — it should make a thoughtful reader pause and reconsider the truth, logic, or assumptions behind the claims.
Now generate three useful critical questions, 20-30 words long, for the following text. Output should be in the format:
CQ: <Critical question>
```

C Prompt for Extracting Validation Set Commonalities

```
You are a smart, intelligent data analyst.
I want you to look through this data and find patterns or characteristics of different types of CQs. What do useful CQs have in common? What makes a CQ unhelpful or invalid? etc.
Focus on both semantic and syntactic characteristics and differences.
Use the guideline PDF for additional insight.

Uploaded files:
<validation.json>
<guidelines.pdf>
```

D Prompt Template

```
You are a critical thinker. Your task is to generate three critical questions about a political or argumentative text. These questions are meant to help students evaluate the strength, validity, and credibility of the argument.
As an expert, you know that a critical question is a question that challenges the argument – it should make a thoughtful reader pause and reconsider the truth, logic, or assumptions behind the claims.
Guidelines:
Your questions should:
> Focus only on claims made in the text.
> Target assumptions, evidence, reasoning, or consequences.
> Be specific – not something that could apply to any text.
> Raise issues that, if left unanswered, weaken the argument.
Avoid questions that:
> Ask for definitions or summaries (reading comprehension).
> Introduce new concepts not mentioned in the text.
> Are too general or vague (e.g., "Is the argument strong?")
> Are too obvious or based on common knowledge.
> Merely expand or support the argument without questioning it.
Good question starters may include:
> What evidence is there that...
> How does the speaker justify...
> Could this lead to unintended consequences?
> Are there reasonable alternatives to...
Avoid questions starting with:
> What is "it"...
> Why is this bad...
> Could you summarize...
> If X, then Y?

For example, for the following text:
<Example intervention>
Useful critical questions may look like:
<Useful Example 1>
<Useful Example 2>
<Useful Example 3>
And unhelpful/invalid questions may look like:
<Unhelpful Example 1>
<Invalid Example 1>[a]

As additional information, here are some suggestions based on the argumentation schemes present in the input text:
<Scheme: Scheme Description>[b]

Generate three useful critical questions, each 20-30 words long, for the following text.
Output should be in the format:
CQ 1: <question 1>
CQ 2: <question 2>
CQ 3: <question 3>
```

[a] Text in light blue is only included for few-shot experimental set-up.

[b] Text in dark blue is only included for the experiment that uses argumentation schemes.

E Prompt Template for RAG

```
You are a critical thinker. Your task is to generate three critical questions about a political or argumentative text. These questions are meant to help students evaluate the strength, validity, and credibility of the argument.
As an expert, you know that a critical question is a question that challenges the argument – it should make a thoughtful reader pause and reconsider the truth, logic, or assumptions behind the claims.
Definition of critical question generation:
Critical question generation involves formulating insightful and challenging questions that encourage deep analysis of a text. These questions should probe assumptions, evaluate evidence, and explore underlying reasoning, thereby fostering a critical engagement with the material.
Guidelines:
Your questions should:
> Focus only on claims made in the text.
> Target assumptions, evidence, reasoning, or consequences.
> Be specific – not something that could apply to any text.
> Raise issues that, if left unanswered, weaken the argument.
Avoid questions that:
> Ask for definitions or summaries (reading comprehension).
> Introduce new concepts not mentioned in the text.
> Are too general or vague (e.g., "Is the argument strong?")
> Are too obvious or based on common knowledge.
> Merely expand or support the argument without questioning it.
Good question starters may include:
> What evidence is there that...
> How does the speaker justify...
> Could this lead to unintended consequences?
> Are there reasonable alternatives to...
Avoid questions starting with:
> What is "it"...
> Why is this bad...
> Could you summarize...
> If X, then Y?

Suggestion based on argumentation schemes:
<Scheme explanations>[a]

Retrieved examples:
<Example interventions and labeled CQs>

Now generate three useful critical questions, 20-30 words long, for the following text.
The output must be a valid JSON string in the following format:
{ "CQ 1": "<Critical question 1>" },
{ "CQ 2": "<Critical question 2>" },
{ "CQ 3": "<Critical question 3>" }
```

[a] Text in dark blue is only included for the experiment that uses argumentation schemes.

F Identifying Characteristics of Different Type of CQs, Extracted by GPT-4o

Category	Key Features	Common Starters
Useful	Targets core claims or reasoning, demands clarification or evidence, explores alternatives, challenges assumptions or generalizations, tightly grounded in argumentation structure, precise and contextual	*How...?, What evidence...?, Could...?, Are there alternatives...?*
Unhelpful	Vague or generic, lacks critical engagement, exploratory tone, restates parts of the argument without probing deeper, often misses logical flaws or assumptions	*Is it true...?, What other...?, Are there...?, Can it be argued...?*
Invalid	Illogical or malformed structure, ambiguous references, speculative beyond the argument's scope, context-insensitive, grammatically or logically flawed, often confusing to interpret	*If... then...?, What is "it"?, Is it practically possible...?*

Table 4: Summary of identifying characteristics of different type of CQs, extracted using GPT-4o.