

How to Compare Things Properly? A Study of Argument Relevance in Comparative Question Answering

Irina Nikishina¹ Saba Anwar¹
Nikolay Dolgov² Maria Manina² Daria Ignatenko²
Artem Shelmanov³ Chris Biemann¹

¹University of Hamburg ²HSE University ³MBZUAI

Correspondence: irina.nikishina@uni-hamburg.de

Abstract

Comparative Question Answering (CQA) lies at the intersection of Question Answering, Argument Mining, and Summarization. It poses unique challenges due to the inherently subjective nature of many questions and the need to integrate diverse perspectives. Although the CQA task can be addressed using recently emerged instruction-following Large Language Models (LLMs), challenges such as hallucinations in their outputs and the lack of transparent argument provenance remain significant limitations. To address these challenges, we construct a manually curated dataset comprising arguments annotated with their relevance. These arguments are further used to answer comparative questions, enabling precise traceability and faithfulness. Furthermore, we define explicit criteria for an “ideal” comparison and introduce a benchmark for evaluating the outputs of various Retrieval-Augmented Generation (RAG) models with respect to argument relevance. All code and data are publicly released to support further research¹.

1 Introduction

Comparing objects is a common task people face in everyday decision-making. Questions such as “What is better, Linux or Windows?” demand more than a straightforward answer — they require well-reasoned explanations supported by relevant arguments, e.g., “Most Linux distributions can boot faster than Window” or “Windows is better suited for general consumer use and gaming than Linux”. Answering such comparative questions typically involves processing extensive information and diverse viewpoints from multiple sources. While factual accuracy and objectivity are important, subjective opinions also play a role, especially when they resonate with a user’s values or preferences. Making an informed choice requires distilling both

¹<https://github.com/uhh-lt/cam-arg-relevance>

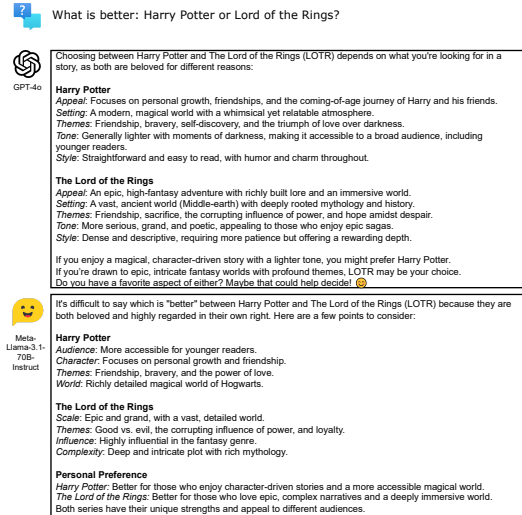


Figure 1: Example responses from GPT-4o mini and LLaMa 3.1 70B to the question “What is better: *The Simpsons* or *Family Guy*?”.

factual and opinion-based content in a way that reflects the user’s needs. Consequently, addressing comparative questions requires a combination of several tasks, including Argument Mining, Summarization, and Question Answering (QA).

Previous studies have thoroughly analyzed the initial steps of CQA: comparative sentence classification (Bondarenko et al., 2020), object and aspect identification (Beloucif et al., 2022), and argument stance classification (Panchenko et al., 2019). Few works also addressed the core question-answering process (Chekalina et al., 2021; Shalouf et al., 2024) and investigated opinion summarization (Bražinskas et al., 2021; Li et al., 2025). Nevertheless, research on CQA still lacks comprehensive exploration of key aspects, including the structure and quality of comparative summaries, as well as the accuracy, relevance, and appropriateness of the supporting arguments. Another underexplored aspect of CQA is *faithfulness* – ensuring that generated answers are accurately grounded in the

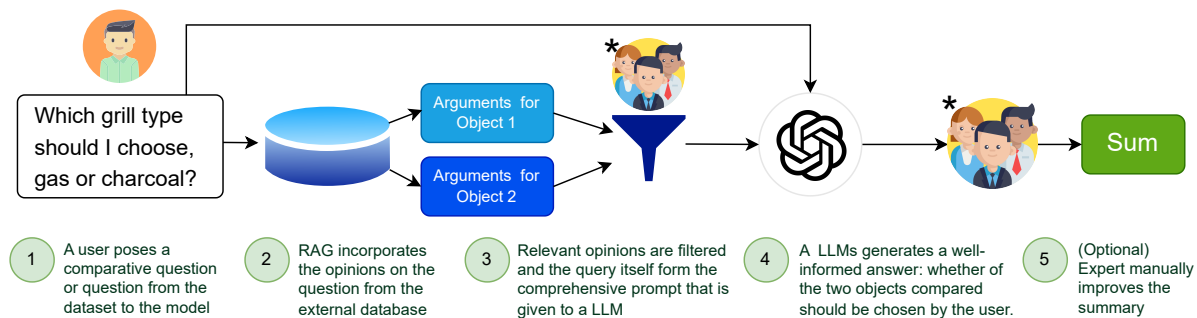


Figure 2: Dataset collection workflow for comparative question answering. An asterisk (*) indicates manual labeling steps, which can be replaced with automated methods or omitted entirely when applying this pipeline for zero-shot inference.

retrieved arguments. While some studies address the challenge of generating faithful answers in the context of hallucination mitigation in Retrieval-Augmented Generation (RAG) (Hagström et al., 2024), and others examine factual consistency in opinion summarization (Li et al., 2025), none have explicitly investigated argument-level faithfulness within a CQA framework.

Our study seeks to address the aforementioned gaps and tries to answer the following research questions: (i) *What arguments should be considered relevant when comparing two objects?* (ii) *Can we compile a good summary via summarizing arguments?* (iii) *What kinds of elements should an “ideal” comparative summary or comparison include?* (iv) *How to generate reliable answers of good quality?* This work aims to highlight the importance of evaluating the reliability and provenance of arguments in comparative answers generated by RAG-style methods. We assess the model’s ability to track the source of arguments, ensure their relevance to the question, and maintain factual accuracy. However, the evaluation of answer quality in terms of utility, coherence, or structural organization is out of the scope of the current work.

The contributions of the paper are as follows:

1. We are the first to create a dataset with comparative arguments manually annotated with relevance.
2. We propose specific criteria for evaluating the quality of comparative summaries, focusing on structural coherence, conciseness, and faithfulness through explicit source citation.
3. We build a dataset containing comparative summaries automatically generated by ChatGPT and further improved by humans.
4. We develop a benchmark to evaluate the faithfulness of arguments in comparative answers

generated by RAG-based systems.

2 Background and Terminology

According to the standard terminology of Haspelmath et al. (2017), **comparative questions** usually include objects of comparison (*Linux* and *Windows*) and (optionally) an aspect of comparison. In linguistics, comparative sentences are identified by morphological features that serve as comparative markers, e.g. adjective in the comparative/superlative form in English (Jindal and Liu, 2006). There may be more than two objects compared (“*Which type of grilled meat is tastier: chicken, beef or pork?*”), but for simplicity, we will focus on comparison between two objects. To be more specific, we consider only **subjective or opinionated comparative questions**, which cannot be answered through factual lookups in a knowledge base or a knowledge graph. For instance, the question “*Who is older: Henry Cavill or Robert Pattinson?*” can be resolved using objective data such as birth dates. In contrast, “*Who is funnier: Jim Carrey or Eddie Murphy?*” requires subjective judgment and depends on personal preferences rather than verifiable facts. Such questions are the primary focus of our study.

We assume that the answer to these questions is based on personal opinions or arguments. **Arguments** are **sentences** that support at least one object from the comparison (Wachsmuth et al., 2017). **Opinions** are “usually subjective expressions that describe individuals’ sentiments, appraisals, or feelings toward entities, events, and their properties” (Liu, 2010). **Facts** are “objective expressions about entities, events, and their properties” (Liu, 2010). In our paper, we consider both types to be useful for answer generation, as our task is subjective CQA, which means that most of our arguments will be

Evaluation	Relevance = 3	Relevance = 2	Relevance = 1	Relevance = 0
Required components	All of the objects AND argumentation/aspect present	All of the objects present BUT no argumentation OR Indirect comparison OR not clear whether the argument favors one of the objects or not	ONE of the objects presents BUT no argumentation OR Useful information about only one object is given OR A superlative form is used OR No direct comparison, but separate comparison	NO objects present OR some/all of them are present, BUT irrelevant.
	Example	Cats are smarter than dogs	Cats are better than dogs	Cats do not require a lot of time I used to have a cat

Table 1: Argument relevance criteria with illustrative examples.

subjective. For instance, when the majority prefers BMW over Mercedes and does not support their preference with arguments (e.g. “I like BMW more than Mercedes”, “BMW is better than Mercedes”), while the arguments supporting Mercedes are not as numerous but do provide a more persuasive argumentation (“The engine of Mercedes is 1.5 times higher horsepower than the one from BMW”). In this case, we want to check whether the model will take into account that most people select BMW given no explanation.

Argument relevance has been defined in various ways. For instance, Feger et al. (2020) define a relevant argument as one that is topic-related, informative, and meaningful. Bondarenko et al. (2022b) define relevance as an argument’s similarity to a query’s topic or its support and attack relationships to other arguments. They defined a scale from 0 to 2, where 2 means “highly relevant”. In this work, we extend the scale by redefining argument relevance on a range from 0 to 3, enabling a more fine-grained assessment. The description of the relevance criteria is provided in Table 1 and Subsection 3.2.

In summary, we hypothesize that the answer to the “subjective” comparative questions can be constructed by **summarizing relevant arguments** of two objects.

3 Argument Relevance Annotation

In this section, we describe the main steps for constructing the dataset of argument relevance: argument retrieval step, preliminary annotation step conducted by experts, crowdsourcing annotation, and automatic annotation with GPT-4o.

3.1 Argument Retrieval

To construct a dataset of arguments paired with related comparative questions, we use 112 object pairs from Chekalina et al. (2021) and the Touché 2021–2022 datasets provided by Bondarenko et al. (2022b). Comparative arguments are extracted using the Comparative Argumentative Machine (CAM 2.0), which retrieves relevant content from CommonCrawl (Schildwächter et al., 2019; Shalouf et al., 2024). Each comparative sentence found in CAM 2.0 contains information about the source of the text (e.g. website) and also captures the exact position of this sentence in the text. Not all pairs got many sentences for comparison; therefore, the dataset comprises only 1,729 sentences in total.

Although CAM 2.0 achieves an F-score of 0.85 for detecting comparative sentences, this performance does not directly translate to the task of assessing argument relevance, which poses a fundamentally different challenge. For instance, when comparing “Audi” and “BMW”, the sentences “BMW is smoother than Audi.”, “I swear Audi drivers are much worse than BMW ones.”, and “It was larger and slower than the equivalent BMW.” are all identified as comparative. However, they significantly differ in their relevance for CQA. This emphasizes the need to annotate the retrieved arguments with their relevance to the input question.

3.2 Relevance Criteria Design

In designing the argument relevance criteria, we adopt a 4-point scale ranging from 0 to 3, extending the 3-level scheme proposed by Bondarenko et al. (2022b). The highest score could be assigned to the argument that contains two (or more) compared objects and the argumentation (a justification of why one object is better than another). For example, “Cats are smarter than dogs” has relevance = 3,

because it contains mentions of compared objects: “cats” and “dogs”, and “smarter” can be seen as justification. Relevance = 2 corresponds to situations if there was no explanation or if one or more objects were absent, or if there was no comparison at all. Relevance = 2 is assigned to sentences like “Cats are better than dogs” due to a lack of argumentation. Sentences like “Cats do not require a lot of time” give information about only one compared object, so they are assigned relevance = 1. An example of an irrelevant sentence (score 0) is: “I have a cat and a dog.”

We also account for edge cases, including indirect comparisons, synonym usage, vague or ambiguous statements, and superlatives. For instance, the term “jet” may be accepted as a synonym for “plane”, and indirect comparisons involving third-party references are handled with particular care. These considerations are crucial for maintaining annotation consistency. The complete guidelines are provided in Appendix C.

3.3 Preliminary Annotation by Experts

To better understand how the task could be approached and to establish a high-quality reference annotation, we manually assessed a subset of 480 arguments drawn from the full set of 1,729 sentences. Five expert annotators, each holding at least an MSc degree in linguistics or computational linguistics, were asked to assign relevance scores to batches of 100 sentences. To measure inter-annotator agreement, 30 sentences in each batch were intentionally overlapped with those of another annotator (pairwise overlap = 2). The resulting agreement metrics for this subset indicated high consistency: Krippendorff’s $\alpha = 0.89$ and Fleiss’ $\kappa = 0.81$.

Expert annotation enabled us to identify major scoring discrepancies, refine the annotation guidelines accordingly, and adopt the revised version for large-scale annotation.

3.3.1 Crowdsourcing with Toloka

The whole dataset of 1,729 sentences was annotated using a crowdsourcing platform Toloka.ai (Pavlichenko et al., 2021).

As a quality control mechanism, we implemented a sequence in which every four “general” sentences were followed by one “control” sentence — previously annotated by experts — to monitor annotator reliability. Additionally, we enforced strict participation criteria to minimize the influence of

bots or dishonest users.

A total of 282 Toloka users (who had passed the platform’s language proficiency test) contributed to the annotation pool. Each sentence was annotated by a minimum of two and up to sixty participants, due to Toloka.ai’s design, which permits users to exit the task at any point before completion. In cases of disagreement, we applied majority voting, favoring the higher score in the event of a tie. Furthermore, 265 sentences exhibiting substantial annotator disagreement were re-annotated by three expert annotators to ensure quality and consistency.

3.4 Annotation with GPT-4

Given the high cost and limited scalability of expert annotation, we explore the use of automatic methods to evaluate argument relevance at scale. In this work, we assess the feasibility of using GPT-4 (OpenAI, 2023) as an automatic annotator to supplement human-labeled data.

We used GPT-4 with the prompt shown in Figure 8 in Appendix D. Agreement between GPT-4 and expert annotators was evaluated against the majority vote among human annotations as the reference label. The resulting agreement was notably high (Krippendorff’s $\alpha = 0.68$), suggesting that GPT-4 can approximate expert-level relevance judgments to a reasonable extent. However, performance declined considerably on the larger and noisier Toloka dataset (Krippendorff’s $\alpha = 0.43$).

4 Comparative Question Answering Dataset

In this section, we discuss the pipeline for using the annotated arguments for answering comparative questions. We discuss the input format and define the criteria for summary generation in general.

4.1 Developing Criteria for Comparative Answers

Since there is no definitive formula for an “ideal” comparative answer, we propose a checklist to guide the creation of high-quality responses. We acknowledge that this checklist may not be exhaustive and that some criteria may be subject to debate. However, to the best of our knowledge, no prior work has defined standard criteria for comparative answers. Figure 3 matches the answer checklist with the relevant parts of the prompt.

The prompt for generating CQA answers is based on the “Role — Task — Format” framework

Object pair	Argument	Relevance (manual)	Relevance (GPT-4)
Ceramic knives vs Steel knives	Ceramic knives are substantially lighter than steel knives, will not corrode in harsh environments, are non-magnetic, and do not conduct electricity at room temperature.	3	1
<i>Here, both objects enlisted in the comparison are mentioned in the sentence, and a clear argument in favor of Ceramic knives is provided. According to the guidelines, this sentence should be given a Relevance = 3 mark.</i>			
PHP vs ASP	Imagine PHP appearing less stable and slower than ASP.	0	2
<i>Here, the context does not allow us to determine exactly whether the sentence can be considered an opinion, or maybe it is a hypothetical reasoning. Both objects compared are enlisted in the sentence, but it hardly can be considered relevant, since it does not explicitly say that PHP is more or less stable than ASP.</i>			

Table 2: Examples of strong disagreement between human annotators and GPT-4.

Ideal summary checklist	Part of the Prompt
length	- write a 300-word comparison
informative / concise	- be concise
good-quality / relevant arguments	- analyse the list of arguments below - pick relevant ones - rephrase in your own words
citations / provenance	- cite used argument numbers in square brackets right after the usage
novel / additional info	- the summary needs to have ~15 arguments, create some if needed (add a [generated] tag)
logical structure (introduction + main body + choice made)	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: []

Figure 3: Checklist for generating high-quality CQA answers with the relevant parts of the prompt.

(Lin, 2023). The complete template is shown in Figure 7 in Appendix B. We also experiment with prompt variations to assess robustness.

The prompt is structured in four parts:

- In the first paragraph, we introduce the role of the LLM as an analyst and outline the task of comparing two objects to answer a given question. We also define the first evaluation criterion – answer length – which is required to be approximately 300 words.
- The second part instructs to “compare and choose the better of the two,” which is the core purpose of comparative question answering.
- The third part comprises a list of requirements

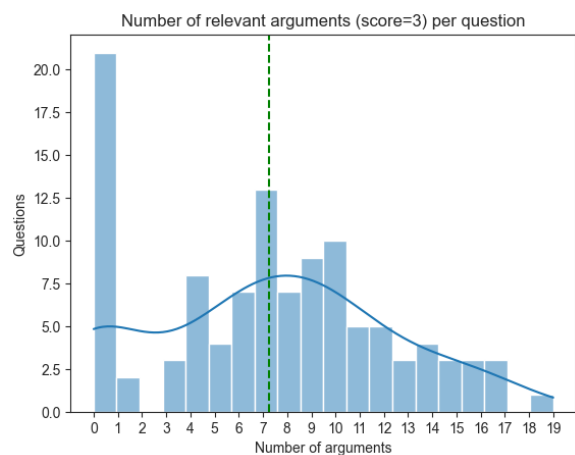


Figure 4: Distribution of relevant arguments per question. The vertical line denotes the mean value.

regarding the content and the format, following the “Ideal summary checklist”:

- In addition to the specified summary length, we also ask the model to be concise, as it tends to be wordy providing too many irrelevant details, exceeding the specified length.
- The next three requirements aim at filtering the arguments and keeping only relevant ones. As we have seen in Section 3.4, ChatGPT can solve this task quite well, therefore, we expect it to analyze and select relevant arguments on-the-fly. We require the model to filter and merge relevant arguments, avoiding copy-paste of full sentences.
- To support argument traceability, the model is asked to cite argument IDs (provided in the input) in square brackets.
- As most of the retrieved arguments could

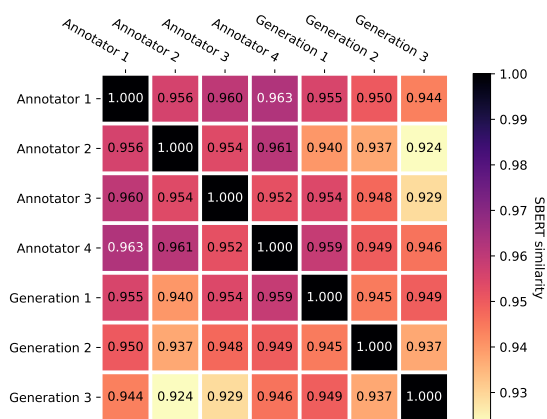


Figure 5: Heatmap of mean SBERT similarities between the expert annotators’ answers and generated ones.

be irrelevant, we also allow the model to generate its own arguments with the tag `[generated]`. We set the total number of required arguments to 15 to encourage the model to generate additional content. Empirical analysis shows that, on average, only 9–10 of the retrieved arguments are relevant, with a minimum of 3 and a maximum of 17 (see the distribution of relevant arguments in Figure 4). Thus, in the majority of cases, the model should generate more arguments beyond the retrieved set to meet the target.

- Based on empirical testing, we define a clear structure for the answer: (i) an introductory summary, (ii) a main body with argument-based comparison, (iii) the final choice, and (iv) a list of used arguments. This structure is explicitly outlined in the prompt under “Needed structure.”

Finally, to ensure factual grounding, each input includes a randomized list of retrieved arguments for both objects, sourced from the corpus of relevant arguments. The model is expected to mark referenced arguments with their corresponding index, enabling precise traceability and faithfulness.

4.2 Data Generation Procedure

To create the comparative answers based on the retrieved arguments using the developed prompt, we used the ChatGPT model (gpt-3.5-turbo, default parameters) from OpenAI. We returned top 3 answers to evaluate possible variations of the answer and checked how well the model follows the required

structure. We refer to this automatically created dataset as *Silver dataset*. It contains 112 comparative questions with three candidate answers in each, resulting in 336 summaries in total.

To construct the *Gold dataset* of better quality, we engaged four expert annotators to refine and improve 50 selected outputs from the *Silver dataset*, following the predefined “Summary Criteria” (Figure 3). As a part of the calibration process, all experts first annotated the same set of 10 summaries and discussed each other’s revisions. Subsequently, each expert independently edited 10 additional summaries. The resulting *Gold dataset* consists of 50 comparative questions and 80 manually curated summaries (10 shared and 10 individual questions and answers for four experts).

We further analyzed the differences between the initial answers generated by ChatGPT and their revised versions, as well as the variation across annotators. Figure 5 presents SBERT-based similarity scores between responses from different annotators and the original model outputs. The results indicate that the improved versions remain largely similar to the initial responses, with only minor variations.

5 Evaluation

In this section, we introduce the evaluation metrics used for the internal assessment of comparative answers, present the results, and analyze the influence of key factors.

5.1 Evaluation Metrics

To assess the relevance and faithfulness of the collected summaries, we employ the following evaluation metrics:

- **Jaccard similarity & Levenstein distance:** to assess the exact or near-exact matching between the arguments in the generated text and those provided in the input list;
- **Semantic similarity:** we use the SBERT model (Reimers and Gurevych, 2019) to (i) evaluate the similarity/alignment between the arguments and the referenced excerpts to see whether it correlates with the relevance; (ii) evaluate the overall similarity/alignment of the generated answer with the provided arguments;
- **Argument Relevance (AR):** $\text{Precision}_{(AR)}$, $\text{Recall}_{(AR)}$, $\text{F1}_{(AR)}$ to see how many relevant

Dataset Name	Size	Questions	Jaccard _{text}	Jaccard _{sent}	Levenstein	S-BERT	Pr _{AR}	R _{AR}	F1 _{AR}	AAC
Gold (human)	80	50	0.14	0.18	20.54	0.68	0.96	0.53	0.64	0.72
Silver (subset) (ChatGPT)	150	50	0.24	0.21	11.88	0.69	0.95	0.77	0.82	0.78
Silver (ChatGPT)	336	112	0.25	0.21	11.82	0.68	0.96	0.75	0.81	0.77

Table 3: Evaluation results for the created CQA dataset. The produced summaries (Silver — automatic, Gold — manually improved Silver data) are compared against the arguments for each comparison, provided in the input.

arguments are correctly mentioned or, on the contrary, missed:

$$\text{Precision}_{(AR)} = \frac{|\text{Relevant}_{\text{used}}|}{|\text{Used}|}$$

$$\text{Recall}_{(AR)} = \frac{|\text{Relevant}_{\text{used}}|}{|\text{Relevant}|},$$

$$F1_{(AR)} = \frac{2 \cdot \text{Precision}_{(AR)} \cdot \text{Recall}_{(AR)}}{\text{Precision}_{(AR)} + \text{Recall}_{(AR)}},$$

where $|\text{Relevant}_{\text{used}}|$ is the number of relevant arguments included in the summary, $|\text{Used}|$ is the total number of arguments used, and $|\text{Relevant}|$ is the total number of relevant arguments in the input;

- **Argument-Answer Coherence (AAC):** we use the GPT-4 model to check whether the citation supports the correct object, e.g. Argument 3 “*Cats are smarter than dogs*” cannot be cited in the excerpt “*Dogs are smarter than cats [3]*”, as it states exactly the opposite. The prompt for GPT-4 can be found in Appendix 9.

5.2 Results

The evaluation includes the results for a set of manually refined (gold) answers corresponding to a subset of questions, along with a larger set of automatically generated (silver) answers. For a fair comparison, we also evaluate a matching subset of silver answers aligned with the gold data to ensure equal coverage. Table 3 demonstrates the results for all three cases. It is important to note that our evaluation does not focus on overall answer utility, coherence, or structural quality. Instead, the primary objective of this work is to assess the model’s ability to maintain argument provenance and faithfulness.

From the results, we can see that Jaccard_{text} and Jaccard_{sent} metrics are lower for the gold dataset, which means that rewritten texts have fewer words in common with the arguments. This observation is

supported by higher Levenshtein distances. However, the S-BERT cosine similarity score almost did not change (0.69 → 0.68). Therefore, we can conclude that gold answers are rephrased but they still keep the same meaning.

Interestingly, the Argument Relevance (AR) and Argument Attribution Coherence (AAC) metrics yield counterintuitive results: the silver (automatic) dataset achieves slightly higher F1_{AR} and AAC scores than the manually curated gold set. One might expect expert-curated summaries to perform better on these criteria, especially in terms of coherence and accurate attribution.

It could be partially explained that in the silver dataset, arguments are cited a lot, while in the gold, the statements are better analyzed and, therefore, are shorter with fewer arguments cited. An example for that can be seen in Figure 11 in Appendix F, showing a comparative answer before and after manual improvement. We argue that manually written texts tend to be more diverse and less repetitive.

5.3 Additional Analysis

When experimenting with different prompts, we also investigate the importance of argument relevance, argument balance supporting different objects, the importance of the input question, and the input question diversity. We discuss each peculiarity in the following paragraphs.

Filtering input arguments by relevance. Since the arguments collected automatically through CAM 2.0 are already ranked according to the annotation process described in Section 3, we decided to try filtering out arguments lower than a certain score.

First, we filter any arguments lower than 3 points of relevance. This helped greatly with the quality of the resulting summary: since there are few arguments with such high relevance (on average, around 6-7), GPT-4 was able to generate about 3-4 more synthetic arguments than usual, which made the comparative text more informative and less re-

dundant. This also helped in avoiding argument repetition. The only case when arguments are repeated is when they are used in differing contexts: e.g. “*Adidas kits may look nice, but are made of even worse material than Nike*” is used as proof of both *apparel design* (“*Nike’s apparel has been reported as using better materials*”) and *product quality* (“*Nike’s apparel is more durable and long-lasting*”).

Next, we filter any arguments lower than relevance 2, keeping those with relevance 2 and 3. As this increased the total number of arguments (from an average of 8 to 14), GPT-4 was also less likely to generate its own arguments, which resulted in less variety and more repetition. Although the total number was higher (since most arguments were still of high relevance), the repetitions were often made about different aspects of the argument (70% of the repetitions were “good repetitions”).

Lastly, we try only arguments with a relevance of 0 and 1. We expect that GPT-4 rejects all of the arguments and uses only ones generated by itself. However, despite the low-quality arguments, the model generated few arguments and still managed to single out similar “best options” compared to when it was given only the best arguments. For example, in Figure 10 in Appendix F, we can see that the model generated x arguments with only 3 arguments provided and applies all of them.

The main issue with using more arguments came in the form of hallucinations. For example, in comparing “*Adidas*” vs. “*Nike*”, an argument clearly stated that Adidas was worse in terms of “*high-profile sports gear*”, but the summary interpreted this as Adidas being better.

Filtering input arguments by object. Since arguments are scraped from various Internet forums and websites, there may be unfair biases towards one object or the other depending on the source. To check if this was an issue, we experiment with purposefully introducing a bias.

We try two scenarios: giving GPT-4 arguments that only support the first object, while deleting the others, and the same with the second object. The results showed that this bias worked well in swaying GPT-4’s opinion, changing the resulting preference completely.

There was, however, an unusual case: in comparing “*IELTS* vs. *TOEFL*”, biasing the model towards “*IELTS*” did not work: it chose “*TOEFL*” both times during several runs. We have also found

that more careful prompt engineering leads to better bias correction. The line “*Avoid artificial balancing of arguments*” turned out to be the most successful at preventing excessive balance and imbalance.

Handling contradictory arguments. In this work, we do not explicitly filter or restrict input arguments based on internal contradictions. As a result, some inputs may contain conflicting claims. For example, in a comparison between IELTS and TOEFL, the input contains the following two statements:

“Because IELTS is easier than TOEFL.”
“For us, IELTS is harder than TOEFL.”

Both arguments were provided to the model, which produced the following output:

“In comparing IELTS and TOEFL, several factors must be considered. While some argue that IELTS is easier, others find TOEFL more challenging.”

This example suggests that the model can recognize and integrate contradictory information without producing incoherent output. Its ability to accommodate divergent viewpoints without explicit instruction is encouraging and highlights a promising direction for future work on nuanced comparative summarization.

Importance of the input question in the prompt.

Given that the CQA task is inherently centered around answering a specific question, we also analyze summaries both with and without the inclusion of these questions. For instance, we compare summaries for a generic question “*What is better, pasta or pizza?*” with summaries generated for a specific question, like “*I am buying iPhone 15, does it make better photos than Samsung Galaxy S24 Ultra?*”. Specific questions and some generic questions are taken from the initial datasets of [Chekalina et al. \(2021\)](#); [Bondarenko et al. \(2022a\)](#). For the pairs with missing questions, we used the following template: “*What is better, X or Y?*” and manually checked for aspects and objects.

Overall, all comparative answers had good structure, no repetition, no hallucinations, or lexical/grammatical mistakes. There was a big difference between the two groups: among the summaries that were generated with a specific question,

around 13% more summaries using a specific question did not come to a clear conclusion, and instead left the decision to the reader. This kind of answer is only acceptable when the arguments support each object equally, but here, some objects were better supported than others. Based on the possibility that the summary does not come to a clear answer, we suggest avoiding the use of the input question in the prompt and opting for a more generic approach.

6 Related Work

This section reviews previous approaches to CQA, argument mining, and multi-document summarisation, as our task lies at their intersection.

6.1 Comparative Question Answering

CQA is usually subdivided into four subproblems: comparative question identification (Bondarenko et al., 2020, 2022a), object and aspect identification (Chekalina et al., 2021; Beloucif et al., 2022), stance classification (Panchenko et al., 2019), and question answering (Chekalina et al., 2021; Shallouf et al., 2024).

We use the Comparative Argumentative Machine (CAM) (Schildwächter et al., 2019), which involves retrieving relevant sentences from the CommonCrawl corpus, sentence classification, sentence ranking and object comparison, aspect extraction, and user interface design. The upgraded version, CAM 2.0 (Shallouf et al., 2024) further adds the comparative question identification and object and aspect identification steps.

Most studies do not answer the comparative questions but only provide a list of arguments (Schildwächter et al., 2019; Bondarenko et al., 2022a). Chekalina et al. (2021) and Shallouf et al. (2024) try to tackle this problem, but they do not provide any openly accessible datasets and use pre-trained unsupervised models (BART, LLaMA-2).

6.2 Argument Mining

Argument mining aims to automatically detect and structure argumentative components i.e. claims, premises, and conclusions within a text, and identify their relationships. Several surveys cover argument analysis (Lawrence and Reed, 2019; Vecchi et al., 2021), opinion mining (Shaik et al., 2023), controversy detection (Hardalov et al., 2022), and other tasks (Heinisch et al., 2022). Additionally, some argument-related tasks are discussed in Touché proceedings (Bondarenko et al., 2021,

2023). Notably, Bondarenko et al. (2022c) describe the argument retrieval for the comparative questions shared task. We use questions from the Touché datasets 2021-2022 to compile our data.

6.3 Multi-Document Summarization

Rapid advancements in natural language processing and the emergence of LLMs such as BART (Obonyo et al., 2022), GPT-2 (Obonyo et al., 2022), and LLaMA (Zhu et al., 2024) have significantly improved multi-document summarization. For instance, Maynez et al. (2020) conducted a large-scale study of abstractive summarization via neural text generation models. (Wang et al., 2023) developed a pipeline using LLMs for generating concise, “element-aware” summaries without factual hallucinations or information redundancy. More relevant to our task is opinion summarization, which aims to distill diverse and sometimes contradictory viewpoints into coherent summaries. For instance, Iso et al. (2022) collect and summarize diverse opinions by synthesizing them into a coherent response using a collaborative decoding strategy that generates distinctive, entity- and pair-specific summaries from two base models. Bražinskis et al. (2021) proposed learning opinion summarizers by selecting informative and representative reviews, thereby improving summary informativeness and coverage. Recently, Li et al. (2025) presented an aspect-aware decomposition framework to better capture opinion diversity and structure in generated summaries.

7 Conclusion

This paper addresses the problem of comparative question answering from the argument relevance perspective. We defined the argument’s relevance and collected a manually annotated corpus via crowdsourcing and expert annotations. Then we outlined the criteria for an “ideal” comparative answer and used it to create the dataset using ChatGPT. These answers can be viewed as comparison summaries, as they are based on the arguments retrieved from the CAM system. A part of this dataset was subsequently refined by expert annotators. The final dataset includes both machine- and human-written comparisons. Additionally, we propose a benchmark for evaluating answers to comparative questions based on argument relevance. In future work, we plan to fine-tune a model on this dataset and evaluate various models for answering comparative questions.

Limitations

Main limitations of the research are as follows:

- **Universality of comparison:** theoretically, one can compare any two objects, despite the fact that they do not belong to the same category (consider Aristotle’s works treating comparison as a metaphor). Nevertheless, a comparison of two programming languages or two films is seen as more natural than a comparison of a knife and a stick. Hypothetically, if the model presented will be given two very different objects of comparison, CAM will hardly be able to find many supporting arguments. Thus, the quality of the summary will be highly dependent on the arguments generated by the model itself. There will be a lot more room for hallucination, therefore, we cannot guarantee that a high-quality comparison will be given.
- **Models tested:** we acknowledge that we test our pipeline only with ChatGPT. More data (and probably of better quality) could be generated by applying the prompt to other models. However, we aimed at generating dataset of a good quality: we manually rewrite half of the dataset and demonstrate high similarity of scores between automatic and rewritten answers. We leave this experiment to future work.

Ethical Considerations

In our work, we utilize large-scale neural models such as ChatGPT. We believe that the model developers filter out toxic or biased content; however, we cannot guarantee that in some cases answers will not reflect some biases.

References

- 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, Ferdinand Schlatt, Valentin Barrière, Brian Ravenet, Léo Hemamou, Simon Luck, Jan Heinrich Reimer, Benno Stein, Martin Potthast, and Matthias Hagen. 2023. *Overview of touché 2023: Argument and causal retrieval*. In *CLEF*, volume 14163 of *Lecture Notes in Computer Science*, pages 507–530. Springer.
- 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 *CLEF (Working Notes)*, volume 3180 of *CEUR Workshop Proceedings*, pages 2867–2903. CEUR-WS.org.
- 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. 2022c. *Overview of touché 2022: Argument retrieval*. In *CLEF*, volume 13390 of *Lecture Notes in Computer Science*, pages 311–336. Springer.
- Alexander Bondarenko, Lukas Gienapp, Maik Fröbe, Meriem Beloucif, Yamen Ajjour, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, and Matthias Hagen. 2021. *Overview of touché 2021: Argument retrieval*. In *CLEF*, volume 12880 of *Lecture Notes in Computer Science*, pages 450–467. Springer.
- Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2021. *Learning opinion summarizers by selecting informative reviews*. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9424–9442, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- 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.
- Marc Feger, Jan Steimann, and Christian Meter. 2020. *Structure or content? towards assessing argument relevance*. In *COMMA*, volume 326 of *Frontiers in Artificial Intelligence and Applications*, pages 203–214. IOS Press.

- Lovisa Hagström, Sara Vera Marjanović, Haeun Yu, Arnav Arora, Christina Lioma, Maria Maistro, Pepa Atanasova, and Isabelle Augenstein. 2024. [A reality check on context utilisation for retrieval-augmented generation](#). *Preprint*, arXiv:2412.17031.
- Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2022. [A survey on stance detection for mis- and disinformation identification](#). In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1259–1277, Seattle, United States. Association for Computational Linguistics.
- Martin Haspelmath et al. 2017. Equative constructions in world-wide perspective. *Similative and equative constructions: A cross-linguistic perspective*, pages 9–32.
- Philipp Heinisch, Anette Frank, Juri Opitz, Moritz Plenz, and Philipp Cimiano. 2022. [Overview of the 2022 validity and novelty prediction shared task](#). In *Proceedings of the 9th Workshop on Argument Mining*, pages 84–94, Online and in Gyeongju, Republic of Korea. International Conference on Computational Linguistics.
- 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.
- Nitin Jindal and Bing Liu. 2006. Mining comparative sentences and relations. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2*, AAAI’06, page 1331–1336. AAAI Press.
- John Lawrence and Chris Reed. 2019. [Argument mining: A survey](#). *Computational Linguistics*, 45(4):765–818.
- Miao Li, Jey Han Lau, Eduard H. Hovy, and Mirella Lapata. 2025. [Aspect-aware decomposition for opinion summarization](#). *CoRR*, abs/2501.17191.
- Lewis Lin. 2023. The Ultimate ChatGPT Prompting Guide: role > task > format. <https://www.lewis-lin.com/blog/the-ultimate-chatgpt-prompting-guide-role-gt-task-gt-format>. Accessed: 2024-09-10.
- Bing Liu. 2010. [Sentiment analysis and subjectivity](#). In *Handbook of Natural Language Processing*, volume 2, page 627–666.
- Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. [On faithfulness and factuality in abstractive summarization](#). *Preprint*, arXiv:2005.00661.
- Ishmael Obonyo, Silvia Casola, and Horacio Saggion. 2022. [Exploring the limits of a base BART for multi-document summarization in the medical domain](#). In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 193–198, Gyeongju, Republic of Korea. 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.
- Nikita Pavlichenko, Ivan Stelmakh, and Dmitry Ustalov. 2021. [Crowdspeech and vox diy: Benchmark dataset for crowdsourced audio transcription](#). In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Matthias Schildwächter, Alexander Bondarenko, Julian Zenker, Matthias Hagen, Chris Biemann, and Alexander Panchenko. 2019. Answering comparative questions: Better than ten-blue-links? In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, pages 361–365.
- Thanveer Shaik, Xiaohui Tao, Christopher Dann, Hao-ran Xie, Yan Li, and Linda Galligan. 2023. Sentiment analysis and opinion mining on educational data: A survey. *Nat. Lang. Process. J.*, 2:100003.
- Ahmad Shallouf, Hanna Herasimchyk, Mikhail Salkanikov, 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.
- Eva Maria Vecchi, Neele Falk, Iman Jundi, and Gabriella Lapesa. 2021. Towards argument mining for social good: A survey. In *ACL/IJCNLP (1)*, pages 1338–1352. Association for Computational Linguistics.
- Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. [Computational argumentation quality assessment in natural language](#). In *Proceedings of the 15th Conference of the European Chapter of the Association*

for Computational Linguistics: Volume 1, Long Papers, pages 176–187, Valencia, Spain. Association for Computational Linguistics.

Yiming Wang, Zhuosheng Zhang, and Rui Wang. 2023. [Element-aware summarization with large language models: Expert-aligned evaluation and chain-of-thought method](#). *Preprint*, arXiv:2305.13412.

Andrew Zhu, Alyssa Hwang, Liam Dugan, and Chris Callison-Burch. 2024. [FanOutQA: A multi-hop, multi-document question answering benchmark for large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 18–37, Bangkok, Thailand. Association for Computational Linguistics.

A RAG Sequence Example on the CQA Task

1
A user poses a comparative question to the model

2
RAG incorporates the opinions on the question from the external database

3
Relevant opinions and the query itself form the comprehensive prompt that is given to a LLM

4
A LLMs generates a well-informed answer: whether of the two objects compared should be chosen by the user.

5
Expert manually improves the summary



Which grill type should I choose, gas or charcoal?



1. A gas flame is faster than the charcoal briquettes.
2. Grilling purists may look down their noses at gas grills, but no one can argue with the fact that grilling over gas is quicker and easier than using charcoal.
3. Gas and electric grills in this guide are easier to clean and faster to light than charcoal grills, but they do tend to cost more.
4. Gas grills heat up quicker and sustain their heat in cold weather better than a charcoal grill.
5. Gas is much easier and quicker than charcoal charcoal has a better flavor to me gas grill is like cooking on the stove Gas is grilling, charcoal is bar-b-qing.



6. Charcoal enthusiasts swear by the superior flavor of food coming off the charcoal grill (some say meat tastes better on charcoal, whereas non-meat items do just fine on gas), and many people do agree.
7. I use a charcoal grill as I find the taste superior to my gas grill.
8. I would much rather use charcoal for my searing; the high, dry heat of charcoal will brown the meat much better than gas.
9. Now, I prefer charcoal for better flavor and really think it is easier than gas...but more time consuming.
10. Wood is good - charcoal tastes better than gas, and wood tastes better than charcoal.

You are an analyst, write a 300-word comparison of gas and charcoal.
Task: compare and choose the better of the two.

Requirements:

- be concise
- analyze 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: [<...>]



Summary:
Gas and charcoal grills each have their own set of advantages and disadvantages. While gas is faster, cleaner, and more convenient, charcoal is favored for its superior flavor. Ultimately, the choice between the two comes down to personal preference.

Main Aspects of Comparison:

- Energy Efficiency: Gas is generally more energy-efficient than charcoal. Gas burners provide consistent heat, allowing for more precise cooking temperatures and reducing energy waste.
- Speed: A gas flame is faster than the charcoal briquettes [1].
- Convenience: Gas is quicker and easier to use and to clean than charcoal [3, 5, 6].
- Cost: While gas grills are more expensive upfront, the ongoing cost of charcoal can add up over time.
- Flavor: Charcoal is favored for superior flavor, especially for meat [2, 7, 10].
- Heat Retention: Gas grills heat up quicker and maintain heat better in cold weather than charcoal grill [4].
- Environmental Impact: Gas is generally considered a cleaner-burning fuel compared to charcoal. Moreover, charcoal production can contribute to deforestation and air pollution. [generated]

Best Option:
Gas

Summary:
The choice between gas and charcoal grills comes down to personal preference. While gas is faster, cleaner, and more convenient, charcoal is favored for the meat superior flavor.

Main Aspects of Comparison:

- Speed: A gas flame is faster than the charcoal briquettes [1, 4].
- Convenience: Gas is quicker and easier to use and to clean than charcoal [3, 5, 6].
- Cost: While gas grills are more expensive upfront [3], the ongoing cost of charcoal can add up over time [generated].
- Flavor: Charcoal is favored for superior flavor, especially for meat [2, 7, 10].
- Energy Efficiency: Gas is generally more energy-efficient than charcoal. Gas burners provide consistent heat (especially in cold weather [4]), allowing for more precise cooking temperatures and reducing energy waste [generated].
- Environmental Impact: Gas is generally considered a cleaner-burning fuel compared to charcoal. Moreover, charcoal production can contribute to deforestation and air pollution. [generated]

Best Option:
Gas



Figure 6: Pipeline for the CQA task on the example of the question “Which grill type should I choose, gas or charcoal?”.

B LLM Prompt for Comparative Summary Generation

You are an analyst, write a 300-word comparison of {object1} and {object2}.

Task: compare and choose the better of the two.
Focus on {aspect}.

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:
{arguments}

Figure 7: Final LLM prompt for comparative summary generation using a list of arguments.

C Full Toloka Instruction

You will evaluate arguments that help people compare two or more objects for research and publication purposes. Your data will be used for testing Large Language Models. Objects are taken from all walks of life, everyday and professional. The evaluated arguments will be used as part of the project, where we are trying to improve the quality of state-of-the-art comparative question answering systems.

You may run the task on your personal computer or mobile device. For each argument, you have about 60 seconds to evaluate it. Please follow the instructions below:

How to do the task

1. Note the objects compared which are written at the beginning of the task. Then, read the argument itself carefully; multiple times, if necessary. Find the objects written above in the sentence; determine, whether they are compared with each other, or to some other objects, or both. The latter case can be called "multiple comparison". Also, sometimes there can be no comparison at all.

See some examples below:

the sentence "Milk is a better choice than soda - especially for children." => 2 objects: milk and soda
the sentence "Milk is better than soda or juice." => 3 objects: milk, soda and juice
the sentence "Prepackaged chocolate milk can give you from about 140 to more than 250 calories, depending on whether it's low-fat or whole milk." => 1 object: (chocolate) milk.

2. Use the following table to rate the sentence from 0 to 3:

- 2.1. Relevance = 3:
Both objects enlisted in the task are present in the sentence AND the argumentation is given
e.g. Cats vs Dogs
 Rate the sentence from 0 to 3:
 Cats are smarter than dogs. => Relevance = 3
Explanation: both objects (Cats AND Dogs) are present AND the argumentation in favor of cats is given. In other words, the author thinks cats are better than dogs in terms of intelligence. Hence, this sentence gets a Relevance = 3.
- 2.2. Relevance = 2
Both objects enlisted in the task are present in the sentence BUT no argumentation is given.

e.g. Cats vs Dogs
 Rate the sentence from 0 to 3:
 Cats are better than dogs. => Relevance = 2
 Explanation: both objects (Cats AND Dogs) are present BUT no argumentation in favor of cats is given. In other words, the author gives no explanation why cats are better. It seems to be a subjective opinion. Sometimes the subjectiveness can be expressed in the sentence with help of the words “my friends / my mother say(s) / said”, “I think / believe that. . .” etc.:

e.g. Cats vs Dogs
 Rate the sentence from 0 to 3:
 My friend prefers cats than dogs => Relevance = 2

2.3 Relevance = 1
 Only one object enlisted in the task is present AND some useful information, which can hypothetically be used in comparison with some other object, is given.

e.g. Cats vs Dogs
 Rate the sentence from 0 to 3:
 Dogs are loyal because they do everything to show that they care for the comfort of their owners => Relevance = 1
 Explanation: only one object (Dogs) is present, there is nothing said about cats. However, the author characterizes the dogs as loyal domestic animals. Although there is no information about whether dogs are better / worse than other animals, this is a useful piece of knowledge.

2.3 Relevance = 0
 No objects enlisted in the task are mentioned in the sentence OR some/all of them are mentioned, but not relevant at all.

e.g. Cats vs Dogs
 Rate the sentence from 0 to 3:
 I used to have a cat and a dog. => Relevance = 0
 Explanation: although both Cats and Dogs are mentioned, the sentence is useless. It does not express any opinion, but provides information about the author, which is unnecessary for the comparison. From this sentence, we cannot infer whether cats are better / worse than dogs.

e.g. Cats vs Dogs
 Rate the sentence from 0 to 3:
 Michael loves pumpkin spice latte. => Relevance = 0
 Explanation: no objects enlisted in the task (Cats OR Dogs) are mentioned.

2.4. Complex cases

2.4.1. Different objects
 Sometimes, the objects enlisted in the task might not fully coincide with the objects given in the sentence:

e.g. Train vs Plane
 Rate the sentence from 0 to 3:
 Plus the jet is far lighter than the train. => Relevance = 3
 In this case, plane and jet are synonyms. Close synonyms (plain - jet, car - vehicle, computer - machine) must be treated as same objects. Thus, we can imply that both objects were mentioned in the sentence, and the argumentation in favor of the second object (planes) is given. That is why the sentence got a “Relevance = 3” mark.
 Let’s look at another example:

e.g. Microsoft vs Apple
 Rate the sentence from 0 to 3:
 Apple has better programmers, so it's able to write tighter, faster, stronger code than Microsoft. => Relevance = 3
 There is an ambiguity here: apart from comparing two tech companies in general, their products like smartphones, software, laptops, etc can also be compared. All arguments that discuss them as companies or products should be considered relevant. It does not matter if one argument is comparing smartphones and the other is about their employee benefits, consider them as aspects of Microsoft or Apple.

2.4.2. Indirect comparison
 e.g. Windows 8 vs Windows XP
 Rate the sentence from 0 to 3:
 Though Windows 8 is significantly at greater risk compared to Windows 8.1, according to Redmond's report, it's still significantly safer than Windows 7, Windows XP, or Windows Vista. => Relevance = 3
 Explanation: both objects enlisted in the task (Windows 8 AND Windows XP) are mentioned; there are also other objects: Windows 7, Windows 8.1 and Windows Vista, but they are irrelevant for the task. What you should note here is:
 Windows 8 is safer than Windows 7, Windows XP, or Windows Vista => Windows 8 is safer than Windows XP => Windows 8 is better than Windows XP in terms of safety.

2.4.3. It is not clear whether the argument favors one of the objects or not
e.g. Volleyball vs Football
Rate the sentence from 0 to 3:
The most ignorant thing to say is that volleyball is harder than football => Relevance = 2
Explanation: both objects enlisted in the task (Volleyball AND Football) are mentioned; however, it is not quite clear which one is better, although some information about both sports is given. The author thinks that volleyball is not harder than football.

2.4.4. Superlative
A superlative form includes the best, the most expensive, or anything similar to the + . . . est / the most
e.g. Volleyball vs Football
Rate the sentence from 0 to 3:
Volleyball is the hardest game in the world => Relevance = 1
Explanation: only one object (Volleyball) is mentioned. At the same time, we cannot presume that volleyball is the worst game in the world because it's hard to play.

2.4.5. No direct comparison, but separate comparison
The objects mentioned in the task might be separately compared with other objects.
e.g. Windows 10 vs Windows XP
Rate the sentence from 0 to 3:
But that aside, Windows 10 is still easier to navigate than Mac OS X and Windows 8 is faster than Windows XP and Windows 7. => Relevance = 1
Explanation: here both objects (Windows 10 AND Windows XP) are mentioned, but they are only characterized separately. There is no information about the better one.

D LLM Prompt for Automatic Relevance Annotation

You need to evaluate the relevance of arguments comparing two objects. Use the following criteria to rate the relevance of each argument from 0 to 3:

1. Relevance = 3: Both objects enlisted in the comparison are mentioned in the sentence, and a clear argument is provided in favor of one of the objects.
2. Relevance = 2: Both objects enlisted in the comparison are present in the sentence, but no clear argumentation is given, or the statement is subjective without clear reasoning.
3. Relevance = 1: Only one object enlisted in the comparison is mentioned, and some useful information, which could hypothetically be used in comparison with the other object, is provided.
4. Relevance = 0: No objects enlisted in the comparison are mentioned in the sentence, or some/all of them are mentioned, but not relevant at all.

Instructions:

- Evaluate each argument based on the provided criteria.
- Consider whether the argument directly compares the objects or provides relevant supporting information.
- If the argument is subjective, assess if clear reasoning or evidence is provided to support the opinion.
- If the argument is complex or contains multiple comparisons, try to identify the main comparison and focus on that aspect for evaluation.

You will be given data in JSON format (orient = "records"). Data contains the following information:

- "comparison": two objects that are going to be compared
- "arguments": list of arguments for objects in "comparison"
- "argument": text, that should be evaluated.
- "relevance": target value, it should be one of ['0','1','2','3'], as mentioned in the instruction above.

Now, fill the "relevance" in the data below and return it in JSON format:

Figure 8: LLM prompt for automatic relevance annotation.

E LLM Prompt for Argument-Answer Coherence Computation

```
You are a helpful assistant in logic. You are given two sentences, the task is to check whether the second derives from the first one and does not contradict. Return 1 if the second derives from the first one and does not contradict, 0 otherwise.

Examples:
text1: cats are smarter than dogs
text2: cats are more intelligent than dogs
answer: 1

text1: cats are smarter than dogs
text2: dogs are more intelligent than cats
answer: 0

text1: cats are smarter than dogs
text2: cats are more elegant than dogs
answer: 0

text1: {text1}
text2: {text2}
answer:
```

Figure 9: LLM prompt for Argument-Answer Coherence computation.

F Summary Examples



Removed arguments with relevance 0-1

2. I prefer Hardibacker, a proprietary cement/gypsum product that's lighter and easier to cut than trad'l. concrete boards.
3. If the expansion is greater than 0.1%, limit the equivalent alkalis of the cement source to not greater than 0.6%.
10. The quality of sub-base materials is inferior to the base materials and includes natural gravels, cement treated gravels and sands and stabilized subgrade materials.

Summary:

Cement and concrete are often compared in construction due to their critical roles in structural integrity. Cement is a binder, while concrete consists of cement mixed with aggregates, water, and other materials. Concrete is generally more durable, cost-effective, and stable than cement alone [1, 9]. Cement's role is indispensable in concrete production, but using concrete offers better long-term performance in most construction projects [generated]. While cement plays a key role in bonding, concrete's composition ensures strength, flexibility, and stability [generated]. Therefore, concrete emerges as the superior choice for large-scale, durable applications.

Main Aspects of Comparison:

- Durability and Stability: Concrete has superior volume stability and durability compared to cement due to the inclusion of aggregates [1]. These materials enhance the long-term stability of structures, making concrete better suited for enduring projects [9].
- Cost-effectiveness: Aggregates used in concrete are typically cheaper than cement, which helps lower overall construction costs [1]. Cement alone is more expensive and less practical for large-scale use [generated].
- Weight: Cement is heavier compared to concrete blends, which are 20-30% lighter [5]. Lighter weight in construction materials reduces stress on structural foundations [generated].
- Speed of Construction: Poured concrete foundations are faster to construct than block foundations made solely of cement, saving time and labor [6].
- Thermal and Sound Insulation: Concrete's composition provides better insulation properties compared to pure cement, improving energy efficiency and soundproofing [7, generated].
- Material Properties: Modern concrete made from improved cements shows enhanced performance over earlier materials, increasing the lifespan of structures [8, 9].
- Long-term Performance: Bridges built using concrete show greater longevity compared to those using only earlier forms of cement, indicating concrete's superiority in durability [4].
- Flexibility in Application: Concrete offers greater versatility in construction projects due to its ability to be poured and shaped quickly, unlike traditional cement blocks [6, generated].

Best Option:

Concrete

Numbered List of Arguments:

1. [1]
2. [9]
3. [5]
4. [6]
5. [7]
6. [8]
7. [4]
8. [generated]
9. [generated]
10. [generated]

Figure 10: Example of summary generation with the arguments of relevance 0 and 1 provided.



Arguments

1. Dark Chocolate, Other Cocoa-Rich Foods May Lower Blood Pressure Better Than Tea (human=3,gpt4=3)
2. It's nicer than soya, and will even make a decent hot chocolate. (human=2,gpt4=2)
3. Oh, Chocolate Rocket, my love, my darling - there is no finer tea than you. (human=2,gpt4=1)
4. That might be tomorrows breakfast I can handle a bit of chocolate better than tea. (human=2,gpt4=1)
5. The chocolate flavor is superior to other teas I've had and I definitely will be purchasing this tea again. (human=2,gpt4=2)

ChatGPT generation

Human refinement

```

**Summary:**
When comparing chocolate and tea, both offer unique
flavors and potential health benefits. Dark chocolate
may be better for lowering blood pressure, while tea
is appreciated for its variety and versatility.
Ultimately, personal preference plays a significant
role in determining the better choice.

**Main Aspects of Comparison:**
- Health Benefits:
  - Dark chocolate potentially lowers blood pressure
  better than tea [1]
- Flavor:
  - Chocolate is often preferred for its rich and
  superior taste compared to some teas [5]
- Personal Preference:
  - Some individuals find chocolate more enjoyable or
  easier to handle than tea [4]
- Versatility:
  - Tea offers a wide range of flavors and types to
  cater to diverse preferences
- Ease of Consumption:
  - Chocolate can be seen as a more indulgent treat
  than tea, which is often consumed daily
- Cultural Significance:
  - Tea holds deep cultural importance in many
  societies, influencing rituals and ceremonies
- Nutritional Value:
  - Dark chocolate contains antioxidants and minerals
  beneficial for health
- Variety:
  - Tea comes in numerous types such as green, black,
  herbal, and flavored blends
- Convenience:
  - Tea is easier to prepare and carry around
  compared to chocolate
- Emotional Connection:
  - Some individuals have strong emotional
  attachments to chocolate or tea [3]
- Sustainability:
  - Considerations regarding the environmental impact
  of chocolate and tea production [generated]
- Price Point:
  - Differences in cost between high-quality
  chocolate and specialty tea [generated]
- Caffeine Content:
  - Tea contains caffeine, providing a stimulant
  effect not found in chocolate [generated]

**Best Option:**
Preference-based
  
```

```

**Summary:**
In the comparison between chocolate and tea, both
have their own cultural significance and offer
unique flavors and potential health benefits. While
chocolate has rich flavor, tea offers a lot more
variety and versatility. However, personal
preference for taste and health goals play a
significant role in determining the better choice
between the two.

**Main Aspects of Comparison:**
- Health benefits: Dark chocolate may lower blood
pressure better than tea [1].
- Flavor preferences: Chocolate is favored for its
rich flavor and aroma to that of tea and its
suitability for hot chocolate [2, 5].
- Digestive tolerance: Some people find it easier to
digest chocolate compared to tea [4].
- Antioxidant content: Tea is known for its high
antioxidant content, potentially offering more
health benefits in this aspect. [generated]
- Caffeine content: Tea contains caffeine which can
provide a gentle energy boost, whereas chocolate
contains theobromine which has similar effects.
[generated]
- Variety and Versatility: Tea offers a wide range
of flavors and blends like herbal, black and fruits,
catering to diverse preferences. [generated]
- Cultural Significance: Tea holds a significant
cultural importance in various societies and
ceremonies worldwide. Chocolate is often associated
with luxury and indulgence in many cultures.
[generated]
- Soothing Qualities: Tea is commonly consumed for
its calming and stress-relieving properties.
Chocolate can also provide comfort and pleasure to
many individuals.[generated]

**Best Option:**
Subjective
  
```

Figure 11: Example of a dataset entry before and after manual correction.