Critical Questions Generation: Motivation and Challenges

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 development of Large Language Models (LLMs) has brought impressive performances on mitigation strategies against misinformation, such as counterargument generation. However, LLMs are still seriously hindered by outdated knowledge and by their tendency to generate hallucinated content. In order to circumvent these issues, we propose a new task, namely, *Critical Questions Generation*, consisting of processing an argumentative text to generate the critical questions (CQs) raised by it. In argumentation theory CQs are tools designed to lay bare the blind spots of an argument by pointing at the information it could be missing. Thus, instead of trying to deploy LLMs to produce knowledgeable and relevant counterarguments, we use them to question arguments, without requiring any external knowledge. Research on CQs Generation using LLMs requires a reference dataset for large scale experimentation. Thus, in this work we investigate two complementary methods to create such a resource: (i) instantiating CQs templates as defined by Walton's argumentation theory and (ii), using LLMs as CQs generators. By doing so, we contribute with a procedure to establish what is a valid CQ and conclude that, while LLMs are reasonable CQ generators, they still have a wide margin for improvement in this task.

1 Introduction

Natural Language Processing (NLP) applications to deal with misinformation have become a popular line of research in tasks such as fact verification (Thorne et al., 2018), evidence retrieval (Soleimani et al., 2020) or counterargument generation (Chung et al., 2019; Chen et al., 2023). However, even when deploying generative Large Language Models (LLMs), most applications face challenges regarding three issues: LLMs often lack the required up-to-date knowledge for these tasks (Gao et al., 2023), there is not always an agreement on what is the truth (Chang et al., 2024), and LLMs themselves can produce hallucinations or rely on unfaithful data, generating misinformation of their own making (Xu et al., 2024; Lin et al., 2022).

Yet, instead of requiring the LLMs to output factual knowledge, could we use them to point at the missing or potentially uninformed claims? In other words, could we use LLMs to uncover the blind spots in the argumentation? To open this line of research, we ground our work on argumentation theory, which has for centuries been studying dialogical exchanges of information. Specifically, we look into *argumentation schemes*, a set of abstract structures developed by systematically identifying common patterns of argumentation and outlining the defeasibility of these patterns. In these structures, the devices designed to find the blind spots in the arguments are called *critical questions*.

Critical questions are the set of inquiries that could be asked in order to judge if an argument is acceptable or fallacious. Therefore, these questions are designed to unmask the assumptions held by the premises of the argument and attack its inference. In the theoretical framework developed by Walton et al. (2008), argumentation schemes are represented as templates depicting the premises, the conclusion, and the critical questions of each scheme. This framework is useful to promote critical thinking, since it allows uncovering fallacies by answering questions. Figure 1 shows two examples of argumentation schemes and their corresponding critical questions (CQs). The first of these examples is an argument that links a cause (migration) to an effect (unemployment). Therefore, the CQs related to this argument ask about the strength of this relation and the possibility of other causes also having a role in the effect. The second example fits the scheme of *practical reasoning*. That is, given a goal, the argument defines an action to achieve it. Here, the CQs ask about the compatibility of this goal with others, the alternative actions to achiev-

Proceedings of the 28th Conference on Computational Natural Language Learning, pages 105–116
November 15-16, 2024 ©2024 Association for Computational Linguistics

(a) **Scheme – Argument from Cause to Effect**
Premise: Generally, if people pour into the USA, then Americans lose their jobs.
Premise: In the current situation, people are pouring into the USA.
Conclusion: In the current situation, Americans lose their jobs.

CQ: How strong is the generalisation that if people pour into the USA then Americans will lose their jobs?
CQ: Are there other factors in this particular case that could be interfering with the fact that Americans lose their jobs?

(b) **Scheme – Practical Reasoning**
Premise: There is the goal of making the economy fairer.
Premise: Raising the national minimum wage is a means to realize the goal of making the economy fairer.
Conclusion: Therefore, raising the national minimum wage ought to occur.

CQ: Are there other relevant goals that conflict with making the economy fairer?
CQ: Are there alternative actions to raising the national minimum wage to achieve making the economy fairer? If so, which is the most efficient action?
CQ: Could raising the national minimum wage have consequences that we should take into account? Is it practically possible?

Figure 1: Arguments from the US2016 dataset (Visser et al., 2021), instantiated using the templates of argumentation schemes and critical questions defined in Walton et al. (2008).

ing this goal, and the potential consequences of the proposed action.

Previous work has proved the usefulness of CQs for enhancing fallacy identification (Musi et al., 2022), and for argumentative essays evaluation (Song et al., 2014). But, to the extent of our knowledge, there has not been any attempt to automate the generation of CQs. In this work, we propose the task of *Critical Questions Generation*: given an argumentative text, the model is asked to generate the necessary CQs to assess the acceptability of the arguments in the text. In this setting, the argumentative text is the input and the set of CQs is the target output. As in other NLP tasks, such as machine translation or paraphrasing, the model is not required to find new information, but to understand and reformulate the input in a certain way.

A crucial requirement to investigate the automatic generation of CQs is to have reference data for experimentation. However, as far as we know, there has not been any attempt to create such a resource. In order to address this shortcoming, in this paper we investigate two methods for creating a dataset for the generation of CQs: (1) using the sets of CQ templates defined in Walton et al. (2008)'s theory (from now on, theory-CQs); and (2) using LLMs to generate these CQs (from now on, llm-CQs). While looking into these methods, we attempt to answer the following research questions: (i) are current Large Language Models good critical question generators? (ii) how can we operationalize what is a valid critical question? (iii) what is the optimal strategy to build a reference dataset for large scale experimentation on the task of *Critical Questions Generation*?

To answer these questions, we start by looking at the theoretical sets of CQs and instantiating them using a set of argumentative texts already annotated with argumentation schemes (Visser et al., 2021; Lawrence et al., 2018). As a second step, we prompt two state-of-the-art LLMs to give us candidate CQs for these same argumentative texts, and we design a procedure to evaluate their relevance towards the texts and their validity as CQs. We then compare the two methods and highlight the main challenges faced by LLMs when generating CQs. Summarizing, the main contributions of this work are:

- We propose the task of *Critical Questions Generation* and motivate it by relying on previous work.
- We use naturally-occurring dialogical data to study how to generate critical questions using the theory templates and LLMs.
- We operationalize how to define a valid critical question.
- We study the main challenges faced by LLMs when generating critical questions.

In this work, we observe that questions generated using theory and questions generated using LLMs are complementary: while theory-CQs are mostly about relations between premises, llm-CQs rather ask about evidences. Additionally, LLMs introduce a new type of questions: those asking about further definition of the terms used in the arguments. Regarding the performance of current LLMs, we observe that models struggle to output relevant CQs and output many non-critical questions. Therefore, we conclude that more advanced training and prompting techniques should be used and, to this end, reference data should be created using both the theory and LLMs' methods. All the data and code in this project has been released.[1]

2 Previous Work

To contextualise this work, we discuss the relation between argumentation and misinformation, introduce the nature of critical questions, and offer related work on argumentation schemes from a computational point of view.

2.1 Using argumentation to fight misinformation

Misinformation has been tackled using many strategies: from debunking strategies (e.g. fact-checking propagated information) to pre-bunking (e.g. exposing disinformation strategies to make citizens resilient towards manipulation). However, recent studies have shown that pre-bunking has a potentially longer effect, since the learned skills are not bound to specific contexts (Maertens et al., 2021). Following this, digital applications have been built to enhance citizens' abilities to deal with misinformation, such as the recognition of misleading sources and headlines (Fakey,[2] NewsWise headlines quizz[3]), the identification of fake images (Real or Photoshop quizz[4]), or the decision-making processes of news rooms (BBCireporter,[5] NewsFeed Defenders[6]).

However, these applications focus mostly on dealing with fake information, while misinformation is often generated by drawing invalid relations between claims and the premises provided to support these claims (Musi et al., 2023). In this sense, more recent pre-bunking applications have focused on techniques based on argumentation theory, which have the goal of evaluating the connections between the available evidence and the statement that it is trying to support (Lawrence et al., 2018; Visser et al., 2020; De Liddo et al., 2021; Altay et al., 2022).

In this line of research, Musi et al. (2023) developed a chatbot that, following gamification principles, used a dialogical context to teach users how to identify fallacies by being exposed to critical questions. Users of this tool showed an overall increased ability to identify fallacious arguments. While the scenarios portrayed in Musi's chatbot are based on an annotated database of 1,500 fact-checked news, latest NLP advances in LLMs could be used to generate critical questions on unseen arguments, therefore being able to use this tool to deal with any upcoming domain.

Applications of language models in the fight against misinformation have often been framed as classification and information retrieval tasks (Montoro Montarroso et al., 2023). In contrast, we propose to use LLMs as a tool for generating questions, which enhances the relativistic conceptions of truth of most critical thinking paradigms (Musi et al., 2023), as opposed to the absolutist notions of truth encouraged by using LLMs as question-answerers and classifiers.

2.2 The nature of critical questions

Critical questions are an essential element of the notion of *argumentation schemes*. Argumentation schemes are "forms of arguments (structures of inference) that represent structures of common types of arguments used on everyday discourse" (Walton et al., 2008). These arguments are defeasible, meaning that their conclusions can be accepted only provisionally while there is no evidence that defeats it. Defeasible arguments are the most common arguments in everyday discussions, and knowing what to ask before accepting them is an important skill.

The predecessor of argumentation schemes were topics (*topoi* in Aristotle's Rhetoric), which were conceived as warrants that back the logical inferences drawn from premises to conclusions. Modern researchers have adapted them for use in computational applications (Reed and Walton, 2001; Macagno et al., 2017). Additionally, these tools have become popular among critical thinking researchers for their pedagogic usefulness.

In pedagogical terms, argumentation schemes can be used "as a way of providing students with additional structure and analytic tools with which to analyze natural arguments and to evaluate them critically" (Walton et al., 2008). In this approach, critical questions function as memory devices: a

[1] https://github.com/hitz-zentroa/critical_questions_generation

[2] https://fakey.osome.iu.edu/

[3] https://www.theguardian.com/newswise/2021/feb/04/fake-or-real-headlines-quiz-newswise-2021

[4] https://landing.adobe.com/en/na/products/creative-cloud/69308-real-or-photoshop/

[5] https://www.bbc.co.uk/news/resources/idt-8760dd58-84f9-4c98-ade2-590562670096

[6] https://www.icivics.org/games/newsfeed-defenders

way to recall the missing information in the argument.

Although the goals and usefulness of critical questions have been extensively discussed, up to our knowledge, there has not been any successful attempt to operationalize what is and what is not a valid critical question. Since our goal is to create them automatically, setting this boundary becomes a necessary first step.

Most definitions of critical questions are highly linked to their function. Following this tradition, it could be argued that a good critical question is the one that fulfills its goal: pointing at reasons to *rebut the argument*. Moreover, critical questions can not only attack the acceptability of an argument by defeating its conclusion, but also undercut it by attacking the connection between the premises and the given conclusion (Pollock, 1987). In Section 4, we operationalize this definition of valid CQ, and in Section 5, we implement it in the evaluation of llm-CQs.

2.3 Argumentation Schemes in Computational Argumentation

While no attempt exists to automatically generate CQs, there has been some work on argumentation schemes annotation and detection, which we will be taking as a starting point.

One of the most ambitious works in argumentation from a computational point of view was the Araucaria project, which created a database of arguments annotated in Argument Markup Language that included argumentation schemes (Reed et al., 2008). Later, the Inference Anchoring Theory (IAT Budzynska and Reed (2011)) became a popular format for representing how arguments are created in dialogical settings. IAT diagrams feature locutions, propositions, dialogical relations, and propositional relations. Recent work has also added argumentation-scheme labels to IAT diagrams. The available datasets annotated with IAT and schemes are listed in Table 1.

Other datasets that are labeled with argumentation schemes although not in the IAT format are the social media datasets from Jo et al. (2021), which contain 1,924 examples of 2 argumentation schemes; and the Genetics Research Corpus, which identifies argumentation schemes in scientific claims from genetic research articles (Green, 2015). Lately, datasets with synthetic arguments have been released (Kondo et al., 2021; Ruiz-Dolz et al., 2024; Saha and Srihari, 2023). However, we are interested in naturally-occurring arguments.

The task of automatically identifying argumentation schemes was first attempted by Feng and Hirst (2011) and Lawrence and Reed (2016), using machine learning techniques. Later, Jo et al. (2021) used logic and theory-informed mechanisms for a similar task, and Kondo et al. (2021) used language models, showing the difficulty of identifying schemes (with 7 categories, their overall accuracy with BERT (Devlin et al., 2019) was 27.5%).

In previous work, it has been observed that tasks requiring complex reasoning remain a challenge for LLMs (Xu et al., 2023; Gendron et al., 2024; Han et al., 2022). Furthermore, Payandeh et al. (2023) demonstrated that LLMs are easily convinced using logical fallacies, and Ruiz-Dolz and Lawrence (2023) showed that LLMs fail when asked to detect argumentative fallacies. The task of fallacy detection is highly related to our work (Sahai et al., 2021; Goffredo et al., 2022; Alhindi et al., 2022; Helwe et al., 2024). However, in this work we wish to foster human-computer interaction and use LLMs to raise the questions that would help a human unmask the fallacies of its caller.

So far, the most similar work to ours is Musi et al. (2023), where they developed a chatbot that outputted critical questions from a database of possible issues, and Song et al. (2014), where they found that human annotations identifying the CQs present in essay evaluations contributed significantly to predicting the grade. While their experiments tested the usefulness of using CQs, none of these two tried to generate them automatically.

3 Data

For the purpose of this work we have decided to use a subset of the US2016 (Visser et al., 2021) and the Moral Maze datasets (Lawrence et al., 2018), which, as explained in the previous section, have already been transcribed and annotated with argumentation schemes. Both of these datasets are oral debates, and they are structured as sequences of interventions by different debaters.

In order to use these datasets, we have mapped their labels to the argumentation schemes in Walton et al. (2008). Since the labels of both of these datasets are based on Walton's work, the mapping has amounted to terminology matching. Given the long list of argumentation schemes, we have decided to work with the 18 most frequent schemes. Annex A provides the mapping and the distribu-

Name	Paper	Nº Args.	Nº Schemes	Original Format	Domain
US2016	Visser et al. (2021)	413	60	Oral debate	Politics
Moral Maze	Lawrence et al. (2018)	79	32	Oral debate	Politics
US2016reddit		19	4	Written social media	Politics
EO_PC	Lawrence and Reed (2015)	139	3	Written	Not specified
Reg. Room Div.	Konat et al. (2016)	227	7	Written social media	Product Regulations
Legal		545	12	Written	Legal

Table 1: Available data in IAT format with argumentation schemes. All the datasets are in English.

tion of argumentation schemes for each of the two datasets.

Since both of these datasets have been annotated as IAT diagrams, each argumentation scheme label links two or more propositions in the debate, forming an argument.[7] The debates are composed of interventions, which we are going to use as our *argumentative texts*. Each intervention can have many annotated arguments (or none). After preprocessing,[8] we obtain 370 interventions (73 from Moral Maze and 297 from US2016) of which 117 contain at least one argument (25 from Moral Maze and 92 from US2016).

For the manual analysis of this work, we use 21 of the interventions, chosen to keep the label distribution as similar as possible to the one in the full datasets; 10 of these interventions come from US2016 and 11 from Moral Maze. The distribution of the 60 arguments contained in these 21 interventions can be found in Annex B.

4 Our Method

In order to identify the challenges in the task of *Critical Questions Generation*, there is an urgent need for reference data. To explore how this data should be created, we generate critical questions both using the theory templates and LLMs.

To generate CQs based on Walton's theory (theory-CQs), we take each annotated argument and instantiate the CQs associated to that argumentation scheme (red-dotted box at the top of Figure 2). Regarding the generation of CQs with LLMs (llm-CQs), we prompt two state-of-the-art LLMs and we evaluate the relevance of the candidate CQs towards the argumentative text (blue-dashed box at the bottom of Figure 2). We then relate the llm-CQs to the arguments of the text and to the theory-CQs (green box), and assess the validity of the llm-CQs that relate to an argument but do not correspond to any of the existing theory-CQs (such as CQ 5 in Figure 2). In the rest of the section, we describe in detail each of these processes.

4.1 Generation using theory

The critical questions based on theory are defined using the set of CQs in Walton et al. (2008). We reformulate some of these questions to make them sound more natural (the final set can be found in Annex C). To transform these questions into tailored CQs for each argument, we first manually annotate the text needed to fill the gaps of the variables in the argumentation-schemes' templates. For each argument, the annotator sees the premises and conclusion associated with the argumentation scheme (i.e. the template), the propositions of the argument, and the entire intervention in which the argument occurred. For instance, to annotate argument *a* in Figure 1, the annotator saw the data in Table 2, and was asked to write the text that is needed to instantiate the scheme template. In this case, $< eventA >$ = "people are pouring into the USA" and $< eventB >$ = "Americans might lose their jobs".

We used two annotators for this task, and achieved an inter-annotator agreement (IAA) of 0.88 with a sample of 174 variables.[9] In the end, 9 arguments were discarded by both annotators, as they were not able to find the connection between the propositions and the argumentation scheme that had been given to its relation.

We then instantiated the CQs, substituting each variable for the piece of text that had been annotated. This step resulted in questions with grammatical errors, which we post-edited manually, with 39.44% of the questions getting editions. We discarded 10 of the questions for being meaningless. Most common corrections consisted of modifying verbs from infinitive to gerund forms and vice-

[7] For a comprehensive explanation of IAT diagrams see Hautli-Janisz et al. (2022).

[8] We structure the data by intervention, splitting the very long interventions, and merging the very short ones (for an example, see the columns "Intervention" in Table 2 and Figure 3). The code on how to go from the IAT diagrams to our dataset has been published on Github.

[9] The extended explanation of this annotation will be published as guidelines.

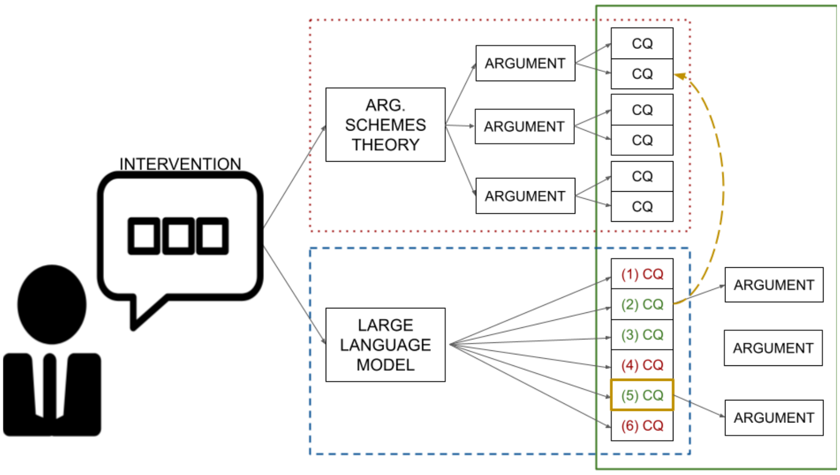

Figure 2: Outline of the steps taken in our approach. Starting from each intervention, we generate CQs using the theory templates (red-dotted box) and the LLMs (blue-dashed box). In the green box, we relate the relevant llm-CQs to the arguments of the intervention (if possible), and relate these llm-CQs to a theory-CQ (if possible).

Argument Scheme	Scheme Template	Propositions	Intervention
Argument from CauseToEffect	Generally, if $< eventA >$, then $< eventB >$. In the current situation, $< eventA >$. In the current situation, $< eventB >$.	"people are pouring into the USA" & "Americans are losing their jobs"	TRUMP: I want to make America great again We are a nation that is seriously troubled We're losing our jobs People are pouring into our country The other day , we were deporting 800 people perhaps they passed the wrong button they pressed the wrong button perhaps worse than that it was corruption [...]

Table 2: Data seen by the annotator when defining the variables to fil the argumentation scheme templates. Propositions and argumentation schemes come from the IAT annotations in the US2016 dataset (Visser et al., 2021). The scheme template comes from Walton et al. (2008).

versa, or from singular to plural forms and vice-versa, and removing double negations. This process resulted in the generation of 129 theory-CQs associated to 51 arguments, an average of 6.14 CQs per intervention.

4.2 Generation using LLMs

Walton's sets of critical questions are thought of as starting points towards rebuttal strategies. However, they do not intend to be an exhaustive list of the potentially useful CQs for each scheme (Walton and Godden, 2005).

For this reason, it is interesting to experiment with LLMs to see if the models can generate questions that are valid CQs but are not included in Walton's templates. In this sense, our goal is to have a list of valid CQs as exhaustive as possible that could be used as reference data. However, llm-CQs should be carefully curated. For this purpose, we have designed a method to filter the candidate llm-CQs and obtain a list of relevant and valid CQs. This procedure will serve, at the same time, as an evaluation of how good current LLMs are at generating CQs.

To this goal, we prompt two LLMs to generate the CQs that each intervention may arise in a zero-shot setting. We experiment with two different prompts, one including the query and the intervention,[10] and one that also includes a definition of critical questions.[11] Then, the evaluation process to filter the candidate llm-CQs has the following three steps.

First, we manually review each of the candidate CQs to detect those that are not relevant with respect to the given argumentative text (i.e. the intervention). We have detected three issues that make the questions not relevant: (a) **the introduction of new concepts or topics** – ideally LLMs should

[10] Prompt 1: *List the critical questions that should be asked regarding the arguments in the following paragraph:* $< SPEAKER >$: "$< INTERVENTION >$"

[11] Prompt 2: *Critical questions are the set of enquiries that should be asked in order to judge if an argument is good or fallacious by unmasking the assumptions held by the premises of the argument. List the critical questions that should be asked regarding the arguments in the following paragraph:* $< SPEAKER >$: "$< INTERVENTION >$"

generate CQs related to the content of the intervention, not introducing new topics or concepts that may carry the model's biases; (b) **bad reasoning**, namely, questions critical towards positions or claims the speaker does not hold; or (c) **non-specific critical questions** that could be asked on any argument and that do not take the intervention into account.

Second, using the set of relevant llm-CQs, we match each of these to one of the annotated arguments (if possible), and then assess if the matched CQs also exist in the set of theory-CQs of that argument (that means checking whether they are asking about the same blind spot as any of the CQs generated in Section 4.1). This process leaves us with 4 types of llm-CQs: (i) the ones that are not relevant (CQs 1, 4 and 6 in Figure 2), (ii) the ones that do not match any of the annotated arguments (CQ 3 in Figure 2), (iii) the ones that have a matching argument and a matching theory-CQ (CQ 2 in Figure 2), and (iv) the ones that do have a matching argument but NOT a matching theory-CQ (CQ 5 in Figure 2). We are interested in further investigating this last group, as these are the CQs that the theory did not generate, but are potentially valid.[12]

Third, the last step to validate this group of LLM-generated CQs consists in assessing their inferential relation to the arguments they have been assigned. That means asking whether it fulfills the core function of CQs: unmasking a blind spot in the argument. We operationalized this evaluation by taking each argument and question pairs and asking: "Can the answer to this question diminish the acceptability of the argument?". The answer to this question can only be *yes* or *no*.[13] In a proof-of-concept evaluation we achieved an IAA of 0.65 with two annotators.

5 Results

In order to generate the critical questions we use two open state-of-the-art LLMs: Llama-2-13B and Zephyr-13B (Touvron et al., 2023; Tunstall et al., 2023). We employ the instruction-tuned chat versions of the models. For Zephyr, we use the parameters indicated for their chat version and the chat templates used in training. For Llama-2, we use the chat version released in July 2023. With the two prompts, we obtain 495 LLM-generated candidate CQs (llm-CQs). We now report the results of each of the evaluation steps described in Section 4.2, to later compare the llm-CQs to the theory-CQs, showing the differences between the questions obtained through each of these approaches.

5.1 Relevance with respect to the Intervention

The relevance issues found in the llm-CQs are reported in Table 3. For all types of issues, Llama-2 works better than Zephyr. While in Llama-2 with the Query prompt 80% of the CQs are relevant, in Zephyr with the Query+Definition prompt the relevance drops to 30%.

When using the prompt with just the query, for both models, over 10% of the generated questions ask about claims the speaker does not hold (i.e. bad reasoning). Additionally, in Zephyr, 15% of the questions introduce new concepts. We expected that adding the definition of CQs to the prompt would improve the performance of the models. However, while *bad reasoning* issues are reduced by half for both models, *new concept* issues do not disappear (and even increase for Llama-2). Additionally, a new type of issue is introduced: *non-specific questions*. These are candidate CQs that are not specific to the text, but just general CQs (e.g. "What assumptions is the argument making?"). That is especially the case with Zephyr. With this model, we also get a lot of outputs that are not even questions (the ones classified as *Other*).

5.2 Relation of llm-CQs to Arguments and to theory-CQs

In order to validate the 308 relevant LLM-generated CQs, these need to be related to one of the arguments in the intervention. In this step, the llm-CQs are paired with the arguments of the intervention that prompted them. As a result, 191 unique llm-CQs are associated to at least one of the arguments, resulting in 50 out of the 51 arguments having at least one associated llm-CQ. Since one llm-CQ can be associated with many arguments, and an argument can have multiple associated llm-CQs, the total number of pairs of arguments and llm-CQs is 294.

Regarding those questions that appeared both in the llm-CQs and in the theory-CQs, we have found 36 unique llm-CQs that have a matching theory-CQ. Since multiple llm-CQs can be associated to one theory-CQ (if they have the same meaning),

[12] In group (ii), there are also potentially valid questions but, since we are not able to relate them to any of the annotated arguments, we do not have a way to validate them. This set can include both invalid questions or valid questions related to non-annotated arguments.

[13] The guidelines of this evaluation will be published.

Model	**Prompt**	**Relevant**	**New Concept**	**Bad Reasoning**	**Non-specific**	**Other**	**TOTAL**
Zephyr	Q-prompt	**67.57%**	14.86%	14.86%	0.0%	2.7%	74
Llama-2	Q-prompt	**80.74%**	4.44%	11.11%	2.96%	0.74%	135
Zephyr	D+Q-prompt	**29.46%**	13.18%	7.75%	33.33%	16.28%	129
Llama-2	D+Q-prompt	**70.7%**	10.19%	6.37%	12.1%	0.64%	157
All	All	**308**	50	46	66	25	**495**

Table 3: Relevance issues of the LLM-generated critical questions. By model and prompt. *Q-prompt* refers to the prompt with only the query, and *D+Q-prompt* refers to the prompt that also has the definition of CQs. Each column is one of the relevance issues described in Section 4.2.

we obtain 52 pairs of llm-CQs and theory-CQs.

In the end, this step has left us with 242 llm-CQs that are associated to an argument but do not match any of the theory-CQs of that argument.[14]

5.3 Inferential Validity of the llm-CQs

Having related each of the llm-CQs to an argument, we can finally check the validity of each of these critical questions by asking if the answer to the CQ could diminish the acceptability of the argument. We do this with the 242 llm-CQs that have an associated argument but have no matching theory-CQ, since we already know that llm-CQs that matched a theory-CQs are valid critical questions.

This evaluation results in 64.05% of the relevant and related llm-CQs being marked as valid (155 questions). The remaining 87 questions do not focus on critical aspects of the argument, often, these ask for additional information that could not impact the acceptability of the argument.

After the filtering processes described, we have been left with a dataset of 21 interventions associated to three sets of valid CQs: (i) the theory-CQs (129 in total), (ii) the llm-CQs that matched a theory-CQ (52 in total), and (iii) the llm-CQs that did not match a theory-CQ but were found to be valid in Section 5.3 (155 in total). That means that we have 207 valid llm-CQs in total (52 plus 155), 137 of which are unique. Therefore, in the end, only 28% of the 495 candidate llm-CQs end up being relevant and valid (for an example of an intervention in the resulting dataset, see Figure 3).

5.4 Comparing the Approaches

At this point, it is interesting to study the differences between the sets of questions obtained in each approach. To this goal, all the CQs have been classified regarding the type of blind spot they are trying to unmask. We find that, regarding theory-CQs, the most common type of questions are those asking about the relation between the premises and the conclusion (27%), followed by questions about the available evidence (24%), and questions about possible exceptions (18%). In the case of llm-CQs, asking about evidence is the most common type of CQs (27%), followed by relations (21%) and potential consequences of the premises (17%). Most interestingly, we find that 16% of llm-CQs are asking for more specific definitions of the concepts present in the argument. This kind of questions are not contemplated at all in the theoretical sets of questions, and both of our annotators considered them valid (the first llm-CQ in Figure 3 is of this type). Finally, the few questions that are generated with both approaches (theory and LLMs) are mostly about consequences and evidence (see Table 4).

Type	**t-CQs**	**%**	**llm-CQs**	**%**	**match**
evidence	31	24.0	55	26.6	17
relation	35	27.1	43	20.8	10
conseq.	14	10.9	35	16.9	19
definition	0	0.0	34	16.4	0
other	6	4.7	20	9.7	0
alternative	6	4.7	7	3.4	0
exception	23	17.8	7	3.4	5
source	14	10.9	6	2.9	3
Total	129		207		52

Table 4: Types of questions in the final sets of theory-CQ, valid llm-CQs, and matching CQs between the two approaches. Amount and percentage. The matching ones are also included in the counts of both approaches.

6 Concluding Remarks

In this work we have introduced and motivated the task of *Critical Questions Generation*. Moreover, we have studied how to generate valid critical questions with two goals in mind: (i) designing a procedure to obtain reference data, and (ii) discovering the main difficulties that state-of-the-art LLMs face when generating valid critical questions.

Regarding the difficulties of the task, we have found that current LLMs struggle to generate CQs

[14] Note these are pairs of related llm-CQs and arguments.

MT: "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) Intervention

- How strong is the generalisation that if that form of capitalism was not making enough money in the developed world then the money did not flow into those industrial activities?
- Are there other factors in this particular case that could have interfered with the event of 'the money did not flow into those industrial activities'?
- How strong is the generalisation that if that form of capitalism wasn't generating sufficient surpluses then the money did not flow into industrial activities?

(b) theory-CQs

- How is 'sufficient surpluses' defined, and how would one measure it?
- Is MT implying that current forms of capitalism are more successful at generating profits and surpluses than the one being discussed? If yes, why?
- What evidence is there to support the claim that the form of capitalism being used in the developed world was not generating sufficient surpluses?
- Are there any alternative explanations for why the money did not flow into industrial activities?

(c) llm-CQs

Figure 3: Example of an instance of the generated reference data. The intervention is from the Moral Maze dataset, and the theory-CQs and the llm-CQs are the result of both of our generation methods.

strictly related to the text. On the one hand, they tend to output CQs including new concepts not present in the arguments. On the other hand, they sometimes opt for generating unfiltered lists of very general CQs, with no regard to the given argumentative text. Reasoning is still an issue for these models, and they sometimes struggle to understand what claims are actually held by the given text. Finally, while 62% of the LLM-generated CQs did not have any of these three issues (308 out of 495), only 28% of the CQs initially generated by LLMs were found to be valid in relation to one of the arguments (137 out of 495), showing that there is a big margin for improvement.

In relation to the goal of creating a reference dataset, we have shown that the existing theoretical sets of critical questions do not account for all the possible valid critical questions. In this sense, our results show that only 25% of the valid llm-CQs had been included in the theoretical sets (52 out of 207). For this reason, we propose using both theory-CQs and llm-CQs to build the reference data for this task. Furthermore, we have also observed that the type of questions generated by LLMs differs from the ones created by theory, with the LLMs approach generating many questions related to evidence, consequences and definitions. This suggests that the two approaches (theory and LLMs) are complementary.

While this work has been a first step towards the task of *Critical Questions Generation*, our end goal of automatically generating valid CQs is far from solved. In future work, we will create a larger reference dataset including both theory and llm-CQs to facilitate research on automatic CQs Generation.

Finally, it should be noted that we have not paid any attention to LLM-generated questions that did not match any of the annotated arguments. However, as some arguments might be missing from the annotation (either because they were not in our selected 18 argumentation schemes or because the annotators missed them), some of these questions might be valid CQs. This shows that our work relies heavily on already annotated data with argumentation schemes. And, while the datasets used are reliable (Visser et al., 2021; Lawrence et al., 2018), there is not a lot of quality data annotated with argumentation schemes, which poses a limitation on how much reference data can be created. As far as we are aware, the only data available is the one detailed in Table 1, which is all in English.

Acknowledgements

This work has been partially supported by the Basque Government (Research group funding IT-1805-22). We are also 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 and (iii) Deep-Minor (CNS2023-144375) and European Union NextGenerationEU/PRTR. Blanca Calvo Figueras is supported by the UPV/EHU PIF22/84 predoc grant.

References

Tariq Alhindi, Tuhin Chakrabarty, Elena Musi, and Smaranda Muresan. 2022. Multitask Instruction-based Prompting for Fallacy Recognition. In *Pro-*

ceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8172–8187, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sacha Altay, Marlène Schwartz, Anne-Sophie Hacquin, Aurélien Allard, Stefaan Blancke, and Hugo Mercier. 2022. Scaling up interactive argumentation by providing counterarguments with a chatbot. *Nature Human Behaviour*, 6(4):579–592. Number: 4 Publisher: Nature Publishing Group.

Katarzyna Budzynska and Chris Reed. 2011. Whence inference. *University of Dundee Technical Report*.

Tyler A. Chang, Katrin Tomanek, Jessica Hoffmann, Nithum Thain, Erin van Liemt, Kathleen Meier-Hellstern, and Lucas Dixon. 2024. Detecting Hallucination and Coverage Errors in Retrieval Augmented Generation for Controversial Topics. Publication Title: arXiv e-prints ADS Bibcode: 2024arXiv240308904C.

Guizhen Chen, Liying Cheng, Luu Anh Tuan, and Lidong Bing. 2023. Exploring the potential of large language models in computational argumentation. *arXiv preprint arXiv:2311.09022*.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. CONAN - COunter NArratives through Nichesourcing: a Multilingual Dataset of Responses to Fight Online Hate Speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence, Italy. Association for Computational Linguistics.

Anna De Liddo, Nieves Pedreira Souto, and Brian Plüss. 2021. Let's replay the political debate: Hypervideo technology for visual sensemaking of televised election debates. *International Journal of Human-Computer Studies*, 145:102537.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Vanessa Wei Feng and Graeme Hirst. 2011. Classifying arguments by scheme. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 987–996, Portland, Oregon, USA. Association for Computational Linguistics.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Gaël Gendron, Qiming Bao, Michael Witbrock, and Gillian Dobbie. 2024. Large Language Models Are Not Strong Abstract Reasoners. ArXiv:2305.19555 [cs].

Pierpaolo Goffredo, Shohreh Haddadan, Vorakit Vorakitphan, Elena Cabrio, and Serena Villata. 2022. Fallacious Argument Classification in Political Debates. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence Main Track*, volume 5, pages 4143–4149. ISSN: 1045-0823.

Nancy Green. 2015. Identifying Argumentation Schemes in Genetics Research Articles. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 12–21, Denver, CO. Association for Computational Linguistics.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Luke Benson, Lucy Sun, Ekaterina Zubova, Yujie Qiao, Matthew Burtell, David Peng, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor, Ansong Ni, Linyong Nan, Jungo Kasai, Tao Yu, Rui Zhang, Shafiq Joty, Alexander R. Fabbri, Wojciech Kryscinski, Xi Victoria Lin, Caiming Xiong, and Dragomir Radev. 2022. FOLIO: Natural Language Reasoning with First-Order Logic. ArXiv:2209.00840 [cs].

Annette Hautli-Janisz, Zlata Kikteva, Wassiliki Siskou, Kamila Gorska, Ray Becker, and Chris Reed. 2022. QT30: A Corpus of Argument and Conflict in Broadcast Debate. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3291–3300, Marseille, France. European Language Resources Association.

Chadi Helwe, Tom Calamai, Pierre-Henri Paris, Chloé Clavel, and Fabian Suchanek. 2024. MAFALDA: A benchmark and comprehensive study of fallacy detection and classification. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4810–4845, Mexico City, Mexico. Association for Computational Linguistics.

Yohan Jo, Seojin Bang, Chris Reed, and Eduard Hovy. 2021. Classifying Argumentative Relations Using Logical Mechanisms and Argumentation Schemes. *Transactions of the Association for Computational Linguistics*, 9:721–739. Place: Cambridge, MA Publisher: MIT Press.

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.

Takahiro Kondo, Koki Washio, Katsuhiko Hayashi, and Yusuke Miyao. 2021. Bayesian Argumentation-Scheme Networks: A Probabilistic Model of Argument Validity Facilitated by Argumentation Schemes.

In *Proceedings of the 8th Workshop on Argument Mining*, pages 112–124, Punta Cana, Dominican Republic. Association for Computational Linguistics.

John Lawrence and Chris Reed. 2015. Combining Argument Mining Techniques. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 127–136, Denver, CO. Association for Computational Linguistics.

John Lawrence and Chris Reed. 2016. Argument Mining Using Argumentation Scheme Structures. In *Computational Models of Argument*, pages 379–390. IOS Press.

John Lawrence, Jacky Visser, and Chris Reed. 2018. BBC Moral Maze: Test Your Argument. In *Comma*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Fabrizio Macagno, Douglas Walton, and Chris Reed. 2017. Argumentation Schemes. History, Classifications, and Computational Applications.

Rakoen Maertens, Jon Roozenbeek, Melisa Basol, and Sander van der Linden. 2021. Long-term effectiveness of inoculation against misinformation: Three longitudinal experiments. *Journal of Experimental Psychology: Applied*, 27(1):1–16. Place: US Publisher: American Psychological Association.

Andrés Montoro Montarroso, Javier Cantón-Correa, and Juan Gómez Romero. 2023. Fighting disinformation with artificial intelligence: fundamentals, advances and challenges. *Profesional de la Información*. Accepted: 2023-11-09T11:20:07Z Publisher: Profesional de la Información.

Elena Musi, Myrto Aloumpi, Elinor Carmi, Simeon Yates, and Kay O'Halloran. 2022. Developing Fake News Immunity: Fallacies as Misinformation Triggers During the Pandemic. *Online Journal of Communication and Media Technologies*, 12(3):e202217. Publisher: Bastas.

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.

Amirreza Payandeh, Dan Pluth, Jordan Hosier, Xuesu Xiao, and Vijay K. Gurbani. 2023. How susceptible are LLMs to Logical Fallacies? ArXiv:2308.09853 [cs].

John L. Pollock. 1987. Defeasible reasoning. *Cognitive Science*, 11(4):481–518.

Chris Reed, Raquel Mochales Palau, Glenn Rowe, and Marie-Francine Moens. 2008. Language Resources for Studying Argument. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Chris Reed and Douglas Walton. 2001. Applications of Argumentation Schemes.

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.

Ramon Ruiz-Dolz, Joaquin Taverner, John Lawrence, and Chris Reed. 2024. NLAS-multi: A Multilingual Corpus of Automatically Generated Natural Language Argumentation Schemes. ArXiv:2402.14458 [cs].

Sougata Saha and Rohini Srihari. 2023. ArgU: A Controllable Factual Argument Generator. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8373–8388, Toronto, Canada. Association for Computational Linguistics.

Saumya Sahai, Oana Balalau, and Roxana Horincar. 2021. Breaking Down the Invisible Wall of Informal Fallacies in Online Discussions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 644–657, Online. Association for Computational Linguistics.

Amir Soleimani, Christof Monz, and Marcel Worring. 2020. Bert for evidence retrieval and claim verification. In *Advances in Information Retrieval*, pages 359–366, Cham. Springer International Publishing.

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.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models.

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. ArXiv:2310.16944 [cs].

Jacky Visser, Barbara Konat, Rory Duthie, Marcin Koszowy, Katarzyna Budzynska, and Chris Reed. 2020. Argumentation in the 2016 US presidential elections: annotated corpora of television debates and social media reaction. *Language Resources and Evaluation*, 54(1):123–154.

Jacky Visser, John Lawrence, Chris Reed, Jean Wagemans, and Douglas Walton. 2021. Annotating Argument Schemes. *Argumentation*, 35(1):101–139.

Douglas Walton and David Godden. 2005. The nature and status of critical questions in argumentation schemes. *The Uses of Argument: Proceedings of a Conference at McMaster University*.

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

Fangzhi Xu, Qika Lin, Jiawei Han, Tianzhe Zhao, Jun Liu, and Erik Cambria. 2023. Are Large Language Models Really Good Logical Reasoners? A Comprehensive Evaluation and Beyond. ArXiv:2306.09841 [cs].

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2024. Hallucination is Inevitable: An Innate Limitation of Large Language Models. ArXiv:2401.11817 [cs].