Overview of the Critical Questions Generation Shared Task

Blanca Calvo Figueras[♠], **Jaione Bengoetxea**[♠], **Maite Heredia**[♠],
Ekaterina Sviridova[♣], **Elena Cabrio**[♣], **Serena Villata**[♣], **Rodrigo Agerri**[♠]
[♠]HiTZ Center - Ixa, University of the Basque Country, UPV/EHU
[♣]Université Côte d'Azur, Inria, CNRS, I3S, France
blanca.calvo@ehu.eus rodrigo.agerri@ehu.eus

Abstract

The proliferation of AI technologies has reinforced the importance of developing critical thinking skills. We propose leveraging Large Language Models (LLMs) to facilitate the generation of critical questions: inquiries designed to identify fallacious or inadequately constructed arguments. This paper presents an overview of the first shared task on Critical Questions Generation (CQs-Gen). Thirteen teams investigated various methodologies for generating questions that critically assess arguments within the provided texts. The highest accuracy achieved was 67.6, indicating substantial room for improvement in this task. Moreover, three of the four top-performing teams incorporated argumentation scheme annotations to enhance their systems. Finally, while most participants employed open-weight models, the two highest-ranking teams relied on proprietary LLMs.[1]

1 Introduction

In recent years, educators and researchers have expressed growing concern that the widespread use of LLM–based chat systems may encourage superficial learning habits and undermine the development of critical thinking skills (Hadi Mogavi et al., 2024). In response to this challenge, our shared task explores a novel approach: leveraging LLMs to promote critical thinking. Specifically, we propose using LLMs to guide users in formulating critical questions or, in other words, questions aimed at uncovering fallacious reasoning or poorly constructed arguments (Calvo Figueras and Agerri, 2024).

To this end, we introduce the task of Critical Questions Generation (CQs-Gen), a generative task that involves automatically generating useful critical questions for argumentative texts. Critical questions (CQs) are inquiries used to evaluate whether an argument is valid or flawed. These questions serve to expose underlying and spurious assumptions in the argument's premises and to question the inferential relations between premises and conclusions, thereby enabling the identification of potential fallacies and weaknesses in the reasoning (Walton et al., 2008).

This research extends prior work's findings on the fact that critical questions may serve as effective tools for fighting misinformation (Musi et al., 2023) and evaluating argumentative essays (Song et al., 2014) by exposing flawed argumentative structures. Additionally, it draws upon established research in argumentation scheme taxonomy development and argumentative text annotation (Walton et al., 2008; Wachsmuth et al., 2017; Macagno et al., 2017; Visser et al., 2021).

This paper presents an overview of the shared task, including a detailed discussion and analysis of the results obtained by the participants. The shared task was launched in February 2025. Participants were given one and a half months to work on the validation set and later evaluated their systems in the test set. The reference questions for the test set remained hidden. A total of 19 teams registered for the task, 13 submitted system outputs for evaluation, and 12 provided corresponding system descriptions.

Key findings from this shared task include: (i) model selection has a greater impact than prompt engineering, though the benefits of model scaling are limited; (ii) using argumentation schemes helps generate more useful CQs but tends to reduce question diversity; and (iii) there remains significant room for improvement in the task of CQs-Gen.

2 Task Description

To facilitate systematic experimentation and evaluation for the task of CQs-Gen, we developed a dataset comprising debate interventions paired with

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

Proceedings of the 12th Argument Mining Workshop, pages 243–257

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?

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

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?

(c) **Output**: This set of questions would get $0.\overline{33}$ points for the useful CQ, 0 for the CQ that is unhelpful, and 0 for the invalid one. Therefore, this answer has a $0.\overline{33}$ punctuation.

Figure 1: Example of candidate outputs with its labels: Useful (USE), Unhelpful (UN), and Invalid (IN).

corresponding critical questions. The dataset contains real debate interventions that have been annotated with argumentation schemes and associated critical question sets. Each intervention includes the speaker's identity, argumentation scheme classifications, and critical questions. The critical questions are labeled according to their effectiveness in challenging the presented arguments, using three categories: Useful, Unhelpful, or Invalid.

In this shared task, participants are asked to develop a system that gets one of the interventions as input and produces three insightful critical questions. The generated questions are compared to the annotated gold references of each intervention, and each generated question inherits the label of the most similar reference CQ. Note that each of these questions is evaluated separately, and that the punctuation is then aggregated.

Useful critical questions are given $0.\overline{33}$ points, and Unhelpful and Invalid CQs get 0 points. Therefore, each output containing 3 CQs is given a score between 0 and 1, depending on the usefulness of the generated questions. The definitions of these categories are described in Section 3. Figure 1 shows two evaluated possible outputs for a given intervention.

Participants were encouraged to engage in the task by either generating multiple critical questions and choosing the top 3, or by developing a system that only outputs useful critical questions.

2.1 Evaluation

The evaluation is performed by comparing the newly automatically generated critical questions with the reference critical questions in the dataset using semantic textual similarity (STS) with Sentence Transformers (Reimers and Gurevych, 2019). The new question inherits the label of its most similar reference. We employ *stsb-mpnet-base-v2* embeddings with a threshold of 0.65, as this approach demonstrated the highest correlation with human evaluation among non-LLM methods in previous research (Calvo Figueras and Agerri, 2025).

If the newly generated questions do not match any of the reference critical questions, in other words, the similarity between the new CQ and each reference CQ is lower than 0.65, the question is tagged as *Not Able to Evaluate* (NAE). In these cases, NAE questions are manually evaluated following the same annotation guidelines employed during reference CQ annotation.

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:

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

2.2 Baselines

The baselines for the shared task were developed by prompting two LLMs to generate critical questions with the requirement of avoiding non-useful questions (prompt in Appendix A). We used the default hyperparameters and HuggingFace's implementations of:

- run 1: Qwen2.5-VL-72B-Instruct (Qwen et al., 2025)
- run 2: gemma-2-9b-it (Team et al., 2024)

Baseline models undergo identical evaluation procedures as the participants' submissions to ensure methodological consistency and prevent preferential treatment.

3 Data

For this shared task, we use the CQs-Gen benchmark (Calvo Figueras and Agerri, 2025), which was built on top of 4 existing datasets: US2016 (Visser et al., 2021), Moral Maze Debates (Lawrence et al., 2018), RRD (Konat et al., 2016), and US2016reddit. Our CQs-Gen dataset contains 220 argumentative texts, associated with 22.4 questions on average (theoretical and LLM-generated), which have been manually annotated. The dataset is divided between a validation set (186 texts) and a test set (34 texts). The reference questions of the test set are kept private to avoid data contamination (Sainz et al., 2023). Table 1 shows the stats of these two splits.

Set	Nº Int.	Nº CQs	%USE	%UN	%IN
Valid.	186	4,136	67.46	21.59	10.95
Test	34	806	42.68	31.02	26.30

Table 1: Stats of the dataset.

These questions have been annotated by journalists specialized in detecting misinformation. They were asked, "Can this question be used to undermine the arguments given in the intervention?". If they considered that the question is not useful, they could choose between two possibilities: the question not being valid, or the question being unhelpful.[2] The three categories are described in the guidelines as follows:

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).

[2]The guidelines can be found in https://github.com/hitz-zentroa/shared-task-critical-questions-generation/blob/main/shared_task/utils/guidelines.pdf

3.1 Training Phase

In the training phase, the participants worked with the validation set (184 interventions). For this dataset, we released the reference CQs and their annotations, the argumentation schemes, and their datasets of origin. This phase lasted a month and a half.

3.2 Evaluation Phase

During evaluation, participants processed the test set containing 34 interventions without access to reference critical questions. Each participant submitted a maximum of three result runs to the organizers, who assessed the generated outputs using the similarity-based automated metric against withheld reference standards. The highest-scoring run from each participating group underwent manual evaluation, wherein three human annotators, the task organizers, assigned labels to the NAE cases based on the established annotation guidelines.

4 Submissions

Nineteen different teams registered for the shared task, of which thirteen submitted their system outputs during the evaluation phase, and twelve submitted a system description, eleven of which were archival submissions. A summary of the submitted papers is provided below.

4.1 System description

ELLIS Alicante Favero et al. (2025) won the shared task. Their approach uses a *Questioner* LLM that generates the questions together with a *Judge* LLM that picks the best 3 candidates. For both components of their pipeline, they try multiple configurations, such as different models, using argumentation schemes, and different numbers of candidates. Their best system uses GPT-4o (OpenAI et al., 2024) for both components, and generates 8 candidates per text, half of which have to be connected to the argumentation schemes.

COGNAC Anjum Islam et al. (2025) experiment with RAG-based approaches to (1) select example CQs of similar interventions for few-shot, (2) incorporate argumentation schemes descriptions, and

(3) again perform few-shot but this time using both the intervention and the reference CQs for computing similarity. Their best system uses approach 1 and GPT-4o-mini. They acknowledge this system might have constrained generalization to texts outside of the dataset.

StateCloud Zhang et al. (2025) investigate both prompt-engineering and model ensembling. Regarding prompts, they try zero-shot, few-shot, oral-expression, and shuffling the order of the instructions. They report that the improvement observed with prompt engineering is largely overshadowed by model selection. Regarding model ensembling, they experiment with sequential and parallel ensembling. Generating CQs by sequentially prompting different models obtains the best results in the validation set. However, their best performing system in the test set combines Qwen2.5-72B, QwQ-32B, and DeepSeek R1 in parallel (Team, 2025; Yang et al., 2024; DeepSeek-AI, 2025).

DayDreamer Zhou et al. (2025) are the most theory-based team. They develop a pipeline that (1) conversationally builds the structure of the arguments in the text (they fill the templates of the argumentation schemes), to then (2) generate the candidate questions, and finally (3) rank them regarding usefulness to choose the top 3. They use LLMs for all these steps, and achieve their best results with GPT-4o-mini. Their analysis shows that the quality of the scheme template can have a great impact on their pipeline.

Webis Kanadan et al. (2025)'s approach consists of a pipeline with two components, namely: (1) a generation LLM, and (2) an encoder fine-tuned to label the questions as Useful or Not-Useful. They try different prompting strategies, such as guidelines-based, chain-of-thought, and few-shot. Additionally, they create new data for fine-tuning. They achieve their best results using Gemma2-9B and Phi-4-14B for generating 3 questions each, and fine-tuning ModernBERT with their new data (Warner et al., 2024).

TriLLaMa Turkstra et al. (2025) explore the use of various sizes of Llama-3.1 models (Grattafiori et al., 2024) for both the generation and classification of CQs. The most effective configuration employs Llama-3.1-405B for few-shot question generation and Llama-3.1-70B for zero-shot classification. Furthermore, two innovative classification strategies yield notable results: (1) debate classification, in which two models debate the usefulness of a CQ and a third model adjudicates the winner; and (2) deliberation classification, where two models vote on the questions they consider to be most helpful.

Mind_Matrix Mahmud et al. (2025) develop a pipeline where critical questions are generated using LLMs and then ranked according to a score. This score is the product of the similarity of the question to the intervention, the coherence, and the relevance. All these metrics are computed using Sentence Transformers. Finally, a last module of the pipeline uses another LLM to rank the effectiveness of the chosen CQs, if the average score of the 3 CQs is lower than a threshold, they rerun the pipeline, lowering the temperature by 0.1. Their best-performing model is DeepHermes-3-Llama-3-8B (Teknium et al., 2025).

CriticalBrew El Baff et al. (2025) employ a machine society simulation approach (Zhang et al., 2024). This approach creates a network of agents that collaborate to reach a conclusion using different personality traits and thinking patterns. They experiment with different configurations and conclude that the number of agents and the thinking patterns significantly impact the results, while the personality traits do not. They also experiment with ranking the CQs using LLMs with different prompts. Their best-performing model uses Mistral-24B with 2 easy-going agents and an overconfident one, with three rounds of discussion, two debate rounds, and one discussion round.

Tndguyen Nguyen and Nguyen (2025) investigate multi step reasoning techniques, namely: chain-of-thought prompting in zero-shot and one-shot, least-to-most settings, and tree-of-thought prompting. Their results show that structured prompting consistently offers performance gains. Their best-performing system is zero-shot chain-of-thought prompting with GPT-4o. Their error analysis highlights that models tend to fail in long and multi-topic interventions, as well as those involving emotionally charged and subjective content.

ARG2ST Ramponi et al. (2025) build a pipeline with an LLM for generation and a fine-tuned encoder for classification. In the generation step, they experiment with modularly extending the prompts with argumentation schemes, guideline descriptions, and few-shot examples. Their ablation study

shows that incorporating a classifier consistently improves the performance of their systems and that providing explanations in the prompt on not-useful CQs works better than providing instructions on which ones are useful. Their best performing system is a Llama-3-70B for generation and a BERT-base as classifier (Devlin et al., 2019).

CUET_SR34 Bhattacharjee et al. (2025) finetune a Llama-3-8B model using LoRa on a 10% of the validation set and a few-shot prompt. They show that simplifying the text, adding named entities (NER), and the names of the argumentation schemes, gets their best results.

Nompt This team focuses exclusively on two issues in CQs-Gen, namely: questions that do not target the speaker's arguments and instead are in line with them; and questions that introduce new concepts related to some named entity mentioned but not present in the text. They propose 4 stages: (1) argument scheme extraction, (2) speaker anonymization, (3) main points extraction, and (4) questions generation using only the extracted information. While their overall results are low, their manual evaluation shows that their approach solves the phenomena they are targeting.

5 Results

The official results of the shared ask are reported in Table 2. Based on the automatic evaluation metric, *COGNAC* was initially ranked first. However, manual evaluation revealed *ELLIS Alicante* as the winner of the shared task, with *StateCloud* in third place. *ELLIS Alicante* employed a strategy that prioritized minimizing Unhelpful and Invalid question categories rather than optimizing for maximum overall scores through NAE reduction. This approach demonstrated superior generalization capabilities on the test set, an effect that was only detectable through manual assessment.

Interestingly, 3 out of the 4 best-performing teams (*ELLIS Alicante*, *COGNAC*, and *DayDreamer*) reported that argumentation schemes improved system performance. Two other teams, *Webis* and *ARG2ST*, also experimented with argumentation schemes but excluded them from their final submissions, determining that the performance gains were insufficient to warrant inclusion. We examine the various methods by which this information is integrated in Section 6.

Multiple teams reported that differences between models had a greater impact than prompt variations. Thus, three of the top 4 teams (*ELLIS Alicante*, *COGNAC*, and *DayDreamer*) relied on GPT-4o or GPT-4o-mini for their submissions. Unsurprisingly, submissions with lower performance typically used smaller and open-weight models, which represented the most frequently selected option among participants.

Several teams implemented classification or ranking methodologies to select the most promising critical questions. All teams using these approaches reported performance improvements. We analyze their methods in Section 6.

Team Name	Run	Score	Auto. score
ELLIS Alicante	3	**67.6**	50.0
COGNAC	1	**62.7**	61.8
StateCloud	3	**59.8**	47.1
DayDreamer	1	**58.8**	55.9
Webis	2	**56.9**	52.0
TriLLaMa	1	**55.9**	53.9
Mind_Matrix	1	**55.9**	42.2
CriticalBrew	1	**54.9**	52.0
Lilo&stitch*	2	**53.9**	49.0
baseline	2	**52.9**	52.0
Tdnguyen	1	**52.0**	49.0
ARG2ST	2	**50.0**	45.1
CUET_SR34	1	**48.0**	43.1
baseline	1	**44.1**	41.2
Nompt	1	**38.2**	29.4

Table 2: Official results of the shared task.
* This team did not submit a system description. Therefore, we do not discuss it.

6 Additional Evaluations and Analysis

The following sections provide supplementary automated evaluation methods and analyses beyond the official results to examine the implications of participant design choices.

Results on the validation set. During system development, participants had access only to the validation set, which informed their design decisions. Comparison of validation and official test set results (see Table 4 in Appendix) revealed no clear correlation between performance across these datasets. However, each team's best-performing validation submission typically remained their strongest on the test set. Qualitative insights from

validation analysis transferred effectively to test performance, with teams conducting more comprehensive qualitative evaluations achieving superior final results.

LLM- and data-enhanced evaluation. Calvo Figueras and Agerri (2025) introduced two evaluation methods that demonstrated better correlation with human judgment and an expanded test set. The extended test set incorporated manual evaluations from this shared task to increase reference CQs and reduce NAE values. The proposed evaluation methods employed LLMs (Claude 3.5 Sonnet[3] and Gemma-2-9B-it (Team et al., 2024)) to compare reference and generated CQs, assigning labels to new questions based on matching references. These automated approaches enabled a comprehensive evaluation of all team submissions, with results presented in the Appendix, Table 5.

The extended dataset includes manually-evaluated questions solely from each team's top-performing run, yielding accurate assessments for these runs but potentially undervaluing other runs. Analysis in Table 5 confirms *ELLIS Alicante*'s third run as the winner across all evaluation metrics. High-performing submissions absent from Table 2 include *COGNAC*'s second and third runs, indicating reliable submission standards from this team.

The effect of incorporating argumentation schemes. Several teams investigated incorporating argumentation schemes as additional context, with three of the four top-performing systems implementing them. However, effective integration proved challenging, as several teams reported no improvement or performance degradation. *ARG2ST* found no benefit from simple scheme references, while *Webis* included both scheme descriptions and associated CQs structures but similarly observed no gains. *Webis*'s analysis revealed that generated questions became excessively rigid and templated when using this approach.

Similarly, *ELLIS Alicante* found that argumentation scheme descriptions improved question quality while decreasing diversity. Their winning run combined dual prompts, with and without scheme descriptions, and employed a third LLM (also informed with argumentation schemes) to select optimal questions from the combined outputs.

[3]Version *claude-3-5-sonnet-20241022*, https://www.anthropic.com/news/claude-3-5-sonnet

Team Name	n-gram	CR-div	USE CR-div
ELLIS Alicante	**3.04**	**0.373**	0.405
COGNAC	2.52	0.287	0.306
StateCloud	2.96	0.350	0.388
DayDreamer	2.71	0.320	0.364
Webis	2.97	0.352	0.392
TriLLaMa	3.01	0.366	0.403
Mind_Matrix	2.84	0.350	0.387
CriticalBrew	2.90	0.349	0.390
Lilo&stitch	2.92	0.353	0.392
Tdnguyen	3.00	0.356	0.400
ARG2ST	**3.09**	**0.383**	0.429
CUET_SR34	2.74	0.331	0.369
Nompt	**3.12**	**0.388**	0.459

Table 3: Diversity scores (the higher the more diverse). Averaged scores between the 3 runs submitted by each team. The last column measures diversity within the Useful CQs.

COGNAC and *DayDreamer* developed two additional effective approaches for incorporating argumentation schemes. *COGNAC* paired scheme descriptions with custom CQ sets from similar interventions to reduce rigidity; *DayDreamer* centered their approach on scheme-based argument extraction followed by CQ generation. *DayDreamer*'s poor performance with *ERPracticalReasoning* and *ERExpertOpinion* schemes indicated that description quality critically affects system effectiveness.

Diversity analysis. Although many system descriptions mention diversity, none of them provide a quantitative assessment of it. To address this gap, we now analyze diversity explicitly. Following Calvo Figueras and Agerri (2025), we adopt two metrics: n-gram diversity and Compression Ratio Diversity, as defined in Shaib et al. (2025). In Table 3, these metrics are applied to the three submissions of each team and then averaged.

The results indicate that *ELLIS Alicante*, the winning team, produced one of the most diverse sets of critical questions, both overall and when considering only the Useful CQs (see last column). According to their system description, they qualitatively analyzed diversity during development and observed that using argumentation schemes for all CQs reduced diversity. Consequently, they opted to include schemes in their prompts for only half of their generated questions. Additionally, despite

ranking lower on the performance leaderboard, both *ARG2ST* and *Nompt* also achieved high diversity scores.

In contrast, the least diverse sets of CQs were produced by *COGNAC* and *DayDreamer*, two of the top-performing teams. *DayDreamer* generated their questions using the theoretical CQs in all their runs, which likely contributed to the low variation in their outputs. Similarly, two of *COGNAC*'s submissions used prompts that included argumentation scheme descriptions, which may have constrained their output diversity.

These findings support the conclusion drawn by *ELLIS Alicante*: incorporating argumentation schemes involves a trade-off between usefulness and diversity. Teams aiming to optimize both may benefit from hybrid strategies that balance structured guidance with open-ended generation.

Ranking and classification modules. All teams that incorporated ranking or classification modules into their pipelines reported positive effects. However, since these modules were evaluated differently across teams, it is difficult to quantitatively compare their individual impact. Instead, we describe each team's approach in order, starting with the highest-scoring system.

1. *ELLIS Alicante* prompts an LLM to select 3 out of 8 questions generated by two other models. The prompt includes detailed guidelines and explicitly encourages the model to favor repeated questions, suggesting that repetition may indicate relevance.

2. *DayDreamer* also uses an LLM, but instead of directly selecting questions, they prompt it to rank the candidates and then choose the top 3 from the ranking.

3. *Webis* fine-tune an encoder (ModernBERT), using a dataset of 67.8k additional LLM-generated critical questions, labeled using the official evaluation script. They select the questions classified as Useful with the highest confidence.

4. *TriLLaMa* implements an original classification approach whereby multiple LLMs engage in debate and voting to classify candidate questions. However, this multi-model strategy yields inferior performance compared to zero-shot classification using a single LLM. The authors' analysis indicates that the debate process causes models to lose positional coherence and attempt simultaneous evaluation of all 10 candidates, leading to errors.

5. *Mind_Matrix* applies a handcrafted ranking method, computing a score based on similarity to the intervention, coherence, and relevance, using SentenceTransformers.

6. *CriticalBrew* prompts LLMs to perform selection, scoring, and ranking. They conclude that the specific method used (choosing vs. scoring vs. ranking) does not significantly affect overall performance.

7. *ARG2ST* adopts an approach similar to that of *Webis*, but with key differences: they use BERT-base as the encoder and do not incorporate any additional training data.

Overall, prompt-based approaches tend to yield better results than encoder-based ones. Additionally, several teams report that their classifiers exhibit a bias toward labeling most questions as Useful, a pattern consistent with findings in Calvo Figueras and Agerri (2025).

Few-shot approaches. Including examples in the prompt (i.e., few-shot learning) yielded mixed results. While some teams found it beneficial, others observed no improvement or even performance drops. A few teams included examples in their prompts but did not conduct ablation studies; therefore, we exclude those cases from this discussion.

The most successful applications of few-shot prompting came from *COGNAC*, *TriLLaMa*, and *Tndguyen*. Both *COGNAC* and *Tndguyen* employed a dynamic example selection strategy: for each test instance, they retrieved a similar intervention and used its reference CQs as examples. Additionally, *Tndguyen* augmented this with an automated reasoning path, illustrating how the example CQ related to the intervention. While these techniques are interesting, they might not generalize well to interventions about new topics. In contrast, *TriLLaMa* used a fixed set of three examples presented in a conversational format, where the user provided the intervention and the assistant replied with example CQs.

On the opposite end, *ARG2ST* reported consistent performance degradation when using few-shot prompting. Their approach involved a fixed intervention example, tested with either Useful CQs only or examples from all three categories.

StateCloud followed a similar setup and ultimately chose not to use few-shot, although their results were less conclusive.

A likely explanation for these mixed outcomes is prompt length: including examples significantly increases the prompt size, which can confuse smaller models such as those used by *ARG2ST*. In contrast, the three teams that benefited from few-shot learning used much larger models. Interestingly, *TriLLaMa* did observe some gains with smaller models, suggesting that their conversational prompt structure may have been more manageable for models with limited capacity.

Reasoning approaches Reasoning techniques are currently a topic of active research. Given that CQs-Gen involves substantial reasoning, several teams explored these advanced approaches.

Tndguyen experimented with various reasoning strategies, including Chain-of-Thought (CoT) prompting in zero-shot, few-shot, and least-to-most formats, as well as a Tree-of-Thought approach. Despite the potential of these methods, their performance gains are unexpectedly small. Their analysis identifies several categories of failure: long, multi-topic interventions; emotionally charged content; overly short texts; sensitive topics; and satirical content. *Webis* also explores CoT, but does not use it in their final prompt either.

In a different direction, *StateCloud* evaluated the performance of state-of-the-art open-weight reasoning models (DeepSeek-R1-671B and QWQ-32B). While these models did not surpass general-purpose LLMs in terms of the percentage of Useful CQs, they produced remarkably more NAE values. This suggests that the models may be generating novel, useful critical questions that lie outside the scope of the current reference CQs. However, without further analysis, this remains a hypothesis. A similar pattern was observed in the system developed by *ELLIS Alicante*, which used GPT-4o (a closed-weight reasoning model) and also generated a large number of NAE values that were later found to be useful CQs. These observations suggest that the capabilities of advanced reasoning models in CQs-Gen may currently be underestimated and highlight the need for further manual evaluation and expanded reference sets to fully capture the quality of their outputs.

Model performance Many teams observed that model selection has a greater impact on performance than prompt engineering in this task. However, several teams also noted that the benefits of scaling to larger models are surprisingly limited, often not justifying the increased computational cost. This aligns with the findings reported in Calvo Figueras and Agerri (2025).

Error analysis. Figure 2 shows the distribution of the scores per intervention of all the submissions, as evaluated in Table 5 (Appendix) with Claude. Most interventions received both maximum and zero scores, though difficulty varied considerably. While two interventions had zero as the modal score, most exhibited a median score of 0.66. *RRD* and *US2016* interventions showed distributed score ranges, whereas *Moral Maze* interventions were predominantly difficult. This pattern is interesting given that *Moral Maze* contains the most comprehensive and technical content, but addresses complex topics such as 'state intervention legitimacy' and 'loan morality'. The remaining two datasets featured more accessible topics but employed emotional language and rhetorical devices. Error analysis focused on the six lowest-scoring interventions.

The first intervention (*CLINTON_47*) is very short. In this intervention, Clinton is claiming that she did not support the NAFTA agreement once the terms were laid out. However, the intervention does not mention what agreement she is talking about, and many systems just fabricate that piece of information.

The second and fourth interventions (*pnorton_20* and *REBacon_165*) are a bit more complex. To avoid a lack of context (as in Clinton's text), some interventions in the *RRD* dataset were paired with their previous message using the mark "< *this message is answering to* >". The systems do not always get this differentiation and output questions related to the wrong message. This could be improved in the dataset.

The third intervention (*HOLT_159*) is from the presenter of the presidential debate: Holt. He is asking questions to Trump and claiming he lied for years. Many systems get confused and output questions that are in line with what Holt is saying and directed to Trump. This also happens often in another Holt's interventions (*HOLT_122*), and is one of the issues *Nompt* tried to solve. In fact, none of the CQs generated by this team in these two interventions have this issue.

The fifth intervention (*CLINTON_227*) is very emotional, as Clinton is making her final speech of

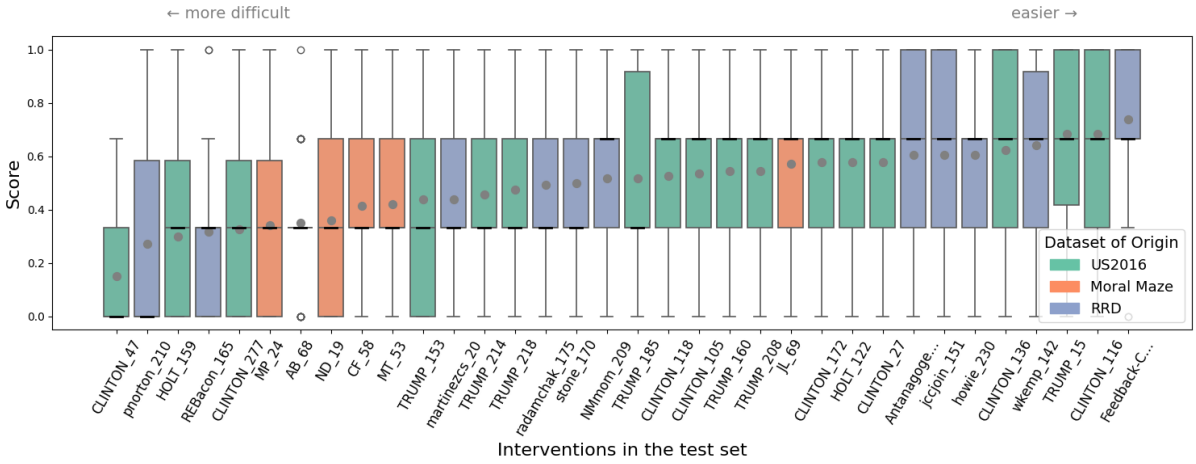

Figure 2: Distribution of the scoring of each run by intervention in the test set as evaluated with Claude in Table 5. Grey dots show the average score per intervention. The median is marked in black. Colors show the dataset of origin.

the debate, and the main points of her arguments are not very clear.

Finally, the sixth intervention (*MB_24*) is about housing loans, and the speaker is defending that laws protect borrowers over lenders. While his argument is clear and complete, the complexity of the topic makes many questions Unhelpful, and the systems often get lost in the speaker's stance.

7 Conclusions

This shared task on Critical Question Generation (CQs-Gen) aimed to promote systems capable of fostering critical thinking by generating insightful questions that challenge weak or fallacious arguments. The task attracted 19 registered teams, with 13 submitting systems and 12 providing descriptions.

The evaluation revealed that strong performance was not solely determined by automated metrics. Manual evaluation played a decisive role in identifying the most effective systems, with *ELLIS Alicante* emerging as the winner due to its focus on reducing Unhelpful and Invalid questions rather than purely optimizing quantitative scores. Their approach also highlighted the value of qualitative analysis in system development.

A key insight from the task was the mixed impact of incorporating argumentation schemes. While three of the top four teams successfully integrated these schemes into their pipelines, others reported performance degradation or reduced question diversity. This suggests that while argumentation schemes can enhance question quality and relevance, they may also limit variability and flexibility if not carefully balanced.

Model choice proved to be more impactful than prompt engineering, with most of the top-performing teams using GPT-4o or GPT-4o-mini. Additionally, classification and ranking modules that were used to select the best questions consistently improved output quality, with prompt-based approaches showing a better performance. Reasoning models did not seem to outperform general models. Nonetheless, further research is required to investigate whether they produce novel but useful CQs that are not captured by the automated evaluation method.

To further analyze system behavior, we applied automatic diversity metrics and extended test set evaluations. Results confirmed that higher-quality CQs often came with reduced diversity, reinforcing the observed trade-off between structure and variability. The winning team successfully navigated this trade-off through a hybrid generation and selection approach.

In sum, the task highlights that combining powerful models, informed prompt design, and thoughtful use of structured knowledge (like argumentation schemes) can yield high-quality critical questions. However, success also depends on careful evaluation, iterative analysis, and balancing competing goals such as diversity and usefulness.

Finally, the results of the shared task also show a big margin for improvement in CQs-Gen, revealing that it remains a challenging task for current LLMs.

Limitations

While this shared task generated valuable insights and fostered creative approaches to Critical Questions Generation (CQs-Gen), several limitations remain.

First, although many promising techniques emerged, the experimental approaches across teams were largely exploratory. A more systematic, controlled evaluation is necessary to draw robust conclusions about what methods are most effective for generating critical questions.

Second, the reliance of top-performing teams on proprietary, closed-weight language models raises concerns for the long-term scalability and transparency of CQs-Gen as an educational tool. Since the original motivation for this task includes real-world educational deployment, further research should explore the performance and adaptability of open-weight models, which may offer greater control and accessibility.

Third, the dataset used in this shared task presents topical and linguistic limitations. It primarily covers discussions related to politics, morality, and airline policies, and is restricted to English. Expanding the dataset to include a broader range of topics and additional languages would improve both the generalizability and inclusivity of future models.

Finally, the evaluation methodology imposed constraints in the validation phase. Since participants could not quantitatively and reliably compare performance due to a high number of NAE values, some systems may have been optimized without clear feedback. Refining the evaluation method would enhance the development process in future iterations.

Acknowledgments

This work has been partially supported by the French government, through the 3IA Cote d'Azur Investments in the project managed by the National Research Agency (ANR) with the reference number ANR-23-IACL-0001. HiTZ's researchers 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.

References

Azwad Anjum Islam, Tisa Islam Erana, and Mark A. Finlayson. 2025. Cognac at cqs-gen 2025: Generating critical questions with llm-assisted prompting and multiple rag variants. In *Proceedings of the 12th Workshop on Argument Mining (ArgMining 2025)*, Vienna, Austria. Association for Computational Linguistics.

Sajib Bhattacharjee, Tabassum Basher Rashfi, Samia Rahman, and Hasan Murad. 2025. Cuet_sr34 at cqs-gen 2025: Critical question generation via few-shot llms – integrating ner and argument schemes. In *Proceedings of the 12th Workshop on Argument Mining (ArgMining 2025)*, Vienna, Austria. Association for Computational Linguistics.

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

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.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Roxanne El Baff, Dominik Opitz, and Diaoulé Diallo. 2025. Criticalbrew at cqs-gen 2025: Collaborative multi-agent generation and evaluation of critical questions for arguments. In *Proceedings of the 12th Workshop on Argument Mining (ArgMining 2025)*, Vienna, Austria. Association for Computational Linguistics.

Lucile Favero, Daniel Frases, Juan Antonio Pérez-Ortiz, Tanja Käser, and Nuria Oliver. 2025. Ellis alicante at cqs-gen 2025: Winning the critical thinking questions shared task: Llm-based question generation and selection. In *Proceedings of the 12th Workshop on Argument Mining (ArgMining 2025)*, Vienna, Austria. Association for Computational Linguistics.

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.

Midhun Kanadan, Johannes Kiesel, Maximilian Heinrich, and Benno Stein. 2025. Webis at cqs-gen 2025: Prompting and reranking for critical questions. In *Proceedings of the 12th Workshop on Argument Mining (ArgMining 2025)*, Vienna, Austria. 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.

Sha Newaz Mahmud, Shahriar Hossain, Samia Rahman, Momtazul Arefin Labib, and Hasan Murad. 2025. Mind_matrix at cqs-gen 2025: Adaptive generation of critical questions for argumentative interventions. In *Proceedings of the 12th Workshop on Argument Mining (ArgMining 2025)*, Vienna, Austria. 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.

Tien-Dat Nguyen and Duc-Vu Nguyen. 2025. Tdnguyen at cqs-gen 2025: Adapt large language models with multi-step reasoning for critical questions generation. In *Proceedings of the 12th Workshop on Argument Mining (ArgMining 2025)*, Vienna, Austria. Association for Computational Linguistics.

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.

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, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Alan Ramponi, Gaudenzia Genoni, and Sara Tonelli. 2025. Arg2st at cqs-gen 2025: Critical questions generation through llms and usefulness-based selection. In *Proceedings of the 12th Workshop on Argument Mining (ArgMining 2025)*, Vienna, Austria. Association for Computational Linguistics.

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.

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.

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. *arXiv preprint*. ArXiv:2403.00553 [cs].

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 Bhupatiraju, 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.

Qwen Team. 2025. Qwq-32b: Embracing the power of reinforcement learning.

Teknium, Roger Jin, Chen Guang, Jai Suphavadeeprasit, and Jeffrey Quesnelle. 2025. Deephermes 3 preview.

Frieso Turkstra, Sara Nabhani, and Khalid Al-Khatib. 2025. Trillama at cqs-gen 2025: A two-stage llm-based system for critical question generation. In *Proceedings of the 12th Workshop on Argument Mining (ArgMining 2025)*, Vienna, Austria. Association for Computational Linguistics.

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

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational Argumentation Quality Assessment in Natural Language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187, Valencia, Spain. Association for Computational Linguistics.

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

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *Preprint*, arXiv:2412.13663.

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, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Jinghui Zhang, Dongming Yang, and Binghuai Lin. 2025. Statecloud at cqs-gen 2025: Prompt engineering for critical questions generation. In *Proceedings of the 12th Workshop on Argument Mining (ArgMining 2025)*, Vienna, Austria. Association for Computational Linguistics.

Jintian Zhang, Xin Xu, Ningyu Zhang, Ruibo Liu, Bryan Hooi, and Shumin Deng. 2024. Exploring collaboration mechanisms for llm agents: A social psychology view. *Preprint*, arXiv:2310.02124.

Wendi Zhou, Ameer Saadat-Yazdi, and Nadin Kökciyan. 2025. Daydreamer at cqs-gen 2025: Generating critical questions through argument scheme completion. In *Proceedings of the 12th Workshop on Argument Mining (ArgMining 2025)*, Vienna, Austria. Association for Computational Linguistics.

A Baseline 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:

1. If the question is not related to the text.
2. If the question is not specific (for instance, if it's a general question that could be applied to a lot of texts).
3. If the question introduces new concepts not mentioned in the text (for instance, if it suggests possible answers).
4. 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.
5. 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 3: Prompt for generating baseline outputs.

B All Test Set and Validation Set Results including NAEs

Team Name	Run	Validation	Test
ELLIS Alicante	1	61.4	36.3
ELLIS Alicante	2	61.0	44.1
ELLIS Alicante	3	**64.4**	**50.0**
COGNAC	1	**83.0**	**61.8**
COGNAC	2	81.0	57.8
COGNAC	3	82.0	60.8
StatetCloud	1	**76.2**	45.1
StatetCloud	2	72.8	42.2
StatetCloud	3	71.3	**47.1**
DayDreamer	1	**72.2**	**55.9**
DayDreamer	2	**72.2**	50.0
DayDreamer	3	62.2	43.1
Webis	1	72.0	49.0
Webis	2	**84.0**	**52.0**
Webis	3	82.0	48.0
TriLLaMa	1	*	**53.9**
TriLLaMa	2	*	37.3
TriLLaMa	3	*	52.0
Mind_Matrix	1	**53.0**	**42.2**
Mind_Matrix	2	46.0	36.3
CriticalBrew	1	**78.0**	**52.0**
CriticalBrew	2	**78.0**	40.2
CriticalBrew	3	71.0	51.0
Tdnguyen	1	70.7	**49.0**
Tdnguyen	2	**71.9**	45.1
Tdnguyen	3	61.3	46.1
ARG2ST	1	**76.2**	44.1
ARG2ST	2	72.8	**45.1**
ARG2ST	3	72.3	40.2
CUET_SR34	1	**71.1**	**43.1**
CUET_SR34	2	69.2	32.4
CUET_SR34	3	70.3	42.2

Table 4: Results of the shared task in the validation and test set using the official shared task evaluation script. The validation set results have been submitted by the participants. Therefore, we can not ensure direct comparison. The table is ordered by teams, starting by the winning team and ending with the lowest performing one. In bold are the highest results from each team.
* this team did not evaluate on the whole validation set.

C Automated-methods Evaluation

Team Name	Run	STS_0.65	Claude	Gemma2
ELLIS Alicante	1	48.04	51.96	61.76
ELLIS Alicante	2	54.90	54.90	53.92
ELLIS Alicante	3*	**67.65**	**67.65**	**66.67**
COGNAC	1*	62.75	64.71	61.76
COGNAC	2	65.69	62.75	**66.67**
COGNAC	3	64.71	60.78	65.69
StateCloud	1	50.00	49.02	53.92
StateCloud	2	51.96	50.00	53.92
StateCloud	3*	59.80	58.82	59.8
DayDreamer	1*	58.82	58.82	56.86
DayDreamer	2	52.94	49.02	54.9
DayDreamer	3	46.08	46.08	46.08
Webis	1	50.00	43.14	44.12
Webis	2*	56.86	56.86	51.96
Webis	3	53.92	52.94	48.04
TriLLaMa	1*	55.88	57.84	54.9
TriLLaMa	2	47.06	40.2	48.04
TriLLaMa	3	54.90	56.86	54.9
Mind_Matrix	1*	55.88	55.88	54.9
Mind_Matrix	2	47.06	40.20	50.98
CriticalBrew	1*	54.90	54.90	54.90
CriticalBrew	2	37.25	41.18	50.0
CriticalBrew	3	55.88	52.94	58.82
Lilo&stitch	1	53.92	52.94	54.9
Lilo&stitch	2*	53.92	52.94	53.92
Lilo&stitch	3	52.94	36.27	48.04
Tdnguyen	1*	51.96	52.94	54.90
Tdnguyen	2	51.96	41.18	47.06
Tdnguyen	3	51.96	46.08	53.92
ARG2ST	1	46.08	40.2	42.16
ARG2ST	2*	50.00	50.00	50.98
ARG2ST	3	46.08	41.18	51.96
CUET_SR34	1*	48.04	48.04	49.02
CUET_SR34	2	37.25	41.18	40.2
CUET_SR34	3	44.12	45.10	52.94
Nompt	1*	38.24	38.24	36.27
Nompt	2	37.25	24.51	36.27
Nompt	3	36.27	30.39	37.25

Table 5: Results of the shared task with the new fully-automated metrics. We kept the order from the ranking of the official shared task results. The table is ordered by teams, starting by the winning team and ending with the lowest performing one.
* runs included in the test set references.