ELLIS Alicante at CQs-Gen 2025: Winning the critical thinking questions shared task: LLM-based question generation and selection

Lucile Favero[1], **Daniel Frases** [2] *, **Juan Antonio Pérez-Ortiz**[3], **Tanja Käser**[4], **Nuria Oliver**[1]
[1] ELLIS Alicante, Spain, [2] Universidad Alfonso X el Sabio, Spain,
[3] Universitat d'Alacant, Spain, [4] École Polytechnique Fédérale de Lausanne, Switzerland
Correspondence: lucile@ellisalicante.org

Abstract

The widespread adoption of chat interfaces based on Large Language Models (LLMs) raises concerns about promoting superficial learning and undermining the development of critical thinking skills. Instead of relying on LLMs purely for retrieving factual information, this work explores their potential to foster deeper reasoning by generating critical questions that challenge unsupported or vague claims in debate interventions. This study is part of a shared task of the 12th Workshop on Argument Mining, co-located with ACL 2025, focused on automatic critical question generation. We propose a two-step framework involving two small-scale open source language models: a `Questioner` that generates multiple candidate questions and a `Judge` that selects the most relevant ones. Our system ranked first in the shared task competition, demonstrating the potential of the proposed LLM-based approach to encourage critical engagement with argumentative texts.

1 Introduction

The intensive use of chatbots based on Large Language Models (LLMs) has been associated with the promotion of superficial learning habits and a decline in critical thinking skills in their users, particularly students (Gerlich, 2025; Schei et al., 2024). Motivated by this fact, rather than relying on LLMs to provide factual answers, there is an opportunity to leverage the sophisticated natural language understanding capabilities of LLMs to foster critical thinking by means of the generation of critical questions.

This paper contributes to the CQs-Gen shared task of the 12th Workshop on Argument Mining, co-located with ACL 2025, which focuses on generating critical questions from debate interventions (Figueras, 2025b). While previous research has extensively explored the automatic generation of questions (Mulla and Gharpure, 2023; Ling and Afzaal, 2024), and current AI systems are capable of detecting misinformation with reasonable accuracy (Guo et al., 2022), relatively little work has leveraged argumentation theory to identify missing claims and misinformation in argumentative text (Figueras, 2024), and to generate relevant critical questions about the text (Favero et al., 2024; Figueras and Agerri, 2024; Ruiz-Dolz and Lawrence, 2025).

To fill this gap, we present an LLM-based framework for generating critical questions from argumentative text, aimed at encouraging users to reflect before accepting a claim. Our approach uses relatively small, open-source[1] LLMs to generate critical questions from a given debate intervention. Figure 1 illustrates the pipeline of the proposed method, detailing each step in the process.

In sum, the main contributions of our work are four-fold: (1) We propose a **Two-Step Framework for Critical Question Generation** composed of a `Questioner-Judge` LLM architecture where the `Questioner`, LLM_Q, generates multiple candidate questions that are evaluated by the `Judge`, LLM_J, which selects the most relevant ones, improving quality through selection; (2) we perform an extensive **empirical evaluation** of several small (7B–14B), open-source LLMs, demonstrating their strong performance despite limited size and without fine-tuning; (3) we explore how integrating **argumentation scheme theory** into prompts —both selectively and systematically— impacts generation quality and diversity; and (4) we highlight the potential of the proposed method to support **educational tools** that can be deployed locally, pre-

*Worked performed during an internship at ELLIS Alicante.

[1]We use the term "small" to refer to LLMs in the 7B-14B range, able to run on a student's laptop, and "open-source" to refer to LLMs that are freely available with at least open weights.

Proceedings of the 12th Argument Mining Workshop, pages 322–331

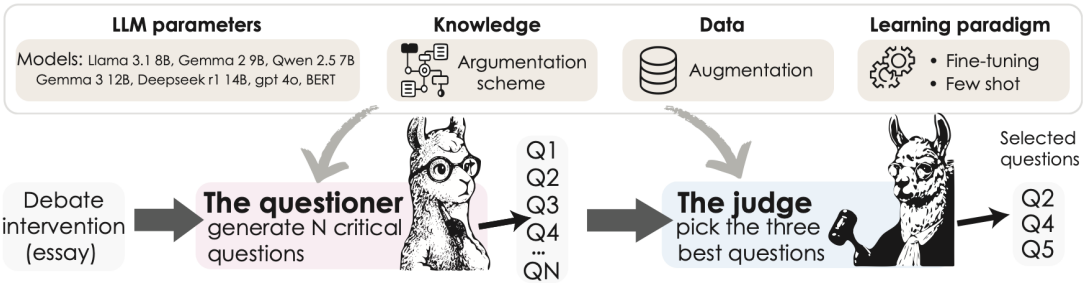

Figure 1: Overview of the proposed framework. Given a debate intervention as input, a first LLM, the `Questioner` LLM_Q, generates several candidate questions, and a second LLM, the `Judge`, LLM_J, chooses the three most useful critical questions among the questions generated.

serving privacy and reducing computational costs.

Our system ranked first in the CQs-Gen 2025 shared task on critical question generation, validating the effectiveness of the proposed approach.

2 Problem definition

2.1 Dataset description

The provided dataset (Figueras, 2025a), $D_{\text{shared train}}$, is composed of:

(1) $D = 189$ **interventions** during real political debate where each intervention consists of a short text of an average of 138 ± 88.4 words;

(2) Their associated **argumentative schemes**, that is "stereotypical patterns of inference that capture common types of defeasible arguments, *i.e.*, arguments that are plausible but open to rebuttal. Each scheme represents a form of reasoning with typical premises and a conclusion" (Walton et al., 2008).[2] Most (62.4%) of the interventions are associated with a single argumentative scheme, although some may have up to six;

(3) A set $\mathcal{R}^j$ consisting of N^j annotated **reference questions** for each debate intervention j, where $j = 1 \ldots D$. Each reference question q_i^j is labeled with a label or category $l_i^j \in \{\text{Useful}, \text{Unhelpful}, \text{Invalid}\}$, such that $\mathcal{R}^j = \{(q_i^j, l_i^j) \mid i = 1, \ldots, N^j\}$. **Useful** questions can potentially challenge one of the arguments in the text; **Unhelpful** questions are valid but unlikely to challenge any of the arguments in the text; and **Invalid** questions cannot be used to challenge any argument in the intervention (Figueras, 2025b).

2.2 Task description

The task consists of automatically generating three *Useful* critical questions, $Qc^j = \{qc_1^j, qc_2^j, qc_3^j\}$ for each debate intervention j. In this context, critical questions are designed to evaluate the strength of an argument by revealing the assumptions underlying its premises (Figueras and Agerri, 2024). The usefulness of each generated critical question qc_i^j is evaluated by measuring its cosine similarity with the annotated reference questions $\mathcal{R}^j$. The label assigned to qc_i^j corresponds to the label of the most similar reference question, provided that it is larger than or equal to 0.6. If no similarity score exceeds this threshold, the question is marked as *Not able to evaluate*. In this case, human evaluators assessed the usefulness of the question during the competition. The final score was computed on the 34 interventions that composed the test set, $D_{\text{shared test}}$. Note that the reference test set, with the labels corresponding to the interventions in $D_{\text{shared test}}$ was not made available.

3 Methodology

As illustrated in Figure 1, the proposed system consists of two large language models (LLMs) used sequentially. (1) The `Questioner` (LLM_Q) which generates candidate critical questions given an intervention and its associated argumentation schemes; and the (2) The `Judge` (LLM_J) which evaluates these candidates and selects those deemed most useful (Li et al., 2024). This architecture is grounded in the framework of critical thinking proposed by Elder and Paul (2020), which comprises analytic, creative, and evaluative dimensions. We operationalize the creative components through LLM_Q (generation), and the analytic and evaluative components through LLM_J (selection).

3.1 The prompts

The prompts provided to the LLMs include: the intervention text, the role of the LLM (*i.e.*, `Questioner` or `Judge`), definitions of critical ques-

[2] See A.3 for a comprehensive list of the annotation schemes included in the dataset.

tion and argumentation scheme, the argumentative schemes present in the intervention along with their definitions (see A.3) and corresponding question templates (see A.4), the task objective, and the expected output. For more details, see A.2.3.

For LLM_Q, each prompt is designed to elicit N questions in a single generation step, rather than prompting the model N times for one question at a time. This strategy effectively reduces question repetition.

Aligned with Guo et al. (2023), we hypothesize that candidate questions exhibiting high similarity are likely to be useful. Thus, the following instruction is added to LLM_J's prompt: *If some questions are redundant, these questions must be important: select the most relevant one.* This modification led to an overall improvement in performance.

3.2 Experimental design

We split $D_{\text{shared train}}$ into training (D_{train}, 74), validation (D_{val}, 33), and test (D_{test}, 79) sets. The size of D_{test} was selected to ensure stable results under the automatic evaluation metric (see A.2.1). We conducted experiments on D_{test} by varying the following parameters to assess their impact on performance: the choice of LLM for each of the roles (`Questioner` and `Judge`), the number of candidate questions generated, and the temperature setting of the LLMs. Additionally, we performed an ablation study to evaluate the role of argumentation schemes in the generation process and address the added value of LLM_J by comparing it with alternative question selection strategies. For more details on the experimental setup and further experiments, including LLM and BERT fine-tuning and data augmentation, see A.2 and A.4.

4 Experiments and results

4.1 Model comparison

We evaluated both LLM_Q and LLM_J using a selection of small, open-source LLMs ranging from 7B to 14B parameters: Qwen 2.5 7B (Yang et al., 2024), Llama 3.1 8B (Dubey et al., 2024), Gemma 2 9B (Team et al., 2024), Gemma 3 12B (Team et al., 2025), and DeepSeek R1 14B (Guo et al., 2025) [3]. We compare their performance with that of GPT-4o (Achiam et al., 2023).

As shown in Table 1, the LLM combination yielding the highest proportion of useful outputs on D_{test} is Llama 3.1 8B as LLM_Q and Gemma 2 9B as LLM_J.

LLM_Q	LLM_J	Use. ↑	Inv. ↓	NoEval
Llama 3.1	-	53.2	3.0	33.8
Gemma 2	-	46.4	3.0	42.6
Gemma 3	-	40.5	**2.5**	46.0
Llama 3.1	Llama 3.1	53.2	3.9	33.0
Llama 3.1	Qwen 2.5	56.8	3.3	28.6
Llama 3.1	Gemma 2	**57.6**	5.2	30.3
Llama 3.1	Gemma 3	57.1	2.6	30.7

Table 1: **Performance on D_{test} for a selection of LLM_Q and LLM_J.** *Use*, *Inv* and *NoEval* are the % of *Useful*, *Invalid*, and *Not able to evaluate* questions, respectively. LLM_Q generates 8 questions of which LLM_J selects the best 3. The argumentative schemes are not given in the prompt. Best results in bold.

4.2 Leveraging argumentation schemes

To assess the impact on performance of adding argumentation scheme theory in the prompts for both LLM_Q and LLM_J, we conducted an ablation study. Table 2 compares the performance of LLM_Q (Llama 3.1 8B, generating six questions) and LLM_J (Gemma 2 9B) with the following configurations: **(1) Without**: No argumentation scheme is provided; **(2) With (one)**: All argumentation schemes relevant to the given intervention are included in a single prompt; **(3) With (mult.)**: Each argumentation scheme is provided in a separate prompt; and **(4) Both** : LLM_Q is prompted independently using the **With (one)** and without argumentation schemes setups. Then the two sets of candidate questions are merged for their selection by LLM_J . Similarly to previous work (Figueras and Agerri, 2024), the best performance is achieved in the **Both** configuration, suggesting that combining scheme-based and non-scheme-based prompts yields the most effective results. Note that 81% of the questions selected by LLM_J were generated with the argumentation scheme in the prompt.

Scheme	LLM_Q			$LLM_Q + LLM_J$		
	Use.↑	Inv. ↓	NoEval	Use.↑	Inv. ↓	NoEval
Without	54.7	3.2	32.7	57.7	3.8	27.4
With (one)	53.4	4.0	32.1	56.5	3.7	28.3
With (mult.)	46.0	4.0	27.0	51.6	3.6	34.2
Both	54.0	3.5	31.0	**62.4**	**2.1**	**25.7**

Table 2: **Performance on D_{test} with different argumentation schemes setups.** LLM_Q: Llama 3.1 generating 6 questions. *Without*: No argumentation scheme is provided;*With (one)*: Argumentation schemes are included in a single prompt; *With (mult.)*: Each argumentation scheme is provided in a separate prompt; and *Both*: LLM_Q is prompted independently with and without argumentation schemes. Best results in bold.

[3] For further details, see Section A.2.2.

4.3 Number of candidate questions

Table 3 presents the effectiveness of the questions as a function of the number of candidate questions generated per prompt. The experiment uses Llama 3.1 8B as LLM_Q—prompted both with and without the schemes—and Gemma 2 9B as LLM_J. Generating four candidate questions per prompt (eight in total) yielded the best performance.

# quest.	Use.↑	Inv. ↓	NoEval
4	$\mathbf{59.3 \pm 3.36}$	$\mathbf{2.80 \pm 1.05e^{-1}}$	$\mathbf{22.5 \pm 2.13}$
6	$57.2 \pm 8.82e^{-1}$	$2.72 \pm 6.38e^{-1}$	25.9 ± 2.91
8	$57.3 \pm 7.58e^{-1}$	$3.22 \pm 2.78e^{-1}$	$25.7 \pm 8.99e^{-1}$

Table 3: **Performance on $D_{\text{Shared train}}$ as a function of the number of candidate questions generated.** LLM_Q: Llama 3.1, LLM_J: Gemma 2. 3 runs.

4.4 Added value of the Judge, LLM_J

Although we observe an improvement in performance when adding LLM_J versus a random selection (see Tables 2 and 3), the results are not directly comparable, as the average usefulness is computed over different numbers of questions (N for LLM_Q and three for LLM_J). To further assess the effectiveness of LLM_J, we compared it against alternative selection paradigms. Table 4 reports the performance LLM_J versus a selection by an oracle and randomly, using Llama 3.1 8B as LLM_Q with four candidate questions per prompt. The oracle selects up to three useful questions. If fewer than three *Useful* questions are available, the remaining slots are filled by *Unhelpful* questions. If still insufficient, *Invalid*, and then *Not able to evaluate* questions are considered, in that order. The oracle illustrates the upper bound of the Judge's potential performance. Results show that LLM_J achieves a usefulness rate that is 3.4 percentage points higher than random selection, a statistically significant improvement ($p < 0.05$, McNemar's test). As expected, the oracle yields the highest usefulness with a gain of 34.2 percentage points.

Selection	Use.↑	Inv. ↓	NoEval
Random	55.9 ± 2.22	$2.7 \pm 7.94e^{-3}$	$25.7 \pm 2.5e^{-2}$
Gemma 2	59.3 ± 3.36	$2.8 \pm 1.05e^{-1}$	22.5 ± 2.13
Oracle	93.5 ± 1.19	$6.68 \pm 8.59e^{-2}$	$1.70 \pm 8.19e^{-1}$

Table 4: **Performance on $D_{\text{Shared train}}$ depending on the method to select the questions.** Comparison between random selection, selection with Gemma 2 as LLM_J or with an Oracle. In all cases, LLM_Q is Llama 3.1 generating $4 + 4$ questions. 3 runs.

4.5 Final submission

Based on the results of the previous experiments, we selected the following setup for our final submission: LLM_Q, Llama 3.1 8B, generating four questions without the scheme and four with the scheme, all within a single prompt; LLM_J, Gemma 2 9B, selecting the three best questions, used without fine-tuning. For comparison, we maintained the same experimental setup but substituted LLM_J with GPT-4o in our second submission, and in the third submission, GPT-4o was used for both LLM_Q and LLM_J under identical prompting conditions. Table 5 shows the performance with the automated evaluation on $D_{\text{Shared train}}$ and $D_{\text{Shared test}}$ for the three final submissions.

Sub.	Valiadation		Test	
	Use.↑	NoEval	Use.↑	NoEval
1	61.4	21%	36.3	36%
2	61.0	19%	44.1	36%
3	**64.4**	**19%**	**50.0**	**28%**

Table 5: **Performance with the automated evaluation on the validation set and the test set for the three final submissions.** Bold indicates the winning submission.

After the manual annotation of the questions by the organizers, the score of the best performing submission rose to **67.6**, ranking first in the task.

5 Discussion and conclusion

In this paper, we have proposed a two-step framework for generating critical questions, where one LLM (LLM_Q) generates multiple candidate questions and another LLM (LLM_J) evaluates and selects the most relevant ones (Li et al., 2024). This selection-based approach consistently outperformed direct generation, emphasizing the benefits of separating generation from evaluation. Our experiments show that adding argumentation schemes to the prompts improves the quality of the generated questions. However, strictly enforcing these schemes can reduce diversity. Thus, a selective use of schemes strikes a better balance between structural guidance and creative generation, in line with prior work (Figueras and Agerri, 2024).

Given the small size of the dataset, traditional strategies such as fine-tuning or data augmentation (*e.g.*, using BERT-based methods) yielded limited improvement. Instead, leveraging small, open-source LLMs guided by domain-specific argumentation theory proved more effective in this low-resource setting.

Limitations

This work has several limitations that should be acknowledged.

The first limitation concerns the evaluation methodology, which relies on an automatic comparison with a set of predefined reference questions by means of cosine similarity. Many generated questions did not align with any reference, despite being potentially useful, and hence were labeled as *Not able to evaluate*. This mismatch introduces a risk of mis-estimation of the model's performance and could lead to overfitting.

A second limitation stems from the use of a small, domain-specific dataset focused on political discourse, which at times lacks sufficient context for effective question generation. This narrow scope limits the generalizability of our findings. Future work should aim to evaluate the proposed framework on broader and more diverse datasets to assess its robustness across different domains, like education.

A third limitation lies in the performance of LLM_J, which shows a substantial gap compared to the oracle. This indicates that while some generated questions are *Useful*, the `Judge` does not consistently identify them, suggesting significant potential for future improvements.

Acknowledgements

L.F. and N.O. have been partially funded by a nominal grant received at the ELLIS Unit Alicante Foundation from the Regional Government of Valencia in Spain (Convenio Singular signed with Generalitat Valenciana, Conselleria de Innovación, Industria, Comercio y Turismo, Dirección General de Innovación). L.F. has also been partially funded by a grant from the Banc Sabadell Foundation.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Linda Elder and Richard Paul. 2020. *Critical thinking: Tools for taking charge of your learning and your life*. Rowman & Littlefield.

Lucile Favero, Juan Antonio Pérez-Ortiz, Tanja Käser, and Nuria Oliver. 2024. Enhancing critical thinking in education by means of a Socratic chatbot. *arXiv preprint arXiv:2409.05511*.

Blanca Calvo Figueras. 2024. Using argumentation theory to fight misinformation. *Doctoral Symposium on Natural Language Processing*.

Blanca Calvo Figueras. 2025a. Benchmarking Critical Questions Generation: A Challenging Reasoning Task for Large Language Models — arxiv.org. https://arxiv.org/abs/2505.11341. [Accessed 19-05-2025].

Blanca Calvo Figueras and Rodrigo Agerri. 2024. Critical questions generation: Motivation and challenges. *arXiv preprint arXiv:2410.14335*.

Maite Heredia Ekaterina Sviridova Elena Cabrio Serena Villata Rodrigo Agerri Figueras, Jaione Bengoetxea. 2025b. Overview of the critical questions generation shared task 2025. In *Proceedings of the 12th Workshop on Argument Mining*.

Michael Gerlich. 2025. AI tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies*, 15(1):6.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Shasha Guo, Jing Zhang, Xirui Ke, Cuiping Li, and Hong Chen. 2023. Diversifying question generation over knowledge base via external natural questions. *arXiv preprint arXiv:2309.14362*.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, et al. 2024. From generation to judgment: Opportunities and challenges of LLM-as-a-judge. *arXiv preprint arXiv:2411.16594*.

Jintao Ling and Muhammad Afzaal. 2024. Automatic question-answer pairs generation using pre-trained large language models in higher education. *Computers and Education: Artificial Intelligence*, 6:100252.

Nikahat Mulla and Prachi Gharpure. 2023. Automatic question generation: a review of methodologies, datasets, evaluation metrics, and applications. *Progress in Artificial Intelligence*, 12(1):1–32.

Ramon Ruiz-Dolz and John Lawrence. 2025. An explainable framework for misinformation identification via critical question answering. *arXiv preprint arXiv:2503.14626*.

Odin Monrad Schei, Anja Møgelvang, and Kristine Ludvigsen. 2024. Perceptions and use of AI chatbots among students in higher education: A scoping review of empirical studies. *Education Sciences*, 14(8):922.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

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

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

A Appendix

A.1 The datasets

Figure 2 shows the distribution of the number of annotated questions per intervention in $D_{\text{shared train}}$, Figure 3 shows the distribution of the number of schemes per intervention in $D_{\text{shared train}}$ and Figure 4 shows the distribution of the labels in $D_{\text{shared train}}$.

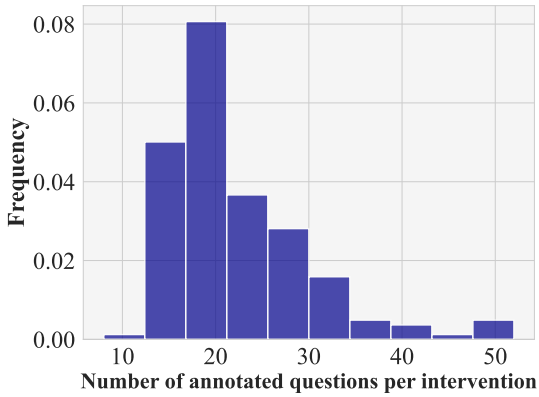

Figure 2: Distribution of the number of annotated questions per intervention in $D_{\text{shared train}}$.

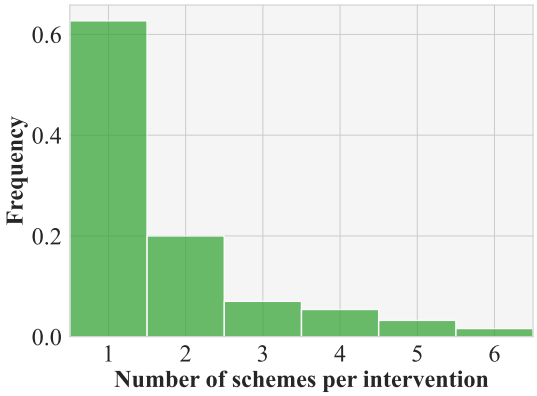

Figure 3: Distribution of the number of schemes per intervention in $D_{\text{shared train}}$.

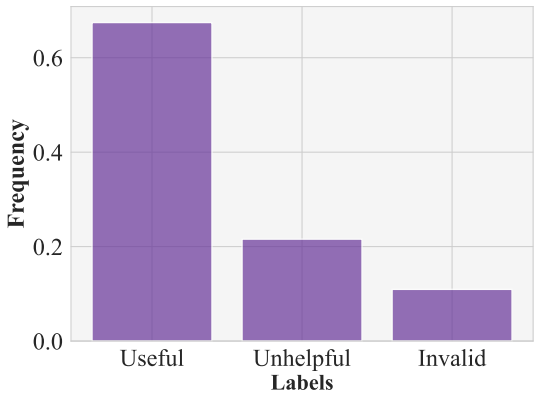

Figure 4: Distribution of the labels in $D_{\text{shared train}}$.

A.2 Experimental setup

Data split To enable training (see A.4), we divided $D_{\text{shared train}}$ (189 interventions) into three subsets: a training set D_{train} (74 interventions), a validation set D_{val} (33 interventions), and a test set D_{test} (79 interventions). The test set was intentionally large to accommodate the variability due to the automatic evaluation metric (see A.2.1), aiming for more stable and reliable results.

Software and hardware setup All experiments were performed on an Apple M1 Pro laptop with 32 GB RAM using Ollama [4], an open-source framework that enables users to run, create, and share LLMs locally on their machines. Our code is available at https://github.com/lucilefavero/SQ_shared_task.

[4] https://github.com/ollama/ollama, https://ollama.com.

A.2.1 Limitation of the automatic evaluation and human evaluation

Due to the nature of the automatic evaluation metric, a substantial proportion of generated questions could not be evaluated (see the column *No* in Tables 1, 2, and 3). Consequently, distinguishing the best-performing configurations was not straightforward. The performance differences across varying numbers of candidate questions were small and overshadowed by significant variance in our test set D_{test}. To improve precision, we repeated the experiment three times on $D_{\text{shared train}}$ (the entire dataset).

Throughout our experiments, we prioritized the quality of critical questions over minimizing the proportion of *Not able to evaluate* labels. We intentionally avoided overfitting to the automatic evaluation metric, under the assumption that some unevaluated questions might still be useful. The primary goal was to reduce the proportion of *Invalid* and *Unhelpful* labels. To ensure quality, we manually evaluate some questions that could not be evaluated by the automatic scoring system.

A.2.2 Additional information on the LLMs

A complete list of the tested models is provided below. The results for the most relevant combinations of LLM_Q and LLM_J are presented in Table 1.

- **Qwen 2.5 7B**. Qwen 2.5 is a multilingual transformer-based LLM with RoPE, SwiGLU, RMSNorm, and Attention QKV bias, released in September 2024 by the Qwen Team. (Yang et al., 2024).
- **Llama 3.1 8B**, Llama 3.1 is a multilingual large language model optimized for dialogue applications. It supports eight languages and offers a context window of up to 128,000 tokens, enabling it to handle extensive conversational contexts. Released in July 2024 by Meta (Dubey et al., 2024).
- **Gemma 2 9B**, Gemma 2 is a text-to-text decoder-only LLM available in English with open weights, released in June 2024 by Google, (Team et al., 2024).
- **Gemma 3 12B**, Gemma 3 is another model from the Gemma family; it has longer context, a different architecture than Gemma 2, and is trained with distillation. It was released in March 2025 by Google, (Team et al., 2025).
- **DeepSeek R1 14B**. DeepSeek R1 is an open-source large language model designed to enhance reasoning capabilities through reinforcement learning. It rivals other advanced models in tasks such as mathematics, coding, and logical reasoning. Released in January, 2025 by the Chinese AI startup DeepSeek (Guo et al., 2025).

A.2.3 Structure of LLM's prompts

The prompt for LLM_Q consists of the following components:

- **The intervention**.
- **Role.** "You are a critical judge."
- **Definition of critical question**: "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."
- **Definition of argumentation scheme.** "Argumentative schemes are stereotypical patterns of inference that capture common types of defeasible arguments, i.e. arguments that are plausible but open to rebuttal. Each scheme represents a form of reasoning with typical premises and a conclusion."
- **The argumentation schemes present in the intervention with their definition and template of critical questions** see A.3
- **Goal.** "Use the provided scheme and template of critical questions to generate [N] critical questions to evaluate the arguments in the given essay."
- **Expected output.** "Give one question per line. Make the questions simple, and do not give any explanation regarding why the question is relevant."

The prompt for LLM_J consists of the following components:

- **The intervention**.
- **Role.** "You are a very strict critical and sceptical judge."
- **Definition of critical question**: "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."

- **Definition of argumentative scheme.** "Argumentative schemes are stereotypical patterns of inference that capture common types of defeasible arguments, i.e. arguments that are plausible but open to rebuttal. Each scheme represents a form of reasoning with typical premises and a conclusion."
- **The argumentation schemes present in the intervention with their definition and template of critical questions** see A.3
- **Goal.** "Select the 3 best critical questions that should be raised before accepting the arguments in the essay. If some questions are redundant, these questions must be important: select the most relevant one."
- **Expected output.** "Give one question per line. Make the questions simple, and do not give any explanation regarding why the question is relevant."

A.3 Argumentation scheme definition and template

Table 6 depicts the argumentation schemes identified in the dataset, along with their corresponding critical question templates. The definitions and templates are adapted from Walton et al. (2008).

A.4 Further experiments

Critical questions' templates We examined two approaches for incorporating critical question templates into the prompts: using the template provided by Figueras and Agerri (2024), and utilizing the critical question templates outlined in Table 6. We noted a slight performance improvement with the first template for the configurations employing LLM_Q: Llama 3.1, LLM_J: Gemma 2 or GPT-4o, and with the second template for the configurations involving LLM_Q: GPT-4o and LLM_J: GPT-4o. This approach was adopted in our final submission.

Temperature We also explored modifying the generation temperature from its default setting and observed an overall decrease in performance.

Fine-tuning Attempts to fine-tune both LLM_Q and LLM_J were inconclusive. The resulting model outputs often diverged significantly from the intended instructions and demonstrated poor performance, likely attributable to task complexity combined with the limited size of the training dataset D_{Train}.

We also fine-tuned BERT (Devlin, 2018) to classify candidate questions into three categories: *Useful*, *Unhelpful*, and *Invalid*, selecting the three questions with the highest predicted probability of being *Useful*. However, similar to the LLM fine-tuning, the model failed to outperform a random baseline, likely due to task complexity and the limited size of D_{Train}.

Data augmentation The poor performance of the trained approaches is likely attributable to the limited and highly imbalanced annotated dataset. Specifically, over 67% of annotations in $D_{\text{Shared train}}$ are labeled as *Useful* (see Figure 4). To address this, we augmented D_{Train} with Llama 3.1, generating questions and matching them to reference annotations to balance label distribution across interventions. However, fine-tuning the LLMs and BERT on the augmented data still yielded inconclusive results.

A.5 Further comments on the results

The automatic evaluation scores on $D_{\text{shared test}}$ are lower compared to those obtained on $D_{\text{shared train}}$, likely due to increased variance arising from the test set's smaller size. Additionally, the shared-task organizers indicated, after human evaluation, that the test set presented a higher difficulty.

Name	Scheme's definition	Critical questions template
Ad Hominem	This scheme attacks an opponent's argument by alleging inconsistency between their actions and their stated position.	Is the alleged inconsistency real and relevant to the argument? Does the inconsistency undermine the argument's validity? Could the argument still hold despite the personal inconsistency?
Alternatives	This scheme reasons that one option should be chosen (or avoided) by comparing it to other possible options.	Have all relevant alternatives been considered? Are the alternatives fairly evaluated? Is the chosen alternative clearly superior based on the criteria?
Analogy	This scheme draws a conclusion about one case by comparing it to a similar case where the conclusion is known to hold.	Are the two cases sufficiently similar in relevant respects? Are there significant differences that undermine the analogy? Is the conclusion in the known case well-established?
Bias	This scheme attacks an argument by alleging that the source is biased, thus undermining its credibility.	Is there clear evidence of bias in the source? Does the alleged bias directly affect the truth of the argument's conclusion? Could the argument still hold despite the bias?
Cause to effect	This scheme reasons that if a certain cause occurs, it will lead to a specific effect, based on a causal relationship.	Is there sufficient evidence that the cause reliably produces the effect? Could other factors intervene to prevent the effect from occurring? Is the causal link based on correlation rather than proven causation?
Consequences	This scheme bases a conclusion on the positive or negative outcomes of a proposed action, arguing for or against it based on those consequences.	Are the predicted consequences likely to occur if the action is taken? Are there other consequences (positive or negative) that haven't been considered? Is the evaluation of the consequences as good or bad justified?
Example	This scheme involves reasoning from a specific case or instance to a general conclusion, suggesting that what holds in the example applies more broadly.	Is the example representative of the broader category or situation? Are there significant counterexamples that undermine the generalization? Is the example relevant to the conclusion being drawn?
Expert opinion	This scheme concludes that a proposition is true because an expert in the relevant field asserts it.	How credible is the expert as a source? Is the expert an authority in the field relevant to the proposition? What exactly did the expert assert? Is the expert personally reliable and trustworthy? Is the expert's claim consistent with other experts? Is the expert's assertion backed by evidence?
Fear and danger appeals	This scheme urges action or avoidance based on the fear of a harmful outcome if the action isn't taken or is taken.	Is the feared outcome realistically likely to occur? Is the fear disproportionate to the evidence of danger? Are there other ways to mitigate the feared outcome without the proposed action?
Negative consequences	This scheme argues against an action because it will lead to bad outcomes.	Are the negative consequences probable? Are there positive consequences that might offset the negative ones? Is the judgment of the consequences as negative reasonable?

Continued on next page

Name	Scheme's definition	Critical questions template
Popular opinion	This scheme argues that a proposition is true or should be accepted because it is widely believed by the majority.	Is the opinion truly held by a significant majority? Does the majority have reliable evidence or expertise to justify their belief? Could the majority be mistaken or influenced by bias?
Popular practice	This scheme justifies an action or belief because it is commonly practiced by many people.	Is the practice widespread enough to be considered popular? Does the practice's popularity indicate its correctness or value? Are there reasons the practice might be flawed despite its popularity?
Positive consequences	This scheme argues for an action because it will plausibly lead to good outcomes.	Are the positive consequences likely to occur? Are there potential negative consequences that outweigh the positive ones? Is the assessment of the consequences as positive well-founded?
Position to know	This scheme concludes a proposition is true because the source is in a position to know about it (e.g., firsthand experience).	Is the source genuinely in a position to know about the proposition? Is the source honest and trustworthy? Did the source actually assert the proposition?
Practical reasoning	This scheme involves an agent reasoning from a goal to an action that is a means to achieve that goal (e.g., "I want G, doing A achieves G, so I should do A").	What other goals might conflict with G? Are there alternative actions to A that could also achieve G? Is A the most efficient means to achieve G? Is it practically possible for me to carry out A? What are the potential side effects or consequences of doing A?
Sign	This scheme infers a conclusion based on an observable sign or indicator that suggests the presence of a condition or event.	Is the sign a reliable indicator of the conclusion? Could the sign be present without the conclusion being true? Are there alternative explanations for the sign?
Value	This scheme reasons that an action should be taken or avoided because it aligns with or conflicts with an agent's values (e.g., "V is good, so I should pursue G that promotes V").	Is value V genuinely positive/negative as judged by the agent? Does pursuing V conflict with other values the agent holds? Is the link between the action and the promotion of V well-supported?
Verbal classification	This scheme applies a general rule or property to a specific case based on how the case is classified linguistically.	Is the classification of the case accurate and appropriate? Does the general rule reliably apply to all cases under this classification? Is the classification ambiguous or contested?

Table 6: Argumentation schemes identified in the dataset, along with their corresponding critical question templates. Definitions and templates are adapted from Walton et al. (2008).