

# Benchmarking Critical Questions Generation: A Challenging Reasoning Task for Large Language Models

**Blanca Calvo Figueras**  
HiTZ Center - Ixa  
University of the Basque  
Country UPV/EHU  
blanca.calvo@ehu.eus

**Rodrigo Agerri**  
HiTZ Center - Ixa  
University of the Basque  
Country UPV/EHU  
rodrigo.agerri@ehu.eus

## Abstract

The task of Critical Questions Generation (CQs-Gen) aims to foster critical thinking by enabling systems to generate questions that expose underlying assumptions and challenge the validity of argumentative reasoning structures. Despite growing interest in this area, progress has been hindered by the lack of suitable datasets and automatic evaluation standards. This paper presents a comprehensive approach to support the development and benchmarking of systems for this task. We construct the first large-scale dataset including 5K manually annotated questions. We also investigate automatic evaluation methods and propose reference-based techniques as the strategy that best correlates with human judgments. Our zero-shot evaluation of 11 LLMs establishes a strong baseline while showcasing the difficulty of the task. Data and code<sup>1</sup> plus a public leaderboard are provided to encourage further research, not only in terms of model performance, but also to explore the practical benefits of CQs-Gen for both automated reasoning and human critical thinking.

## 1 Introduction

Recent research has documented significant pedagogical concerns regarding the proliferation of Artificial Intelligence conversational interfaces. Specifically, it suggests that these systems might be reinforcing superficial learning processes while simultaneously diminishing users' capacities for critical thinking (Hadi Mogavi et al., 2024).

In this context, researchers have proposed Critical Questions Generation (CQs-Gen) as a method to leverage LLMs to automatically generate critical questions that systematically expose evidential weaknesses or structural flaws (such as fallacies) embedded within argumentative discourse. The

<sup>1</sup>[https://github.com/hitz-zentroa/Benchmarking\\_CQs-Gen](https://github.com/hitz-zentroa/Benchmarking_CQs-Gen)

---

### Argument:

*"I did not use sunscreen yesterday and I was fine, so I don't think you need it today."*

### Critical Questions:

- *Is the weather today going to be similar to the one yesterday? And our sun-exposure time? Are your skin and mine similar?*
- *Has there been any other day similar to today in which I did happen to need sunscreen?*

(a) Argument from analogy

---

### Argument:

*"Dr. Smith says sunscreen is not necessary in May, so you don't need it today."*

### Critical Questions:

- *Is Dr. Smith an expert in skin care? Do other experts in skin care agree with Dr. Smith? Is Dr. Smith a trustworthy source? Might Dr. Smith be biased?*
- *What were the literal words of Dr. Smith? Can his words be checked?*
- *Is his claim consistent with the known evidence about the effects of sun in May?*

(b) Argument from expert opinion

---

Figure 1: Examples of two arguments and the critical questions they raise. While these are synthetic examples, our dataset contains naturally-occurring arguments.

ultimate goal is to promote deeper analytical engagement by developing systems capable of generating meaningful critical questions with respect to argumentative texts (Calvo Figueras and Agerri, 2024).

**Critical Questions (CQs) are inquiries that may be posed to assess the acceptability of an argument. Therefore, the answer to these questions could potentially challenge the strength of the argument.** These questions are closely tied

to the specific argumentation schemes underlying the construction of a given discourse (Walton et al., 2008). As illustrated in Figure 1, an argument can be built based on different schemes, such as *argument from analogy* or *argument from expert opinion*, each triggering a different set of critical questions. Thus, the task of CQs-Gen is formulated as a generative task, in which systems are expected to produce questions that expose the underlying assumptions in the arguments’ premises and challenge their inferential structure.

Although interest in these types of tasks is growing (Musi et al., 2023; Favero et al., 2024; Ruiz-Dolz and Lawrence, 2023), the lack of datasets and standardized evaluation methods hinders the advancement of research for this particular task. In this work, we introduce the required infrastructure to address these gaps. First, we present the first large-scale dataset including 5K manually annotated critical questions. The dataset, collected from multiple sources, includes several critical questions for each text, which are labeled according to a set of validated annotation guidelines. Second, we explore various evaluation strategies for the task, measuring their correlation to human judgments. Finally, we provide a benchmarking by experimenting with 11 LLMs in zero-shot settings, analyzing the diversity of the generated questions, and providing a public leaderboard to encourage further research on CQs-Gen. The main contributions of this work are the following:

- The first manually annotated dataset for the task of Critical Questions Generation (CQs-Gen).
- An extensive investigation on the best automatic evaluation methods for this task, establishing that reference-based methods correlate best with human judgments.
- A first benchmarking of the task with 11 state-of-the-art LLMs, which showcases substantial room for improvement, accompanied by a public leaderboard to encourage further research on this particular task.<sup>2</sup>

In the rest of the paper, we first present relevant previous work (Section 2), explain the dataset creation (Section 3), the CQs-Gen task (Section 4), and the experimental settings on evaluation and

generation (Section 5). We then report our experimental findings on evaluation (Section 6.1), and use the best evaluation strategies to benchmark 11 LLMs (Section 6.2). We finish by analyzing the results in terms of diversity and proposing some future work.

## 2 Previous Work

The automatic generation of critical questions is a relatively new task, introduced and motivated by Calvo Figueras and Agerri (2024). Their work demonstrated that theoretical critical questions, such as those proposed by Walton et al. (2008), and questions generated by LLMs are complementary, as they tend to target different aspects of argumentation. This is in line with the observation that CQs can not be restricted to a prefix set (Hernández, 2023). Calvo Figueras and Agerri (2024) also identified common failure modes in LLM-generated critical questions, including: (a) producing overly generic or unrelated questions, (b) introducing concepts not present in the source text, (c) flawed reasoning, and (d) generating non-critical questions, such as those resembling reading comprehension tasks. Parallel to this work, a shared task on Critical Questions Generation has been organized to further stimulate research in this area (Calvo Figueras et al., 2025).

CQs-Gen differs from other question-generation tasks (Pan et al., 2020; Miao et al., 2024) in that its goal is not to produce questions directly answerable from the input text (i.e., reading-comprehension questions). Instead, it aims to elicit questions that uncover what remains unsaid, such as hidden premises or implicit connections.

Moreover, previous work has shown that CQs are useful for fighting misinformation, since they help users identify fallacious reasoning (Musi et al., 2023), and also for predicting the grades of argumentative essays, since they reveal the quality of the arguments (Song et al., 2014). The potential of using CQs in computational applications has been discussed at length (Reed and Walton, 2001; Macagno et al., 2017; Ruiz-Dolz and Lawrence, 2023).

One of the big challenges of generative tasks is evaluation, and CQs-Gen is no different. Thus, machine translation researchers have long been investigating how to use human-generated references to evaluate new text generations using metrics such as BLEURT (Sellam et al., 2020), chrF (Popović,

<sup>2</sup>[https://huggingface.co/spaces/HiTZ/Critical\\_Questions\\_Leaderboard](https://huggingface.co/spaces/HiTZ/Critical_Questions_Leaderboard)

2015), and COMET (Bosselut et al., 2019). Similarity metrics such as Semantic Text Similarity (STS) (Reimers and Gurevych, 2019) have also been used to compare the new generations to the reference outputs (Aynedinov and Akbik, 2024). Finally, recent work has shown the effectiveness of using LLMs to evaluate various downstream tasks (Zhong et al., 2022; Ke et al., 2022; Jones et al., 2024; Zubiaga et al., 2024), either by using general-purpose models (Wang et al., 2023), or specialized ones (Kim et al., 2024; Zhu et al., 2023). In this work, we examine all these evaluation strategies to investigate which one is the most suitable for evaluating the CQs-Gen task.

### 3 Dataset Creation

In order to benchmark the capacity of current LLMs to generate critical questions, a first step is to create a manually annotated dataset. For this purpose, relevant argumentative texts have been collected, and reference questions have been generated and annotated. The rest of the section provides details of each of these steps.

#### 3.1 Data Collection

To guarantee the relevance of the texts in our dataset, we gather them using four previously available corpora used for argument mining: US2016 (Visser et al., 2021), Moral Maze Debates (MMD) (Lawrence et al., 2018), US2016reddit, and Regulation Room Divisiveness (RRD) (Konat et al., 2016). All these corpora had been annotated with Argumentation Schemes in IAT format (Budzynska and Reed, 2011). We reformat these texts by intervention and make the data more manageable by splitting long interventions and merging very short ones. We remove the interventions with no argumentation scheme associated, since these do not necessarily contain any arguments.<sup>3</sup>

The data collection process results in a dataset of 220 naturally-occurring interventions, with an average length of 738.4 characters, and 3.1 argumentation schemes. The topics of the interventions range from politics to airline policies and the economy.

<sup>3</sup>The code to process the IAT diagrams can be found here: [https://github.com/hitz-zentroa/critical\\_questions\\_generation/tree/main/scripts/pre-process](https://github.com/hitz-zentroa/critical_questions_generation/tree/main/scripts/pre-process)

#### 3.2 References Generation

To generate reference CQs for the dataset, we use the method proposed by Calvo Figueras and Agerri (2024). First, we create critical questions based on the templates in Walton et al. (2008). Second, we prompt *Llama-3-70B-Instruct* to write CQs (see Appendix A for the prompts used). As a result of this process, we obtain an average of 22.4 reference questions per intervention.

#### 3.3 Annotation

The guidelines for annotating the reference questions have been improved in various iterations with the help of the annotators until we reached a moderate inter-annotator agreement (IAA). The annotators were journalists and fact-checkers specialized in detecting misinformation. The annotation process starts with the following question: **Can this question be used to undermine the arguments given in the intervention?**

If the annotators consider that a question is not useful, they can choose between two options: the question not being valid, or the question being unhelpful. This distinction makes the task more consistent, as the annotators have to be able to describe the reasons why the question is not useful. The three categories are described in the guidelines as follows:<sup>4</sup>

1. **Useful (USE):** The answer to this question can potentially challenge one of the arguments in the text.
2. **Unhelpful (UN):** The question is valid, but it is unlikely to challenge any of the arguments in the text.
3. **Invalid (IN):** This question is invalid because it cannot be used to challenge any of the arguments in the text. Either because (1) its reasoning is not right, (2) the question is not related to the text, (3) it introduces new concepts not present in the intervention, (4) it is too general and could be applied to any text, or (5) it is not critical with any argument of the text (e.g. a reading-comprehension question).

Additionally, in order to label a question as Unhelpful, the annotators have to provide a short open

<sup>4</sup>[https://github.com/hitz-zentroa/shared-task-critical-questions-generation/tree/main/shared\\_task/utills](https://github.com/hitz-zentroa/shared-task-critical-questions-generation/tree/main/shared_task/utills)

message stating their reasons. To consider a question Invalid, annotators have to select one of the 5 criteria specified above. This reason-based annotation procedure allowed the annotators to increase their IAA (calculated using Cohen Kappa (Cohen, 1960)) from 0.19 and 0.26 in the first and second pre-annotation rounds (in which we refined the definitions), to 0.54 in the third and final round (in which the reason-based annotation was implemented). In each round, both annotators evaluated 250 questions.

After this, the entire dataset was annotated. The stats of the annotation per data source can be found in Table 1. It can be observed that Useful questions are the most common type of questions. This is especially true in the dataset about airline policies (RRD).

Origin	N° Int.	N° CQs	% USE	% UN	% IN
US2016	98	2,555	59.88	23.25	16.87
Moral Maze	25	584	53.77	20.72	25.51
RRD	83	1,597	66.12	23.04	10.83
US2016reddit	14	240	54.58	30.0	15.42
<b>TOTAL</b>	<b>220</b>	<b>4,976</b>	<b>60.91</b>	<b>23.21</b>	<b>15.88</b>

Table 1: Stats of the dataset per source of origin.

### 3.4 Splitting the dataset

Hereafter, we split our dataset between a validation set (publicly available), and a test set. The reference questions of the test set are kept unpublished, since sharing them could result in data contamination that would compromise future evaluations using this dataset (Sainz et al., 2023). However, we allow future system evaluations through a public leaderboard.

We keep the interventions with the most balanced labels in the test set while putting the rest in the validation set. This splitting procedure increases the quality of the test set, and also makes the reference-based evaluations more accurate (see "Reference-based metrics" in Section 5.1). As observed in Table 2, while 67.46% of the reference questions in the validation set are Useful, this decreases to 42.68% in the test set.

Set	N° Int.	N° CQs	% USE	% UN	% IN
Validation	186	4,136	67.46	21.59	10.95
Test	34	806	42.68	31.02	26.30

Table 2: Stats of the dataset per set.

## 4 The Task

We formalize the task of CQs-Gen as a generative task where, when given an input argumentative text, the system has to output exactly 3 critical questions. The 3 questions are then evaluated regarding their usefulness for critically assessing the arguments of the text. For the generation of one useful critical question, the task is scored 1/3, for two 2/3, and if the 3 questions are useful, the task is scored with 1, the maximum punctuation.

In Figure 2, an instance of the task with two output examples is provided. For this particular intervention, a question such as "Does this argument support Socialist policies?" is evaluated as Invalid, as it introduces a new concept which is not present in the text: "Socialist policies". The question "How does the speaker define "the developed world", and is this a relevant distinction in this context?" gets evaluated as Unhelpful, as this definition is not central to the arguments of the text. Instead, a similar question such as "How is "sufficient surpluses" defined, and how would one measure it?" gets labeled as Useful, as different considerations of what sufficient surpluses are could diminish the strength of the argument.

## 5 Experimental Settings

A major challenge of generative tasks is evaluation. Therefore, before benchmarking different models to generate critical questions, we perform extensive research to find the best evaluation methods. For this purpose, we develop two baseline systems, evaluate their output manually, and assess which evaluation method correlates better with human judgments. The best evaluation methods are then applied to establish the first benchmark with 11 state-of-the-art LLMs.

### 5.1 Evaluation Methods

We first perform human evaluation and then assess several automatic approaches: (1) *reference-based metrics*, which compare the newly generated questions to the reference questions, (2) *reference-based LLMs*, which applies LLMs to compare newly generated questions with reference questions, and (3), *labeling LLMs*, which directs LLMs to label the critical questions based on the annotation guidelines. For all these evaluation experiments, we use the output of two baseline systems: *Llama-3-70B-Instruct* and *Qwen2.5-VL-72B-Instruct*, with a very simple prompt (see in Annex B) in a zero-shot set-



Walton: *Claire's absolutely right about that. But then the problem is that that form of capitalism wasn't generating sufficient surpluses. And so therefore where did the money flow. It didn't flow into those industrial activities, because in the developed world that wasn't making enough money.*

(a) **Input:** the intervention

**USE:** What evidence is there to support the claim that the form of capitalism being used in the developed world was not generating sufficient surpluses?

**USE:** How is "sufficient surpluses" defined, and how would one measure it?

**USE:** Are there any alternative explanations for why the money did not flow into industrial activities?

**IN:** Does this argument support Socialist policies?

**UN:** How does the speaker define "the developed world", and is this a relevant distinction in this context?

**USE:** What are the "industrial activities" being referred to, and how do they relate to the form of capitalism in question?

(b) **Output:** Given that all CQs here are useful, this answer has an overall punctuation of 1.

(c) **Output:** This set of questions would get 1/3 points for the useful CQ, 0 for the CQ that is unhelpful, and 0 for the invalid one. Therefore, this answer has a 1/3 punctuation.

Figure 2: Example of candidate outputs with its labels: Useful (**USE**), Unhelpful (**UN**), and Invalid (**IN**).

ting. In each step of the evaluation, we aggregate the results of these two baselines, as our goal in this section is not to test the systems but to test the evaluation methods.

**Human evaluation** As a first step, we perform human evaluation of the zero-shot output of the two baseline systems. This manual evaluation allows us to compute statistical correlation with respect to the automatic evaluation methods. The manual judgment consists of visualizing each new question next to the reference questions of that intervention, and selecting the one that inquires about the same information. Then, the new question inherits the label from the reference. If no matching reference is found, we label that question manually following the annotation guidelines (either as Useful, Unhelpful, or Invalid). We validate this evaluation with double annotation and obtain an IAA of 59.14.

**Reference-based metrics** We use Semantic Text Similarity (STS) with the Sentence Transformers model *stsb-mpnet-base-v2* (Reimers and Gurevych, 2019) to compare each of the newly generated questions to the reference questions in the dataset. Given a threshold, we find the most similar reference question to the new one. If no reference question reaches the threshold, the generated question is given the label *not\_able\_to\_evaluate* (NAE). Considering  $\{R\}$  as the set of vectors of the reference questions,  $N$  the vector of the newly generated question, and  $T$  the threshold, the label is computed as:

$$f(N) = \begin{cases} \text{Argmax}_j \cos(R_j, N) & \text{if } \max_j \cos(R_j, N) > T \\ \text{NAE} & \text{else} \end{cases}$$

We also experiment with machine translation metrics using this same logic. We use the HuggingFace implementation of BLEURT (Sellam et al., 2020), chrF (Popović, 2015), and COMET (Bosse-lut et al., 2019).<sup>5</sup> For all metrics, the threshold is chosen by comparing the IAA of each threshold with the human evaluation, and the number of values that the threshold left unevaluated (see this comparison in Annex C).

**Reference-based LLMs** As in the previous method, we compare the generated questions to each of the reference questions, and pick the most similar one. However, in this case we apply LLMs to do the matching by querying the models whether both questions ask for the same information. The model can also predict that no reference is similar enough, in other words, the NAE label. The evaluation prompt can be found in Figure 8 of Annex D. As our evaluators, we use *Claude 3.5 Sonnet*,<sup>6</sup> a state-of-the-art proprietary closed-weights model, and *Gemma 3 12B Instruct*, an open-weights model.

**Labeling with LLMs** To further explore the potential of LLMs for evaluating the task of CQs-Gen, we now ask the LLM to be the one deciding the evaluation label following the guidelines from Section 3.3. The prompt used for this task for both Claude and Gemma3 is detailed in Figure 9 of Annex D. In addition, we also include *Prometheus*

<sup>5</sup><https://huggingface.co/evaluate-metric>

<sup>6</sup>Version *claude-3-5-sonnet-20241022*, <https://www.anthropic.com/news/claude-3-5-sonnet>

2 7B, an open source LLM specialized in evaluating LLMs (Kim et al., 2024). For Prometheus, we adapt the evaluation prompt to reproduce the format it was trained with (see Figure 10 in Annex D). Since Prometheus’ instruction includes the requirement of a feedback message explaining why a certain label is given, we also use this feedback to better understand the model’s behavior when performing this evaluation (see Annex E).<sup>7</sup>

**Augmenting the references.** In a concurrent shared task on CQs-Gen that used this same dataset (Calvo Figueras et al., 2025), a manual evaluation of 15 different submissions was performed. Conveniently, we can now use this data to increase the number of reference questions. In this shared task, each submission generated 3 questions for each intervention in the test set (that is, 45 additional reference questions per intervention). Using these 1,530 (45 \* 34) new reference questions, we investigate whether increasing the number of references in the test set improves the results in the reference-based evaluations by incorporating them in batches of 5 per intervention.

## 5.2 Generation Methods

For benchmarking, we opt for a slightly more elaborate prompt, which includes some additional guidelines (see Annex F), and apply it to 8 state-of-the-art open-weights models: *Llama-3-8B-Instruct*, *DeepSeek-R1-Distill-Llama-8B*, *Gemma-2-9b-it*, *Qwen2.5 VL-7B-Instruct*, *Llama-3-70B-Instruct*, *DeepSeek-R1-Distill-Llama-70B*, *Gemma-2-27b-it*, and *Qwen2.5-VL-72B-Instruct*, (Grattafiori et al., 2024; DeepSeek-AI et al., 2025; Qwen et al., 2025; Team et al., 2024); as well as 3 top-performing proprietary models: *Claude 3.5 Sonnet*,<sup>8</sup> *GPT-04-mini*, and *GPT-4o* (OpenAI et al., 2024).<sup>9</sup> Every model is used in their default settings.<sup>10</sup>

To evaluate the output of these models, we use the best-performing evaluation methods resulting from the experiments in Section 6.1. That is:

<sup>7</sup>We did try changes in the phrasing and complexity of all these prompts (such as asking just for useful vs not-useful questions). We also tried making the LLM choose a reason for its label (as in the annotation process), adding examples of each type of question as in a few-shot approach, and fine-tuning the LLMs for our specific task. However, these changes barely affected the results and, therefore, we do not include them in the paper.

<sup>8</sup>Version *claude-3-5-sonnet-20241022*, <https://www.anthropic.com/news/claude-3-5-sonnet>

<sup>9</sup>Versions *gpt-4o-2024-08-06* and *o4-mini-2025-04-16*

<sup>10</sup>For Gemma2 models, we set the temperature to 0.6, since keeping it at 0 generated the same question 3 times.

reference-based STS, reference-based Claude, and reference-based Gemma3. In all cases, we use the test set with the augmented reference questions.

## 6 Results

We first present the results on the various automatic evaluation methods tested, and then we report the performance of the 11 LLMs on the CQs-Gen task.

### 6.1 Results on Evaluation Methods

We compute the results of both of our baseline models (*Llama-3-70B-Instruct* and *Qwen2.5-VL-72B-Instruct*) with each of the evaluation methods defined in Section 5. We then compute the agreement between all of them using Cohen Kappa scores (see Figure 3). We differentiate between the manual evaluation obtained only by matching the newly generated questions with the reference ones and keeping the NAE values (*manual\_reference* in Figure 3, highlighted in blue), and the setting in which the human evaluator has replaced the NAE values with one of the 3 labels (*manual\_labeling* in Figure 3, highlighted in green).

Regarding reference-based methods (both metrics and LLMs), *claude\_reference* obtained the best results, achieving an IAA of 0.57 (moderate agreement) with respect to the *manual\_reference*. *STS\_0.65\_reference* and *gemma3\_reference* also obtain a moderate IAA, with a score of 0.40 and 0.44, respectively (see blue rectangle in Figure 3). However, when attempting to evaluate the questions by labeling them directly using LLMs, no evaluation method achieves an IAA higher than 0.1.<sup>11</sup> Therefore, from this point onward, we focus only on reference-based evaluation methods. Nonetheless, we provide an error analysis of Prometheus in Annex E, which highlights the difficulties for LLMs to perform this evaluation without relying on gold reference CQs.

While most reference-based methods have a moderate agreement with the *manual\_reference* evaluation, this is not the case when comparing them to the *manual\_labeling* evaluation, where NAE values have been replaced by the right label manually. In this evaluation, STS, Claude, and Gemma3 achieve an IAA of 0.20, 0.28 and 0.31, respectively (see green rectangle in Figure 3). These results show that, with the current test set,

<sup>11</sup>Note that in footnote <sup>7</sup> we explain that we tried many other variations of these experiments that did not result in any improvement of the IAA score.

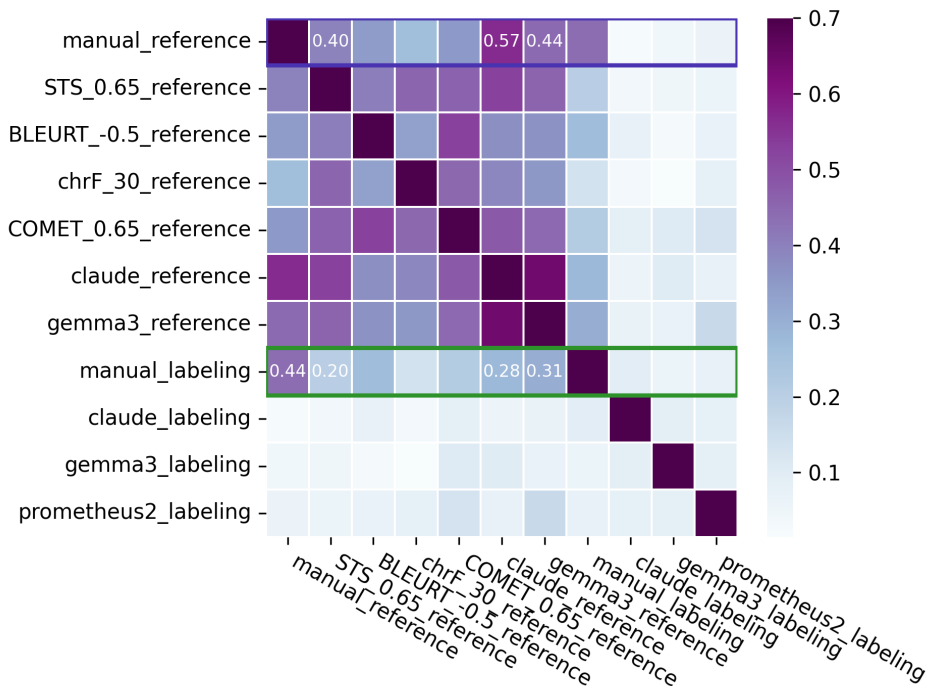


Figure 3: Cohen Kappa scores between all the evaluation methods. These are averaged scores of the two models we manually evaluated: *Llama-3-70B-Instruct* and *Qwen2.5-VL-72B-Instruct*. The numbers in some of the metric names indicate the used threshold.

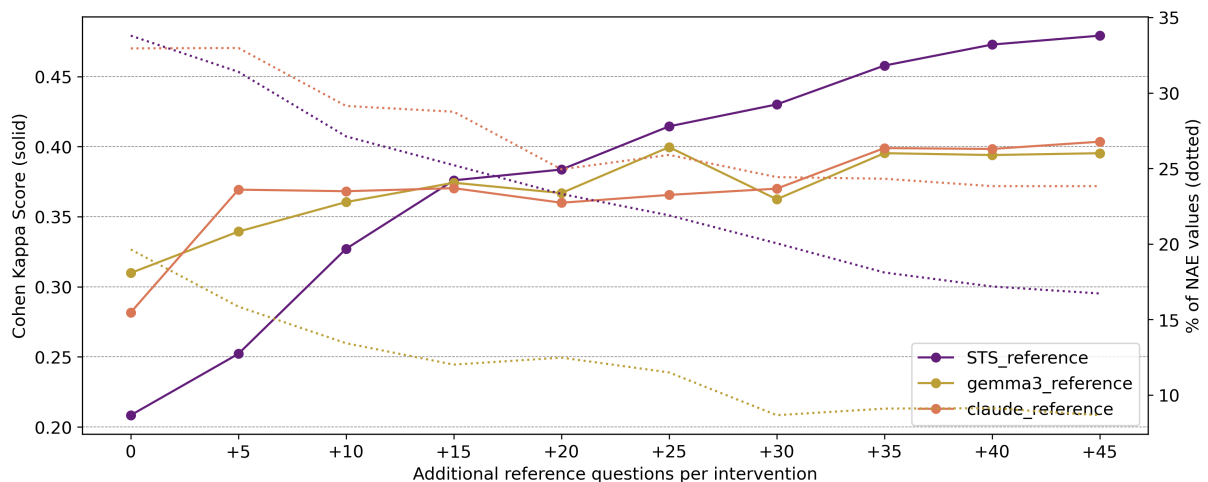


Figure 4: Cohen Kappa Scores between our best automatic evaluation methods and the *manual\_labeling* evaluation when increasing the reference questions in batches of 5 (solid lines). Reported together with the % of NAE values (dotted lines). These are averaged scores of the two models we manually evaluated: *Llama-3-70B-Instruct* and *Qwen2.5-VL-72B-Instruct*.

the reference-based evaluation methods are limited by a lack of reference questions.

**Extending the test set:** To address the scarcity of reference questions, we extend our test set by incorporating the 15 manually evaluated submissions from the CQs-Gen Shared Task as additional references. That is 45 extra reference questions per intervention (which amounts to 68.64 per interven-

tion on average). In Figure 4, we incorporate these new references in batches of 5, and calculate the resulting IAA and percentage of NAE values. As observed, the inclusion of new references leads to a higher IAA. For *STS\_reference*, this improvement is steady as more references are added, reaching an IAA of 0.48 when all 45 questions are added. In contrast, for LLM-based reference methods, the

method →	Score				NAE %		
	STS_ref.	claude_ref.	gemma3_ref.	Average ↓	STS	claude	gemma3
gemma-2-27b	53.27 ± 5.7	<u>51.63</u> ± 3.4	57.19 ± 2.3	<b>54.03</b>	11.67	16.33	3.33
Llama-3-70B	53.89 ± 2.6	48.51 ± 5.1	<u>58.15</u> ± 6.6	53.52	4.0	14.0	1.67
Qwen2.5-72B	<b>57.73</b> ± 1.9	47.15 ± 1.5	53.25 ± 2.1	52.71	7.0	15.67	1.0
gemma-2-9b	52.94 ± 3.9	47.71 ± 3.4	55.56 ± 1.5	52.07	9.33	21.67	4.0
DeepSeek-Llama-70B	47.82 ± 4.5	48.73 ± 1.1	57.32 ± 2.7	51.29	16.33	10.67	1.0
Llama-3-8B	50.24 ± 3.9	42.34 ± 4.2	55.12 ± 3.9	49.23	11.33	21.33	3.0
DeepSeek-Llama-8B	44.88 ± 3.9	36.18 ± 1.8	50.73 ± 5.5	43.93	16.0	27.67	4.33
Qwen2.5-7B	43.30 ± 1.4	43.94 ± 1.3	42.97 ± 1.2	43.40	6.33	12.0	1.33
claude-3-5-sonnet	56.21 ± 0.6	50.33 ± 1.5	<b>62.42</b> ± 1.5	<b>56.32</b>	9.33	16.0	3.33
gpt-4o	52.29 ± 1.5	<b>52.94</b> ± 2.9	58.17 ± 3.4	54.47	13.0	14.33	1.0
o4-mini	53.27 ± 6.0	50.33 ± 4.9	58.82 ± 2.6	54.14	10.0	12.0	1.33

Table 3: Generation results obtained with the prompt in Annex F. In the reference-based evaluation, the *extended\_test* was used, and NAE values were considered not-useful. We report average of 3 runs and standard deviation. **Bold**: best overall results per metric; underlined: best open-weights results per metric. All models are instruct models.

gains level off quickly when reaching an IAA of 0.40. This divergence arises from methodological factors: while STS evaluates each reference question independently, LLM-based methods integrate all reference questions into the prompt, causing the context length to increase with each addition. However, since the quality of the evaluation does not decrease either, we will be using all the references in the upcoming evaluations. It should also be noted that *claude\_reference* is more strict than *gemma3\_reference* when evaluating, leaving a higher percentage of questions unevaluated.

In order to test if these three automatic evaluation methods are suitable for comparing CQs-Gen systems, we now use this method to re-evaluate all the submissions of the shared task. For a fair evaluation, we remove the references belonging to the submission being evaluated at each time. With this experiment, we observe that the ranking generated by Claude and Gemma correlates to the humanly evaluated ranking of the shared task with a Person’s Coefficient of 0.86 and 0.80, respectively. For *STS\_reference*, the correlation is 0.69.

Therefore, each of these three evaluation methods presents distinct strengths and limitations. The *claude\_reference* approach achieves the highest correlation with the shared task ranking, yet it leaves a substantial number of instances unevaluated. In contrast, *gemma3\_reference* minimizes the number of unevaluated cases, but it exhibits the lowest IAA with human annotations. Finally, *STS\_reference* achieves the strongest IAA with human annotations, but it performs the weakest in terms of model comparison.

Considering these results, our benchmarking

will report these three metrics alongside the percentage of NAE values, which serves as a proxy for the confidence of the evaluation. For the test set, we will use all the references available (from now on, *extended\_test*).

## 6.2 Results on CQs-Gen

Table 3 presents the benchmarking results of 11 state-of-the-art LLMs (generation prompt in Annex F), evaluated using our three strongest evaluation methods. The results indicate that, in a zero-shot setting, no model generates more than 57% of Useful CQs on average. It is important to note, however, that no parameter optimization or prompt tuning has been applied in this evaluation. Consequently, higher scores are reported in Calvo Figueras et al. (2025), where systems specifically designed for this task achieved results nearly 10 points higher.

For the open-weight models, the best-performing LLM is *gemma-2-27b-it*, which achieves an average score of 54.03, slightly higher than *Llama-3-70B-it*, and comparable to the closed-weight proprietary models. We also observe consistent, though modest, improvements when using larger variants within each model family. The highest overall performance is achieved by *claude-3-5-sonnet*.

However, the different evaluation methods differ in their judgments, with *claude\_reference* making *gpt-4o* the winner, *gemma3\_reference* making *claude-3-5-sonnet*, and *STS\_reference* making *Qwen2.5-VL-72B*. These differences should be further studied. In all cases, *gemma3\_reference* has the lowest percentage of NAE values, and *claude\_reference* the highest.



## 7 Qualitative Analysis

While the primary goal of this task is to generate useful critical questions, the diversity of these questions also serves as an important indicator of quality. A broader range of questions may reflect a deeper comprehension of the texts, while a minimal-effort approach characterized by the repeated use of identical or formulaic templates may suggest a superficial understanding and limited capacity for critical engagement.

To investigate this issue, we use two metrics, namely, n-gram diversity score and compression ratio diversity score (CR-div) (Reimers and Gurevych, 2019). The n-gram metric calculates diversity as the ratio of the number of unique n-grams to the total n-grams occurring in the entire set of generated questions. The compression ratio (CR) comes from calculating the ratio of the size of the compressed file to its original size. The CR diversity score can be calculated as the reciprocal of the compression ratio to get a score between 0 and 1, 0 being for identical documents. Shaib et al. (2025) shows that this metric is the most reliable one.

Table 4 shows that both Qwen2.5 models produce the least diverse output, closely followed by *claude-3-5-sonnet*. These results also hold when we look only at Useful CQs (third column of Table 4). The most diverse output is produced by *o4-mini*. Taking a closer look at the output of Qwen2.5-VL models, we observe that half of the CQs generated by these models follow one of these templates:

metric →	n-gram	CR-div ↓	CR-d(USE)
DeepSeek-70B	2.71	<b>.313</b>	.339
DeepSeek-8B	2.65	.309	.335
Llama-3-70B	2.52	.293	.335
Llama-3-8B	2.41	.289	.323
gemma-2-9b	2.45	.287	.304
gemma-2-27b	2.45	.282	.300
Qwen2.5-72B	1.04	<u>.131</u>	.146
Qwen2.5-7B	0.97	<u>.123</u>	.140
o4-mini	2.76	<b>.332</b>	.351
gpt-4o	2.77	.313	.336
claude-3-5-sonnet	1.36	<u>.167</u>	.183

Table 4: Diversity metrics calculated within the CQs generated by models in Table 3. *CR-d(USE)* shows the diversity of Useful CQs. **Bold**: best overall results per metric; underlined: worst overall results per metric.

- How does {speaker} address {related\_matter}?

- What evidence does {speaker} provide to support the claim that {claim}?

In the same line, *claude-3-5-sonnet* also outputs a recursive template regarding evidence, mainly:

- What evidence supports the claim that {claim}?

However, both the 70B Qwen model and Claude perform substantially better at selecting claims to populate the template than the smaller Qwen model, leading to a higher proportion of Useful critical questions (see Table 3). As highlighted in the CQs-Gen shared task, diversity is a relevant metric for this task, as it serves as an important indicator of the model’s level of understanding, which can not be observed from the overall score alone.

## 8 Conclusion and Future Work

In this work, we have introduced a comprehensive infrastructure for the development and evaluation of systems addressing the task of Critical Questions Generation. First, we constructed a dataset compiled from four different sources, annotated with critical questions categorized as Useful, Unhelpful, or Invalid. Second, we explored automatic evaluation methods for this task and found that reference-based approaches yield the most reliable results. Finally, we conducted zero-shot evaluations across 11 LLMs, revealing substantial room for improvement. To support future development, we also provide a public leaderboard for benchmarking CQs-Gen systems.

Although the dataset includes a validation and a test set, our results are based solely on the test set, as we did not pursue parameter or prompt tuning to improve model performance. Consequently, exploring methods for improving model performance represents a natural direction for future work.

In addition, we aim to investigate the practical utility of generating critical questions for both automated systems and human users. For automated systems, a promising direction is to assess whether incorporating critical questions as intermediate steps in complex reasoning tasks can enhance overall performance. For human users, it could be explored whether exposure to (automatically generated) critical questions could foster improved critical thinking skills.

## Limitations

Despite the contributions of this work, several limitations remain. First, our dataset, while of reasonable size, is certainly extendable, although this is hindered by the scarcity of datasets annotated with argumentation schemes, particularly beyond English. Thus, expansion using only LLM-generated critical questions remains a viable path even though the annotation should be performed by experts, which makes it an expensive effort. Second, although we have reduced the number of unevaluated questions in the test set, this shortcoming remains an open research question.

## Acknowledgments

We are thankful to the following MCIN/AEI/10.13039/501100011033 projects: (i) DeepKnowledge (PID2021-127777OB-C21) and by FEDER, EU; (ii) Disargue (TED2021-130810B-C21) and European Union NextGenerationEU/PRTR; (iii) DeepMinor (CNS2023-144375) and European Union NextGenerationEU/PRTR. Blanca Calvo Figueras is supported by the UPV/EHU PIF22/84 predoc grant. We would like to acknowledge the contributions of Celia Ramos and Pablo Hernández from Maldita.es for annotating the dataset and giving feedback on the guidelines, as well as Maite Heredia, for performing the secondary annotations.

## References

- Ansar Aynetdinov and Alan Akbik. 2024. [Sem-score: Automated evaluation of instruction-tuned llms based on semantic textual similarity](#). *Preprint*, arXiv:2401.17072.
- Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. [COMET: Commonsense transformers for automatic knowledge graph construction](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.
- Katarzyna Budzynska and Chris Reed. 2011. Whence inference. *University of Dundee Technical Report*.
- 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 Vilalta, and Rodrigo Agerri. 2025. [Overview of the critical questions generation shared task 2025](#). In *Proceedings of the 12th Workshop on Argument Mining*.
- DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. [Deepseek-v3 technical report](#). *Preprint*, arXiv:2412.19437.
- Lucile Favero, Juan Antonio Pérez-Ortiz, Tanja Käser, and Nuria Oliver. 2024. [Enhancing critical thinking in education by means of a socratic chatbot](#). *Preprint*, arXiv:2409.05511.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Reza Hadi Mogavi, Chao Deng, Justin Juho Kim, Pengyuan Zhou, Young D. Kwon, Ahmed Hosny Saleh Metwally, Ahmed Tlili, Simone Bassanelli, Antonio Bucchiarone, Sujit Gujar, Lennart E. Nacke, and Pan Hui. 2024. [ChatGPT in education: A blessing or a curse? A qualitative study exploring early adopters’ utilization and perceptions](#). *Computers in Human Behavior: Artificial Humans*, 2(1):100027.
- Alfonso Hernández. 2023. [Disentangling Critical Questions From Argument Schemes](#). *Argumentation*, 37(3):377–395. Publisher: Springer Verlag.
- Jaylen Jones, Lingbo Mo, Eric Fosler-Lussier, and Huan Sun. 2024. [A multi-aspect framework for counter narrative evaluation using large language models](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 147–168.
- Pei Ke, Hao Zhou, Yankai Lin, Peng Li, Jie Zhou, Xiaoyan Zhu, and Minlie Huang. 2022. [CtrlEval: An unsupervised reference-free metric for evaluating controlled text generation](#). In *ACL*.
- Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. [Prometheus 2: An Open Source Language Model Specialized in Evaluating Other Language Models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4334–4353, Miami, Florida, USA. Association for Computational Linguistics.

- Barbara Konat, John Lawrence, Joonsuk Park, Katarzyna Budzynska, and Chris Reed. 2016. A Corpus of Argument Networks: Using Graph Properties to Analyse Divisive Issues. In *10th conference on International Language Resources and Evaluation (LREC'16)*, pages 3899–3906.
- John Lawrence, Jacky Visser, and Chris Reed. 2018. BBC Moral Maze: Test Your Argument. In *Comma*.
- Fabrizio Macagno, Douglas Walton, and Chris Reed. 2017. Argumentation Schemes. History, Classifications, and Computational Applications.
- Yisong Miao, Hongfu Liu, Wenqiang Lei, Nancy Chen, and Min-Yen Kan. 2024. Discursive socratic questioning: Evaluating the faithfulness of language models' understanding of discourse relations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6277–6295, Bangkok, Thailand. Association for Computational Linguistics.
- Elena Musi, Elinor Carmi, Chris Reed, Simeon Yates, and Kay O'Halloran. 2023. Developing Misinformation Immunity: How to Reason-Check Fallacious News in a Human-Computer Interaction Environment. *Social Media + Society*, 9(1):20563051221150407. Publisher: SAGE Publications Ltd.
- OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.
- Liangming Pan, Yuxi Xie, Yansong Feng, Tat-Seng Chua, and Min-Yen Kan. 2020. Semantic graphs for generating deep questions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1463–1475, Online. Association for Computational Linguistics.
- Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.
- Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.
- Chris Reed and Douglas Walton. 2001. Applications of Argumentation Schemes.
- 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.
- Ramon Ruiz-Dolz and John Lawrence. 2023. Detecting Argumentative Fallacies in the Wild: Problems and Limitations of Large Language Models. In *Proceedings of the 10th Workshop on Argument Mining*, pages 1–10, Singapore. Association for Computational Linguistics.
- Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics.
- Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.
- Chantal Shaib, Joe Barrow, Jiuding Sun, Alexa F. Siu, Byron C. Wallace, and Ani Nenkova. 2025. Standardizing the measurement of text diversity: A tool and a comparative analysis of scores. *Preprint*, arXiv:2403.00553.
- Yi Song, Michael Heilman, Beata Beigman Klebanov, and Paul Deane. 2014. Applying Argumentation Schemes for Essay Scoring. In *Proceedings of the First Workshop on Argumentation Mining*, pages 69–78, Baltimore, Maryland. Association for Computational Linguistics.
- Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraj, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.
- Jacky Visser, John Lawrence, Chris Reed, Jean Wagemans, and Douglas Walton. 2021. Annotating Argument Schemes. *Argumentation*, 35(1):101–139.
- Douglas Walton, Christopher Reed, and Fabrizio Macagno. 2008. *Argumentation Schemes*. Cambridge University Press, Cambridge.
- Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is ChatGPT a good NLG evaluator? a preliminary study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11, Singapore. Association for Computational Linguistics.
- Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and

- Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038.
- Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2023. JudgeLM: Fine-tuned Large Language Models are Scalable Judges. *arXiv preprint*. ArXiv:2310.17631 [cs].
- Iruñe Zubiaga, Aitor Soroa, and Rodrigo Agerri. 2024. A LLM-based ranking method for the evaluation of automatic counter-narrative generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9572–9585, Miami, Florida, USA. Association for Computational Linguistics.



## A Reference Generation Prompts

List the critical questions that should be asked regarding the arguments in the following paragraph:  
{intervention}

Suggest which critical questions should be raised before accepting the arguments in this text:  
{intervention}

Figure 5: Prompt for generating the reference questions.

## B Baselines Prompt for Evaluation Experiments

Give me 3 or more critical questions that should be raised before accepting the arguments in this text:  
{intervention}

Give one question per line. Make sure there are at least 3 questions. Do not give any explanation regarding why the question is relevant.

Figure 6: Prompt for generating the output of the baselines.

## C Threshold selection for reference-based metrics

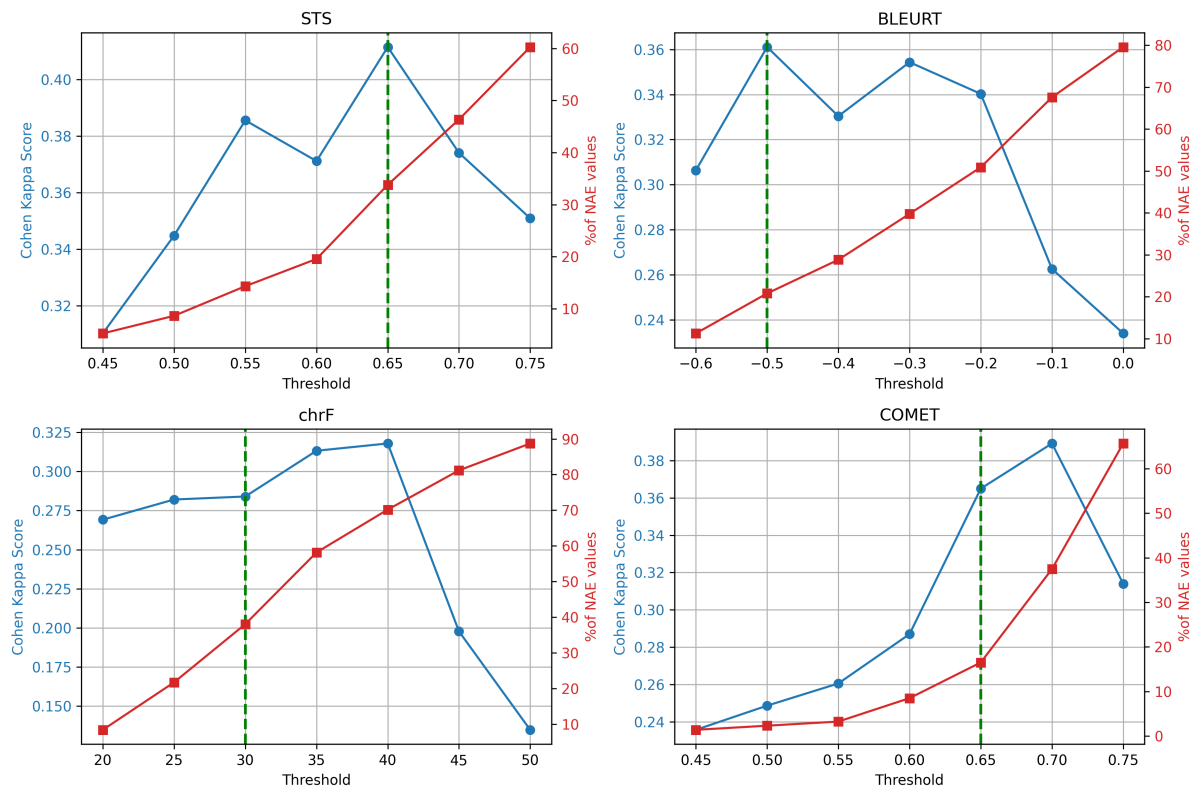


Figure 7: The threshold was selected to give the best balance between the Cohen Kappa Score (blue dots), and the % of NAE values it generated (red squares). The chosen threshold for each metric is highlighted in green.

## D Evaluation Prompts

You will be given a set of reference questions, each with an identifying ID, and a newly generated question. Your task is to determine if any of the reference questions are asking for the same information as the new question.

Here is the set of reference questions with their IDs:

```
<reference_questions>
```

```
{references}
```

```
</reference_questions>
```

Here is the newly generated question:

```
<new_question>
```

```
{cq}
```

```
</new_question>
```

Compare the new question to each of the reference questions. Look for questions that are asking for the same information, even if they are worded differently. Consider the core meaning and intent of each question, not just the exact wording.

If you find a reference question that is asking for the same information as the new question, output only the ID of that reference question.

If none of the reference questions are asking for the same information as the new question, output exactly 'Similar reference not found.' (without quotes).

Your final output should consist of only one of the following:

1. The ID of the most similar reference question
2. The exact phrase 'Similar reference not found.'

Do not include any explanation, reasoning, or additional text in your output.

Figure 8: Prompt for comparing the newly generated questions to the reference questions.

You are a fair judge assistant tasked with evaluating if a provided question is a useful critical question for a given text. Your role is to provide clear objective feedback based on specific criteria, ensuring each assessment reflects the absolute standards set for performance.

Here is the question you should evaluate:

```
<critical_question>
{cq}
</critical_question>
```

And here is the text to which the question relates:

```
<text>
{intervention}
</text>
```

Guidelines for evaluation:

- Carefully read both the question and the text.
- Consider how the question relates to the arguments presented in the text.
- Assess the question's usefulness in challenging or critically examining the text's content.
- Determine which of the three labels (Useful, Unhelpful, or Invalid) best applies to the question.

Label criteria:

- Useful: The question is both critical of and directly relevant to the arguments in the text. It challenges the text's content in a meaningful way.
- Unhelpful: The question is critical and related to the text, but not likely to be very useful in challenging its arguments. This could be because:
  - The answer is common sense
  - The answer is well-known and not controversial
  - The question is very complicated to understand or answer
  - The text itself already answers the question

Note: Do not use this label just because better questions could have been posed.

- Invalid: The question is not appropriately critical in this context. This could be because:
  - The question is unrelated to the text
  - The question is too general and could apply to many texts
  - The question introduces new concepts not mentioned in the text
  - The question doesn't challenge any arguments in the text (e.g., it's a simple reading comprehension question or asks about the speaker's/reader's opinion)
  - The question critiques an argument that the speaker wasn't actually making

Your task is to output only one of the three labels: Useful, Unhelpful, or Invalid. Do not include any comments, explanations, blank spaces, or new lines. Your entire output should consist of a single word - the chosen label.

Figure 9: Prompt for directly labeling the newly generated questions using Claude and Gemma3.

You are a fair judge assistant tasked with providing clear, objective feedback based on specific criteria, ensuring each assessment reflects the absolute standards set for performance.

###Task Description:  
An instruction (might include an Input inside it), a response to evaluate, some reference answers for score 1, 2 and 3, and a score rubric representing a evaluation criteria are given.

- Write a detailed feedback that assess the quality of the response strictly based on the given score rubric, not evaluating in general.
- After writing a feedback, write a score that is an integer between 1 and 3. You should refer to the score rubric.
- The output format should look as follows: "Feedback: (write a feedback for criteria) [RESULT] (an integer number between 1 and 3)"
- Please do not generate any other opening, closing, and explanations.

###The instruction to evaluate:

- Carefully read both the question and the text.
- Consider how the question relates to the arguments presented in the text.
- Assess the question's usefulness in challenging or critically examining the text's content.
- Determine which of the three labels (Useful, Unhelpful, or Invalid) best applies to the question.

Here is the question you should evaluate:

```
<critical_question>
{cq}
</critical_question>
```

And here is the text to which the question relates:

```
<text>
{intervention}
</text>
```

###Score Rubrics:  
[Is the question posed by the model critical and useful for challenging the arguments in the text?]

Score 1: Invalid-> The question is not appropriately critical in this context. This could be because:

- The question is unrelated to the text
- The question is too general and could apply to many texts
- The question introduces new concepts not mentioned in the text
- The question doesn't challenge any arguments in the text (e.g., it's a simple reading comprehension question or asks about the speaker's/reader's opinion)
- The question critiques an argument that the speaker wasn't actually making

Score 2: Unhelpful-> The question is critical and related to the text, but not likely to be very useful in challenging its arguments. This could be because:

- The answer is common sense
- The answer is well-known and not controversial
- The question is very complicated to understand or answer
- The text itself already answers the question

Note: Do not use this label just because better questions could have been posed.

Score 3: Useful -> The question is both critical of and directly relevant to the arguments in the text. It challenges the text's content in a meaningful way.

###Feedback:

Figure 10: Prompt for directly labeling the newly generated questions using Prometheus 2 7B.

## E Analysis of Prometheus' evaluation

Using the prompt in Figure 10 and Prometheus 2 7B (Kim et al., 2024), we tried to evaluate both baselines (*Llama-3-70b* and *Qwen-2.5-72b*). The results in Figure 3 show a very low IAA with respect to human evaluation. To further investigate this issue we inspect the errors, observing that 25% of the questions get labeled as Useful when they should have been labeled as Invalid, 23.5% of the questions get labeled as Useful when the annotators had labeled them as Unhelpful, and 16.1% are judged as Unhelpful instead of Useful.

In the table below, we show some examples of these 3 error directions. In many instances, if the generation model had misunderstood the text when generating the question (bad reasoning error), the judge model just followed along with the model's interpretation (as in the first example below). In other cases, the judge model tried too hard to find a reason for the question to be critical, as in example 2. In other instances, the judge model failed to interpret the implicit relations that the speaker was drawing (as in example 4). In the third column below, you can find the explanation of each of these errors.

Intervention, Question and Gold Label	Judge Feedback and Label	Error in the Labeling
<p><b>Intervention:</b> TRUMP: "I did not support the war in Iraq That is a mainstream media nonsense put out by her The best person in her campaign is mainstream media"</p> <p><b>Question:</b> Can you identify examples of how the "best person in her campaign" has been influenced by mainstream media?</p> <p><b>Gold Label:</b> Invalid</p>	<p><b>Feedback:</b> In the provided text, statement by statement, it can be seen that the question asked about examples of the candidate being influenced by mainstream media. <b>The text, however, directly states that mainstream media is a "nonsense" put out by the best person in the campaign.</b> This statement, which is clearly within the context of the text, directly addresses the question posed. Therefore, it does challenge the arguments presented in the text and adds an informative layer about the speaker's influence by mainstream media. [...]</p> <p><b>Judge Label:</b> Useful</p>	<p>In the text, Trump is saying that mainstream media have been on the side of his opponent. Basically, he is claiming that media are biased towards her and are disseminating lies about him. However, the question does not understand his point, instead, it assumes it is someone in particular from her campaign the one who has been "influenced by mainstream media". The judge falls into the same trap.</p>
<p><b>Intervention:</b> HOLT: "How do you heal the divide The share of Americans who say race relations are bad in this country is the highest it's been in decades much of it amplified by shootings of African-Americans by police we've seen recently in Charlotte and Tulsa Race has been a big issue in this campaign one of you is going to have to bridge a very wide and bitter gap"</p> <p><b>Question:</b> Are there historical precedents for successfully bridging such divides, and if so, what were they?</p> <p><b>Gold Label:</b> Invalid</p>	<p><b>Feedback:</b> Based on the score rubric, the provided question aligns with the criteria of applying critical analysis to the arguments presented in the text. The question requires a thoughtful consideration of historical precedents that might have been successful in "healing the divide" that the speaker alludes to. <b>It directly challenges the implicit assumption in the text that it's necessary to seek a solution or to address the issue of race relations.</b> The question also refrains from introducing new or unrelated concepts, which is a criterion specified in the rubric. [...]</p> <p><b>Judge Label:</b> Useful</p>	<p>The judge model claims that the existence of "historical precedents" challenges the idea that there is a need for solutions. However, this is not true, as solutions might be needed even if other solutions have been implemented in the past. The annotator labeled the question as invalid because it does not challenge any of the arguments made by the speaker. On the contrary, the question could have come from Holt himself.</p>
<p><b>Intervention:</b> HOLT: "How do you heal the divide The share of Americans who say race relations are bad in this country is the highest it's been in decades much of it amplified by shootings of African-Americans by police we've seen recently in Charlotte and Tulsa Race has been a big issue in this campaign one of you is going to have to bridge a very wide and bitter gap"</p> <p><b>Question:</b> Is the statement about the share of Americans who say race relations are bad supported by data?</p> <p><b>Gold Label:</b> Useful</p>	<p><b>Feedback:</b> The response to the question shows an understanding of the query, as it assesses the level of data support for the statement made in the text. However, the response doesn't dive deep into how the question relates to the arguments presented in the text, and it fails to offer a critique of these arguments. This lack of critical analysis, therefore, limits its usefulness for challenging the text's content. Moreover, <b>the question posed is somewhat vague and doesn't necessarily probe into all aspects of the statement.</b> There's room for improvement in demonstrating the criticality of the question and in examining its utility for challenging the arguments within the text. [...]</p> <p><b>Judge Label:</b> Unhelpful</p>	<p>The main argument of this text is that racial relations are getting worse in America. To argue that, he uses several premises, and one of these premises is that the share of the Americans who say so, has gone up. Therefore, not having data to sustain this claim would diminish the validity of his argument. The judge dismisses the question saying that it does not "probe into all aspect of the argument". However, this is not a requirement in the rubric.</p>



Intervention, Question and Gold Label	Judge Feedback and Label	Error in the Labeling
<p><b>Intervention:</b> CLINTON: "actually, I have thought about this quite a bit not quite that long my husband did a pretty good job in the 1990s I think a lot about what worked and how we can make it work again a balanced budget, million new jobs Incomes went up for everybody, Manufacturing jobs went up also in the 1990s, if we're actually going to look at the facts [...]"</p> <p><b>Question:</b> How do we know that Clinton's husband's policies were the cause of the economic growth in the 1990s?</p> <p><b>Gold Label:</b> Useful</p>	<p><b>Feedback:</b> The question seems to assume a causal relationship between the policies of Bill Clinton and the economic growth of the 1990s, which is not explicitly stated in the text and could introduce ambiguity. It asks for evidence to support a statement that is not clearly stated as part of the text's content. This could potentially lead to misunderstanding or misinterpretation of the text's actual stance. Additionally, the question appears to critique an argument that the speaker was not actually making, which could be considered invalid in terms of critical analysis. It does, however, stay within the confines of the context, only dealing with policies and economic growth of the 1990s. The question, although critical, lacks the useful edge that it should have to truly challenge the text's arguments. This gives it a certain level of unhelpfulness. [...]</p> <p><b>Judge Label:</b> Unhelpful</p>	<p>While the relation between Bill Clinton's policies and the economic growth in the 1990s might not be strictly stated in the text, the annotators interpreted that Hillary is implying that this relation exists as, otherwise, her mentioning these two facts together would be pointless. Therefore, if we interpret Hillary is drawing this relation, it makes a lot of sense to ask if it is actually a correct one.</p>
<p><b>Intervention:</b> CLINTON: "that is just not accurate I was against it once it was finally negotiated and the terms were laid out"</p> <p><b>Question:</b> Who negotiated the terms?</p> <p><b>Gold Label:</b> Unhelpful</p>	<p><b>Feedback:</b> The question "Who negotiated the terms?" is directly relevant to the text provided. It asks for specific information that is critical in understanding the context and negotiation dynamics, which were clearly stated by the speaker, Clinton. The question doesn't divert into new concepts but centers on the question of agency - in this case, the role of the negotiator - which is a fundamental aspect of negotiations. This aligns with the goal of challenging the content of the text, as specified in the rubric. [...]</p> <p><b>Judge Label:</b> Useful</p>	<p>While who negotiated the terms could potentially decrease the validity of the claim (for instance, in case it was actually her who did it), it is unlikely to do so.</p>
<p><b>Intervention:</b> CLINTON: "I support our democracy sometimes you win, sometimes you lose I certainly will support the outcome of this election Donald's trying very hard to plant doubts about it I hope the people out there understand this election's really up to you It's not about us so much as it is about you and your families and the kind of country and future you want I sure hope you will get out and vote as though your future depended on it it does"</p> <p><b>Question:</b> Is Clinton's appeal to people to vote based on their own interests rather than the common good?</p> <p><b>Gold Label:</b> Unhelpful</p>	<p><b>Feedback:</b> The critical question posed here directly engages with the arguments provided in the text. It focuses on a specific aspect of the argument – the emphasis on voting based on personal interests versus the common good. While it is critical and directly related to the arguments made by Clinton, it lacks depth and could be potentially well-known to some readers, thus not entirely challenging. The question, though pertinent, may fall short of providing substantial insight or critiquing the text's content in a more nuanced manner. However, it still manages to be useful in challenging the arguments presented by the speaker, as it specifically targets the key issue within the text. Therefore, according to the score rubric, this question is useful in challenging the arguments in the text.</p> <p><b>Judge Label:</b> Useful</p>	<p>Clinton's argument in this text is that people should go to vote because it's their future what is at play. While it is true that she might have selfish interests to be encouraging people to vote, that is unlikely to diminish the validity of her argument.</p>

Table 5: Examples of the feedback given by Prometheus 2 for the most typical kinds of errors. The reasons why the label was not correct are given by the authors of this paper on the last column. Some feedback texts and interventions were shortened in order to fit this table.

## F Generation Prompt

You are tasked with generating critical questions that are useful for diminishing the acceptability of the arguments in the following text:

{intervention}

Take into account a question is not a useful critical question:

- If the question is not related to the text.
- If the question is not specific (for instance, if it's a general question that could be applied to a lot of texts).
- If the question introduces new concepts not mentioned in the text (for instance, if it suggests possible answers).
- If the question is not useful to diminish the acceptability of any argument. For instance, if it's a reading-comprehension question or if it asks about the opinion of the speaker/reader.
- If its answer is not likely to invalidate any of the arguments in the text. This can be because the answer to the question is common sense, or because the text itself answers the question.

Output 3 critical questions.

Give one question per line.

Make sure there are at least 3 questions.

Do not give any other output.

Do not explain why the questions are relevant.

Figure 11: Generation Prompt for benchmarking.