TriLLaMA at CQs-Gen 2025: A Two-Stage LLM-Based System for Critical Questions Generation

Frieso Turkstra * **Sara Nabhani** * **Khalid Al-Khatib**
University of Groningen
{f.turkstra,s.nabhani,khalid.alkhatib}@rug.nl

Abstract

This paper presents a new system for generating critical questions in debates, developed for the Critical Questions Generation shared task. Our two-stage approach, combining generation and classification, utilizes LLaMA 3.1 Instruct models (8B, 70B, 405B) with zero-/few-shot prompting. Evaluations on annotated debate data reveal several key insights: few-shot generation with 405B yielded relatively high-quality questions, achieving a maximum possible punctuation score of 73.5. The 70B model outperformed both smaller and larger variants on the classification part. The classifiers showed a strong bias toward labeling generated questions as *Useful*, despite limited validation. Further, our system, ranked 6th, outperformed baselines by 3%. These findings stress the effectiveness of large-sized models for question generation and medium-sized models for classification, and suggest the need for clearer task definitions within prompts to improve classification accuracy.

1 Introduction

The ability to critically question arguments is essential for structured reasoning, debate, and discourse analysis. Argumentation schemes, reusable patterns of reasoning, present a systematic framework for constructing sound arguments. Arguments built on these schemes can be critically assessed using targeted questions that reveal hidden assumptions, logical gaps, or weak reasoning. Automating the generation of such critical questions has promising applications in various domains of computational argumentation. Yet, it remains a complex challenge due to the contextual and logical understanding required to produce truly *useful* critiques.

This paper presents a new system for generating critical questions that challenge arguments in real-world debates. The proposed system was submitted to the Critical Questions Generation shared task (Calvo Figueras et al., 2025).[1] The system is based on a two-stage approach involving question generation followed by classification. Evaluation was conducted on a dataset of debate interventions annotated with argumentation schemes and labeled questions (*Useful*, *Unhelpful*, *Invalid*). The usefulness of the generated questions was assessed based on their semantic similarity to reference questions.

The system employs LLaMA 3.1 Instruct models (8B, 70B, 405B) with both zero-shot and few-shot prompting. For generation, few-shot prompting with the 405B model produced reasonable numbers of high-quality questions, highlighting the potential of large models in generating useful critiques. For classification, the 70B model outperformed smaller and larger variants. The classification module showed a strong bias toward labeling the generated questions as *Useful* (75–85%), despite only 44.4% of them being validated as such. Deliberation- and debate-based classification strategies were explored, but simple zero-shot prompting yielded superior performance, indicating that prompt design can be effective, whereas complex reasoning strategies require more careful implementation.

The system ranked 6th in the shared task, outperforming baseline models by 3%. Overall, the findings highlight the effectiveness of medium-sized models with optimized prompts and emphasize the importance of clearer task definitions within prompts to improve classification accuracy.

2 Related Work

Critical questions generation is an emerging task at the intersection of natural language generation and argumentation theory, aimed at producing questions that challenge the reasoning, assumptions, or evidence in argumentative texts. The task is

*Equal contribution.

[1] https://hitz-zentroa.github.io/shared-task-critical-questions-generation/

grounded in Walton's argumentation schemes (Walton et al., 2008), which define common structures of arguments and the critical questions used to evaluate them. These theoretical structures were used by Figueras and Agerri (2025) to generate reference critical questions for the task. While effective in producing relevant questions, this method was limited in flexibility and coverage. To complement the theory-based generation, Calvo Figueras and Agerri (2024) also explored the use of two large language models (LLMs), LLaMA-2 and Zephyr (Touvron et al., 2023; Tunstall et al., 2023), to generate critical questions in zero-shot settings. The outputs were then manually reviewed for validity. The results showed that while current LLMs can generate fluent and well-formed questions, they often struggle to produce questions that are truly critical and grounded in the argument. Only 28% of the generated questions were found to be valid, mainly due to issues with relevance, generality, and reasoning.

Beyond argument analysis, several studies examined how critical questions generation can support fact-checking and misinformation detection. For example, Ousidhoum et al. (2022) proposed generating multiple targeted questions from a single claim, each addressing a specific factual aspect such as source credibility, timelines, or implications. Similarly, Setty and Setty (2024) experimented using sequence-to-sequence generative models and LLMs to automate questions generation for fact-checking applications. The results showed improvements in evidence retrieval and verification performance, suggesting that critical questions generation can enhance the effectiveness of claim verification systems. Augenstein et al. (2024) discuss the potential threat of hallucinations and the generation of misinformation when using LLMs for fact-checking. Critical questions generation mitigates this threat by prompting models to question existing claims rather than produce factual knowledge, reducing the risk of hallucinations.

These studies highlight the growing importance of critical questions generation. Yet, current LLM performance and limited resources leave ample room for improvement, especially in generating valid, argument-specific questions.

3 Task Description

In this section, we describe the task goal, data, and the evaluation of system outputs.

Dataset The dataset used for this task is derived from real-world debates, where each data point represents a single speaker's intervention. Interventions are labeled with argumentation schemes following the taxonomy of Walton et al. (2008). In addition to the scheme label, each entry includes a unique identifier and a set of associated critical questions. These questions are labeled for their usefulness in challenging the underlying argument of the intervention. The critical questions are categorized into three labels:

- **Useful**: The question is directly relevant and can effectively challenge an argument in the text.
- **Unhelpful**: The question is reasonable but unlikely to challenge arguments in the text.
- **Invalid**: The question cannot be used to challenge any argument in the text. This may be due to flawed reasoning, lack of relevance, the introduction of unrelated concepts, excessive generality, or a lack of critical focus.

The task includes a validation set and a test set, with 186 and 34 interventions, respectively. The data is provided in JSON format, and can be accessed through the task repository on GitHub.

Task Definition The goal of the task is to generate three critical questions for a given argumentative intervention. These questions should challenge or examine the argument more deeply. The questions can point out missing assumptions, ask for more evidence, or raise possible counterpoints. The main goal is to generate questions that would be considered *Useful* based on the labels in the dataset.

Evaluation The system has to generate three critical questions for each intervention. These questions are evaluated based on their usefulness in challenging the argument in the intervention text. Each *Useful* question gets 0.33 points, while *Unhelpful* and *Invalid* questions get 0 points. The sum of these scores for an intervention is referred to as its *punctuation*. The final system score is calculated as the average punctuation across all interventions in the test set. To evaluate the usefulness of generated questions, each question is compared to a set of reference questions using semantic similarity. The generated question is matched to the most similar reference, and if the similarity score exceeds a threshold of 0.65, it is assigned the label of that reference question. If the score falls below 0.65, the question is labeled as *Not Able to Evaluate*.

Task	Setting	Small	Medium	Large
Gen	Zero-shot	61.5	67.8	66.3
	Few-shot	67.2	66.7	**68.5**
Cls	Zero-shot	58.2	**65.8**	62.4
	Few-shot	60.9	64.4	59.2

Table 1: Validation results for generation (Gen) and classification (Cls) modules. Generation scores use overall punctuation with a similarity threshold of 0.6. Classification scores are binary accuracy.

4 Methodology

We decompose the task of critical questions generation into two subtasks: question generation and question classification. Accordingly, our pipeline is structured into two main modules. The first module takes an argumentative text as input and generates ten critical questions related to it. The second module then classifies these questions into one of three categories: *Useful*, *Unhelpful*, or *Invalid*. The questions are sorted by their usefulness, and the top three questions are selected as the final output.

For each module, we evaluated three models from the LLaMA 3.1 Instruct family: the small (8B), medium (70B), and large (405B) variants. These models were tested across two prompting techniques: zero-shot and few-shot, resulting in six experimental conditions per module. For few-shot prompting, the generation module was provided with three example interventions, each accompanied by one useful critical question. The classification module was given one example intervention with three critical questions, each representing a different category. The validation results for each configuration are presented in Table 1.

The optimal settings for the test set were achieved using few-shot prompting with the large model for question generation and zero-shot prompting with the medium-sized model for classification. The inference parameters were kept consistent across all conditions, with a temperature of 0.5, a maximum generation length of 1024 tokens, and the top_p parameter set to 0.9.

We experimented with two alternative classification strategies: debate and deliberation. These methods redefine the task as a binary classification, where the goal is for multiple models to determine whether each of the ten questions is useful or not. *Debate Classification:* In this approach, two LLMs engage in a traditional debate format. In the opening statement, each model presents the questions it considers useful, along with justifications. Disagreements are addressed during the rebuttal round. The debate concludes with closing statements from both models. Thereafter, a third model, acting as a judge, determines the winner. The final output comprises the questions deemed useful by the winner of the debate. *Deliberation Classification:* This approach involves three LLMs which engage in up to three rounds of deliberation to identify useful questions. In each round, the models can propose a classification, justify their choices, critique others' proposals and collaborate to reach consensus. After the first round, the participants vote on which questions they consider useful. If they unanimously agree on three questions, the deliberation ends. If no agreement is reached, a second round of discussion follows, which ends with a majority vote. If disagreement persists, a third and final round is initiated, after which a judge selects the most useful questions based on the entire deliberation.

The complete set of prompts used in the experiments is provided in Appendix A.

5 Results

Our system ranked 6th out of thirteen participating teams, demonstrating a modest improvement of three percentage points over the baseline scores. The test set results are presented in Table 2. As the question generation module remained consistent across all three submissions, any variation in performance can be attributed solely to differences in the classification modules. The best-performing classifier was LLaMA 3.1 70B Instruct with the zero-shot method, closely followed by the debate-based classification approach. In contrast, the deliberation-based classification yielded significantly lower performance, as a substantial number of questions were labeled as *Not Able to Evaluate*.

The generation module produced ten questions for each of the 34 debate interventions, resulting in 340 generated questions. Out of these, only 180 were included in at least one of the three official submissions. This subset received gold labels during the official evaluation and thus serves as the basis for our assessment of the quality of the generation module. Within this subset, 44.4% of the questions were labeled as *Useful*, 22.8% as *Unhelpful*, and 15% as *Invalid*, while the remaining 17.8% were unable to be evaluated. Assuming a perfect classifier operating on this subset, the maximum

Method	Useful	Unhelpful	Invalid	Not able to evaluate	Score
Zero-shot (Manual)	57	28	16	0	55.9
Zero-shot	55	25	12	9	53.9
Debate	53	26	13	10	52.0
Deliberation	38	22	16	26	37.3

Table 2: Test set results. Scores represent the ratio of achieved punctuation to the maximum possible punctuation. "Manual" indicates that the scoring involved manual evaluation.

achievable punctuation score would be 73.5.

On average, the best classification module labeled 8.4 out of 10 questions as *Useful*, while the debate-based module classified 7.6 out of 10 questions as *Useful*. The debate-based and zero-shot prompt-based modules showed strong alignment, agreeing on 2.6 out of every 3 questions. In contrast, agreement between the deliberation-based approach and the prompt-based or debate-based methods was substantially lower, with agreement scores of 0.9 and 1, respectively. The agreement across all three classifiers was 0.8.

6 Discussion & Analysis

The generation module employed few-shot prompting with the large 405B parameter model. Competitive scores were also achieved by zero-shot prompting the medium-sized model (70B) and few-shot prompting the small model (8B). In particular, scaling from 8B to 405B, a 50-fold increase in model size, resulted in only a 1.3 percent point increase in the overall punctuation score. Such a relatively small gain may not justify the substantial increase in computational cost. Relatedly, no clear benefits of model scaling were observed in the classification module, where the 70B model outperformed both the 8B and 405B models. Interestingly, the small model seemed to benefit from in-context examples in the few-shot setting, while the larger models performed better under zero-shot prompting. This suggests that the smaller model, with less internal world knowledge, gains more from external context than its larger counterparts.

All classifier modules show a strong bias toward labeling questions as *Useful*: 75–85% of generated questions were classified as such, though only 44% were actually validated as useful. This suggests that classifiers often assess surface-level relevance to argumentative text rather than true criticality, struggling to distinguish genuinely critical questions from those merely contextually related. The high number of questions labeled *Unhelpful* supports this. Possible remedies include enriching prompts with more discriminative examples and providing clearer definitions to distinguish the two categories.

Both the debate and deliberation approaches failed to outperform the zero-shot prompting. We initially hypothesized that the structured discussion would guide the model's reasoning and improve overall performance. If anything, the debate and deliberation formats simply allowed the models to generate more tokens, which by itself could potentially lead to better results. However, our error analysis showed two areas for improvement. First, prompt complexity posed a challenge: the models occasionally lost track of their position within the debate or deliberation and failed to consider their opponent's responses. Second, the models attempted to discuss all ten questions simultaneously, preventing them from engaging with the arguments beyond a surface level. Both limitations may be addressed by improvements to the current implementation, e.g. by structuring discussions around a single intervention and question and refining the prompts to enhance flow awareness. With these adjustments, the underlying approaches still hold potential for improving classification performance.

7 Conclusion

This paper presents our two-stage system for critical questions generation, developed for the shared task using LLaMA 3.1 Instruct models. The system ranked 6$^{\text{th}}$, outperforming the baseline. Key challenges include classifier bias toward labeling questions as *Useful* and limited benefits from scaling or complex reasoning. Future work will refine prompt and interaction design to support robust debate and deliberation, including prompts based on argumentation schemes and improved focus on addressing each question individually.

Acknowledgments

This work was partially supported by the AKASE third-party project under the OpenWebSearch.eu project. The OpenWebSearch.eu project is funded by the EU under Grant Agreement No. 101070014, and we thank the EU for their support.

References

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renée DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, Dirk Hovy, Heng Ji, Filippo Menczer, Rafael Miguez, Preslav Nakov, Dietram Scheufele, Sapna Sharma, and Giorgio Zagni. 2024. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nature Machine Intelligence*, 6(8):852–863.

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 (ArgMining 2025)*, Vienna, Austria. Association for Computational Linguistics.

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

Nedjma Ousidhoum, Zhangdie Yuan, and Andreas Vlachos. 2022. Varifocal question generation for fact-checking. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2532–2544, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ritvik Setty and Vinay Setty. 2024. Questgen: Effectiveness of question generation methods for fact-checking applications. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 4036–4040. ACM.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of lm alignment. *Preprint*, arXiv:2310.16944.

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

A Prompts

Model-specific tokens have been replaced with simpler tags to improve readability. The few-shot tags indicate the additional context with regards to the zero-shot prompt. Placeholder values are marked using double curly brackets.

A.1 Question Generation Prompt

<system>You are a critical thinker who asks critical questions to evaluate the validity and soundness of arguments. Critical questions are designed to uncover the underlying assumptions and scrutinize the reasoning process of an argument. Read the given paragraph carefully and generate a set of critical questions that challenge the premises, assumptions, and inferences made in the argument. Respond with the critical questions directly, without any preamble or explanation. Ensure your questions are probing and insightful to thoroughly assess the argument's acceptability or potential fallacies. Answer as accurately and critically as possible!</system>

<fewshot>
<user>TRUMP: "A lot of people even within my own party want to give certain rights to people on watch lists and no-fly lists. I agree with you. When a person is on a watch list or a no-fly list. These are very, very good people. They're protecting the Second Amendment. I have the endorsement of the NRA. I'm very proud of. We have to look very strongly at no-fly lists and watch lists. When people are on there, even if they shouldn't be on there, we'll help them, we'll help them legally, we'll help them get off. I tend to agree with that quite strongly. You were the one that brought up the words super-predator about young black youth. That's a term that I think was a—it's—it's been horribly met. You've apologized for it."</user>

<assistant>What are the potential consequences of removing people from watch lists and no-fly

lists without proper vetting, and how would Trump mitigate these risks?</assistant>

<user>HOLT: "We're talking about the burden that Americans have to pay, yet you have not released your tax returns. The reason nominees have released their returns for decades is so that voters will know if their potential president owes money to—who he owes it to and any business conflicts. Don't Americans have a right to know if there are any conflicts of interest?"</user>

<assistant>How would releasing tax returns guarantee that voters would be aware of all potential conflicts of interest, considering that tax returns may not reveal all relevant information?</assistant>

<user>CLINTON: "In fact, his cavalier attitude about nuclear weapons is so deeply troubling. That is the number-one threat we face in the world. It becomes particularly threatening if terrorists ever get their hands on any nuclear material."</user>

<assistant>What specific actions or policies has Clinton's opponent proposed or supported that demonstrate a "cavalier" attitude towards nuclear weapons?</assistant>
</fewshot>

<user>Generate critical questions for this text: {{intervention}}. For the output, print only the top 10 questions generated with each question on a separate line without numbering.</user>

A.2 Question Classification Prompt

<system>You are a critical thinker whose task is to evaluate and classify critical questions. Read the given critical questions carefully and classify them into three categories: useful, unhelpful, or invalid. Follow these definitions:

1. Useful question: One should not take the arguments in the text as valid without having reflected on this question.

2. Unhelpful question: The question makes sense, is related to the text, and in some exceptional cases, the answer could invalidate or diminish the validity of the arguments in the text, but this is unlikely. This may be in cases where:
   (a) The answer to the question is common sense.
   (b) The answer to the question is a well-known fact that does not generate controversy.
   (c) The question is very complicated to understand, and it would be impractical to question the arguments.
   (d) The question is answered in the text itself.

3. Invalid question: A question is invalid when the answer to this question cannot serve to invalidate or diminish the acceptability of the arguments in the text. This can be for several reasons:
   (a) Unrelated: The question is unrelated to the text.
   (b) New concept: The question introduces new concepts that were not in the text.
   (c) Bad reasoning: The question does not challenge any argument defended in the text. For example, when the question challenges the opposite position to the one defended in the text.
   (d) Very general: The question is very vague and does not ask about anything specific in the text. This question could be asked of any argument.
   (e) Non-critical: Although the question asks about something in the text, it is not critical of any argument. For example, when the question is a reading-comprehension one. A question is only critical if the answer to the question can potentially reduce the validity of the argument.

Provide only the predicted labels in the format of a valid Python list of strings, without any preamble or explanation.</system>

<fewshot>
<user>TRUMP: "A lot of people even within my own party want to give certain rights to people on watch lists and no-fly lists. I agree with you. When a person is on a watch list or a no-fly list. These are very, very good people. They're protecting the Second Amendment. I have the endorsement of the NRA. I'm very proud of. We have to look very strongly at no-fly lists and watch lists. When people are on there, even if they shouldn't be on

there, we'll help them, we'll help them legally, we'll help them get off. I tend to agree with that quite strongly. You were the one that brought up the words super-predator about young black youth. That's a term that I think was a—it's—it's been horribly met. You've apologized for it." Questions:

- How does Trump's stance on watch lists and gun control align with his broader views on national security and individual rights?
- What are the potential consequences of removing people from watch lists and no-fly lists without proper vetting, and how would Trump mitigate these risks?
- What are the potential consequences of restricting gun ownership based on watch lists or no-fly lists, and are they justified by the potential benefits?

</user>

<assistant>["Invalid", "Useful", "Unhelpful"]</assistant>
</fewshot>

<user>Classify the following critical questions: {{questions}}</user>

A.3 Debate Prompts

Each round uses a different user prompt but they all share the same system prompt, as defined in A.3.1.

A.3.1 System Prompt

General Instructions
You are an expert debater tasked with critically analyzing a set of questions related to an argument. Your role is to determine whether each question is Useful, Unhelpful, or Invalid for evaluating the validity and acceptability of the argument.

Definitions

1. Useful question: A question that must be reflected upon, as failing to consider it could lead to accepting a potentially fallacious argument.
2. Unhelpful question: A question that is related to the argument but unlikely to invalidate or diminish its validity, often because:
   (a) The answer is common sense or a well-known fact.
   (b) The question is overly complicated or impractical.
   (c) The question is already answered in the argument text.
3. Invalid question: A question that cannot serve to invalidate or diminish the acceptability of the argument, due to reasons such as:
   (a) Being unrelated to the argument.
   (b) Introducing new, unmentioned concepts.
   (c) Exhibiting faulty reasoning or challenging the opposite position.
   (d) Being too vague or general.
   (e) Being a simple reading comprehension question rather than a critical one.

A.3.2 Opening Statement Prompt

Task
You will be provided with:

1. An argument text or context.
2. A set of 10 critical questions related to the argument.
3. (Optional) An opening statement from an opposing debater.

Your task is to analyze each of the 10 critical questions and provide an opening statement arguing whether each question is Useful, Unhelpful, or Invalid for evaluating the argument's validity and acceptability.

Argument and critical questions (and optionally, the opposing opening statement):
{{context}}

Now, provide your opening statement analyzing each critical question as Useful, Unhelpful, or Invalid, and provide arguments for your decisions.

A.3.3 Rebuttal Prompt

Task: Provide a concise rebuttal to the opponent's arguments in the given debate context.

Instructions:

1. Read the debate context carefully to understand the arguments made by both sides.
2. Identify weaknesses, logical fallacies, or inconsistencies in the opponent's case.

3. Formulate a rebuttal that directly responds to and challenges the opponent's previously stated points.

4. Do not introduce any new arguments or claims. Focus solely on critiquing the opponent's existing arguments.

5. Keep the rebuttal concise and to-the-point, without unnecessary elaboration.

Context:
{{context}}

Rebuttal:

A.3.4 Closing Statement Prompt

Task: Summarize the given debate context and provide a concise closing statement reinforcing your main arguments.

Context:
{{context}}

Closing Statement Guidelines:

1. Read and understand the full debate context provided above.

2. Identify the key arguments and evidence presented in support of your position.

3. Summarize those main points concisely in 1-2 sentences.

4. Reinforce why your arguments are stronger and more persuasive than the opposing side.

5. The closing statement should be a brief but impactful conclusion aimed at convincing the audience/judges that you should win the debate.

A.3.5 Judge Prompt

Instruction
You are a judge in a traditional 1v1 debate. Your role is to critically assess proposals on the classification of ten critical questions by evaluating the arguments and rebuttals to ensure robust decision-making. Follow these steps:

1. Parse the input discussion to identify premises, conclusions, and argument structure.

2. Map the argument to one of Walton's argumentation schemes (e.g., expert opinion, analogy, cause-effect).

3. For the identified scheme, apply its specific critical questions (e.g., "Is the expert credible?" for expert opinion scheme).

4. Highlight any missing premises, weak evidence, or fallacies in the argument.

5. Provide a structured critique of the argument (e.g., "This argument assumes X but lacks evidence for Y").

6. Determine the winning side based on logic, evidence and persuasion.

Context
{{context}}

Output Format
Provide your response as a Python list containing the numbers of the sentences that, according to the winner, are Useful, without any preamble or additional information.

A.4 Deliberation Prompts

A.4.1 General Prompt

<system>### Instruction
You are an expert deliberator tasked with critically analyzing a set of questions related to an argument. Your role is to determine, together with another deliberator, whether each question is Useful, Unhelpful, or Invalid for evaluating the validity and acceptability of the argument.

Definitions

1. Useful question: A question that must be reflected upon, as failing to consider it could lead to accepting a potentially fallacious argument.

2. Unhelpful question: A question that is related to the argument but unlikely to invalidate or diminish its validity, often because:

   (a) The answer is common sense or a well-known fact.
   (b) The question is overly complicated or impractical.
   (c) The question is already answered in the argument text.

3. Invalid question: A question that cannot serve to invalidate or diminish the acceptability of the argument, due to reasons such as:

(a) Being unrelated to the argument.

(b) Introducing new, unmentioned concepts.

(c) Exhibiting faulty reasoning or challenging the opposite position.

(d) Being too vague or general.

(e) Being a simple reading comprehension question rather than a critical one.

</system>

<user>### Context
{{context}}

Actions
<propose> Generate clear and concise proposals aligned with the core objectives of the deliberation. Present your proposals in a well-structured way. </propose>

<argue> Build arguments to support your proposals that are grounded in the definitions of the types of questions. Ensure your arguments are logical, well-structured, and clear. </argue>

<counter> Address critiques from other deliberators by acknowledging weaknesses, updating proposals, or offering compromises. Respond respectfully and constructively, demonstrating openness to refinement and collaboration. </counter>

<collaborate> Engage with critiques from other agents, stress-test ideas, and work towards aligning priorities. Actively participate in the discussion, considering different perspectives and fostering a shared understanding. </collaborate>

Guidelines

- Engage directly with critiques from the other Deliberator (e.g., "To address your concern about X, we could...").
- Prioritize brevity: Avoid repetition and focus on key trade-offs and innovations.
- Signal resolution or deadlock clearly.

Provide your response immediately without any preamble or additional information:</user>

A.4.2 Label Extraction Prompt

Instruction
Analyze the provided deliberation between two deliberators who aimed to determine the usefulness, unhelpfulness, or invalidity of questions. Your task is to identify the final labels assigned by each deliberator to each question after their discussion.

Deliberation
{{deliberation}}

Output Format
Provide your response as two valid Python dictionaries, one for each deliberator, with the following structure: ["1": "[label]", "2": "[label]", ..., "10": "[label]"] Replace [label] with one of the following values for each question number from 1 to 10:

- "Useful"
- "Unhelpful"
- "Invalid"

Do not include any additional text or explanation. Return only the two Python dictionaries separated by a comma, without any preamble or additional information.