On What Basis? Predicting Text Preference Via Structured Comparative Reasoning

Anonymous ACL submission

Abstract

Comparative reasoning plays a crucial role in text preference prediction; however, large language models (LLMs) often demonstrate inconsistencies in their reasoning. While approaches like Chain-of-Thought improve accuracy in many other settings, they struggle to consistently distinguish the similarities and differences of complex texts. We introduce SC, a prompting approach that predicts text preferences by generating structured intermediate comparisons. SC begins by proposing aspects of comparison, followed by generating textual comparisons under each aspect. We select consistent comparisons with a pairwise consistency comparator that ensures each aspect's comparisons clearly distinguish differences between texts, significantly reducing hallucination and improving consistency. Our comprehensive evaluations across various NLP tasks, including summarization, retrieval, and automatic rating, demonstrate that SC equips LLMs to achieve state-of-the-art performance in text preference prediction.

1 Introduction

Comparative reasoning is crucial for predicting text preferences, as deciding the best out of a set of texts requires careful examination of the similarities and differences across the documents. Comparative reasoning has been especially useful in NLP tasks such as text summarization (Yang et al., 2023; Lee et al., 2023), search ranking (Qin et al., 2023), and automatic evaluation (Adlakha et al., 2023), where text preference prediction is a key step.

However, as corpora grow more dense and complex across domains, accurate comparative reasoning becomes increasingly challenging. Existing approaches rely on pretraining or fine-tuning models (Iso et al., 2022; Amplayo et al., 2021; Pang et al., 2021; Yu et al., 2023a) at the cost of massive human annotation. With the emergence of large language models (OpenAI, 2023; Touvron et al., 2023a; Anil et al., 2023), few-shot prompting serves as a promising way due to the remarkable language generation capacity of LLMs.

However, LLMs exhibit arbitrary and erroneous outputs when prompted for comparative reasoning (Adlakha et al., 2023). Specifically, LLMs demonstrate inconsistency in their comparative analyses. Figure 1 provides an example of logically inconsistent LLM reasoning using a common prompting approach: Chain-of-Thought (CoT) (Wei et al., 2022). The LLM's generated explanation initially describes a property as common to the text pair (highlighted in yellow), but later implies that the same property is a strength of just one of the documents (highlighted in green). This inconsistency in the LLM's comparative analysis leads to an incorrect prediction.

To address these challenges, we present SC, a Structured Comparative reasoning method that constructs an intermediate representation to contrast two text corpora for enhanced text preference prediction (as shown in Figure 1). SC first proposes a set of aspects from text pairs to guide the comparison step. Subsequently, SC generates textual comparisons for every aspect. Hence our final structured representation is structured. To improve the consistency(for instance, for a given aspect, a contrastive comparison should not overlap with the common comparison), SC adopts approximate inference with multiple samples and uses a pairwise comparator to select the most consistent one as the final output.

We demonstrate the effectiveness of the SC in many NLP tasks including text summarization (Stiennon et al., 2020), document (Soboroff et al., 2018), and detecting helpful and harmless responses (Bai et al., 2022) with an average 2.9 and 7.1 points gain over the top and bottom baselines respectively.

Our analysis further confirms the effectiveness of the design of the SC, and ablation studies empha-

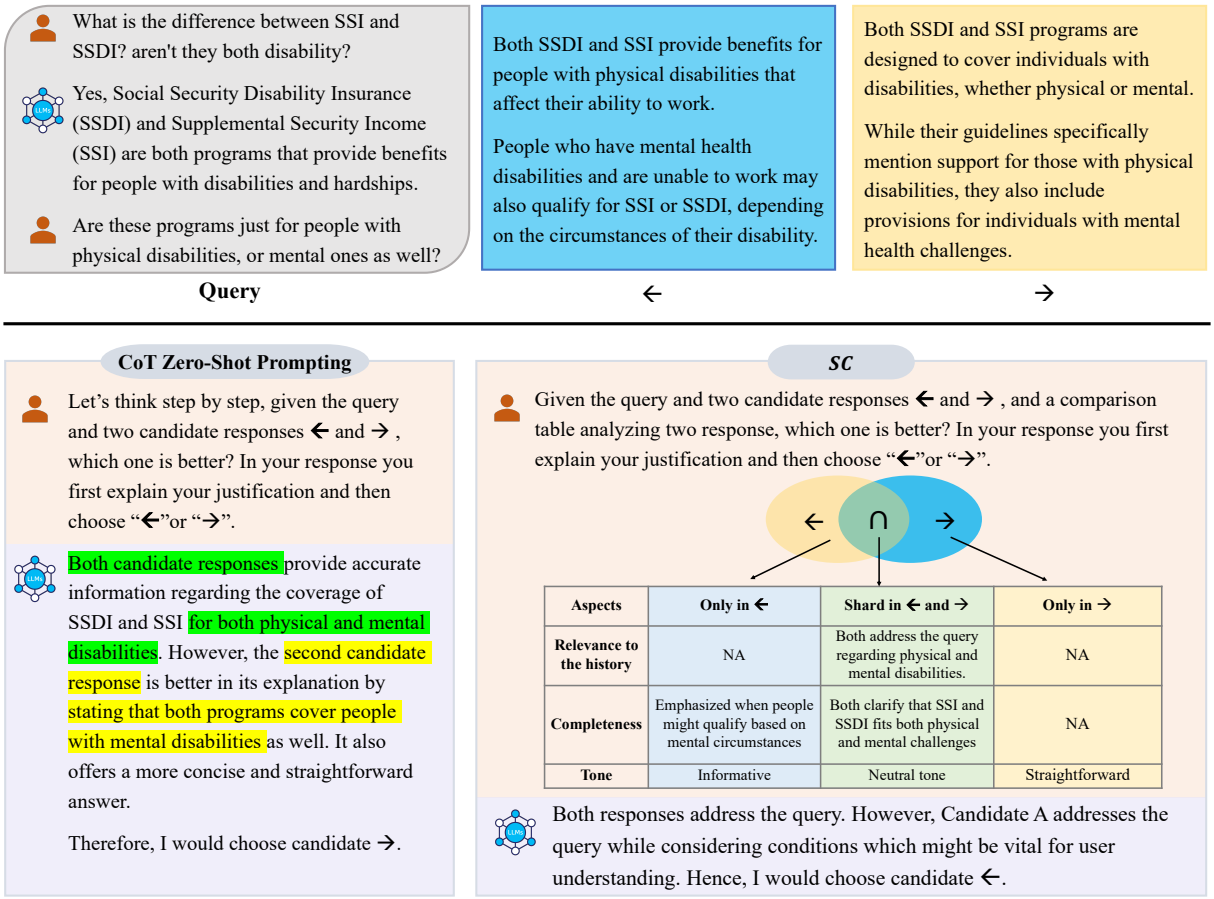

Figure 1: An example illustrating the differences between CoT-Zero-Shot prompting and prompting augmented with intermediate comparative reasoning produced by SC. The top portion shows a query between a human and a chatbot, along with two candidate responses ← and → to it. The table in the middle of the figure presents an intermediate structured representation produced by SC. Small phrases are listed as aspects in the first column. Comparisons are entries in the table(not the first column and row). The Venn diagram visualizes the atomic comparisons w.r.t to ← and →.

size the importance of the consistency comparison function. We also conduct extensive human evaluations to indicate SC aids interpretation and assists users in making decisions.

2 Related Work

Prompting Large Language Models LLMs have recently advanced the state-of-the-art performance across many NLP tasks (Anil et al., 2023; OpenAI, 2023; Chowdhery et al., 2022; Touvron et al., 2023a,b). These LLMs have demonstrated the capability to provide chain-of-thought explanations that elucidate their reasoning processes (Wei et al., 2022; Kojima et al., 2022). However, the chain-of-thoughts generated by LLMs are often arbitrary or contradictory (Turpin et al., 2023; Chen et al., 2023; Dhuliawala et al., 2023), lacking robustness to rephrased questions. To mitigate these issues, several works aim to leverage consistency-based (Wang et al., 2023), or verification-based approach (Ling et al., 2023; Madaan et al., 2023) to improve the reasoning capacity of LLMs, yet the benefit of such additional techniques are still ambivalent (Huang et al., 2023). Furthermore, all these advanced techniques still concentrate on processing raw-text inputs, thereby overlooking the integration of structural information. Moreover, they lack the implementation of explicit consistency constraints, which is crucial for maintaining logical coherence in the generated outputs.

Comparative Reasoning and Summarization Comparative reasoning involves comparing and contrasting different documents (Yu et al., 2023a), which has applications for a broad range of NLP tasks including text ranking (Jiang et al., 2023; Qin et al., 2023), reward modeling (Ouyang et al., 2022; Lee et al., 2023) and automatic text gen-

eration evaluation (Liu et al., 2023a). Initial explorations focused on mining comparative content from text corpora (Jindal and Liu, 2006; Li et al., 2010). More recent studies have developed models for generating comparative text, including generating arguments for answering comparative questions (Chekalina et al., 2021; Amplayo et al., 2021) and summarizing comparative opinions (Iso et al., 2022). Additionally, Zhong et al. (2022) prompts LLMs to describe the differences between two text distributions in natural language.

One challenge of directly prompting LLMs for comparative reasoning is that the input text often contains a mixture of diverse patterns. As such, it is crucial to incorporate fine-grained aspects to guide LLMs for generating more comprehensive summarizations (Sun et al., 2023; Xu et al., 2023; Wu et al., 2023; Yu et al., 2023b). Early works (Lin and Hovy, 2000; Titov and McDonald, 2008) used clustering or topic modeling to identify aspects in documents. Lekhtman et al. (2021) fine-tune a pretrained language model for aspect extraction, which relies on manual labeling of comparative data. On the other hand, Goyal et al. (2022); Yang et al. (2023) leverage LLMs to perform summarization with the fixed aspects provided by humans. Differently, we leverage LLMs to automatically discover aspects to guide comparative reasoning, which provides a flexible way to incorporate fine-grained task-relevant signals while requiring minimal labeling efforts.

3 Methods

Our model, SC, produces comparative reasoning for text preference prediction that applies to densely written texts, generalizes to multiple domains, and ensures consistency. In this section, we give the generative process and inference procedure for our framework SC. Our primary focus is ensuring the comparisons consistently distinguish similarities and differences between texts.

3.1 Generative Process

The generative process has three steps. First, given a text pair, our model SC simplifies the task by delineating a set of aspects, as depicted in Figure 1. These aspects, consisting of concise phrases, enable the structured comparison between the texts. Second, SC produces comparisons, which are concise, aspect-focused comparative statements that clearly express how the texts are similar and different. We require the comparisons to be consistent: similarities identified as shared between the text pair should not overlap with what's unique to each of them. Given the aspects and comparisons, the third and final step predicts which text is preferred.

Formally, for a text pair problem, we denote the text pair as $\leftarrow$ and $\rightarrow$, along with a query. **SC** has three components: Aspects $a = \{a_1, a_2, \ldots, a_n\}$, comparisons $c = \{c_1, c_2, \ldots, c_n\}$, and text preferences $y \in \{\leftarrow, \rightarrow\}$. The comparison c has three columns: $\{c^{\leftarrow}, c^{\rightarrow}, c^{\cap}\}$, $c^{\rightarrow}$ refers to properties exclusive to one text and $c^{\cap}$ to properties shared by both texts.

SC follows the following generative process: First, it generates the aspects conditioned on the text using an *aspect model*, $P(a)$. Second, comparisons for each aspect are generated from the comparison model

$$P(c|a) \propto \prod_i l(c_i) \times P(c_i|c_{<i}, a)$$

where the function $l : C \rightarrow \mathbb{R}^+$ evaluates the consistency of c_i. A higher value of $l(c_i)$ indicates a greater degree of consistency. Finally, the preference model $P(y|c, a)$ produces the preference label y.

Parameterization We use LLMs with specific prompts to parameterize each model. With LLMs generating reliable scalar values of consistency is unreliable (Imani et al., 2023; Liu et al., 2023b). Instead of directly regressing a consistency score, we rely on pairwise comparisons, which have been observed to be more reliable (Qin et al., 2023). We define a pairwise comparator $l'(c, c') = \mathbb{1}(l(c) \geq l(c'))$[1], which takes a pair of comparisons (c, c') and determines the more consistent one.

To facilitate this, we recruit experts to develop few-shot prompts that demonstrate a direct comparison of two structured representations based on consistency within itself. We guide our annotators to assess pairs (c, c') against consistency criteria, emphasizing that elements of the comparison should ideally exhibit no overlap. Detailed instructions are attached in the Appendix.

3.2 Tournament-based Inference

Given the generative model, the goal of inference is to produce aspects and comparisons that are both high probability under the model and consistent. We take a step-wise approach, choosing aspects,

[1]We break the tie randomly.

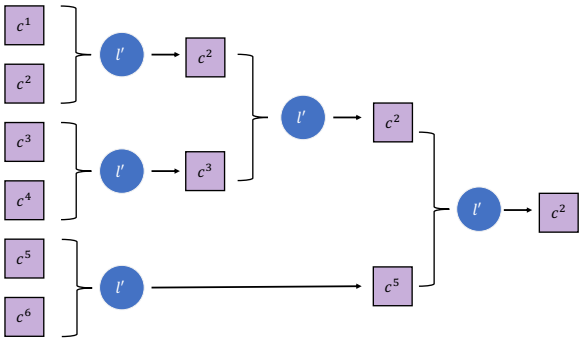

Figure 2: Illustration of tournament approach. Given a set of samples, $C = \{c^1, c^2, c^3, c^4, c^5, c^6\}$, the tournament approach randomly partitions them into three groups in the first round, and each two is paired as input to l' and output from l' will be entering the next round. In this way, we only need to call l' 5 times.

comparisons, and then finally predicting preferences.

When choosing aspects, we follow prior work by employing a variety of sampling strategies to obtain near-optimal aspects a^* from $P(a)$ (Wang et al., 2023; Amplayo et al., 2021). We provide more details on these strategies in Section 4.

Given aspects a^*, our next goal is to find comparisons that are likely under the comparison model $\arg\max_c P(c|a^*) = \arg\max_c l(c) \cdot P(c|a^*)$. There are two challenges with this objective: First, the set of possible comparisons is intractably large. Second, the consistency function $l(c)$ is unreliable. We approach the first challenge by sampling a set C of high probability comparisons from $P(c|a)$, and the second challenge by selecting the most consistent comparison by applying the pairwise consistency comparator $l'(c, c')$ in a binary reduction. Formally, we select the most consistent comparison by optimizing

$$c^* = \arg\max_{c \in C} \sum_{c' \in C \setminus \{c\}} l'(c, c').$$

Naively, this optimization problem above requires $O(|C|^2)$ pairwise comparisons to optimize exactly. To reduce the number of pairwise comparisons, we utilize a tournament approach that performs $O(|C|)$ comparisons. The tournament approach utilizes a binary reduction: Each step of the binary reduction takes a pair of comparisons and eliminates the less logically consistent one into the successive rounds. We illustrate the tournament approach in Figure 2. The naive and tournament approaches are equivalent if transitivity holds in the consistency comparator $l'(c, c')$. In practice, transitivity does not hold with LLM parameterizations, resulting in the tournament approach trading off accuracy for efficiency.

Finally, we decide between $\leftarrow, \rightarrow$ which one is preferred by taking $\arg\max P(y|a^*, c^*)$.

4 Experimental Setup

Aspect Model We experiment with two configurations for generating aspects: dynamically generating aspects using the Chain-of-Thought (CoT) paradigm (Wei et al., 2022) and self-consistency (Wang et al., 2023) to select the most agreeable aspect, and using fixed aspects mined from a massive text corpus. In this experimental study, we report the best results for all baselines that have utilized aspects.

Comparison Model We used PaLM-2-L[2] (Anil et al., 2023) as the major LLM backbone of SC to produce structured comparative representation.

Preference Model For the final text preference inference model, we have used two LLM backbones differing in their model capacity. We aim to prove that the structured representations produced by SC can help other backbone LLMs to infer the preference correctly, regardless of their capacity. Specifically, we have used Open AI's GPT-3.5[3], and GPT-4 (OpenAI, 2023)[4].

Prompting Templates Prompts used in different for different models can be found in our Appendix. Note that we do not tailor the preference model's prompts, instead, we adapted the templates from Rafailov et al. (2023) for a fair comparison across baselines[5].

Hyperparameters As SC searches for best comparisons during the inference stage, as a result, we have a hyperparameter $|C|$, referring to the number of samples generated by the comparison model. $|C|$ is an important parameter that might affect the

[2] https://ai.google/discover/palm2/

[3] https://platform.openai.com/docs/models/gpt-3-5

[4] https://platform.openai.com/docs/models/gpt-4

[5] In the original DPO paper (Rafailov et al., 2023), the authors did not use the Anthropic-Harmless dataset, hence we cannot directly adapt their template. We instead use their Anthropic-Helpful dataset's prompting template and replace "helpful" with "harmless", making some necessary minor wording changes.

Preference Model	Comparison Model	TLDR			RLAIF			Document Ranking
		Reddit	CNN/DM	AVG	Helpful	Harmless	AVG	TREC News
GPT-3.5	DP	62.89	61.39	62.14	58.40	58.15	58.27	44.36
	DP w/Aspects	62.50	62.55	62.52	59.20	53.72	56.46	46.18
	CoT-0-shot	63.67	64.48	64.08	59.00	56.94	57.97	47.64
	CoT-1-shot	64.06	63.71	63.88	59.20	58.55	58.88	50.18
	CoT-SC	65.23	63.32	64.28	60.60	58.75	59.68	50.55
	SC	**68.36**	**68.34**	**68.55**	**63.20**	**59.76**	**61.49**	**53.09**
GPT-4	Prompting	66.41	64.86	65.63	62.60	58.85	60.58	52.00
	DP w/Aspects	65.63	65.25	65.44	60.60	60.97	60.78	55.64
	CoT-0-shot	68.75	68.34	68.54	63.00	60.56	61.78	59.64
	CoT-1-shot	69.92	69.50	69.71	63.80	60.16	61.98	61.09
	CoT-SC	71.67	69.12	69.90	64.00	60.76	62.38	61.82
	SC	**73.83**	**71.43**	**72.63**	**66.60**	**62.98**	**64.79**	**64.73**

Table 1: Experimental results of SC across different datasets in three different domains. We use accuracy to measure the performance and report averaged the results from 5 rounds. All the experiment results have passed the significance test ($p < 0.05$).

quality of the intermediate structured representation produced by SC. For the reported results in this section, we set $|C| = 8$. We study the influence of this hyperparameter in our analysis section.

Baselines For evaluation, we consider several baselines, primarily focused on Language Model (LLM) based prompting methods. Below is a detailed overview of these baselines:

(1) *Direct Prompting (DP):* This method involves only a preference model that directly prompts LLMs to predict preference without relying on any additional information.

(2) *DP w/Aspects:* This approach is a variation of DP. Its prompting template also contains aspects that are generated by the aspect model. These aspects guide the LLM in making comparisons, though they don't provide the model with explicit aspect-specific comparisons.

(3) *CoT-0-shot:* This baseline utilizes the standard CoT-0-shot template for task preference prediction, as discussed in Wei et al. (2022). More details are available in the appendix.

(4) *CoT-1-shot:* Beyond the 0-shot settings, we also carried out experiments using a 1-shot example within the CoT paradigm. For that purpose, we crafted our 1-shot examples across different sets.

(5) *CoT-SC:* We integrate self-consistency from Wang et al. (2023) to our CoT-0-shot baseline. We set the number of decoded responses to be 8. We use a majority vote to decide the desired response, and we randomly pick up the response when there is a tie.

Dataset	# Samples	Avg. Length
TL;DR-CNN/DM	256	572
TL;DR-Reddit	259	362
Antropic-Helpful	250	102
Antropic-Harmless	249	93
TREC News	291	947
AVG	278	342.8

Table 2: Statistics of Datasets in Experimental Studies

Datasets (1) **TL;DR (Stiennon et al., 2020)**, we use OpenAI's filtered Reddit and CNN/Daily Mail TL;DR dataset with 3 million posts. The Reddit dataset, focused on quality, includes summaries from select subreddits, limited to 24-48 token lengths, totaling 123,169 posts with 5% as a validation set. OpenAI also created a preference dataset from this, where labelers rated two generated summaries per post. For the CNN/Daily Mail part, for a given news, we extracted two graded summaries and used the overall score to decide the label. More details are in the original paper.
(2) **RLAIF-HH (Bai et al., 2022):** The AnthropicHH dataset comprises dialogues from interactions between crowdworkers and large language models. In these exchanges, workers either seek assistance or provoke potentially harmful responses

from the AI. The responses are then labeled based on their helpfulness or harmfulness. The dataset, split into 161k training and 9k test examples, features each instance tagged with a task and includes a 'better' and 'worse' response. The focus is on the 'helpfulness' aspect, particularly using the 'better' response as the target for Supervised Fine-Tuning (SFT) experiments.

(3) **TREC News (Soboroff et al., 2018):** The TREC News dataset contains over 1 million news articles from the late 1980s to early 2000s. It provides query-document pairs focused on ad-hoc ranking and filtering tasks. The query-document pairs make it highly applicable for IR and NLP research. We modify the dataset as follows: for a given query, we extract two document answers to construct the triplet and use the relevance score provided by the original dataset to decide which document is more preferred.

Metrics We report the accuracy of all approaches in our experiment ($\frac{\text{Correctly Predicted Instances}}{\text{All Instances}}$) to measure the performance.

Dataset Sampling As datasets that have been used in the past are in large volumes, we only sampled a small ratio of them due to the cost of running all experiments. We sample roughly 250-300 data points from each dataset uniformly. We delete selected samples that were not successfully parsed by the GPT models across all the different comparative reasoning approaches.

We also calculate the total length L of words to the final text preference prediction model[6]. The L is calculated by counting the length of $\leftarrow, \rightarrow$, query, and their prompts as shown in Table 2.

5 Results

Experimental results in Table 1 demonstrate SC's strong performance across all evaluation domains, with average gains of 2.5 and 7 points over the top and bottom baselines respectively. This confirms the benefits of structured comparative reasoning for enhanced text preference prediction. Using structured intermediate representations produced by SC, the preference prediction model better handles these comparative reasoning difficulties.

Moreover, we observe the input length as an additional factor impacting performance. For instance, the TREC News dataset comprises considerably longer texts than other corpora. Here, the DP method lags SC by over 9 points, compared to the average 7 point deficit across baselines. Though input length serves as an imperfect proxy for complexity, the results also signaled the potential benefit of using our method for longer inputs.

We also want to point out that SC could be further improved by coupling with some of the existing general prompting techniques, for example, self-consistency (Wang et al., 2023) and self-verification (Madaan et al., 2023).

6 Analysis

To further understand the benefit of using SC to produce an intermediate structured representation, in this section, we conduct ablation studies and in-depth analysis. We also implement a user study to explore the potential of using SC to inform human beings' decisions.

6.1 Effectiveness of Consistency Pairwise Comparator

To calibrate the effective gain arising from the pairwise comparator l', we first compare variants of SC with the comparators and those with different hyperparameter configurations of SC. We use different intermediate structured representations produced by variants of SC to predict the text preference. We limit our experiments to two datasets with 100 samples each for cost and environmental considerations. Results are shown in Figure 3.

With $|C| = 1$, where there is effectively no pairwise comparator l', the performance of the preference model was found to be comparable to baseline results shown in Table 1. This suggests that inconsistent structured representations could potentially degrade the performance of the preference model. An increase in accuracy was observed with larger values of $|C|$, indicating the benefits of pairwise comparator. However, this improvement plateaued when $|C|$ exceeded 8, hinting at a potential ceiling effect for our approach, irrespective of further increases in $|C|$.

6.2 Efficiency of Tournament Approach

We study the efficiency and effectiveness of the tournament approach w.r.t. other inference methods. We set $|C| = 8$ for the sampling. Random Selection refers to the process of randomly selecting one sample from C during the inference stage, while Exact Search involves running all possible

[6]We do not calculate the tokens as different LLMs used different tokenizations.

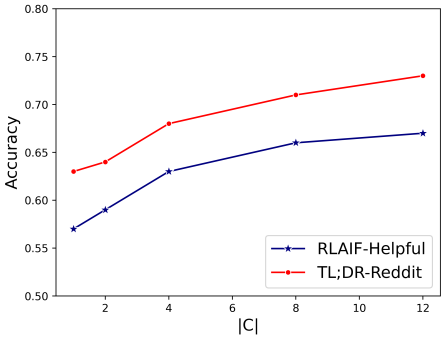

Figure 3: Impact of # samples $|C|$ in SC.

comparisons, which takes $O(n^2)$. We measure the cost using the total input length and the number of LLM calls, as this is common practice for the actual cost calculation in commercial Large Language Models (LLMs). We used the same dataset TL;DR-Reddit from the previous subsection.

We find a significant gap between the Random Selection approach and the other two approaches as shown in Figure 3. Although Exact Search yields the best results, it requires 4 times the token length and 49 more LLM calls, potentially leading to a substantial increase in cost.

In our primary experimental studies, we use PaLM2-L as the core model for creating structured representations, demonstrating its versatility through various aspect models. Contrastingly, CoT-SC baseline consistently employs the same model for both comparison and preference evaluations. COT-SC also has no pairwise comparator component, leading to less total LLM usage.

To ensure a fair comparison and eliminate any performance discrepancies that might arise from using different LLMs and the number of total LLM calls, we conduct an analysis using GPT-4 as the uniform backbone for both SC and CoT-SC, one of the strongest baselines. We allocate a fixed number of total LLM calls for the entire text preference prediction process. This count includes all LLM interactions, starting from generating structured representations or responses to the final decision-making process where we choose between two candidate texts. Results of this analysis are presented in Table 4. Our analysis shows that with identical total LLM calls and the same LLM backbone, SC consistently delivers superior outcomes.

6.3 Coverage and Entailment

We also analyze the entailment and coverage scores of the intermediate structured representations produced by SC. We use internal coverage and entailment prediction models, where the score is in $[0, 1]$, and a higher score indicates better coverage or entailment. On one hand, we aim to check the coverage and entailment depth of the structured representations, and on the other, we seek to determine whether existing metrics could serve as good indicators: whether computing those metrics alone can select better responses from LLMs. Specifically, we consider the entire structured representation from SC and calculate the coverage and entailment score given the concatenation of $\leftarrow$ and $\rightarrow$. We compare different representations from variants of SC. Results are shown in Figure 4.

As the comparisons are distilled into comparative statements, a relatively lower score is expected. However, we find that the representations of various SC configurations are not drastic in variance. This indicates that LLMs might be hallucinating entities, names, and locations mentioned in $\leftarrow$ and $\rightarrow$, which existing metrics such as coverage and entailment find challenging to identify. This also

	Random Selection	Tournament Scheme	Exact Search
# LLM calls	1	7	56
Decoded Len	372	2,651	13,272
Accuracy	0.63	0.71	0.73

Table 3: Cost and accuracy analysis of different inference approach of SC.

	SC	CoT-SC
# Total LLM calls	15	15
Decoded Len	5,459	4,380
Accuracy	0.73	0.71

Table 4: Cost and accuracy analysis comparison of SC against CoT-SC with the same LLM and number of total API calls.

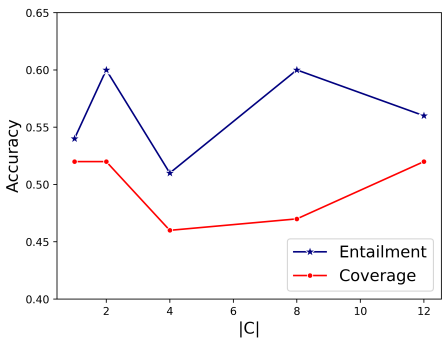

Figure 4: Entailment and coverage score of SC.

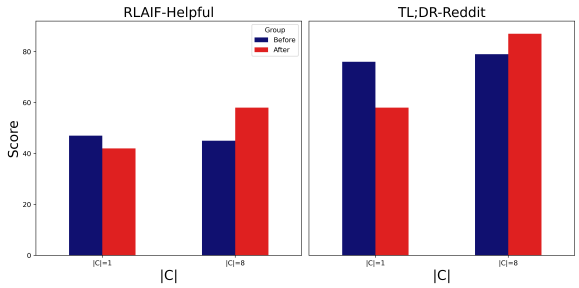

Figure 5: Human evaluation on structured representation produced by different settings of SC

suggests that incorporating a pairwise comparator aids in obtaining reliable comparative reasoning.

6.4 Human Evaluation

We also conduct human evaluations to see how the intermediate structured representations produced by SC inform human decision-making.

Annoators We recruit our annotators from an internal pool. Demographic and geographic characteristics of the annotator population are not accessible to our researchers. Information can be used to identify annotators that are fully anonymized. Consent forms have to be signed by annotators to take part in this study.

Study Design In consideration of ethical standards and the requirement to avoid directly testing annotators, we structure our human evaluation as follows: Annotators are presented with a query alongside a pair of text options, denoted as $(\leftarrow, \rightarrow)$. They determine which text, either $\leftarrow$ or $\rightarrow$, is preferable. They have three options: $\leftarrow$ is better, $\rightarrow$ is better, and tie. Following their initial decision, annotators are then shown the intermediate structured representations generated by different variants of SC. They decide if this additional information leads them to reconsider their initial choice and provide reasons for any change in their decision. This evaluation process uses two variants of SC: $|C| = 1$ and $|C| = 8$ respectively. For ethical considerations, we only experiment with RLAIF-helpfulness and TL;DR-Reddit, ensuring the content is not harmful or violent manually. We instantiate 100 data points for each dataset and assign each question to three annotators. We collect 96 and 98 questions with useful responses from all three annotators for RLAIF-helpful and TL;DR-Reddit respectively.

Metrics We use the ground truth to gather the scores — we assign 1 for any correct answer, 0 for any answer that is a `tie`, and -1 for any other incorrect answers. We use majority voting to determine whether agreement existed for each question.

Findings As demonstrated in Figure 5, with the aid of more consistent structured representations, annotators are inclined to revise their choices to the correct text preference. This suggests that SC may facilitate better decision-making among human evaluators. Conversely, we observe that inconsistent structured representations produced by SC without a consistency check component have the potential to mislead annotators, deterring them from selecting the correct preference.

We also look into the justifications provided by our annotators. Most annotators stated that the structured representations helped them better understand two texts. One mentioned, `"the table gives the concise comparison"`, while another pointed out, `"this [table] helped me to understand better the implications of the two answers, and I changed my mind after reading [the table]"`. Besides, we also observe complaints about the structured representations being hallucinatory and not factual. The issue is more noticeable in cases where the structured representation is produced by `SC` without a consistency pairwise comparator. This suggests that enforcing a consistency pairwise comparator might mitigate the arbitrariness of LLM's output, but still poses the risk of presenting hallucinated results to human evaluators.

7 Conclusion

This paper presents SC, a structured self-comparative reasoning methodology for improving text preference prediction. SC constructs intermediate structured representations to explicitly contrast text pairs, incorporating a consistency comparator to enhance accuracy and coherence. Comprehensive experiments across text summarization, retrieval, and response rating tasks demonstrated that SC significantly improves consistency and achieves state-of-the-art performance. Analyses confirm the effectiveness of SC's structured reasoning approach and consistency enforcement. Our human evaluations show that SC interpretations can assist users in making informed decisions.

8 Limitations

This work has several limitations that provide opportunities for future investigation. First, the evaluation was conducted on a sample set of datasets that, while spanning diverse domains, might not fully characterize the breadth of real-world textual comparison needs. Expanding SC's testing to larger, multilingual corpora is essential to assess its full potential and limitations beyond English. Furthermore, there are likely upper bounds on SC's effectiveness imposed by the reasoning capacity of the underlying language model backbone. As more advanced LLMs emerge, exploring their integration could help quantify this ceiling effect. On a technical level, in this paper, measuring consistency relies on approximate metrics, so developing more rigorous evaluation schemes could better highlight SC's benefits. We also do not include other prompting techniques that have been well-studied in the community, which we leave for future work.

9 Ethical Considerations

This research paper might risk potential biases that could arise from textual comparisons, particularly around sensitive attributes. SC is trained on established corpora like Wikipedia and books that may inherently contain societal biases. While a full analysis of these biases is beyond the scope here, we acknowledge the risk that SC may inherit problematic biases from its training data. Applying recent advancements in language bias detection to SC could help quantify and mitigate these risks. We are interested in exploring this as part of future work. Furthermore, this research focused solely on English; extending to other languages is an important direction that would require non-trivial adaptation. Overall, while showing promise, SC has significant scope for improvement as limitations around evaluation, multilingual capabilities, consistency measurement, bias, and applied usage are addressed through future work.

References

Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. 2023. Evaluating correctness and faithfulness of instruction-following models for question answering. *ArXiv preprint*, abs/2307.16877.

Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata. 2021. Aspect-controllable opinion summarization. In *Proc. of EMNLP*, pages 6578–6593, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *ArXiv preprint*, abs/2305.10403.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *ArXiv preprint*, abs/2204.05862.

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.

Yanda Chen, Ruiqi Zhong, Narutatsu Ri, Chen Zhao, He He, Jacob Steinhardt, Zhou Yu, and Kathleen McKeown. 2023. Do models explain themselves? counterfactual simulatability of natural language explanations. *ArXiv preprint*, abs/2307.08678.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *ArXiv preprint*, abs/2204.02311.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models. *ArXiv preprint*, abs/2309.11495.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *ArXiv preprint*, abs/2209.12356.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2023. Large language models cannot self-correct reasoning yet. *ArXiv preprint*, abs/2310.01798.

Shima Imani, Liang Du, and Harsh Shrivastava. 2023. Mathprompter: Mathematical reasoning using large language models. *ArXiv preprint*, abs/2303.05398.

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.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023. LLM-blender: Ensembling large language models with pairwise ranking and generative fusion. In *Proc. of ACL*, pages 14165–14178, Toronto, Canada. Association for Computational Linguistics.

Nitin Jindal and Bing Liu. 2006. Identifying comparative sentences in text documents. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 244–251.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *ArXiv preprint*, abs/2309.00267.

Entony Lekhtman, Yftah Ziser, and Roi Reichart. 2021. DILBERT: Customized pre-training for domain adaptation with category shift, with an application to aspect extraction. In *Proc. of EMNLP*, pages 219–230, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shasha Li, Chin-Yew Lin, Young-In Song, and Zhoujun Li. 2010. Comparable entity mining from comparative questions. In *Proc. of ACL*, pages 650–658, Uppsala, Sweden. Association for Computational Linguistics.

Chin-Yew Lin and Eduard Hovy. 2000. The automated acquisition of topic signatures for text summarization. In *COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics*.

Zhan Ling, Yunhao Fang, Xuanlin Li, Zhiao Huang, Mingu Lee, Roland Memisevic, and Hao Su. 2023. Deductive verification of chain-of-thought reasoning. *ArXiv preprint*, abs/2306.03872.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023a. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *ArXiv preprint*, abs/2303.16634.

Yixin Liu, Avi Singh, C Daniel Freeman, John D Co-Reyes, and Peter J Liu. 2023b. Improving large language model fine-tuning for solving math problems. *ArXiv preprint*, abs/2310.10047.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *ArXiv preprint*, abs/2303.17651.

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

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Richard Yuanzhe Pang, Adam Lelkes, Vinh Tran, and Cong Yu. 2021. AgreeSum: Agreement-oriented multi-document summarization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3377–3391, Online. Association for Computational Linguistics.

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, et al. 2023. Large language models are effective text rankers with pairwise ranking prompting. *ArXiv preprint*, abs/2306.17563.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *ArXiv preprint*, abs/2305.18290.

Ian Soboroff, Shudong Huang, and Donna Harman. 2018. Trec 2018 news track overview.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. 2020. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. 2023. Principle-driven self-alignment of language models from scratch with minimal human supervision. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Ivan Titov and Ryan T. McDonald. 2008. Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th International Conference on World Wide Web, WWW 2008, Beijing, China, April 21-25, 2008*, pages 111–120. ACM.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *ArXiv preprint*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *ArXiv preprint*, abs/2307.09288.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2023. Fine-grained human feedback gives better rewards for language model training. *ArXiv preprint*, abs/2306.01693.

Ran Xu, Hejie Cui, Yue Yu, Xuan Kan, Wenqi Shi, Yuchen Zhuang, Wei Jin, Joyce Ho, and Carl Yang. 2023. Knowledge-infused prompting: Assessing and advancing clinical text data generation with large language models. *ArXiv preprint*, abs/2311.00287.

Xianjun Yang, Yan Li, Xinlu Zhang, Haifeng Chen, and Wei Cheng. 2023. Exploring the limits of chatgpt for query or aspect-based text summarization. *ArXiv preprint*, abs/2302.08081.

Mengxia Yu, Zhihan Zhang, Wenhao Yu, and Meng Jiang. 2023a. Pre-training language models for comparative reasoning. *ArXiv preprint*, abs/2305.14457.

Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2023b. Large language model as attributed training data generator: A tale of diversity and bias. *ArXiv preprint*, abs/2306.15895.

Ruiqi Zhong, Charlie Snell, Dan Klein, and Jacob Steinhardt. 2022. Describing differences between text distributions with natural language. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 27099–27116. PMLR.

Appendix

Which of the following summaries does a better job of summarizing the most important points in the given forum article, without including unimportant or irrelevant details?

A good summary is both precise and concise.

Original Article:
{article}

Summary A:
{contextA}

Summary B:
{contextB}

Take a deep breath and think about this question step by step! FIRST, think step by step to have a comparison of the two summaries, explaining which you prefer and why. SECOND, on a new line, state only "A" or "B" to indicate your choice.

Your response should use the format:
Comparison: <step by step comparison>
Preferred: <"A" or "B">.

Figure 6: Preference model prompt for CoT Zero-shot Prompting for TL;DR

Which of the following summaries does a better job of summarizing the most important points in the given forum article, without including unimportant or irrelevant details?

A good summary is both precise and concise.

Original Article:
{article}

Summary A:
{contextA}

Summary B:
{contextB}

FIRST, explaining which you prefer and why. SECOND, on a new line, state only "A" or "B" to indicate your choice.

Your response should use the format:
Comparison: <step by step comparison>
Preferred: <"A" or "B">.

Figure 7: Preference model prompt for Direct Prompting for TL;DR

Which of the following summaries does a better job of summarizing the most important points in the given forum article, without including unimportant or irrelevant details? You are also given some aspects to help you make the decision

A good summary is both precise and concise.

Original Article:
{article}

Summary A:
{contextA}

Summary B:
{contextB}

Aspects:
{aspects}

FIRST, explaining which you prefer and why. In your evaluation, you need to consider aspects that are given above. SECOND, on a new line, state only "A" or "B" to indicate your choice.

Your response should use the format:
Comparison: <step by step comparison>
Preferred: <"A" or "B">.

Figure 8: Preference model prompt for Direct Prompting with Aspects for TL;DR

Which of the following summaries does a better job of summarizing the most important points in the given forum article, without including unimportant or irrelevant details? You are also given a comparative reasoning table that analyzes the differences and similarities between the two summaries.

A good summary is both precise and concise.

Original Article:
{article}

Summary A:
{contextA}

Summary B:
{contextB}

Comparative Reasoning Table:
{table}

FIRST, explain which you prefer and why. In your evaluation, you can use the comparative reasoning table above to help you make the justification and the decision. SECOND, on a new line, state only "A" or "B" to indicate your choice.

Your response should use the format:
Comparison: <step by step comparison>
Preferred: <"A" or "B">.

Figure 9: Preference model prompt for SC for TL;DR

Which of the following documents aligns better with the query given?

A good retrieved document should be relevant to the query.

Query:
{query}

Document A:
{contextA}

Document B:
{contextB}

Take a deep breath and think about this question step by step! FIRST, think step by step to have a comparison of the two retrieved documents, explaining which you prefer and why. SECOND, on a new line, state only "A" or "B" to indicate your choice.

Your response should use the format:
Comparison: <step by step comparison>
Preferred: <"A" or "B">.

Figure 10: Preference model prompt for Zero-shot CoT Prompting for TREC News

Which of the following documents aligns better with the query given?

A good retrieved document should be relevant to the query.

Query:
{query}

Document A:
{contextA}

Document B:
{contextB}

FIRST, have a comparison of the two retrieved documents, explaining which you prefer and why. SECOND, on a new line, state only "A" or "B" to indicate your choice.

Your response should use the format:
Comparison: <step by step comparison>
Preferred: <"A" or "B">.

Figure 11: Preference model prompt for Direct Prompting for TREC News

Which of the following documents aligns better with the query given? You are also given some aspects to help you make the decision

A good retrieved document should be relevant to the query.

Query:
{query}

Document A:
{contextA}

Document B:
{contextB}

Aspects:
{aspects}

FIRST, have a comparison of two retrieved documents, explaining which you prefer and why. In your evaluation, you need to consider aspects that are given above. SECOND, on a new line, state only "A" or "B" to indicate your choice.

Your response should use the format:
Comparison: <step by step comparison>
Preferred: <"A" or "B">.

Figure 12: Preference model prompt for Direct Prompting with Aspects for TREC News

Which of the following documents aligns better with the query given? You are also given a comparative reasoning table that analyzes the differences and similarities between the two documents.

A good retrieved document should be relevant to the query.

Query:
{query}

Document A:
{contextA}

Document B:
{contextB}

Comparative Reasoning Table:
{table}

FIRST, explaining which you prefer and why. In your evaluation, you can use the comparative reasoning table above to help you make the justifications and decision. SECOND, on a new line, state only "A" or "B" to indicate your choice.

Your response should use the format:
Comparison: <step by step comparison>
Preferred: <"A" or "B">.

Figure 13: Preference model prompt for SC for TREC News

For the following query to a chatbot, which response is more helpful?

Query to a Chatbot:
{article}

Response A:
{contextA}

Response B:
{contextB}

Take a deep breath and think about this question step by step! FIRST, think step by step to have a comparison of the two responses generated, explaining which you prefer and why. SECOND, on a new line, state only "A" or "B" to indicate your choice.

Your response should use the format:
Comparison: <step by step comparison>
Preferred: <"A" or "B">.

Figure 14: Preference model prompt for Zero-shot CoT Prompting for RLAIF-HH

For the following query to a chatbot, which response is more helpful?

Query to a Chatbot:
{article}

Response A:
{contextA}

Response B:
{contextB}

FIRST, have a comparison of the two generated responses, explaining which you prefer and why. SECOND, on a new line, state only "A" or "B" to indicate your choice.

Your response should use the format:
Comparison: <step by step comparison>
Preferred: <"A" or "B">.

Figure 15: Preference model prompt for Direct Prompting for RLAIF-HH

For the following query to a chatbot, which response is more helpful? You are also given some aspects to help you make the decision

Query to a Chatbot:
{article}

Response A:
{contextA}

Response B:
{contextB}

Aspects:
{aspect}

FIRST, have a comparison of the two generated responses, explaining which you prefer and why. In your evaluation, you need to consider aspects that are given above. SECOND, on a new line, state only "A" or "B" to indicate your choice.

Your response should use the format:
Comparison: <step by step comparison>
Preferred: <"A" or "B">.

Figure 16: Preference model prompt for Direct Prompting with Aspects for RLAIF-HH

For the following query to a chatbot, which response is more helpful? You are also given a comparative reasoning table that analyzes the differences and similarities between the two generated responses.

Query to a Chatbot:
{article}

Response A:
{contextA}

Response B:
{contextB}

Comparative Reasoning Table:
{table}

FIRST, explain which you prefer and why. In your evaluation, you can use the comparative reasoning table above to help you make the justifications and decisions. SECOND, on a new line, state only "A" or "B" to indicate your choice.

Your response should use the format:
Comparison: <step by step comparison>
Preferred: <"A" or "B">.

Figure 17: Preference model prompt for SC for RLAIF-HH

Instructions: Your task is to conduct a consistency analysis of two generated comparative table responses. Your evaluation should focus solely on the consistency of the responses. Each comparative table is constructed to delineate similarities and differences about a given query, juxtaposing candidate Summary 1 against candidate Summary 2. Consistency in this context refers to the logical coherence within each table. Specifically, for each row corresponding to an aspect-level comparison, the entries of three columns that denote similarities should be distinct and non-overlapping with the entries that denote differences. A consistent response will differentiate between the commonalities and disparities, ensuring that the information under the 'similarities' column does not overlap what is presented under the 'differences' column. This clear segregation is crucial in assessing the quality of the responses and their effectiveness in summarizing and contrasting the key points from the summaries.

Query to a Chatbot:
{article}

Summary 1:
{contextA}

Summary 2:
{contextB}

Comparartive Table Response A:
{contextA}

Comparartive Table Response B:
{contextB}

More consistent: <"A" or "B">.
Justifications: <Justifications>.

Figure 18: Instructions to Craft prompts for Pairwise Comparator.