Argument-Based Comparative Question Answering Evaluation Benchmark

Irina Nikishina[1] **Saba Anwar**[1] **Nikolay Dolgov**[2]
Maria Manina[2] **Daria Ignatenko**[2] **Viktor Moskvoretskii**[2,4]
Artem Shelmanov[3] **Tim Baldwin**[3] **Chris Biemann**[1]

[1]University of Hamburg [2]HSE University [3]MBZUAI [4]Skoltech

Correspondence: irina.nikishina@uni-hamburg.de

Abstract

Despite the ability of large language models (LLMs) to generate coherent comparative answers, automatic comparative question answering (CQA) remains challenging due to the absence of standardized evaluation criteria and the high resource demands of manual assessment. To address these problems, this paper proposes a comprehensive evaluation framework designed to assess the quality of CQA summaries using LLMs-as-a-Judge. We formulate 15 evaluation criteria for assessing comparative answers generated by various sources, including LLMs, human experts, and prior work. To capture a diverse range of comparative answers, LLM summaries were generated under various prompting scenarios. We evaluate the effectiveness of our framework using both human assessment and LLMs, demonstrating the consistency between automated and manual evaluations. Finally, we fine-tune Llama-3-8B-Instruct on a dataset generated from the best-performing CQA models in our evaluation benchmark to assess whether open-source models can replace closed-source LLMs. All used data, code, and evaluation results are publicly available[1].

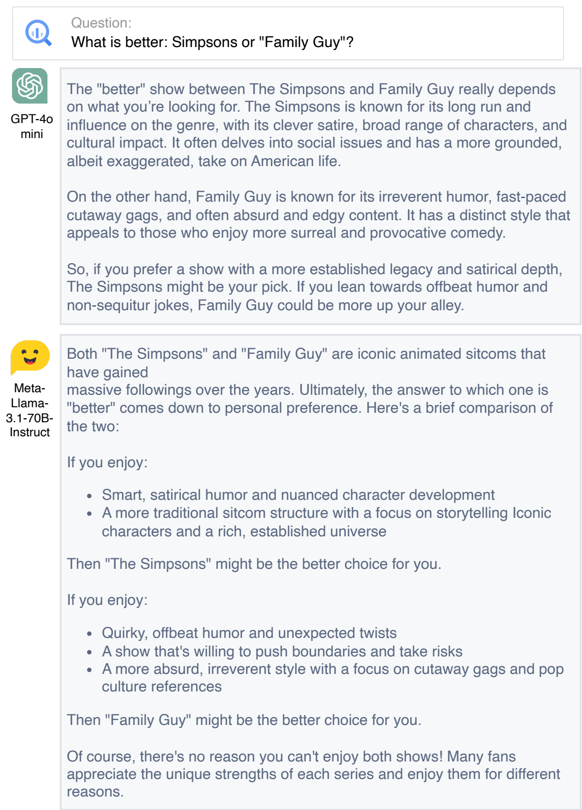

Figure 1: Examples of LLM answers to comparative questions (shortened).

1 Introduction

Are *"Nike"* sneakers better than *"Adidas"*? Is an *"iPhone"* easier to use than an *"Android"* phone? What is faster, *"Windows"* or *"Linux"*? Such questions are common user requests on the web. Making an informed choice requires evaluating and comparing various characteristics of the options under consideration. For instance, when buying a car, it is important to consider the brand reputation, the fuel efficiency, comfort, and general performance. Conducting thorough comparisons manually by scanning different websites requires time and effort, especially given the vast amount of data available online (Iso et al., 2022). In such cases, a comparative answer or summary can serve as an effective and time-saving solution.

Comparative Question Answering (CQA) is a Natural Language Processing task that aims to answer requests containing two or more objects by choosing one of them and supporting the answer with arguments. The final goal is not necessarily to come to a distinct final choice, but to describe each object's advantages and disadvantages, taking into account the specified aspect(s). This task can also be viewed as an abstractive comparative summarization. In this paper, we focus on comparing only two objects, leaving the cases with three or more

[1]`https://github.com/lilaspourpre/cqa-evaluation-framework/README.md`

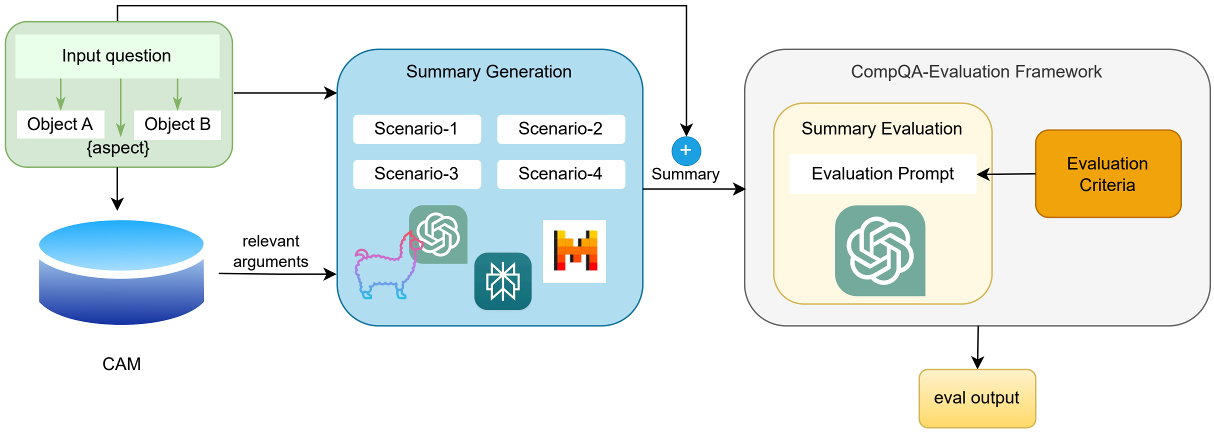

Figure 2: Overall pipeline of the CompQA evaluation dataset creation and evaluation framework.

for future research. Unlike traditional opinion summarization, which focuses on aggregating views about a single object, the task of generating comparative summaries is more complex, as it involves evaluating each object in the context of the other. Traditional approaches to this task focus only on the initial steps of CQA that precede comparative answer generation and only partially address the Argument Summarization task (Chekalina et al., 2021; Shallouf et al., 2024) for answer generation.

Large Language Models (LLMs) have emerged as powerful tools for a wide range of tasks, including search and recommendations. They significantly simplify the process of generating comparative answers, often providing coherent and well-structured responses (see example in Figure 1). However, the quality of these responses varies greatly, posing a persistent challenge in the automated evaluation of such summaries. Prior work suggests that a well-constructed abstractive summary should be *coherent*, *concise*, *factually consistent*, *relevant*, *non-redundant*, *grammatically accurate*, and exhibit *high readability* (Gupta and Gupta, 2019; Kryscinski et al., 2019; Shakil et al., 2024). Achieving this type of qualitative assessment often requires human evaluation, guided by a structured framework that decomposes the evaluation into specific criteria. Importantly, this framework may vary depending on the nature of the summary: for instance, summarizing a Wikipedia article requires different criteria (e.g. factual sequences) than summarizing a comparison between two products. Consequently, there is no consensus regarding the actual criteria a summary must follow. Moreover, manual evaluation is a tedious task; therefore, its automation is highly beneficial for the research and development of CQA systems.

In this paper, we aim to fill the mentioned gaps in benchmarking the CQA systems and answer the following research questions: *(RQ1) Which criteria should be considered for assessing a comparative summary?*, *(RQ2) Can LLMs reliably evaluate comparative summaries with human expert-level quality?*, *(RQ3) How do different LLMs fare against each other in generating high-quality summaries?*

The contributions of this work are as follows:

1. Systemizing previous efforts in CQA, we develop 15 criteria for scoring comparative answers and implement automatic CompQA evaluation pipeline based on a LLM-as-a-Judge approach (Figure 2) that is able to assess CQA summaries on the basis of these criteria. We verify the usefulness of the proposed benchmark by comparing automatic assessments with human judgements. As a part of benchmark, we create two datasets of 50 (main) and 432 (extended) comparative questions with object pairs, aspects, and arguments for model evaluation.

2. Using the developed benchmark, we conduct the first automatic and manual evaluation of several LLMs on the CQA task. We utilize GPT-3.5, GPT-4, Llama-3-8B-Instruct, Llama-3-70B-Instruct, Perplexity, and Mixtral for the main experiments, while Gemma-3-27B, Qwen3-8B, DeepSeek-R1, and Gemini-2.5-Flash are used for additional experiments. To ensure a comprehensive performance evaluation, we use CQA datasets sourced from (Nikishina et al., 2025), CAM 2.0 (Shallouf et al., 2024), and Yahoo! Answers (Chekalina et al., 2021), along with comparative answers gener-

ated by a subset of these LLMs under various prompting scenarios.

3. We demonstrate that the performance gap between open- and closed-source models can be narrowed by fine-tuning Llama-3-8B-Instruct on a high-quality dataset distilled from our evaluation benchmark.

All code, datasets, and evaluation results are available online[1].

2 CompQA Evaluation Framework

Defining criteria for a high-quality answer is crucial, as checking only for coherence and logic—or using metrics like word and vector similarity—is not enough. This is especially relevant to our usecase: when generating answers, we use a list of arguments, and the output is directly tied to the complexity and level of detail of the input prompt (Loya et al., 2023). Thus, the generated answer can vary from an unstructured paragraph to a well-organized summary with clearly defined components, depending entirely on the input. This also means that two summaries comparing the same pair of objects can have a low token-match metric if different arguments were used, despite the fact that both may still be high-quality comparative texts.

Assuming two objects are to be compared, we define a good-quality comparative answer as one that helps the user *decide* between the objects by comparing several *relevant aspects*. These aspects can be general or specified by the user. In establishing our criteria for such an answer, we adopt metrics that align with human perception of quality, requiring that a summary should be *well-structured*, *concise*, *factual*, *relevant*, *coherent*, *informative*, and *non-redundant*. Most of these criteria are primarily adapted from the work of Peyrard (2019) and Kryscinski et al. (2019), as well as recent findings on LLM-based human-like evaluation (Gao et al., 2023; Liu et al., 2023). Thus, to answer the **RQ-1** on developing criteria for the comparative question-answer summary, we define and categorize the following dimensions to assess generated answers:

1. *structure*: a well-defined structure that has: a) a short introductory summary, b) a list of *named aspects* with *short descriptions*, and c) a distinct *final choice*.

2. *relevancy*: arguments are relevant to the aspect of comparison (if given), compare both objects, and are ordered from more generic to specific ones or vice versa.

3. *quality*: should further fulfill the following aspects: *concise* (optimal length focusing on key information, should be between 12 to 20 sentences, aligning with the dataset's argument distribution to ensure comprehensive yet brief coverage), *factual* (no hallucination or contradictory arguments), *informative* (if the summary is required for a specific aspect, then the introduction, arguments, and conclusion should incorporate that, otherwise they should be generic), *coherence* (comparison logic is easy to follow, statements are not self-contradictory or repetitive, the conclusion does not contradict the arguments).

Characterstics	
Total Pairs	50
Pairs with defined aspect	20
Average # of arguments	10
Min # of arguments	3
Max # of arguments	20
Pairs with more than average # of arguments	23

Table 1: Statistics of data retrieved from (Chekalina et al., 2021) and CAM 2.0 (Shallouf et al., 2024).

2.1 Automatic Evaluation using LLM-as-a-Judge

In this section, we present an evaluation framework called CompQA that implements these quality checks with 15 questions and assigns a score between 0 and 19 points. Each question assigns a maximum score of 1 or 2, depending on various scenarios. The criteria are developed to simplify the annotation and increase inter-annotation agreement of human annotators. Figure 3 presents the CompQA framework and explains each scored point for each criterion in detail.

To automatically score the CQA answers, we prompt an LLM. The template is presented in Figure 5. Because of appendix size limitation, the complete prompt scenario can be found in the Github repository[1]. The output of the model is expected to be a JSON dictionary where the keys are criteria numbers and values are the scores assigned by LLM. After preliminary experiments, we have decided not to require models to provide the total score, as the total value is usually not the same as the actual sum. The model provides the scores according to each criterion.

Figure 3: CompQA evaluation framework criteria.

2.2 Question and Argument Collection

We also construct a novel dataset of questions and arguments for our benchmark. First, we randomly select 50 pairs from the Touché dataset (Bondarenko et al., 2022b) and (Chekalina et al., 2021) and retrieve a maximum of 10 arguments for each object from the CAM 2.0 system (Shallouf et al., 2024). More information about these datasets can be found in the Github repository[1]. The exact number varies depending on factors like the popularity of the object, the availability of arguments in its favor on online forums, etc. The pairs from both sources belong to various domains, e.g. companies (*"IBM vs Sony"*), places (*"Virginia vs. Michigan"*), and products of different categories (*"PS3 vs. DS"* and *"tea vs coffee"*), etc. Table 1 provides a few key details about this dataset. On average, each pair contains 10 arguments, with the number ranging from 3 to 20. Only 18 pairs are provided a defined aspect as input. If available, we also extract the questions for the pairs; otherwise, a basic question is used as a default: *"What is better: {object1} or {object2}? Focus on {aspect}."*

Additionally, we have extended the dataset to 432 manually verified questions with extracted objects and aspects using the CompQA dataset of Beloucif et al. (2022). Further benchmark experiments on those datasets are provided in Appendix C.

2.3 Prompt Scenarios for Comparative Answer Generation

We use several prompts with varying levels of complexity and specificity to get summaries from the LLMs for the following reasons:

- check whether our framework can differentiate between good and bad summaries;
- to assess the capability of LLMs to produce comparative answers of good quality with and without provided arguments;
- how the quality of the output answer can vary with prompt-engineering for the same model.

Figure 4 shows the exact content of these scenarios. The **first** scenario is the simplest one that does not include any details from the user other than the objects to compare. It should produce a summary

1st	Compare {*A*} and {*B*}
2nd	You are a helpful assistant. Compare {*A*} and {*B*} using following arguments: {*ARGS*}
3rd	You are an analyst, write a 300-word comparison of {*A*} and {*B*}. Task: compare and choose the better of the two. Requirements: - be concise - think of the most relevant arguments only. Needed structure: - summary (100 words) - bullet-point list of main aspects of comparison (200 words or more) - the best option (1 word)
4th	You are an analyst, write a 300-word comparison of {*A*} and {*B*}. Task: compare and choose the better of the two. Requirements: - be concise - analyse the list of arguments below - pick relevant ones - rephrase in your own words - cite used argument numbers in square brackets right after the usage - the summary needs to have 15 arguments, create some if needed (add a [generated] tag) Needed structure: - summary (100 words) - bullet-point list of main aspects of comparison (200 words or more) - the best option (1 word) - numbered list of used arguments Argument list:{*ARGS*}

Figure 4: Four prompting scenarios used for comparative summary generation. Objects "A" and "B" are extracted from the input question, "ARGS" is the list of arguments extracted from CAM (Shallouf et al., 2024).

with an arbitrary structure. It is important to note that the remaining three scenarios assign roles to the LLM, whereas the first one does not.

The **second** scenario gives the LLM a list of arguments (ARGS) extracted from CAM, to see if the LLM uses the given arguments exclusively or generates new ones.

The **third** and **fourth** scenarios add more instructions to guide the LLM in producing a summary with a specific structure as described in Section 2. The **third** scenario removes CAM arguments, testing the LLM's ability to generate its own arguments. Lastly, the fourth and most comprehensive scenario includes CAM arguments and improves upon the specific instructions. This configuration is based on the methodology established by Nikishina et al. (2025) to generate ideal comparative summaries. Due to their increased complexity, the third and fourth scenarios aim to produce a summary that should score better than other scenarios.

3 Experimental Setup

In this section, we introduce the models we have selected as potential LLM cores for our agent. Our

Summary Evaluation Prompt

You are a helpful assistant.

Task:
- analyze the comparison given
- for each criterion, assign points in the range given
Criteria: {*CRITERIA*}
Output a python dictionary with the structure: {"n": score, "n+1": score}
Write only the dictionary, do not write anything else

Few-shot Examples
Example 1: {*SUMMARY1*}
Scoring 1: {1:score, 2:score, ... ,15:score}
Example 2: {*SUMMARY2*}
Scoring 2: {1:score, 2:score, ... ,15:score}
Example 3: {*SUMMARY3*}
Scoring 3: {1:score, 2:score, ... ,15:score}

Question: What is better {*ASPECT*}: {*A*} or {*B*}? or {*SPECIFIC_QUESTION*}
Comparative answer: {*SUMMARY*}

Figure 5: Evaluation prompt for LLMs based on the CompQA evaluation framework criteria. For few-shot, 2 human answers are added to the placeholder.

methodology employs these models both to generate comparative answer datasets using the prompts shown in Figure 4, and to serve as evaluators for our proposed framework. To ensure a rigorous comparison using the CompQA benchmark, we measure their performance alongside manual human assessments.

Participated CQA Agents. Here are short descriptions of LLMs we used for competition in the CompQA Benchmark. We use the standard configuration and parameters for generation:

- **ChatGPT** (GPT-3.5-turbo), details are described by Brown et al. (2020);
- **GPT-4o**, more details in OpenAI (2023);
- **meta-llama/Meta-Llama-3-8B**, more details in Dubey et al. (2024);
- **meta-llama/Meta-Llama-3-70B**, more details in Dubey et al. (2024);
- **mistralai/Mixtral-8x7B-Instruct-v0.1** is a pretrained generative Sparse Mixture of Experts (Jiang et al., 2024);
- **Perplexity AI**[2] — AI-powered research and conversational search engine that answers queries using natural language predictive text, it uses sources from the web and cites links within the text response;

For each LLM, we generate four summaries as per each prompt template shown in 4. We also return three versions of the answer for the GPT family of models (choices, $n = 3$), which are treated and evaluated as separate summaries. In total, our generative pipeline produced **2,000** comparative

[2] https://www.perplexity.ai

answers (comprising 50 examples across 4 scenarios for the 4 models, and 150 examples across 4 scenarios for 2 GPT-based models).

Additional CQA Datasets. In addition to using the LLM agents to generate comparative answers, we also use the following comparative answer datasets in our evaluation benchmark.

- **CAM** dataset comprises the summaries provided by Shallouf et al. (2024). These were produced using the *"lmsys/vicuna-7b-v1"* model for all **50** pairs used in our dataset.
- **Yahoo!Answers** dataset from Chekalina et al. (2021) comprises questions and answers (written by humans). It only contains **28** object pairs from our dataset.
- **Human** dataset – this dataset is sourced from Nikishina et al. (2025) and consists of **80** comparative summaries. These were initially generated by ChatGPT using *Scenario 4* and subsequently refined by four experts in computational linguistics to ensure maximum quality and factual accuracy.

Obtaining Automatic Assessments using LLMs. We automatically evaluated 2,158[3] answers with all participating LLMs based on the CompQA evaluation framework using the 2-shot prompting.

Obtaining Assessments from Human Annotators. To verify the quality of automatic annotation, a total of *367* answers are randomly selected and manually reviewed by expert annotators (the authors). *123* of them are done with an overlap of two annotators to measure the annotation agreement. Krippendorf's alpha is equal to **0.75** for the final score and **0.71** for all scores on average.

3.1 Comparison of LLM and Human Assessments

In order to understand whether the human annotation can be replaced with an automatic one, we compare the agreement and annotation scores between different LLMs and human answers. Table 2 demonstrates both the agreement score (Krippendorff's alpha) and Spearman correlation scores, following Bavaresco et al. (2024). The results give a positive answer to the **RQ-2**: "LLMs can reliably evaluate comparative summaries with human expert-level quality."

[3] $4scenarios \cdot 150_{GPT-3.5} + 4scenarios \cdot 150_{GPT-4} + 28_{Yahoo} + 50_{CAM} + 80_{Human} + 4scenarios \cdot 50 \cdot 4_{Llama3-8B,Llama3-70B,Mixtral,Perplexity} = 2158$

Model	α	**Spearman's**
GPT-4o, separately	**0.71**	0.69, $p < 0.001$
GPT-4o, final score	**0.58**	0.55, $p < 0.001$
GPT-3.5, separately	0.63	0.63, $p < 0.001$
GPT-3.5, final score	0.31	0.40, $p < 0.002$
Perplexity, separately	**0.71**	0.72, $p < 0.001$
Perplexity, final score	0.39	0.60, $p < 0.001$
Mixtral, separately	0.55	0.69, $p < 0.001$
Mixtral, final score	0.22	0.62, $p < 0.001$
Llama-3-8B, separately	0.48	0.49, $p < 0.001$
Llama-3-8B, final score	0.41	0.46, $p < 0.001$
Llama-3-70B, separately	0.61	**0.76**, $p < 0.001$
Llama-3-70B, final score	0.47	**0.72**, $p < 0.001$

Table 2: Agreement (Krippendorf's α) and correlation scores between human and LLM evaluations. Separate scores are calculated for all scores concatenated for all answers (human annotation against model annotation), the total scores represent the sums of 15 criteria (denoted in Figure 3) for each answer.

From the results in Table 2, we can see that both agreement and correlation scores are much higher for the separate scores than for the summary scores. Moreover, we can also conclude that Llama-3-70B is better according to Spearman's correlation, while GPT-4o and Perplexity show higher agreement according to Krippendorff's α.

4 CQA Systems Performance

In this section, we present the results of existing CQA agents on our benchmark, and answer the **RQ3**: "How do different LLMs fare against each other in generating high-quality summaries?" Table 3 shows the best results obtained by each comparative answer dataset (row) on our benchmark with different evaluators (column). For human evaluations, scores are reported for a randomly sampled dataset of 367 summaries. The results of all LLM evaluators unanimously rate the GPT-4o answers as the first. The second place is shared between Mixtral and Perplexity. Regarding human evaluations, it is evident that the highest quality responses are those generated by humans (**16.69** on average), while the best LLM answers are still generated by GPT-4. Interestingly, the answers ranked as the lowest belong to Yahoo and CAM datasets, primarily because these datasets do not align with the structural criteria used by our evaluation framework. To see the detailed results, refer to Table 5 in Appendix A.

	GPT-4o	GPT-3.5	Perplexity	Mixtral	Llama-3-8B	Llama-3-70B	Human*
GPT-4o	**17.77**$_{\pm 1.06}$	**17.54**$_{\pm 2.14}$	**18.69**$_{\pm 0.69}$	**18.45**$_{\pm 2.64}$	**16.64**$_{\pm 1.86}$	**18.40**$_{\pm 0.70}$	15.96$_{\pm 1.35}$
GPT-3.5	16.66$_{\pm 1.86}$	16.05$_{\pm 2.62}$	16.12$_{\pm 2.64}$	18.05$_{\pm 1.33}$	15.08$_{\pm 3.18}$	16.25$_{\pm 2.32}$	14.58$_{\pm 2.34}$
Perplexity	17.42$_{\pm 1.34}$	16.52$_{\pm 3.00}$	18.47$_{\pm 1.11}$	18.34$_{\pm 0.87}$	16.31$_{\pm 2.06}$	17.87$_{\pm 1.15}$	15.00$_{\pm 1.77}$
Mixtral	17.20$_{\pm 1.31}$	17.05$_{\pm 2.43}$	17.92$_{\pm 1.41}$	18.36$_{\pm 0.94}$	16.29$_{\pm 2.40}$	17.45$_{\pm 1.67}$	13.75$_{\pm 2.63}$
Llama-3-8B	16.58$_{\pm 1.80}$	15.95$_{\pm 2.60}$	17.19$_{\pm 1.81}$	18.08$_{\pm 1.24}$	15.56$_{\pm 2.72}$	16.15$_{\pm 1.90}$	14.77$_{\pm 2.16}$
Llama-3-70B	17.12$_{\pm 1.37}$	16.22$_{\pm 2.76}$	17.93$_{\pm 1.39}$	18.19$_{\pm 1.068}$	15.70$_{\pm 2.95}$	17.27$_{\pm 1.45}$	15.37$_{\pm 1.72}$
CAM (Shallouf et al., 2024)	9.52$_{\pm 3.808}$	8.74$_{\pm 4.56}$	6.46$_{\pm 2.89}$	9.66$_{\pm 4.43}$	6.54$_{\pm 4.21}$	6.40$_{\pm 2.29}$	8.04$_{\pm 2.78}$
Yahoo (Chekalina et al., 2021)	9.32$_{\pm 3.95}$	5.88$_{\pm 4.86}$	5.96$_{\pm 3.68}$	13.32$_{\pm 4.97}$	5.21$_{\pm 5.91}$	5.96$_{\pm 3.64}$	6.00$_{\pm 1.00}$
Human	17.09$_{\pm 1.72}$	15.29$_{\pm 2.68}$	18.33$_{\pm 1.03}$	17.65$_{\pm 1.57}$	14.29$_{\pm 3.23}$	17.96$_{\pm 1.56}$	**16.69**$_{\pm 1.76}$

Table 3: Average scores for all participating models for LLM and human evaluations. Rows are the datasets. Columns represent the evaluation models. Human* evaluations were conducted on a random subset.

Comparison between scenarios. Figure 6 shows average scores for all models. Scenarios 3 and 4 achieved the highest scores from both human and Llama-3-70B evaluators. Conversely, the summaries generated by Scenarios 1 and 2 are consistently ranked lowest. This highlights the fact that the evaluation framework can help to differentiate between summaries of varied structures. The fourth scenario uses arguments from CAM, and they have scored lower than their similar counterpart without these arguments i.e. the third scenario, which is more pronounced in the human evaluation. This discrepancy raises concerns about the quality of CAM arguments, despite the prompt explicitly requiring the selection of only relevant arguments. Thus, we further compare these models for all criteria of our evaluation framework, by grouping them into *structure*, *quality*, and *relevance* in Figure 7. The comparison shows that the summaries generated with the third and fourth scenarios scored higher for all categories of our framework. The second scenario suffered noticeably for *structure*, this difference is less noticeable between the third and fourth scenarios.

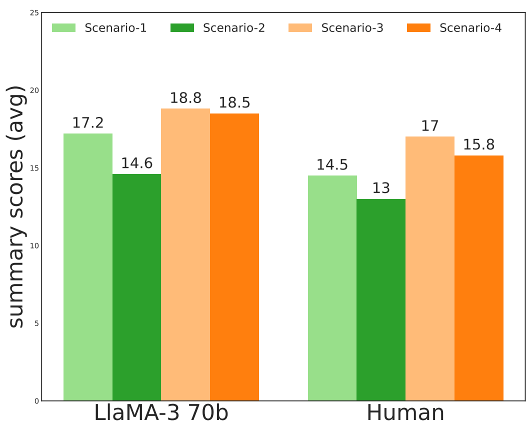

Figure 6: Average scores for each scenario for Llama-3-70B and manual human assessments.

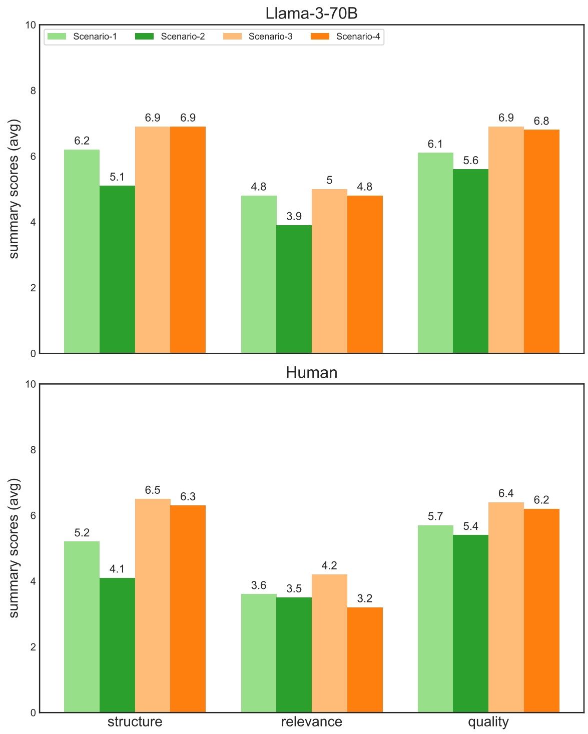

Figure 7: Average scores for each scenario, distributed into three categories of evaluation framework: **structure**–(7 points), **quality**–(5 points), and **relevance**–(7 points).

Comparison between Llama-3-70B and human: Human evaluations are consistently lower than Llama-3-70B. Table 4 shows that Llama-3-70B performs similarly for the questions with and without aspect, while the human scores are lower when the aspect is presented, which implies that Llama-3-70B might not give the importance to the required aspect when evaluating the summary.

Additional analysis is also present in Figure 8: we aim to check, whether the scores assigned by Llama-3-70B to the answers evaluated by humans for the extreme cases are coherent. Human evaluation shows average results (with means at 14-17) for the cases evaluated as good ones by Llama-3-70B, and also low results for the answers assigned

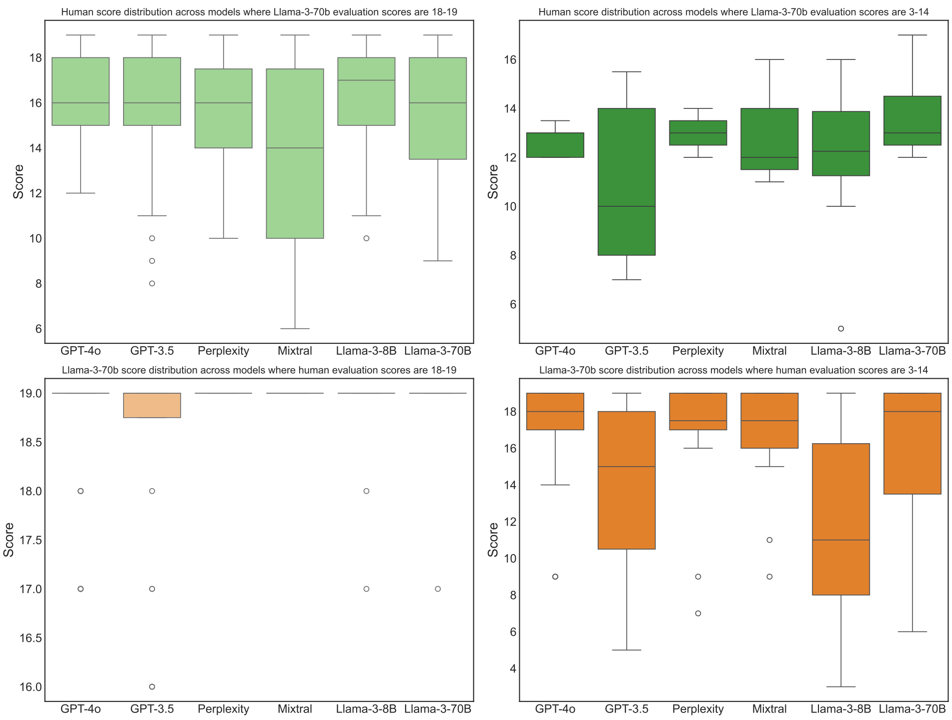

Figure 8: Distribution scores across models for the highest and lowest scores assigned by Llama-3-70B and Human evaluations.

Aspect	Llama-3-70B	Human
yes	$16.43_{\pm 3.52}$	$13.43_{\pm 2.98}$
no	$16.44_{\pm 3.52}$	$15.01_{\pm 2.69}$

Table 4: Average scores of human and Llama-3-70B, for the subset of answers **with** and **without** aspect.

with low scores by the LLM. It means that the model slightly tends to overestimate the performance of the LLMs in comparison to the human evaluation. From the second row, we can see that Llama-3-70B assigns high scores for the answers which are also ranked high by humans (lower left subfigure) and both high and low scores for the answers ranked as "bad" (3-14) by humans (lower right subfigure).

Distilling a Better Open-Source CQA Agent Our initial results indicate that the highest performance on the benchmark is currently dominated by proprietary, closed-source models such as GPT-4o and Perplexity (Table 3). In contrast, Llama-3-8B-Instruct, the smallest LLM used, performs decently. Motivated by its potential, we investigate whether supervised fine-tuning (SFT) with LoRA adapter can close the gap between this open-source baseline and top closed-source models. We compiled the training data using the answers from the models gaining $>=$ 18 points. The training data comprises 1244 question-answer pairs in total. The best SFT variant scored $\mathbf{18.40}_{\pm 1.25}$, which performs on par with GPT-4o; detailed fine-tuning parameters are provided in Appendix B. Overall, the results show that distilling the capabilities of larger models into smaller ones remains promising.

Additional Experiments We have also conducted additional experiments on reasoning LLMs, extended dataset, and possible biases that are presented in Appendix C.

5 Related Work

In this section, we briefly introduce each subtask for Comparative Question Answering and also discuss the existing LLM evaluation benchmarks.

5.1 Comparative Question Answering

Here, we introduce each subtask and list several papers that addressed the topic.

Comparative Question Identification aims at classifying questions into two types: comparative and non-comparative. This classification task is solved with both Encoder and Decoder Transformer models (Bondarenko et al., 2020, 2022a; Shallouf et al., 2024). **Object and Aspect Identification** is a sequence labelling task, aims at finding objects and aspect of comparison in the question. There exist varios datasets and approaches to solve the task, mostly, with Trasnformer models (Chekalina et al., 2021; Beloucif et al., 2022; Bondarenko et al., 2022a; Shallouf et al., 2024). **Stance Classification** is another classification task, that identifies the stance of comparative sentences. Panchenko et al. (2019), Bondarenko et al. (2022a), and Kang et al. (2023) solve the task using standard ML classifier, Encoder-based Transformer, and GPT-4o respectively. **Summary Generation** is only partially tackled by Chekalina et al. (2021) and Shallouf et al. (2024). The closest work on multi-document summarization of differing opinions is by Iso et al. (2022), which focuses on aggregating diverse opinions and synthesizing them into a coherent summary.

5.2 LLM Evaluation Benchmarks

Apart from the well-known benchmarks like SuperGLUE (Sarlin et al., 2020), MTEB (Muennighoff et al., 2023), and SQuAD (Rajpurkar et al., 2016), several more challenging benchmarks have gained prominence: MMLU (Massive Multitask Language Understanding) (Hendrycks et al., 2021), TruthfulQA (Lin et al., 2022), BIG-Bench Hard (Suzgun et al., 2023). The LLM evaluation framework proposed by (Chiang and Lee, 2023) involves presenting a Large Language Model with task instructions, a sample to be evaluated and a question. The researchers use LLM evaluation to score parts of both generated and human-written stories. To facilitate comprehensive evaluations, several initiatives aggregate multiple benchmarks: Hugging Face's Big Benchmarks Collection (a centralized platform for various leaderboards), and LMSys Chatbot Arena (Chiang et al., 2024; Zheng et al., 2023): it also utilizes user ratings and GPT-4o evaluations to assess chatbot performance.

6 Conclusion

We have proposed a comprehensive CompQA evaluation framework based on 15 various criteria to be used in scoring comparative answers, demonstrating that GPT-4 produces the best answers to comparative questions. We have manually evaluated the answers from a range of models and datasets using our framework and compared human scores with LLM evaluations, showing that LLM results have a strong correlation with human expert-level evaluations. Moreover, we have trained a smaller model Llama-3-8B-Instruct to generate summaries at the similar level of quality as closed models.

7 Limitations

The main limitations of the paper are as follows:

- Regarding the human evaluation dataset, we acknowledge that our sample size is quite small and might be extended even further to make comparisons of better quality.
- The arguments used in the strategies are solely derived from the CAM framework. This reliance on a single source may have constrained the retrieval performance, particularly in Scenario 2 and 4, leading to suboptimal results. To address this, future work could involve annotators in crafting or refining arguments, which may enhance the robustness and effectiveness of the strategies.

8 Ethical Considerations

In our benchmark we test multiple LLMs, one key concern is the handling of user data by proprietary models like OpenAI's GPT-4, which are developed and maintained by private companies. These companies often retain the right to use input data for improving their models, as stated in their terms of service and privacy policies. As a result, personal or sensitive information provided by users during interactions with these models could be logged, stored, or analyzed for commercial purposes. While companies may anonymize data, the potential use of personal information in ways that users may not fully understand or consent to raises significant privacy concerns.

References

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, André F. T. Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K. Surikuchi, Ece Takmaz, and Alberto Testoni. 2024. Llms instead of human judges? A large scale empirical study across 20 NLP evaluation tasks. *CoRR*, abs/2406.18403.

Meriem Beloucif, Seid Muhie Yimam, Steffen Stahlhacke, and Chris Biemann. 2022. Elvis vs. M. Jackson: Who has more albums? classification and identification of elements in comparative questions. In *Proceedings of the Thirteenth Language Resources* *and Evaluation Conference*, pages 3771–3779, Marseille, France. European Language Resources Association.

Alexander Bondarenko, Yamen Ajjour, Valentin Dittmar, Niklas Homann, Pavel Braslavski, and Matthias Hagen. 2022a. Towards understanding and answering comparative questions. In *WSDM*, pages 66–74. ACM.

Alexander Bondarenko, Pavel Braslavski, Michael Völske, Rami Aly, Maik Fröbe, Alexander Panchenko, Chris Biemann, Benno Stein, and Matthias Hagen. 2020. Comparative web search questions. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 52–60.

Alexander Bondarenko, Maik Fröbe, Johannes Kiesel, Shahbaz Syed, Timon Gurcke, Meriem Beloucif, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, and Matthias Hagen. 2022b. Overview of Touché 2022: Argument Retrieval. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 13th International Conference of the CLEF Association (CLEF 2022)*, volume 13390 of *Lecture Notes in Computer Science*, Berlin Heidelberg New York. Springer.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *NeurIPS*.

Viktoriia Chekalina, Alexander Bondarenko, Chris Biemann, Meriem Beloucif, Varvara Logacheva, and Alexander Panchenko. 2021. Which is better for deep learning: Python or MATLAB? answering comparative questions in natural language. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 302–311, Online. Association for Computational Linguistics.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *Preprint*, arXiv:2403.04132.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 82 others. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. 2023. Human-like summarization evaluation with chatgpt. *Preprint*, arXiv:2304.02554.

Som Gupta and S. K Gupta. 2019. Abstractive summarization: An overview of the state of the art. *Expert Systems with Applications*, 121:49–65.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *ICLR*. OpenReview.net.

Hayate Iso, Xiaolan Wang, Stefanos Angelidis, and Yoshihiko Suhara. 2022. Comparative opinion summarization via collaborative decoding. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3307–3324, Dublin, Ireland. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, and 7 others. 2024. Mixtral of experts. *CoRR*, abs/2401.04088.

Inwon Kang, Sikai Ruan, Tyler Ho, Jui-Chien Lin, Farhad Mohsin, Oshani Seneviratne, and Lirong Xia. 2023. Llm-augmented preference learning from natural language. *CoRR*, abs/2310.08523.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 540–551, Hong Kong, China. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Manikanta Loya, Divya Sinha, and Richard Futrell. 2023. Exploring the sensitivity of LLMs' decision-making capabilities: Insights from prompt variations and hyperparameters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3711–3716, Singapore. Association for Computational Linguistics.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Irina Nikishina, Saba Anwar, Nikolay Dolgov, Maria Manina, Daria Ignatenko, Artem Shelmanov, and Chris Biemann. 2025. How to compare things properly? a study of argument relevance in comparative question answering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15702–15720, Vienna, Austria. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Alexander Panchenko, Alexander Bondarenko, Mirco Franzek, Matthias Hagen, and Chris Biemann. 2019. Categorizing comparative sentences. In *Proceedings of the 6th Workshop on Argument Mining*, pages 136–145, Florence, Italy. Association for Computational Linguistics.

Maxime Peyrard. 2019. A simple theoretical model of importance for summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1059–1073, Florence, Italy. Association for Computational Linguistics.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. *Preprint*, arXiv:1910.02054.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *Preprint*, arXiv:1606.05250.

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. 2020. Superglue: Learning feature matching with graph neural networks. *Preprint*, arXiv:1911.11763.

Hassan Shakil, Ahmad Farooq, and Jugal Kalita. 2024. Abstractive text summarization: State of the art, challenges, and improvements. *Neurocomputing*, 603:128255.

Ahmad Shallouf, Hanna Herasimchyk, Mikhail Salnikov, Rudy Alexandro Garrido Veliz, Natia Mestvirishvili, Alexander Panchenko, Chris Biemann, and Irina Nikishina. 2024. CAM 2.0: End-to-end open domain comparative question answering system. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2657–2672, Torino, Italia. ELRA and ICCL.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. 2023. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, Toronto, Canada. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

A Complete Results

Table 5 shows the performance of all models, for each scenario.

Model	Scenario	Mixtral	GPT-4o	Human	Llama-3-8B	Llama-3-70B	Perplexity	GPT-3.5
GPT-3.5	1	18.76$_{\pm 0.59}$	16.95$_{\pm 1.44}$	14.93$_{\pm 2.84}$	17.07$_{\pm 2.14}$	16.63$_{\pm 1.27}$	16.84$_{\pm 2.14}$	17.90$_{\pm 1.65}$
GPT-3.5	2	16.77$_{\pm 2.80}$	14.10$_{\pm 3.15}$	11.46$_{\pm 2.93}$	11.86$_{\pm 4.00}$	12.14$_{\pm 4.99}$	11.07$_{\pm 5.79}$	13.39$_{\pm 4.07}$
GPT-3.5	**3**	18.71$_{\pm 0.57}$	18.21$_{\pm 1.02}$	16.92$_{\pm 1.90}$	16.59$_{\pm 3.09}$	18.54$_{\pm 1.02}$	18.44$_{\pm 1.12}$	17.74$_{\pm 1.56}$
GPT-3.5	4	18.02$_{\pm 1.35}$	17.37$_{\pm 1.83}$	15.00$_{\pm 1.70}$	14.81$_{\pm 3.51}$	17.69$_{\pm 2.00}$	18.12$_{\pm 1.53}$	15.17$_{\pm 3.21}$
GPT-4o	1	18.57$_{\pm 0.72}$	17.59$_{\pm 1.00}$	15.22$_{\pm 1.52}$	17.44$_{\pm 1.53}$	17.90$_{\pm 0.86}$	18.41$_{\pm 0.88}$	17.84$_{\pm 2.15}$
GPT-4o	2	17.97$_{\pm 8.75}$	17.22$_{\pm 1.33}$	14.48$_{\pm 1.62}$	14.56$_{\pm 2.78}$	17.75$_{\pm 1.55}$	18.41$_{\pm 1.50}$	16.77$_{\pm 2.76}$
GPT-4o	**3**	18.91$_{\pm 0.31}$	18.23$_{\pm 0.95}$	17.24$_{\pm 1.33}$	17.71$_{\pm 1.53}$	18.95$_{\pm 0.32}$	18.94$_{\pm 0.31}$	18.36$_{\pm 1.14}$
GPT-4o	4	18.38$_{\pm 0.78}$	18.03$_{\pm 0.95}$	16.89$_{\pm 0.94}$	16.83$_{\pm 1.59}$	18.99$_{\pm 0.08}$	18.99$_{\pm 0.08}$	17.21$_{\pm 2.47}$
Llama-3-70B	1	18.74$_{\pm 0.56}$	17.22$_{\pm 1.20}$	14.04$_{\pm 1.59}$	17.34$_{\pm 2.16}$	17.34$_{\pm 1.00}$	18.06$_{\pm 1.00}$	17.38$_{\pm 2.66}$
Llama-3-70B	2	16.44$_{\pm 2.62}$	15.90$_{\pm 1.78}$	13.29$_{\pm 2.40}$	12.70$_{\pm 4.11}$	13.86$_{\pm 4.22}$	15.74$_{\pm 4.30}$	13.69$_{\pm 4.11}$
Llama-3-70B	**3**	18.76$_{\pm 0.59}$	18.06$_{\pm 1.11}$	17.69$_{\pm 1.58}$	16.82$_{\pm 2.67}$	18.96$_{\pm 0.20}$	19.00$_{\pm 0.00}$	17.90$_{\pm 1.31}$
Llama-3-70B	4	18.80$_{\pm 0.49}$	17.30$_{\pm 1.37}$	16.46$_{\pm 1.30}$	15.96$_{\pm 2.86}$	18.90$_{\pm 0.36}$	18.92$_{\pm 0.27}$	15.94$_{\pm 2.95}$
Llama-3-8B	1	18.70$_{\pm 0.71}$	16.78$_{\pm 1.40}$	14.00$_{\pm 2.94}$	16.31$_{\pm 2.36}$	16.52$_{\pm 1.71}$	17.56$_{\pm 0.93}$	17.67$_{\pm 2.15}$
Llama-3-8B	2	16.12$_{\pm 3.15}$	14.10$_{\pm 3.36}$	11.94$_{\pm 2.65}$	11.51$_{\pm 4.90}$	11.04$_{\pm 4.40}$	13.52$_{\pm 5.41}$	11.76$_{\pm 4.83}$
Llama-3-8B	**3**	18.92$_{\pm 0.27}$	18.26$_{\pm 0.88}$	16.50$_{\pm 1.96}$	18.60$_{\pm 0.64}$	19.00$_{\pm 0.00}$	19.00$_{\pm 0.000}$	18.58$_{\pm 0.70}$
Llama-3-8B	4	18.59$_{\pm 0.81}$	17.18$_{\pm 1.57}$	16.68$_{\pm 1.10}$	15.82$_{\pm 2.98}$	18.02$_{\pm 1.49}$	18.68$_{\pm 0.89}$	15.81$_{\pm 2.73}$
Mixtral	1	18.71$_{\pm 0.62}$	16.80$_{\pm 1.20}$	14.67$_{\pm 1.73}$	17.16$_{\pm 1.81}$	16.58$_{\pm 2.11}$	17.48$_{\pm 0.79}$	17.31$_{\pm 3.05}$
Mixtral	2	17.50$_{\pm 1.98}$	16.12$_{\pm 1.76}$	11.54$_{\pm 3.66}$	13.74$_{\pm 3.81}$	15.54$_{\pm 3.48}$	16.54$_{\pm 3.78}$	15.73$_{\pm 3.32}$
Mixtral	**3**	18.86$_{\pm 0.50}$	18.40$_{\pm 0.86}$	16.29$_{\pm 2.73}$	18.02$_{\pm 1.86}$	18.86$_{\pm 0.53}$	18.84$_{\pm 0.51}$	18.06$_{\pm 1.39}$
Mixtral	4	18.38$_{\pm 0.70}$	17.48$_{\pm 1.43}$	12.50$_{\pm 2.39}$	16.22$_{\pm 2.13}$	18.82$_{\pm 0.56}$	18.80$_{\pm 0.57}$	17.12$_{\pm 1.97}$
Perplexity	1	18.78$_{\pm 0.46}$	17.32$_{\pm 1.24}$	12.86$_{\pm 1.48}$	17.28$_{\pm 1.98}$	17.68$_{\pm 0.91}$	18.08$_{\pm 0.99}$	17.22$_{\pm 3.75}$
Perplexity	2	16.84$_{\pm 2.33}$	16.46$_{\pm 2.07}$	14.14$_{\pm 2.00}$	13.88$_{\pm 2.97}$	16.00$_{\pm 2.79}$	18.04$_{\pm 2.18}$	14.52$_{\pm 3.78}$
Perplexity	**3**	19.00$_{\pm 0.00}$	18.68$_{\pm 0.47}$	17.40$_{\pm 1.84}$	18.40$_{\pm 1.03}$	19.00$_{\pm 0.00}$	19.00$_{\pm 0.00}$	18.53$_{\pm 0.96}$
Perplexity	4	18.76$_{\pm 0.69}$	17.20$_{\pm 1.56}$	15.64$_{\pm 1.75}$	15.68$_{\pm 2.27}$	18.78$_{\pm 0.91}$	18.76$_{\pm 1.29}$	15.82$_{\pm 3.53}$
Human		17.65$_{\pm 1.58}$	17.088$_{\pm 1.72}$	16.69$_{\pm 1.76}$	14.29$_{\pm 3.23}$	17.96$_{\pm 1.56}$	18.33$_{\pm 1.03}$	15.29$_{\pm 2.68}$
Yahoo Chekalina et al. (2021)	-	13.32$_{\pm 4.97}$	9.32$_{\pm 3.95}$	6.00$_{\pm 1.00}$	5.21$_{\pm 5.91}$	5.96$_{\pm 3.64}$	5.96$_{\pm 3.68}$	5.88$_{\pm 4.86}$
CAM Shallouf et al. (2024)		9.66$_{\pm 4.43}$	9.52$_{\pm 3.81}$	8.04$_{\pm 2.78}$	6.54$_{\pm 4.21}$	6.40$_{\pm 2.29}$	6.46$_{\pm 2.89}$	8.74$_{\pm 4.56}$

Table 5: CompQA Benchmark leaderboard for all participating models, against each scenario.

B Training Details

In this section, we present the details of supervised fine-tuning. It was performed on 4 NVIDIA A100 GPUs using DeepSpeed with ZeRO Stage 3 (Rajbhandari et al., 2020) and the Adam optimizer. We used a learning rate of 1e-6, warmup ratio 0.03, 3 epochs, a batch size of 32, weight decay 0.01, with a cosine annealing scheduler.

C Additional Experiments

Reasoning LM Performance Additionally, we generated the comparative answers for the following reasoning language models: o3-mini, DeepSeek-R1, Qwen3-8B (with thinking enabled), and Gemini-2.5-Flash. To provide a comprehensive comparison with traditional models, we also include Gemma-3-27B-it, Qwen3-8B (with thinking disabled), along with GPT-4o (drawn from Table 3). For the evaluation, we use the Llama-3-70B model, which is most correlated with human responses. From the results in Table 6, we can see that DeepSeek-R1 achieves the best score, while Qwen3-8B does not perform well in both reasoning and non-reasoning options. In general, we can also see that models with reasoning enabled perform better. Importantly, these findings support the main contributions of this paper: the developed criteria are extensible to newer models.

Results on the Extended Dataset To check, whether the scores for the main benchmark dataset are consistent with the extended one, we have evaluated the best performing LLMs on the 432 questions carefully filtered from the dataset of Beloucif et al. (2022). It is important to note that 173 questions do not have any arguments retrieved from the CAM system; therefore, the scores for the second and fourth scenarios are noisy. From Table 7, it can be seen that for the non-reasoning models, GPT4-o yields the best results, while for the reasoning models, both DeepSeek-R1 and o3-mini perform the best.

model	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Total score
GPT-4o (from Table 3)	$17.90_{\pm 0.86}$	$17.75_{\pm 1.55}$	$18.95_{\pm 0.32}$	$18.99_{\pm 0.08}$	$18.40_{\pm 0.70}$
Gemma3-27B-it	$16.62_{\pm 1.53}$	$14.32_{\pm 5.48}$	$16.96_{\pm 0.98}$	$18.48_{\pm 1.04}$	$16.60_{\pm 3.29}$
Qwen3-8B (disabled thinking)	$14.12_{\pm 4.96}$	$11.02_{\pm 6.74}$	$17.00_{\pm 2.16}$	$16.08_{\pm 5.80}$	$14.68_{\pm 5.74}$
o3-mini	$17.36_{\pm 1.21}$	$17.80_{\pm 1.22}$	$18.96_{\pm 0.28}$	$19.00_{\pm 0.00}$	$18.28_{\pm 1.13}$
DeepSeek-R1	$18.20_{\pm 0.92}$	$18.24_{\pm 1.09}$	$18.94_{\pm 0.31}$	$19.00_{\pm 0.00}$	$\mathbf{18.60}_{\pm 0.81}$
Gemini-2.5-Flash	$18.16_{\pm 1.05}$	$17.14_{\pm 3.19}$	$18.92_{\pm 0.34}$	$18.92_{\pm 0.39}$	$18.29_{\pm 1.85}$
Qwen3-8B (enabled thinking)	$14.98_{\pm 3.75}$	$13.52_{\pm 6.33}$	$17.58_{\pm 3.23}$	$16.08_{\pm 5.96}$	$15.54_{\pm 5.22}$

Table 6: The results on comparison between reasoning LLMs and traditional autoregressive LLMs using the Llama-3-70B model for evaluation.

model	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Total score
Llama-3-70B	$18.88_{\pm 0.90}$	$18.17_{\pm 2.53}$	$18.95_{\pm 0.32}$	$18.88_{\pm 0.58}$	$18.72_{\pm 1.42}$
GPT-4o	$18.98_{\pm 0.21}$	$18.90_{\pm 1.02}$	$18.97_{\pm 0.27}$	$18.92_{\pm 0.75}$	$18.94_{\pm 0.66}$
o3-mini	$18.92_{\pm 1.03}$	$18.81_{\pm 1.36}$	$19.00_{\pm 0.00}$	$18.91_{\pm 0.29}$	$18.91_{\pm 0.87}$
DeepSeek-R1	$18.92_{\pm 0.92}$	$18.81_{\pm 1.27}$	$18.97_{\pm 0.26}$	$18.92_{\pm 0.28}$	$18.91_{\pm 0.81}$
Llama-3-8B LoRA finetuning	-	-	-	$18.59_{\pm 2.10}$	-

Table 7: The results on the extended dataset using the Llama-3-70B model for evaluation.

Possible Biases Inherent to Certain LLMs. Another possible direction that should be worked on to improve upon this research paper is finding any biases that could exist within some LLMs.

An example of this that we encountered while generating summaries was that GPT-4, when asked to compare Google and Yahoo search engines and given arguments favoring the latter, created comparisons strongly for Google (see examples on Google vs. Yahoo in the Github repository. Moreover, the model rewrote the arguments, hallucinating that they were all originally favoring Google. We argue that such cases of bias where the final answer is not derived from the listed arguments could be because of the model inner bias towards a more popular object or subject.

Needless to say, this kind of favoritism embedded in an LLM may heavily impact its objectivity, and further tests need to be conducted to determine the cause and possible solutions to this issue. This is an exception, however, as no other object pair created such a situation; this means that, hopefully, this clear bias is something exceedingly rare.