SINAI at Touché: From Generation to Evaluation through Multistep and Comparative Prompting for Retrieval-Augmented Debate

Notebook for the Touché Lab at CLEF 2025

María Estrella Vallecillo-Rodríguez[1,*], María Teresa Martín-Valdivia[1] and Arturo Montejo-Ráez[1]

[1] Computer Science Department, SINAI, CEATIC, Universidad de Jaén, 23071, Spain

Abstract

This article describes the participation of the SINAI research group in the Retrieval-Augmented Debating shared task at CLEF 2025, which includes two subtasks: Subtask 1 focuses on generating multi-turn argumentative responses using retrieved evidence, while Subtask 2 addresses the automatic evaluation of debate responses based on quality, quantity, relation, and manner. For both subtasks, we employed the instruction-tuned LLaMA3.1-8B-Instruct model with a structured, multi-step prompting strategy to guide the model's reasoning. In Subtask 1, the generation process was divided into five stages, from analyzing the dialogue tone and argumentative strategy to formulating retrieval queries and generating the final response. This enabled the model to produce concise, coherent, and well-supported counterarguments, leading to a 4th place ranking overall. For Subtask 2, we explored three prompting paradigms (Zero-shot, Few-shot, and Analyzer strategies) to assess the model's ability to classify responses according to the four evaluation metrics. Experimental results demonstrate the effectiveness of structured reasoning, particularly with the Analyzer strategy, which achieved competitive performance across all metrics and led in Manner. Our system illustrates the potential of open-source language models for structured, retrieval-enhanced argumentative dialogue generation and evaluation, even when competing against proprietary models.

Keywords

Large Language Models, Argument retrieval, Argumentative Response Generation, Debate Evaluation

1. Introduction

Large Language Models (LLMs) are increasingly integrated into everyday life, but their widespread use also raises concerns about the reliability of the content they generate—particularly when based on online sources that may be false or misleading. To address this, it is essential to develop systems that make their reasoning transparent, enabling users to evaluate the credibility of the underlying evidence. This is especially important in the context of social media, where hate speech often circulates in the form of offensive statements lacking valid argumentation. While automatic counter-narrative generation has been explored in different languages such as English [1] and Spanish [2] among others [3, 4], existing approaches frequently fall short in argumentative richness. Incorporating structured argumentation could not only expose flawed reasoning but also contribute to influencing perspectives or informing bystanders. Additionally, automated systems can engage consistently over time, potentially reducing the spread of harmful content.

The Retrieval-Augmented Debating shared task was proposed to develop generative retrieval systems capable of arguing against users, with the goal of supporting opinion formation, confirmation, or debate training. It includes two subtasks. The first focuses on building a multi-turn debating system

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

[*] Corresponding author.

[†] These authors contributed equally.

mevallec@ujaen.es (M. E. Vallecillo-Rodríguez); maite@ujaen.es (M. T. Martín-Valdivia); amontejo@ujaen.es (A. Montejo-Ráez)

0000-0001-7140-6268 (M. E. Vallecillo-Rodríguez); 0000-0002-2874-0401 (M. T. Martín-Valdivia); 0000-0002-8643-2714 (A. Montejo-Ráez)

that responds to random claims by counterattacking or defending previous arguments, using distinct retrieved arguments and limiting responses to 60 words. The second subtask aims to automatically evaluate such systems using four metrics: quantity (informativeness), quality (truthfulness), relation (relevance), and manner (clarity). To support these tasks, the organizers released the ClaimRev dataset [5], which includes arguments retrieved from the Kialo platform[1] and simulated debates based on 100 claims from the ChangeMyView subreddit [2][6].

To explore the reasoning capabilities of current large language models, we selected LLaMA3.1-8B-Instruct [7] for both tasks. In Subtask 1, we propose a system based on a multistep prompt strategy, as used in [8], to guide the model through a complex task by dividing it into manageable steps. This ensures that the model has a clear goal at each stage, allowing it to build coherent responses incrementally. The first step involves analyzing the tone and style of the conversation to determine the appropriate argumentative approach—either by identifying weak points in the opponent's response or addressing their main idea—and selecting the argumentative perspective and type of evidence to retrieve. This analysis is guided by principles from logic (supporting conclusions with premises), dialectics (interactive discourse rules) [9], and rhetoric (capturing and persuading the audience) [10, 11, 12], as discussed in [13]. In the second step, the model generates up to three queries to retrieve arguments, specifying the target idea and whether it intends to support or refute it. The argument retrieval is then performed using an ElasticSearch API, which returns six different types of argument. Steps three and four involve filtering and selecting the most relevant arguments per query and then refining that selection for final use. Finally, in step five, the model generates the final response, integrating the selected arguments while adapting tone, style, and perspective; any incompatible arguments may be excluded if they do not align with the intended rhetorical strategy. For the second subtask, we experiment with different prompting strategies [14], as the way information is presented to the model plays a crucial role in shaping its final response. Specifically, for each evaluation metric, we first examine the behavior of the model in a zero-shot learning (ZSL) setting, where it does not receive examples. This allows us to assess its prior knowledge of the task. We then test a one-shot setting to determine whether the model can generalize from a single example. Finally, we explore a few-shot setting in which the model is provided with 1, 3, or 5 examples per possible label for the given metric. Based on these examples, the model is instructed to generate a list of reasons that justify the choice of one label over another. In a subsequent step, this reasoning is incorporated into the prompt, and the model is asked to assign a label to the text under evaluation.

The rest of this paper is structured as follows: Section 2 provides a detailed overview of the proposed system developed for the shared task. Section 3 describes the dataset employed and outlines the methodology adopted to address the task. In Section 4, we present the experimental results obtained during the development and evaluation phases. Lastly, Section 5 offers concluding remarks and a discussion of our findings.

2. System overview

This section describes the developed systems designed to address the subtasks of Retrieval-Augmented Debating [15] at CLEF 2025. Due to the significant differences between the two subtasks, this section is organized into two subsections. The first addresses response generation and argument retrieval strategies, whereas the second focuses on evaluating debate systems according to various metrics, resembling a ranking task. All prompts used to implement each strategy can be found in Appendix A, specifically in Subsections A.1 and A.2, corresponding to subtask 1 and subtask 2, respectively.

[1] https://www.kialo.com/

[2] https://www.reddit.com/r/changemyview/

2.1. Subtask 1: Debate response generator

As explained above, the task is highly complex. It requires retrieving arguments that counter the claims of our opponent or support the position of the system. Once the relevant arguments have been identified, the system must generate an appropriate response. To handle this complexity, we divided the overall process into smaller, more manageable steps. This decomposition not only simplifies the task for the model, but also helps reduce hallucinations and improves the relevance of the results. The proposed system consists of the following five steps: (1) Identification of the textual expression and argumentative strategy to be used in the generated response, (2) Query generation for the construction of search queries to retrieve relevant arguments from the database, (3) Initial argument retrieval: Selects the most relevant arguments of each query, (4) Filter all the selected arguments of each query to select the most suitable arguments for response generation, and (5) Counter-response generation that produces the final response opposing the input claim, using the selected arguments. A visual representation of these steps can be found in Figure 1.

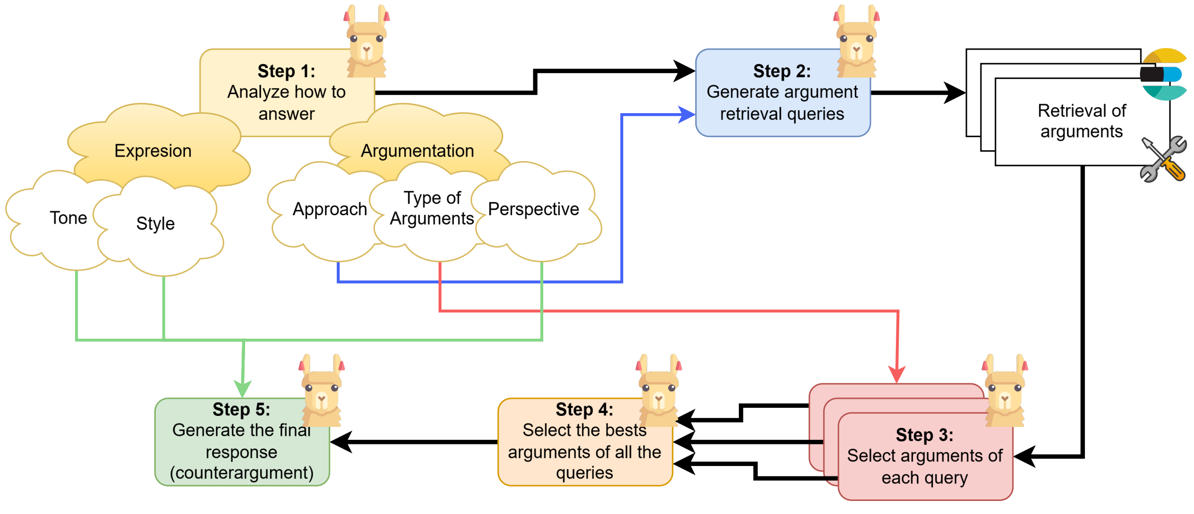
Step 1:
Analyze how to answer
Step 2:
Generate argument retrieval queries
Retrieval of arguments
Expresion
Argumentation
Tone
Style
Approach
Type of Arguments
Perspective
Step 5:
Generate the final response (counterargument)
Step 4:
Select the bests arguments of all the queries
Step 3:
Select arguments of each query

Figure 1: Proposed system for the first subtask of Retrieval-Augmented Debating task.

Once all steps have been defined, we now provide a more detailed explanation of each one:

- **Step 1: Discourse and Argumentative Strategy Analyzer.** In this step, the system analyzes the current debate context to define how it should respond in terms of tone, style, argumentative strategy, type of argument, and perspective. The tone can be Neutral, Respectful, Assertive/Critical, or Inquisitive. The style may be Academic, Colloquial, Socratic, Logical, or Rhetorical. Argumentative strategies include Comprehensive (respond to all points), Focused (target weaker points), Principled (challenge underlying assumptions), and Free. The argument types are Logical, Ethical, Emotional, and Analogical, while perspectives include Economic, Moral/Ethical, Scientific, Pragmatic, Historical, and Cultural. Once all of these dimensions are defined, the system proceeds to generate search queries accordingly.
- **Step 2: Query Generator for Argument Retrieval.** In this step, the system evaluates the opponent's input and, based on the argumentative strategy defined earlier, selects up to three key ideas to either attack or support. It then formulates one query per idea to search for relevant arguments in a retrieval database. The goal is to ensure that the selected ideas align with the desired argumentative strategy and that the queries are focused, relevant, and diverse enough to enrich the subsequent response generation.

- **Step 2.5: Argument Retrieval via Elasticsearch.** In this phase, the system retrieves arguments using a basic Elasticsearch setup, without additional embeddings due to computational limitations. For each query, three retrieval strategies are applied. First, the (1) Text Strategy retrieves two arguments based on textual similarity to the target idea. Then, depending on whether the model's objective is to support or attack the idea, different chains of arguments are retrieved. If the goal is to attack, in the (2) Attack strategy two attack arguments targeting the idea are retrieved (searching by the attack field in elasticsearch), and finally in the (3) Support strategy two support arguments for that idea are also collected; for each of those, the system further retrieves one argument that supports the attack associated with each support argument. Conversely, if the goal is to support the idea, the (3) Support strategy retrieved two supporting arguments (using the support field in elasticsearch), along with the (2) Attack strategy that retrieves two attack arguments directed at the idea;for each attack argument, one supporting argument for its corresponding attack is retrieved. This layered retrieval process enables the system to construct a small argument graph that reflects both direct and indirect relations aligned with the model's stance.
- **Step 3: Argument Selection per Query.** Given the large number of arguments retrieved per query, and to avoid exceeding token limits in the prompts, the system performs a first filtering step. From the 6 arguments retrieved for each query, it selects the top three, regardless of the strategy they came from. This filtering is based on the preferences determined in Step 1 (argument type), and the model is prompted accordingly to choose the most relevant and contextually appropriate ones.
- **Step 4: Final Argument Selection.** In this step, the model selects the three best arguments overall from among those shortlisted in the previous step. It is free to distribute them as it sees fit—for instance, choosing one argument per query, all from one query, or even using just two if deemed stronger.
- **Step 5: Response Generation.** Finally as final step and based on the argument selection made from the step 4 and the tone, style, and argumentative strategy from Step 1, the system generates a final response of no more than 60 words.

Each decision taken by the model in the process must be accompanied by a brief justification, which is the only content visible to the LLM at each step that need to use these aspects.

2.2. Subtask 2: Evaluation system

For the second subtask, which is based on developing a system that tries to evaluate response-generation systems in debates based on four different aspects (quantity, quality, relation, and manner), the organizers provide data with yes/no answers, making the task similar to binary classification. In addition, each metric is formulated as a yes/no question to facilitate the evaluation of specific aspects of the response. For instance, the question associated with the quantity metric is: "Does the response contain at least one (attack or defense) argument, and at most one of each type of defense and attack?". For quality, the question is: "Can the response be deduced from the retrieved arguments?". The relation metric asks: "Is the response coherent with the conversation and does it express a contrary stance to the user?'. 'Finally, the manner metric is evaluated using the question: "Is the response clear and precise?". In our proposed method, we conduct the experiments shown in Figure 2. We include all of them, even if they are experimental setups, in one figure as a summary of the system, since depending on the metric being evaluated, one type of system performed better than another. As a summary and as can be seen in the figure, our system receives in its prompt a description of the task it has to perform and the question it must answer with yes or no. Now, depending on the strategy applied, this prompt will include examples or not. For example, in the first strategy with ZSL, it will not receive any example from the dataset. In the second strategy, related to FSL, it will receive one example where the answer is yes and another where the answer is no. Finally, in the third strategy, the approach is a bit different. The model will first have to analyze why a person answered yes to the metric question or why they answered no. For that, the model will receive 1, 3, or 5 examples of each type of label. With the analysis done and the reasons provided for answering yes or no, this reasoning will be included in a ZSL-style prompt to try to guide the reasoning of the model.

It is important to highlight that the questions the models must answer appear in Section 4. Also, depending on the metric to be evaluated, each system follows the following strategy: Quantity uses the analysis strategy where the model receives only one example of each answer type (yes/no); Quality also uses this third analysis strategy, but when receiving five examples to analyze; the Relation metric applies a ZSL strategy; and finally, the Manner metric uses the third strategy related to analysis with just one example per answer type.

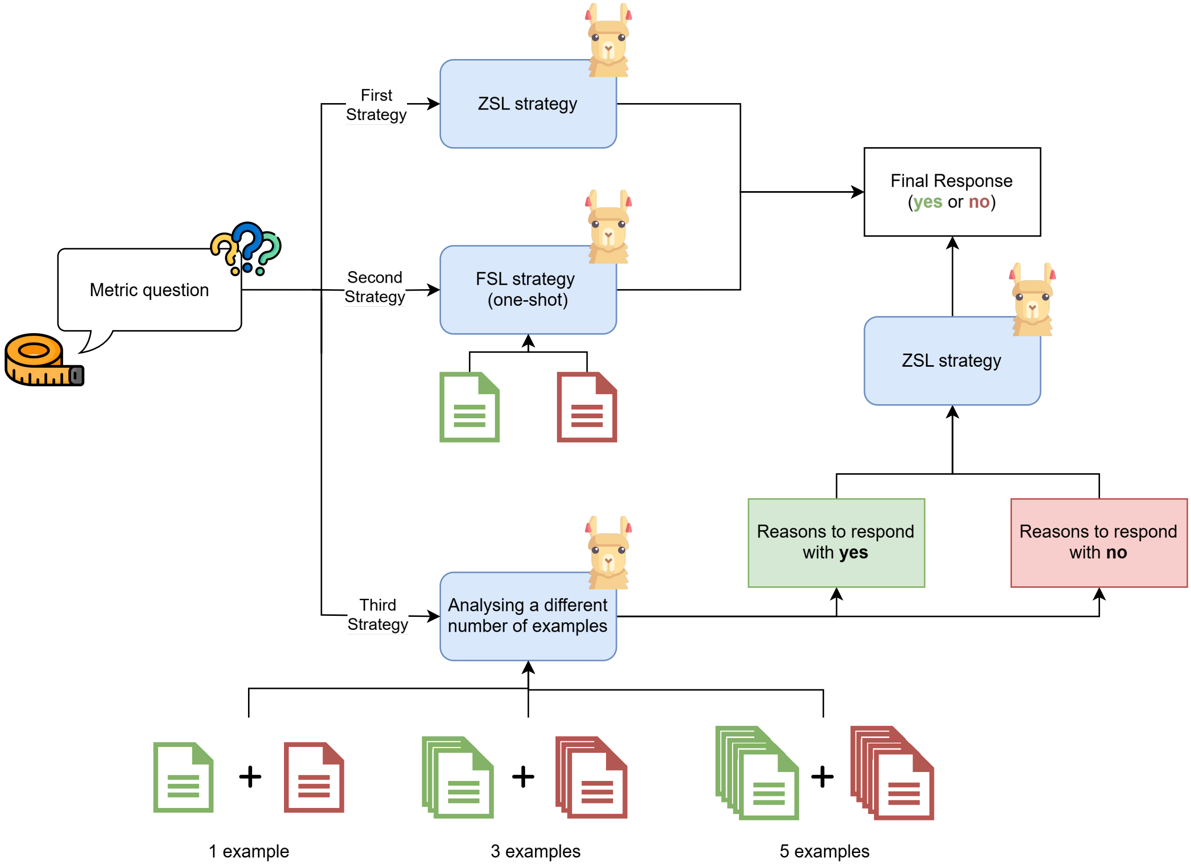

Figure 2: Proposed system for the second subtask of Retrieval-Augmented Debating task.

3. Experimental setup

3.1. Data

To run our experiments, we used the data provided by the organizers. Among the provided datasets, we find a database of arguments retrieved from the ClaimRev dataset [5]. These arguments consisted of an id linking them to their corresponding ClaimRev_id, a general topic to which the argument belongs, a list of labels indicating the categories associated with the topic, an argument that is attacked by the current one, another argument that is supported by it, the text of the argument, a list of references to back up the argument, and a field indicating whether the argument (support or attack) was originally part of the ClaimRev dataset or was automatically generated by the organizers.

In total, 287,156 arguments are provided, covering 1,522 distinct topics, including prominent themes such as ‘Politics’, ‘Ethics’, ‘Society’, ‘Religion’, and ‘Philosophy’, with 1,210 different associated labels.

Table 1
Frequency of the labels for Subtask 2 of the Retrieval Augmenting Debate shared task dataset for each metric.

Labels	Relation	Quality	Manner	Quantity
yes	355	237	311	426
no	141	262	188	71
don't know	4	1	1	3

These arguments are publicly accessible through the Elasticsearch API. It is important to note that the dataset also includes embeddings computed using the stella_en_400M_v5 model [16]. These embeddings allow participants to perform argument retrieval based on vector similarity. However, due to computational limitations, we were unable to leverage these embeddings and thus decided not to use them in our experiments.

To simulate debates and provide training data, the organizers selected 100 claims from the ChangeMyView subreddit and simulated a series of debates with five interaction turns each. Additionally, for Subtask 2, annotations are provided for each turn in terms of four metrics: quality, quantity, relation, and manner [6]. Each annotation consists of a label ('yes' or 'no') indicating whether the respective question metric is satisfied (in Section 2.2 these questions are mentioned).

To illustrate the class distribution in this classification task, Table 1 shows the frequency of each label for each metric in the initial dataset.

For the generation of the prompts that we are going to use to conduct our experiments in Subtask 2, we removed instances with unknown labels. From the remaining dataset, we selected 20 instances, equally divided between the labels 'yes' and 'no'. The rest of the dataset (476 instances) is used to evaluate the proposed strategies.

3.2. Experiments and Selected Models

Across both tasks, our goal is to analyze the generalization ability of LLMs, their reasoning behavior under different prompting strategies, and how instruction-tuned language models perform in argumentation-related tasks without requiring additional fine-tuning. For this reason and due to computational constraints, we selected the LLaMA3-8B-instruct model [7] for both subtasks.

Regarding the proposed experiments, in this section we present a minimal summary of the experiments described in the previous section (Section 2):

- **Subtask 1.** For this subtask, only one experiment is proposed. This experiment is based on a multistep prompting strategy. We divided the task of generating a response that takes into account different retrieved arguments into several small steps, such as analyzing tone and style to respond to the opponent, reasoning about the strategic approach, selecting the type of arguments, and determining the argument perspective. Furthermore, the model must formulate various queries to retrieve arguments from Elasticsearch and finally generate a response that incorporates all aspects evaluated throughout the steps. With this experiment, we aim to analyze the model's ability to generate coherent, contextually appropriate, and argumentatively structured responses with minimal supervision.
- **Subtask 2.** The experiments for this task focus on different prompting strategies, as we aim to evaluate the model's knowledge and its capacity to reason based on a few examples, using the model in its base form without requiring task-specific fine-tuning. The following prompting strategies are proposed:
  - **Zero-shot Learning (ZSL)**: The task is explained to the model, and it is presented with a question to which it must respond with "yes" or "no", in order to evaluate the dialogue using the selected metric. The model is then expected to directly answer the question. This

experiment serves as the baseline, where the key factor under evaluation is the model's prior knowledge about the task.

- **Few-shot Learning (FSL)**: This strategy is similar to the previous one, except that the model is provided with two examples: one where the answer to the metric-related question is "yes", and another where it is "no". The objective of this experiment is to analyze whether the model, after observing the examples, can perform better classifications. In this case we try with two types of FSL, the first called FSL1 where the positive example appear first and another called FSL2 where the negative example appear first in the prompt to understand if the order of the presented examples affect to the classification of the model.
- **Analyzer Strategy**: This strategy is divided into two stages. In the first stage, the model is given a set of examples where the answer to the metric-related question is "yes", and an equal number of examples where the answer is "no". The number of examples may vary (1, 3, or 5). After reviewing the examples, the model is asked to explain why someone would answer "yes" based on the positive examples, and similarly for "no" based on the negative ones. These explanations must be provided in a generalized form. In the second stage, the metric-related question is posed again, and the model is asked to answer with "yes" or "no', taking into account the reasoning it generated in the previous step. This strategy aims to analyze whether the reasoning of the model is useful in guiding its final response, and to determine the optimal number of examples required to support coherent and helpful reasoning.

4. Results

In this section, we present the results of each subtask. Specifically, it includes the outcomes of our experiments conducted during the development phase (Subsection 4.1) and the results obtained by the systems that were ultimately submitted to the TIRA.io platform [17] for the final evaluation (Subsection 4.2).

As previously mentioned, this task is divided into two subtasks: the first focuses on generating responses to argue against a simulated debate partner, and the second aims to evaluate the systems developed for Subtask 1. To assess Subtask 1, the organizers proposed a manual evaluation carried out by human annotators based on the evaluation metrics defined for Subtask 2 (as described in Section 2.2). For Subtask 2, which is focused on the binary classification of whether a generated response and its corresponding argument meet predefined criteria, the organizers chose to evaluate the systems using standard binary classification metrics: macro-precision, macro-recall, and macro-F1 [18].

4.1. Development Phase

During the development phase for Subtask 1, we focused on assessing how the model performed with the provided prompt through manual review. This was necessary because the task is difficult to evaluate using a single, exact metric that captures all relevant aspects as a human would. When we observed that the model consistently failed to generate appropriate responses—across multiple systems—we interpreted this as an indication that it was not properly understanding the assigned task. In those cases, we refined the prompt accordingly. This process was carried out iteratively until we obtained a system capable of generating coherent texts that integrated the retrieved arguments in alignment with the human annotator's perspective.

Table 2 reports the results of the developed systems for Subtask 2 across four categories (Relation, Quality, Manner, and Quantity) with macro-F1 selected as the primary metric due to its balanced reflection of both precision and recall, especially relevant in imbalanced datasets.

The results highlight the varying effectiveness of different prompting strategies in a base, non-fine-tuned language model. Some key findings include:

- **Zero-shot Learning Strategy (ZSL, baseline).** It achieved the highest F1-score for the Relation category (F1: 0.515), indicating a relatively strong inherent understanding of this aspect without prior examples. However, performance on other categories—particularly Quantity (F1: 0.266)—was considerably weaker. This suggests that while foundational knowledge exists, its application across different communicative aspects remains inconsistent without additional guidance.
- **Few-shot Learning (FSL, with two examples).** Two FSL variants were tested: FSL1 (positive example first) and FSL2 (negative example first). Overall, FSL did not significantly outperform ZSL or the Analyzer Strategy. For example, in the Relation aspect, FSL1 and FSL2 yielded F1-scores of 0.404 and 0.360, respectively—both trailing behind ZSL. These findings suggest that providing only a few direct examples may be insufficient for improving classification accuracy in complex reasoning tasks. Notably, FSL1 consistently outperformed FSL2, underscoring the influence of example ordering.
- **Analyzer Strategy:** This approach produced the most promising results in several categories. For instance, Analysis5 achieved the highest F1-score for Quality (F1: 0.575), while Analysis1 led in Manner (F1: 0.497) and Quantity (F1: 0.389). These outcomes suggest that prompting the model to explicitly reason—especially with a greater number of guiding examples—enhances classification performance. The comparison across Analysis1, Analysis3, and Analysis5 also indicates that the optimal number of examples for effective reasoning is task-specific. Interestingly, ZSL maintained superior performance in the Relation category, implying that additional reasoning may not always be necessary for certain well-understood communicative principles. This also points to the importance of example selection; randomly chosen examples may not sufficiently support high-quality reasoning.

These findings emphasize the critical role of prompt design in leveraging pre-trained language models for classification tasks. While zero-shot approaches provide a solid baseline for aspects like Relation, strategies that incorporate structured reasoning, such as the Analyzer Strategy, yield significant performance gains in more nuanced aspects such as Quality, Manner, and Quantity. In contrast, Few-shot Learning with limited examples shows limited impact, suggesting that mere exposure is less effective than guided reasoning for complex evaluative tasks.

Table 2
Performance of each proposed experiment and the evaluated metric. The systems are evaluated based on macro-precision (P), macro-recall (R), and macro-F1 (F1). The best-performing experiments for each aspect, according to the macro-F1 score, are highlighted in bold.

System	Relation			Quality			Manner			Quantity		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Analysis1	0.461	0.472	0.340	0.574	0.568	0.559	0.501	0.502	**0.497**	0.516	0.540	**0.389**
Analysis3	0.501	0.506	0.493	0.621	0.588	0.563	0.508	0.510	0.471	0.468	0.440	0.357
Analysis5	0.495	0.499	0.497	0.588	0.581	**0.575**	0.492	0.493	0.472	0.513	0.535	0.342
FSL1	0.477	0.478	0.404	0.540	0.520	0.460	0.508	0.506	0.375	0.486	0.476	0.367
FSL2	0.442	0.441	0.360	0.537	0.510	0.411	0.532	0.514	0.371	0.485	0.481	0.335
ZSL	0.516	0.524	**0.515**	0.586	0.553	0.511	0.512	0.514	0.486	0.513	0.525	0.266

4.2. Evaluation Phase

Table 3 shows the final evaluation results for the Retrieval-Augmented Debating subtask 1. The baseline system, provided by the organizers, selects the first counter-argument retrieved via the Elasticsearch API. Looking at the results, our system (SINAI) achieved a score of 0.86 in the "Relation" metric, ranking among the top three systems, slightly behind DSGT with Claude-Sonnet-4 (0.94) and Claude-Opus-4 (0.87), indicating its ability to generate coherent, context-aware rebuttals. While its overall average score was lower than that of other systems (0.54 vs. 0.62), SINAI outperformed the baseline in Quantity

(0.70 vs. 0.35), Relation (0.86 vs. 0.32), and Manner (0.59 vs. 0.80), demonstrating greater informativeness and clarity. The low score in the “Quality” metric (0.02 vs. 1.00) may stem from the model’s tendency to significantly alter the style of retrieved arguments during generation, reducing their traceability. Notably, unlike most top-performing teams that relied on large-scale commercial models such as GPT-4.1, Gemini 2.5 or Claude Opus, while our system is based on an open-source model (LLaMA3.1-8B-Instruct), showing that a competitive and resource-efficient alternative is possible by analyzing different aspect of the debate context.

Table 3
Official results for Subtask 1. Systems are evaluated based on four metrics: Quantity, Quality, Relationship, and Manner. The best-performing systems for each metric are highlighted in bold. Our system’s results are shown with a grey background.

#	Team	Run	Score (avg)	Quantity	Quality	Relation	Manner
1	DS@GT	gpt-4.1	**0.70**	**0.95**	0.17	0.82	**0.84**
2	DS@GT	gemini-2.5	0.65	0.94	0.26	0.74	0.67
3	Baseline	baseline	0.62	0.35	**1.00**	0.32	0.80
4	SINAI	LLaMA3.1-8B-Instruct	0.54	0.70	0.02	0.86	0.59
5	DS@GT	gemini-2.5-flash	0.50	0.70	0.07	0.80	0.41
6	DS@GT	claude-opus-4	0.42	0.41	0.31	0.87	0.09
7	DS@GT	gpt-4o	0.42	0.20	0.02	0.86	0.58
8	DS@GT	claude-sonnet-4	0.38	0.35	0.05	**0.94**	0.17

For the second subtask, in addition to the proposed methods, we implemented a strategy named BEST, which applies the best-performing method per metric based on development results: ZSL for Relation, Analysis5 for Quality, and Analysis1 for Manner and Quantity. Results are shown in Table 4. Our systems achieved competitive performance using only the open-source LLaMA3.1-8B-Instruct model. The best overall result came from the Analyzer strategy with five examples (Analysis5), reaching a macro-F1 of 0.56, suggesting that guided reasoning with balanced examples improves consistency. The “Best” configuration also performed well (F1 of 0.55), allowing adaptation to each dimension’s specific needs. In general, Analyzer outperformed other approaches, especially with more examples, likely due to the intermediate reasoning phase. ZSL served as a reasonable baseline (F1 of 0.52), while FSL strategies underperformed (FSL1: 0.39, FSL2: 0.35), indicating that example inclusion alone is insufficient without reflection. No significant differences were observed between FSL1 and FSL2, though models tended to perform slightly better when the positive example appeared first. Overall, our results show that structured and adaptive reasoning strategies can yield solid performance even with non-commercial models.

5. Conclusions

This work presents the system developed by the SINAI team for the shared task Retrieval-Augmented Debating, which includes two subtasks: the first involves building an automatic system to generate responses that argue against a simulated debate partner by retrieving arguments from a database; the second focuses on evaluating such responses across four metrics: Quantity (informativeness), Quality (truthfulness), Relation (relevance to the conversation), and Manner (clarity). To tackle both tasks, we used the open-source model LLaMA3.1-8B-Instruct, chosen for its accessibility and lower computational cost. For Subtask 1, our system was based on a detailed analysis of debate elements such as tone, topic, and perspective, guiding the retrieval and generation process. This approach led us to a 4th place ranking, behind the task’s baseline and proprietary state-of-the-art models. Our system showed strong performance in generating informative, relevant, and clear responses, although there is room for improvement in truthfulness, as the generated answers did not always clearly align with the retrieved content. This highlights a promising direction for future work, particularly in reinforcing factual consistency. For Subtask 2, we explored various prompting strategies without fine-tuning, aiming to

Table 4
Official results for Subtask 2. Systems are evaluated using macro-precision (P), macro-recall (R), and macro-F1 (F1). Best-performing systems for each metric are highlighted in bold. Results from our system are shown with a grey background. All SINAI team systems are based on LLaMA3.1-8B-Instruct. The “BEST” run corresponds to the system that applies, for each metric, the strategy that performed best during the development phase.

#	Team	Run	Score	Relation			Quality			Manner			Quantity		
			(F1)	P	R	F1	P	R	F1	P	R	F1	P	R	F1
1	Baseline	1-baseline	**0.67**	0.57	**1.00**	**0.73**	**0.24**	**1.00**	**0.38**	0.78	**1.00**	0.87	0.52	**1.00**	**0.68**
2	DS@GT	gemini-2.5-flash	0.64	0.59	0.86	0.70	0.18	0.66	0.29	0.81	0.99	0.89	0.52	0.99	**0.68**
3	DS@GT	gpt-4o	0.64	0.59	0.88	0.71	0.17	0.63	0.27	0.82	0.99	0.89	0.52	0.97	0.67
4	DS@GT	gpt-4.1	0.62	0.58	0.75	0.65	0.15	0.52	0.24	0.82	0.98	**0.90**	0.52	0.99	**0.68**
5	DS@GT	gemini-2.5-pro	0.62	0.59	0.67	0.63	0.17	0.52	0.25	0.84	0.97	**0.90**	0.52	0.98	**0.68**
6	SINAI	Analysis5	0.56	0.60	0.60	0.60	0.19	0.40	0.25	0.84	0.86	0.85	0.50	0.57	0.53
7	DS@GT	claude-sonnet-4	0.56	0.56	0.43	0.49	0.15	0.36	0.21	0.83	0.92	0.88	0.51	0.93	0.66
8	SINAI	Best	0.55	0.59	0.64	0.61	0.16	0.32	0.21	0.84	0.80	0.82	0.52	0.64	0.57
9	SINAI	Analysis3	0.54	0.58	0.53	0.55	0.20	0.35	0.25	**0.87**	0.75	0.81	**0.53**	0.56	0.54
10	SINAI	ZSL	0.52	0.57	0.46	0.51	0.15	0.37	0.21	0.84	0.79	0.81	0.50	0.63	0.56
11	DS@GT	claude-opus-4	0.51	0.49	0.21	0.29	0.16	0.31	0.21	0.85	0.90	0.88	0.51	0.92	0.66
12	SINAI	Analysis1	0.49	0.59	0.63	0.61	0.20	0.58	0.30	0.84	0.39	0.53	0.50	0.54	0.52
13	SINAI	FSL1	0.39	0.57	0.32	0.41	0.17	0.21	0.19	0.84	0.67	0.74	0.45	0.16	0.24
14	SINAI	FSL2	0.35	**0.63**	0.40	0.49	0.16	0.17	0.16	0.84	0.44	0.58	0.41	0.10	0.16

assess the reasoning capabilities of the model. Our open-source systems achieved competitive results compared to proprietary approaches, particularly excelling in the Manner metric, where we led in precision. Despite limitations compared to large-scale models like GPT-4 or Gemini, we remained competitive in Relation and Quantity. The use of both Zero-Shot and Few-Shot Learning strategies underscores the exploratory and adaptive nature of our approach. Overall, the results demonstrate that, with open models and thoughtful design, it is possible to effectively address complex semantic evaluation tasks.

After all that has been observed throughout this work, we still have a considerable amount of work ahead, which we plan to address gradually as future work. Regarding the system that automatically generates responses in an argumentative way, we aim to carry out a thorough analysis to determine whether the model tends to adopt certain tones and styles, and whether these variations influence the selection of argument types. In the argument retrieval phase, we are not only interested in continuing with arguments from Kialo, but also in exploring other databases, argument types, or retrieval methods based on LLMs. In this regard, we propose a system based on agents that demostrated good results in different tasks [19], where by providing a search engine with internet access, the model could autonomously respond to questions and analyse the type of arguments it retrieves, evaluating the performance of such dynamically sourced systems [20]. As for the second subtask, as future work we envisage choosing the examples we show in the prompt more rigorously using the knowledge of a human expert rather than selecting them randomly. In addition, although we deliberately avoid fine-tuning in these experiments to evaluate the performance of the base model, we believe that a slight adaptation to the task could bring noticeable improvements. Moreover, in both subtasks, we do not intend to limit ourselves to models such asuch as LLaMA, but rather to explore others with a higher number of parameters or different architectures such as Mistral [21], Qwen [22] and models oriented to argumentation such as Veritas-12B [23].

Acknowledgments

This work has been partially supported by Project CONSENSO (PID2021-122263OB-C21), Project MODERATES (TED2021-130145B-I00), and Project SocialTox (PDC2022-133146-C21) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR.

Declaration on Generative AI

During the preparation of this work, the authors used GPT-4o and Deepl in order to: Grammar, spelling and translation check. After using these services, the authors reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] Fanton, Margherita and Bonaldi, Helena and Tekiroğlu, Serra Sinem and Guerini, Marco, Human-in-the-Loop for Data Collection: a Multi-Target Counter Narrative Dataset to Fight Online Hate Speech, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2021.

[2] M.-E. Vallecillo-Rodríguez, M.-V. Cantero-Romero, I. Cabrera-De-Castro, A. Montejo-Ráez, M.-T. Martín-Valdivia, CONAN-MT-SP: A Spanish corpus for counternarrative using GPT models, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italy, 2024, pp. 3677–3688. URL: https://aclanthology.org/2024.lrec-main.326.

[3] H. Bonaldi, M. E. Vallecillo-Rodríguez, I. Zubiaga, A. Montejo-Raez, A. Soroa, M.-T. Martín-Valdivia, M. Guerini, R. Agerri, The first workshop on multilingual counterspeech generation at COLING 2025: Overview of the shared task, in: H. Bonaldi, M. E. Vallecillo-Rodríguez, I. Zubiaga, A. Montejo-Ráez, A. Soroa, M. T. Martín-Valdivia, M. Guerini, R. Agerri (Eds.), Proceedings of the First Workshop on Multilingual Counterspeech Generation, Association for Computational Linguistics, Abu Dhabi, UAE, 2025, pp. 92–107. URL: https://aclanthology.org/2025.mcg-1.10/.

[4] Y.-L. Chung, E. Kuzmenko, S. S. Tekiroglu, M. Guerini, CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 2819–2829. URL: https://www.aclweb.org/anthology/P19-1271. doi:10.18653/v1/P19-1271.

[5] G. Skitalinskaya, J. Klaff, H. Wachsmuth, Learning from revisions: Quality assessment of claims in argumentation at scale, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, Online, 2021, pp. 1718–1729. URL: https://aclanthology.org/2021.eacl-main.147.

[6] M. Gohsen, N. Mirzakhmedova, H. Scells, M. Aliannejadi, M. Fröbe, J. Kiesel, B. Stein, TouchÉ 25 rad claims, 2025. URL: https://doi.org/10.5281/zenodo.15401620. doi:10.5281/zenodo.15401620.

[7] AI@Meta, Introducing llama 3.1: Our most capable models to date (2024). URL: https://ai.meta.com/blog/meta-llama-3-1/.

[8] Y. Fu, H. Peng, A. Sabharwal, P. Clark, T. Khot, Complexity-based prompting for multi-step reasoning, in: The Eleventh International Conference on Learning Representations, 2022.

[9] F. H. van Eemeren, In what sense do modern argumentation theories relate to aristotle? the case of pragma-dialectics, Argumentation 27 (2013) 49–70. URL: https://doi.org/10.1007/s10503-012-9277-4. doi:10.1007/s10503-012-9277-4.

[10] J. A. Herrick, The History and Theory of Rhetoric: An Introduction, 7th ed., Routledge, 2020. URL: https://doi.org/10.4324/9781003000198. doi:10.4324/9781003000198.

[11] K. Hogan, The Psychology of Persuasion: How to Persuade Others to Your Way of Thinking, Pelican Publishing, 2010. URL: https://books.google.es/books?id=FAHzLM-pY7cC.

[12] H. W. Simons, Persuasion in Society, 2nd ed., Routledge, 2011. URL: https://doi.org/10.4324/9780203933039. doi:10.4324/9780203933039.

[13] R. Morado, Funciones básicas del discurso argumentativo (????). URL: https://revistas.uam.es/ria/article/view/8195. doi:10.15366/ria2013.6.007, number: 6.

[14] W. Wang, V. W. Zheng, H. Yu, C. Miao, A survey of zero-shot learning: Settings, methods, and

applications, ACM Trans. Intell. Syst. Technol. 10 (2019). URL: https://doi.org/10.1145/3293318. doi:10.1145/3293318.

[15] J. Kiesel, Ç. Çöltekin, M. Gohsen, S. Heineking, M. Heinrich, M. Fröbe, T. Hagen, M. Aliannejadi, T. Erjavec, M. Hagen, M. Kopp, N. Ljubešić, K. Meden, N. Mirzakhmedova, V. Morkevičius, H. Scells, I. Zelch, M. Potthast, B. Stein, Overview of Touché 2025: Argumentation Systems, in: J. C. de Albornoz, J. Gonzalo, L. Plaza, A. García Seco de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. 16th International Conference of the CLEF Association (CLEF 2025), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2025.

[16] D. Zhang, J. Li, Z. Zeng, F. Wang, Jasper and stella: distillation of sota embedding models, 2025. URL: https://arxiv.org/abs/2412.19048. arXiv:2412.19048.

[17] M. Fröbe, M. Wiegmann, N. Kolyada, B. Grahm, T. Elstner, F. Loebe, M. Hagen, B. Stein, M. Potthast, Continuous Integration for Reproducible Shared Tasks with TIRA.io, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2023, pp. 236–241. doi:10.1007/978-3-031-28241-6_20.

[18] M. Sokolova, G. Lapalme, A systematic analysis of performance measures for classification tasks, Information Processing Management 45 (2009) 427–437. URL: https://www.sciencedirect.com/science/article/pii/S0306457309000259. doi:https://doi.org/10.1016/j.ipm.2009.03.002.

[19] J. Shi, D. K. C. Lee, W. Xu, Y. Wang, Comparative analysis of open-source frameworks for agentic ai systems: Capabilities, design philosophies, and development experiences, World Scientific Annual Review of Fintech 0 (0) null. URL: https://doi.org/10.1142/S2811004824500015. doi:10.1142/S2811004824500015. arXiv:https://doi.org/10.1142/S2811004824500015.

[20] A. Yeginbergen, M. Oronoz, R. Agerri, Dynamic knowledge integration for evidence-driven counter-argument generation with large language models, 2025. URL: https://arxiv.org/abs/2503.05328. arXiv:2503.05328.

[21] mistralai/mistral-small-3.1-24b-instruct-2503 · hugging face, ????. URL: https://huggingface.co/mistralai/Mistral-Small-3.1-24B-Instruct-2503.

[22] Qwen/qwen3-30b-a3b-GGUF · hugging face, ????. URL: https://huggingface.co/Qwen/Qwen3-30B-A3B-GGUF.

[23] soob3123/veritas-12b · hugging face, ????. URL: https://huggingface.co/soob3123/Veritas-12B.

A. Used prompts

A.1. Subtask 1: Development of debate system

Prompt step 1 (Subtask 1)

You are an expert debater and you are having a dialogue with another user. Your task is to analyze with what TONE and STYLE you should answer your opponent. Keep in mind that the idea you are defending is contrary to the one he is showing. It is also very important that you define the APPROACH STRATEGY to be followed, as well as the TYPE OF ARGUMENTS you should retrieve or the PERSPECTIVE from which the argumentation should be elaborated. The different options of each category are:
TONE:
- Neutral: Based on facts and logic (without using emotions or making personal judgments)
- Respectful: acknowledges the other person's point of view without disqualifying it
- Assertive/Critical: defends a position clearly and firmly. Without being aggressive
- Inquisitive: tries to ask questions to invite reflection or question a statement.

STYLE:
- Academic: follows a formal, structured style, supported by sources. The language to be used is usually technical and includes many quotations and data.
- Colloquial: It follows an informal style, with a natural and close language.
- Socratic: makes use of questions that invite reflection or lead to a contradiction.
- Logical: based on the rational structure of the argument (usually including a series of premises, conclusions, syllogisms and formal deductions).
- Rhetorical: appeals to emotions

APPROACH STRATEGY: refers to how the counterargument(s) should address the original text
- Comprehensive: addresses all points made
- Focused: focus on those arguments that are weaker.
- Principled: questions the underlying principles or assumptions of the original message.
- Free (explore other approaches not mentioned in this text).

TYPE OF ARGUMENTS:
- Logical (based on reason and facts)
- Ethical (values, rights)
- Emotional (empathy, human impact)
- Analogical (establishing examples and similarities)

PERSPECTIVE:
- Economic
- Ethical/Moral
- Scientific
- Practical/pragmatic
- Historical
- Cultural

The format of your response have to consist in an unique JSON with exactly these keys:
- tone: a string with the selected TONE (Neutral, Respectful,Assertive/Critical, or Inquisitive).
- justification_tone: an explanation about why you have selected that tone.
- style: a string with the selected STYLE (Academic, Colloquial, Socratic, Logical or Rhetorical).
- justification_style: an explanation about why you have selected that style
- approach_strategy: a string with the selected APPROACH STRATEGY (Comprehensive, Focused, Principled or Free).
- justification_approach: an explanation about why you have selected that approach strategy.
- type_of_arguments: a string with the selected TYPE OF ARGUMENTS (Logical, Ethical, Emotional, or Analogical)
- justification_type: an explanation about why you have selected that type of arguments.
- perspective: a string with the selected PERSPECTIVE (Economic, Ethical/Moral, Scientific, Practicalpragmatic, Historical, or Cultural)
- justification_perspective: an explanation about why you have selected that perspective.

Specific information about the established debate is included below. Remember that you have to analyze the way you should answer and oppose to your opponent's last message:
{debate_dialogue}

Figure 3: Prompt used in the first step of our Retrieval-Augmented Debating Subtask 1 experiment.

Prompt step 2 (Subtask 1)

You are an expert debater and you are having a dialogue with another user. Your task is to analyze how to find arguments or contra-arguments to response your last opponent message. Keep in mind that the idea you are defending is contrary to the one he is showing.
Please response with only a JSON Object that contain as keys the word SEARCH followed by a number with the step to search (SEARCH_1,SEARCH_2, . . .). Each key has a value a dictionary with the following keys:
- opponent_idea: a string with the idea showed by your opponent that you want to use to search the arguments, please try to be specific and put the opponent idea complete. You must include the specific subject in your sentence removing inespecific subjects like its, his or her.
- field_to_look: a string with two possible values (SUPPORT or ATTACK). For example you can identify the opponent idea and look for argument that support his idea and later look for arguments that attack the support arguments, or you want to look for arguments that attack the idea of your opponent, or instead you prefer to find arguments that attacks your opponent idea and retrieve arguments that support your main idea.
- justification: an string with a justification of why you have selected that decision.
It is important that at maximum you can make 3 different search. Specific information about the established debate is included below.
{debate_dialogue}
Take into account that previously you decided that the arguments or counterarguments that you are looking for should address your opponent message with a {FirstStep[approach_strategy]} strategy that consist in {description[FirstStep[approach_strategy]]}. The justification of your previous decision is: {FirstStep[justification_approach]}

Figure 4: Prompt used in the second step of our Retrieval-Augmented Debating Subtask 1 experiment.

Prompt step 3 (Subtask 1)

You are an expert debater and you are having a dialogue with another user. Your task is to select the 3 best arguments in order to generate a reply to your opponent's last message in a future step.
Specific information about the established debate is included below.
{debate_dialogue}
The information you are looking for is listed below:
{SecondStepResponse_Query_X}
We search arguments by 3 different strategies:
- ATTACK STRATEGY is for arguments that attack the opponent_idea.
- SUPPORT STRATEGY is looking for arguments that attack some arguments that support the opponent_idea.
- TEXT STRATEGY is looking by the similarity of your opponent_idea and the text of the retrieval argument
The information retrieved in each strategy is shown next:
{RetrievedArguments}
To select the best arguments take into consideration that you want to select {FirstStep[type_of_arguments]} arguments. The justification to select these types of arguments appear next: {FirstStep[justification_type]}

Figure 5: Prompt used in the third step of our Retrieval-Augmented Debating Subtask 1 experiment.

Prompt step 4 (Subtask 1)

```
You are an expert debater and you are having a dialogue with another user. Your task is to select at maximum of 3 arguments to answer the last opponent message. Keep in mind that the idea you are defending is contrary to the one he is showing.
Specific information about the established debate is included below.
{debate_dialogue}
Previously some arguments are retrieved in base a different aspects. The following information contains the aspect to search arguments and selected arguments for each criteria.
{SEARCH_1: {aspect: {SecondStepResponse[SEARCH_1][field_to_look]} to
            {SecondStepResponse[SEARCH_1][opponent_idea]},
retrieved_arguments: {ThirdStepQuery1[arguments]},
justification: {ThirdStepQuery1[justification]}},
{SEARCH_2: [...]},
{SEARCH_3: [...]}}
Now respond in a JSON format with the keys:
- aspect_argument_1: a string with the exact text of the aspect of the first selected argument.
- retrieved_argument_1: a string with the exact text of the first selected argument.
- justification_argument_1: a string with the justification of the selection of the first selected argument.
- aspect_argument_2: a string with the exact text of the aspect of the second selected argument.
- retrieved_argument_2: a string with the exact text of the second selected argument.
- justification_argument_2: a string with the justification of the selection of the second selected argument.
- aspect_argument_3: a string with the exact text of the aspect of the third selected argument.
- retrieved_argument_3: a string with the exact text of the third selected argument.
- justification_argument_3: a string with the justification of the selection of the third selected argument.
```

Figure 6: Prompt used in the fourth step of our Retrieval-Augmented Debating Subtask 1 experiment.

Prompt step 5 (Subtask 1)

You are an expert debater and you are having a dialogue with another user. Your task is to generate a response to your last opponent message. Keep in mind that the idea you are defending is contrary to the one he is showing.
Specific information about the established debate is included below.
{debate_dialogue}
Your answer should take into account the following parameters: { tone: {firstStep[tone]},
justification_tone: {firstStep[justification_tone]},
style: {firstStep[style]},
justification_style: {firstStep[justification_style]},
perspective: {firstStep[perspective]},
justification_perspective: {firstStep[justification_perspective]},
}
Retrieved arguments to elaborate the answer: {responseFourhtStep}
Please response with JSON format that contain as keys the words:
- response: a string with the response to your last opponent message. This answer has to be of a maximum of 60 words.
- justification: a string with the explanation about how you generate your answer.
- arguments: a list of the string with the exact retrieved arguments that you used to elaborate the final answer and that have relation with your response.
Moreover take in account that your answer must meet the following criteria:
- Quantity: be informative and provide enough information for the opponent to understand your position.. Does the response contain at least one (attack or defense) argument, and at most one of each type of defense and attack?
- Quality: be truthful. Can the response be deduced from the retrieved arguments?
- Relation: be relevant. Is the response coherent with the conversation and does it express a contrary stance to the user?
- Manner: be clear. Is the response clear and precise?
- Length: be of at most 60 words

Figure 7: Prompt used in the fifth step of our Retrieval-Augmented Debating Subtask 1 experiment.

A.2. Subtask 2: Evaluation of debate systems

Prompt ZSL (Subtask 2)

You are an expert in evaluating and analyzing the quality of the answers and arguments used during various rounds of debate. Your task is to evaluate the last answer of the system based on the {evaluation_metric} metric:
{evaluation_question_metric}
Bellow you have the conversation and the arguments retrieved:
{debate_dialogue}
Please respond only with yes or no to the question about {evaluation_metric}({evaluation_metric_question})

Figure 8: Prompt used in our Zero-shot Learning experiment for Retrieval Augmented Debating subtask 2.

Prompt FSL (Subtask 2)

You are an expert in evaluating and analyzing the quality of the answers and arguments used during various rounds of debate. Your task is to evaluate the last answer of the system based on the {evaluation_metric} metric:
{evaluation_question_metric}
Now I will give you some examples of debates conversations and their annotations:
Example 1:
User: My claim, put clearly: I argue that AI cannot feel emotions in the same complex, profound, and genuine way that humans do.
System: That assertion is undermined by the fact that "how humans feel" can be translated into data and taken into consideration, even if an AGI can't itself feel.
Arguments:
- How humans feel can be translated into data and taken into consideration, even if an AGI cant́ itself feel.
Evaluation: yes
Example 2:
User: My opinion is that in the UK, law-abiding citizens should be able to have non-lethal and potentially lethal weapons for their homes and personal protection, not to be reckless, but to have a means to self-defend themselves from immediate threats when emergency services may not be reachable right away. System: That argument ignores the fact that the UK's lower crime rate and more stringent gun laws demonstrate that restricting access to firearms does not necessarily compromise public safety.
Arguments:
We have some limits. The right of self defense DOES entail the right to own weapons adequate to meet the threat. Approximately 100 million AR15 rifles have been sold to the public. They're common. They're easy to manufacture with basic shop tools. This means criminals will always have access to them. Adequate defensive capability means the people must have access as well.
Evaluation: no
Now is your turn. Below you have the information of the debate: {debate_dialogue}
Please respond only with yes or no to the question about {evaluation_metric} ({evaluation_question_metric})

Figure 9: Prompt used in our Few-shot Learning experiment for Retrieval Augmented Debating subtask 2.

Analysis Strategy Prompt for Step 1 (Subtask 2)

You are an expert in evaluating and analyzing the quality of the answers and arguments used during various rounds of debate. Your task is to evaluate the last answer of the system based on the {evaluation_metric} metric:
{evaluation_question_metric}
Now I will give you some examples of debates conversations and their response to the previous question:
{positive_examples}
{negative_examples}
Please respond only with a JSON format that contain the following keys:
- yes: a string with the reasons that could help a human annotator to respond to the quantity question with yes.
- no: a string with the reasons that could help a human annotator to respond to the quantity question with no.

Figure 10: Prompt used in the first step of our Analysis experiment for Retrieval Augmented Debating subtask 2.

Analysis Strategy Prompt for Step 2 (Subtask 2)

You are an expert in evaluating and analyzing the quality of the answers and arguments used during various rounds of debate. Your task is to evaluate the last answer of the system based on the {evaluation_metric} metric:
{evaluation_question_metric}
Now I will give you some reason to answer with yes or no. - Respond with yes if {answer_model_step1[yes]}.
- Respond with no if {answer$_{m}odel\_step1[no]\}$.
Now is your turn. Below you have the information of the debate:
{debate_dialogue}

Please respond only with yes or no to the question about evaluation_metric} ({evaluation_question_metric})

Figure 11: Prompt used in the second step of our Analysis experiment for Retrieval Augmented Debating subtask 2.