"This Suits You the Best": Query Focused Comparative Explainable Summarization

Anonymous ACL submission

Abstract

Product recommendations inherently involve comparisons, yet traditional opinion summarization often fails to provide holistic comparative insights. We propose the novel task of generating **Q**uery-**F**ocused **C**omparative **E**xplainable **S**ummaries (QF-CES) using **M**ulti-**S**ource **O**pinion **S**ummarization (M-OS). To address the lack of query-focused recommendation datasets, we introduce MS-Q2P, comprising $\mathbf{7,500}$ queries mapped to $\mathbf{22,500}$ recommended products with metadata. We leverage Large Language Models (LLMs) to generate tabular comparative summaries with query-specific explanations. Our approach is personalized, privacy-preserving, recommendation engine-agnostic, and category-agnostic. M-OS as an intermediate step reduces inference latency approximately by $\mathbf{40\%}$[1] compared to the direct input approach (DIA), which processes raw data directly. We evaluate open-source and proprietary LLMs for generating and assessing QF-CES. Extensive evaluations using **QF-CES-PROMPT** across 5 dimensions (`clarity`, `faithfulness`, `informativeness`, `format adherence`, and `query relevance`) showed an average Spearman correlation of $\mathbf{0.74}$ with human judgments, indicating its potential for QF-CES evaluation[2].

1 Introduction

E-commerce platforms host a vast array of products, but users face challenges in decision-making despite recommendation systems. Users, each with unique quality preferences, budget constraints, and desired features, often find themselves sifting through specifications and reviews of multiple (but often quite similar) products. While recommendation systems, match products to queries, they lack comparative insights crucial for informed decisions. Users struggle to understand how recommended items stack up against each other in ways that matter most to their individual needs and specific queries. Tay (2019) highlight that opinion summarization approaches condense reviews by frequently emphasizing recurring aspects, potentially introducing bias and giving users a skewed perception of their importance. However, this approach overlooks valuable information embedded within product metadata, highlighting the need for M-OS. Im et al. (2021); Li et al. (2020a) explored methods incorporating reviews, images, and metadata to provide users with informative summaries that capture both subjective opinions and objective product attributes, as demonstrated by Siledar et al. (2023). These approaches generate single-product summaries without user query context or cross-product comparisons, forcing users to manually compare items, leading to decision fatigue and a sub-optimal shopping experience.

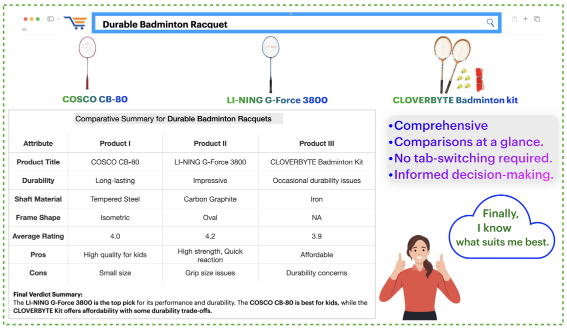

Figure 1: QF-CES enables quick comparison of top-3 recommended products for confident decisions without tab-switching. Check Figure 2 for details.

We propose Query-Focused Comparative Explainable Summarization QF-CES to address these

[1]This percentage reflects the average time reduction across 50 distinct summaries, each generated 50 times for reliability. M-OS averaged 9.99 seconds per summary, compared to 16.55seconds for DIA.

[2]All prompts used in this work are available at https://github.com/annnoonnn-uuxxx/QF-CES.

limitations. QF-CES provides targeted, comparative insights for recommended products in one place, as shown in Figure 2, facilitating informed decision-making. It generates a comparative summary in a tabular format, complemented by a Natural Language Explanation (NLE) as a final verdict that directly addresses the user's query.

Problem Statement:

Input: Query and top-k ($k = 3$) recommended products

Output: QF-CES with tabular comparison and final verdict explanation.

Our contributions are:

1. **QF-CES:** A novel task using LLMs to generate Query-Focused Comparative Explainable Summaries. It leverages Multi-Source Opinion Summaries (M-OS) as an intermediate step, reducing inference latency by **40**% (Section 6) compared to raw data input.

2. **MS-Q2P:** A new dataset featuring queries with top-3 recommended products and associated metadata (Section 4.1).

3. **CES-EVAL:** An QF-CES evaluation benchmark dataset (Section 4.2), with 2,500 summary annotations, assessing 10 comparative summaries for 50 queries from the MS-Q2P. The evaluation covers **5 dimensions**- `clarity`, `faithfulness`, `informativeness`, `format adherence`, and `query relevance` (Appendix A).

4. **QF-CES-PROMPT:** A set of dimension-dependent prompts enables comparative summary generation and evaluation of all the aforementioned 5 dimensions. To the best of our knowledge, we are the first to create a structured tabular comparison with a final verdict summary that directly addresses the user's specific query.

5. Benchmarking of 9 recent LLMs (closed and open-source) on the aforementioned 5 dimensions for the task of comparative summaries, which to the best of our knowledge is first of its kind (Table 4, Section 6).

6. Detailed analysis, comparing an open-source LLM against **4** closed-source LLMs as evaluators for automatic evaluation of comparative summaries on 5 dimensions. Analysis indicates that QF-CES-PROMPT emerges as a good alternative for reference-free evaluation of comparative summaries showing a high Spearman correlation of **0.74** on average with humans (Table 3).

2 Related Work

Explainable Recommendation has been an active area of research in recent years, with early contributions from Chen et al. (2018) and Wang et al. (2018). Li et al. (2020b) and Yang et al. (2021) furthered the field, leading to PETER, a personalized transformer for explainable recommendation by Li et al. (2021). Colas et al. (2023) introduced KnowRec, a knowledge-grounded model, and Wang et al. (2023b) enhanced explanations by extracting comparative relation tuples. Gao et al. (2024) aligned LLMs for recommendation explanations, and Peng et al. (2024) leveraged LLMs to generate explanations. (Ni et al. (2019), Tan et al. (2021), Li et al. (2020c)) Generate templatized explanations using item attributes and sentiment from reviews.

Comparative Summarization has received limited attention. Iso et al. (2022) generated contrastive summaries and a common summary from user reviews, Yang et al. (2022) review-based explanations for recommended items, Echterhoff et al. (2023) generated aspect-aware comparative sentences, while Le and Lauw (2021) proposed a framework incorporating comparative constraints into recommendation models.

LLM-based Evaluators Traditional metrics like ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002) often misalign with human judgments for opinion summaries. Recent NLP advancements, particularly in LLMs, offer promising alternatives. Studies have explored LLM-based evaluation methods (Fu et al., 2023; Chiang and Lee, 2023a,c; Wang et al., 2023a; Kocmi and Federmann, 2023), including Chain of Thought approaches (Liu et al., 2023; Wei et al., 2023) and reference-free evaluation (Chiang and Lee, 2023b). Siledar et al. (2024) proposed two prompt strategies for opinion summary evaluation on 7 metrics.

Our work differs from the existing work through (*1*) **Consolidated Comparison** of three products simultaneously; (*2*) **Query-Based Personalization**, preserving privacy; (*3*) **Dynamic Attribute Gen-**

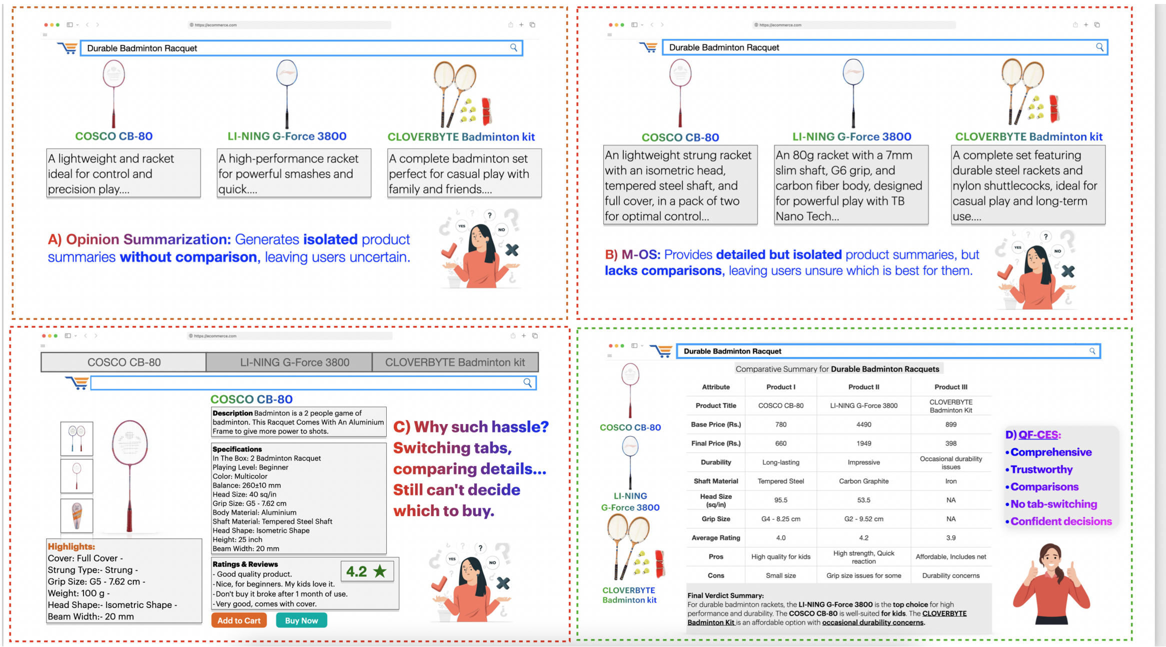

Figure 2: Comparison of approaches: (A) Traditional opinion summaries, (B) M-OS, (C) Single-product views with tab navigation, and (D) textscQF-CES. Unlike traditional methods with isolated summaries, textscQF-CES offers side-by-side comparisons and a final verdict, eliminating tab-switching and enhancing decision-making confidence.

eration tailored to user queries; *(4)* **Category-Agnostic** approach applicable across product domains; *(5)* **Recommendation-Engine Agnostic**, functioning with any ranking system; and *(6)* **Multi-Source Integration**, generating comprehensive summaries beyond user reviews. These features collectively offer a more versatile, privacy-conscious, and informative comparative summarization solution.

3 Methodology

Our study investigates LLMs' capabilities in generating and evaluating comparative summaries, an underexplored area, despite their success in various NLG tasks. We leverage the MS-Q2P dataset (Section 4.1), enabling a thorough assessment of LLMs in comparative summarization.

3.1 Multi-Source Opinion Summary (M-OS)

We developed M-OS using an LLM ensemble, integrating diverse product attributes including product title, descriptions, key features, specifications, reviews, and average ratings for comprehensive representation. This approach establishes a robust foundation for QF-CES generation while reducing inference latency (Section 6). Our prompting-based methodology avoids fine-tuning overhead, offering an efficient solution. We evaluate M-OS quality by adapting the OP-PROMPT framework Siledar et al. (2024) across 7 dimensions (`fluency`, `coherence`, `relevance`, `faithfulness`, `aspect coverage`, `sentiment consistency`, and `specificity`), (Section 5.2), identifying the top-performing LLM (Table 5) for subsequent QF-CES production.

3.2 Query-Focused Comparative Explainable Summary (QF-CES)

QF-CES generation utilizes M-OS that achieved the highest average score across all 7 dimensions. Our approach uses sophisticated, prompt engineering, guiding the LLM through a detailed, step-by-step process. The LLM assigned an expert role, dynamically selects relevant product attributes based on user queries, product specifics, and attribute importance. The resulting QF-CES presents a tabular comparison of top-k ($k = 3$) recommended products, including titles, prices, ratings, selected attributes, and user-derived pros/cons. Missing attributes are marked *"NA"*. A concise explanation provides a final verdict that directly addresses the user's query-specific needs, providing a personalized, query-focused comparison tailored to the user's requirements.

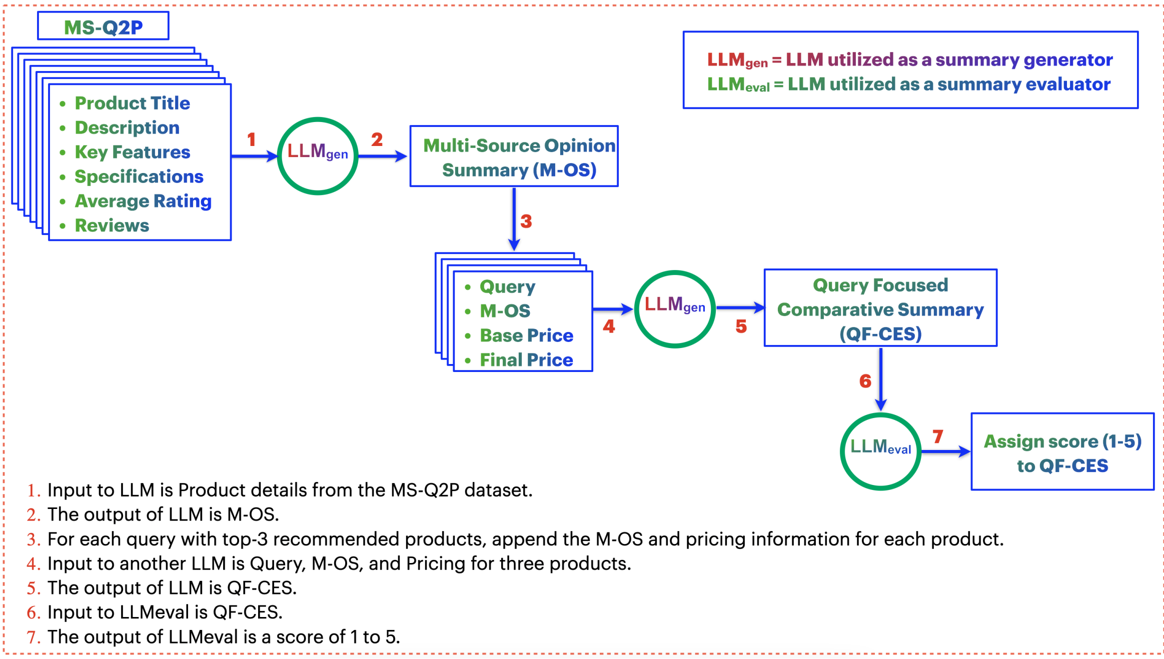

Figure 3: A Multi-phase Pipeline for Generating QF-CES using M-OS and Large Language Models (LLMs). The pipeline involves using LLMs both as summary generators (LLMgen) and summary evaluators (LLMeval) to create and assess QF-CES across various dimensions, incorporating product details from the MS-Q2P dataset. The points 1 through 7 describe the flow of inputs and outputs between the LLMs, from generating M-OS to evaluating QF-CES.

3.3 Evaluation of QF-CES

To the best of our knowledge, we are the first to present an evaluation of comparative summaries using reference-free metrics with both open- and closed-source LLMs as evaluators. Our QF-CES-PROMPT introduces metric-dependent prompts assessing generated QF-CES across 5 dimensions: `clarity`, `faithfulness`, `informativeness`, `format adherence`, and `query relevance`. These dimension-specific prompts are applied to various LLMs for QF-CES evaluation. Additionally, we introduce CES-EVAL, a benchmark dataset (Section 4.2) for QF-CES evaluation.

3.4 Prompt Design Consideration

Our QF-CES-PROMPT design comprises three key components: (*1*) **Generation Prompt:** providing step-by-step instructions for comprehensive, query-relevant tabular comparisons and a final verdict summaries; (*2*) **Evaluation Prompts:** for each dimension, featuring detailed criteria and scoring guidelines $(1-5)$, that require explanations to enhance LLM response quality; and (*3*) **System Message:** defining the LLM's role as a dimension-specific expert. This structured approach ensures high-quality, impartial assessments, improving the quality and relevance of comparative summaries across all dimensions.

3.5 Scoring Function

Liu et al. (2023) proposed a weighted average approach to address discrete LLM scoring limitations. The final score is computed as:

$$o = \sum_{k=1}^{j} p(s_k) \times s_k \tag{1}$$

where s_k are possible scores and $p(s_k)$ their LLM-determined probabilities. $p(s_k)$ is estimated by sampling n outputs ($\mathtt{n} \approx 100$) per input, effectively reducing scoring to a mean calculation. This method aims to enhance scoring nuance and reliability.

3.6 Evaluation Approach

For each query q_i in dataset $\mathcal{D}$, $i \in \{1, ..., \mathcal{Q}\}$, we have $\mathcal{N}$ QF-CES from different models. Let s_{ij} denote the j^{th} QF-CES for query q_i, $\mathcal{M}_m$ denote the m^{th} evaluation metric and $\mathcal{K}$ denote the correlation measure. Bhandari et al. (2020) defines the summary-level correlation as:

$$\mathcal{R}(a,b) = \frac{1}{\mathcal{Q}} \sum_i \mathcal{K}([\mathcal{M}_a(s_{i1}), ..., \mathcal{M}_a(s_{i\mathcal{N}})], [\mathcal{M}_b(s_{i1}), ..., \mathcal{M}_b(s_{i\mathcal{N}})]) \quad (2)$$

Where: $\mathcal{Q}$ is the total number of queries s_{ij} is the QF-CES generated for query q_i by model j $\mathcal{M}_a$ and $\mathcal{M}_b$ are two different evaluation metrics.

4 Dataset

We present two datasets: (*1*) **MS-Q2P**, a novel proprietary dataset. (*2*) **CES-EVAL**, a benchmark dataset for evaluating the QF-CES on 5 dimensions. In this section, we discuss the dataset used, QF-CES evaluation metrics, annotation details, and its analysis.

4.1 MS-Q2P Dataset

MS-Q2P[3] (Multi-Source Query-2-Product) comprises of a user query and the top-k ($k = 3$) recommended products. Each product entry includes diverse attributes: title, description, key features, specifications, reviews, average rating, and pricing details. MS-Q2P consists of products from various domains like electronics, home & kitchen, sports, clothing, shoes & jewelry etc. Detailed statistics of MS-Q2P can be found in Table 1.

Statistic	Value
# of unique queries	7752
Total # of products	23256
Average # of reviews per product	10
Average length of specifications per product (words)	242.6
Average length of reviews per product (words)	17.99
Average length of description per product (words)	105.79
Average length of key features per product (words)	24.64

Table 1: MS-Q2P dataset statistics.

4.2 CES-EVAL Dataset

We developed the CES-EVAL benchmark dataset to evaluate QF-CES across 5 dimensions (detailed in Appendix A). The dataset comprises annotations for 10 model-generated summaries per product for 50 products from MS-Q2P, totaling 7,500 ratings (3 raters × 50 instances × 10 summaries × 5 dimensions). Summaries were evaluated on a 5-point Likert scale by three experienced students (Master's, Pre-Doctoral, Doctoral) with expertise in opinion summarization. This choice of expert raters over crowd workers was based on findings from Gillick and Liu (2010) and Fabbri et al. (2021). We employed a two-round annotation process, with discrepancies of 2 or more points re-evaluated through discussion. Raters (male, aged $24-32$) were given comprehensive guidelines and model identities were concealed to prevent bias. Raters were compensated commensurate with their contributions to the task. Inter-rater correlations are reported in Table 6.

[3]The MS-Q2P, a comprehensive dataset, was provided by an e-commerce company. Details withheld for anonymity during review

4.3 Annotation Analysis

We evaluated the inter-rater agreement for the 3 raters using Krippendorff's alpha coefficient (α) (Krippendorff, 2011). For Round-I, we found the coefficient to be 0.50 indicating *moderate aggrement* ($0.41 \leq \alpha \leq 0.60$). For Round-II, the coefficient increased to 0.80, indicating *substantial agreement* ($0.61 \leq \alpha \leq 0.80$). Table 2 reports the dimension-wise agreement scores for both rounds. Dimension-wise analysis revealed consistently higher agreement for `format adherence`: and `faithfulness`: consistently scoring higher in both rounds, likely due to the clear identification criteria based on format adherence. Post Round-II, `query relevance`: and `informativeness`: show the most disagreement between raters, indicating challenges in consistent assessment.

	Round-I ↑	Round-II ↑
`clarity`	0.55	0.78
`faithfulness`	0.57	0.81
`informativeness`	0.44	0.79
`format adherence`	0.55	0.82
`query relevance`	0.38	0.81
AVG	0.50	0.80

Table 2: Krippendorff's alpha coefficient (α) for Round-I and Round-II on 5 dimensions. As expected, Round-II shows an improvement in (α) scores.

The average frequency of scores given by human raters across 5 dimensions is shown in Figure 4.

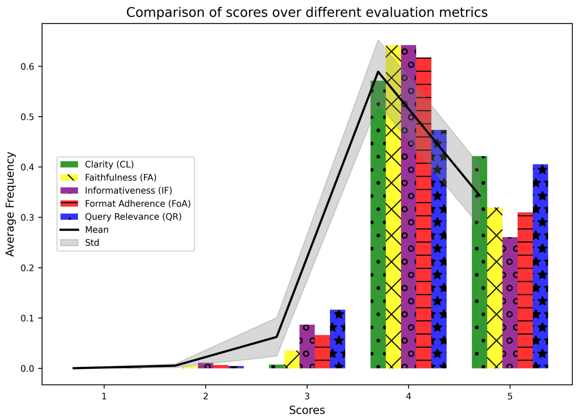

Figure 4: Ratings Distribution. We plot the average frequency of scores obtained by human raters across 5 dimensions. A score of 4 or 5 is mostly preferred.

Query: Best Adidas Sports Shoes for Men Above 5000

Attribute	Product I	Product II	Product III
Product Title	ADIDAS RACER 21 Running Shoes For Men	ADIDAS ADISTAR TD Running Shoes For Men	ADIDAS Faztcort Tennis Running Shoes For Men
Base Price (Rs.)	5999	7999	3999
Final Price (Rs.)	1961	3999	1959
Upper Material	Textile	At least 50% recycled materials	Mesh
Midsole	Cloudfoam	Lightweight cushioning	EVA
Weight (per shoe)	NA	850g	Varies by size
Eco-Friendliness	50% recycled content	At least 50% recycled materials	NA
Average Rating	4.1	4.0	4.0
Pros	Comfortable, stylish, eco-friendly	Comfortable, eco-friendly, high-quality	Comfortable, stylish, high-performing
Cons	Quality issues	May not be as durable as other options	Weight may vary by size

Final Verdict Summary

For men seeking the best Adidas sports shoes above 5000, the ADIDAS ADISTAR TD Running Shoes For Men offer a great balance of comfort, quality, and eco-friendliness. However, if you prioritize style and affordability, the ADIDAS RACER 21 Running Shoes For Men may be a better fit. If you're a tennis player, the ADIDAS Faztcort Tennis Running Shoes For Men provide excellent performance and comfort. Ultimately, consider your specific needs and preferences to make an informed decision.

Figure 5: QF-CES generated by `Qwen2-7B-instruct`

5 Experiments

We present the generation and evaluation of M-OS for QF-CES, using LLMs as baseline metrics, followed by implementation details.

5.1 M-OS Models

We use a custom prompt for the LLMs to generate M-OS. These models were not fine-tuned specifically for multi-source opinion summarization. We use the HuggingFace library (Wolf et al., 2020) to access these 6 open-source models: `LLaMA-3.1.8B-Instruct` (AI@Meta, 2024), `Mistral-7B-Instruct-v0.2` (Jiang et al., 2023), `Mistral-7B-Instruct-v0.3` (Jiang et al., 2023), `Gemma-1.1-7b-it` (Team et al., 2024) , `vicuna-7b-v1.5` (Chiang et al., 2023), `zephyr-7b-beta` (Tunstall et al., 2023)

5.2 M-OS Evaluation

For evaluation of M-OS, we used `LLaMA-3.1.70B-Instruct` model (AI@Meta, 2024) as our evaluator model for these reasons: (*a*) it has have outperformed `GPT-3.5-Turbo` and `GPT-4o` on *IFEval* Benchmark (AI@Meta, 2024) (*b*) it is ranked best amongst the open-source models on the `lmsys/chatbot-arena-leaderboard`, (*c*) we found its instruction following-ness to be better than alternatives, (*d*) its 70B size ensures easy replication compared to `LLaMA-3.1.405B-Instruct`.

5.3 QF-CES Models

Baselines: In our experiments, we adopt a range of recent widely-used LLMs. For close-sourced LLM (accessible through APIs), we evaluate OpenAI's `GPT-4` (OpenAI, 2023). For open-source LLMs, use the HuggingFace library (Wolf et al., 2020) to access these models and experimented with `LLaMA-3.1.70B-Instruct` (AI@Meta, 2024), `LLaMA-3.1.8B-Instruct` (AI@Meta, 2024),`Gemma-1.1-7b-it` (Team et al., 2024), `Gemma-2-9b-it` (Team et al., 2024), `Mistral-7B-Instruct-v0.2` Jiang et al. (2023), `Mistral-7B-Instruct-v0.3` (Jiang et al., 2023), `Qwen/Qwen2-7B-Instruct` (Yang et al., 2024) and `mistralai/Mixtral-8x22B-Instruct-v0.1` (Jiang et al., 2024) as baselines.

5.4 QF-CES Evaluation

For evaluation of QF-CES, we used one proprietary LLM which is OpenAI's `GPT-4o` (OpenAI, 2023) and 4 open source LLMs as evaluators: `LLaMA-3.1.70B-Instruct` (AI@Meta, 2024), `LLaMA-3.1.8B-Instruct` (AI@Meta, 2024), `Mistral-7B-Instruct-v0.2` (Jiang et al., 2023) and `Mistral-7B-Instruct-v0.3` (Jiang et al., 2023). These models were chosen based on their performance on benchmarks, ranking on `lmsys/chatbot-arena-leaderboard`, instruction-following capabilities, replicability, and popularity on Hugging Face.

5.5 Implementation Details

In **summary generation** (M-OS & QF-CES), we configure open-source LLMs with `top_k=25, top_p=0.95, number of beams=3`

	Evaluator LLM	CL ↑		FA ↑		IF ↑		FoA ↑		QR ↑	
		ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ
MS-Q2P	LLaMA-3.1.8B-Instruct	0.60	0.50	0.60	0.43	0.68*	0.54	0.58*	0.39	0.67	**0.59***
	Mistral-7B-Instruct-v0.2	0.67	0.50	0.68*	0.57*	0.68*	0.57*	0.69*	**0.54**	0.67	0.55*
	Mistral-7B-Instruct-v0.3	0.68*	0.50*	0.61	0.46*	0.67	0.48*	0.67	0.50*	0.67	0.55
	LLaMA-3.1.70B-Instruct	0.70*	0.56*	**0.77***	**0.63***	**0.82***	**0.65***	**0.73***	**0.54**	**0.68***	0.46
	GPT-4o	**0.77***	**0.61***	0.75*	**0.63***	0.67	0.57*	0.68*	0.50*	**0.68***	**0.59***

Table 3: *Spearman* (ρ) and *Kendall Tau* (τ) correlations at summary-level on 5 dimensions: clarity (CL), faithfulness (FA), informativeness (IF), format adherence (FoA) and query relevance (QR) for the MS-Q2P dataset. LLAMA-3.1-70B-INSTRUCT demonstrates the highest correlations across most dimensions, indicating strong agreement with human evaluations, followed by GPT-4O. Best performing values are boldfaced, and the second best underlined. $*$ represents significant performance (p-value < 0.05) Refer Figure 8 for graphical representation of model-wise performance across different evaluators.

LLM	CL ↑	FA ↑	IF ↑	FoA ↑	QR ↑	Average ↑
Gemma-1.1-7b-it	4.35	4.39	3.95	3.58	3.79	4.01
LLaMA-3.1.8B-Instruct	4.54	4.25	4.17	4.39	4.24	4.32
Mistral-7B-Instruct-v0.2	4.60	4.24	4.06	4.22	4.45	4.31
Mistral-7B-Instruct-v0.3	4.28	4.28	4.19	4.25	4.43	4.29
Qwen2-7B-instruct	4.47	4.41	**4.58**	4.36	**4.63**	4.49
Gemma-2-9b-it	4.19	4.00	4.19	4.17	4.37	4.18
Mixtral-8x7B-Instruct-v0.1	4.35	4.29	4.17	4.35	4.43	4.32
LLaMA-3.1.70B-Instruct	4.55	4.36	4.47	4.41	4.03	4.36
GPT-4	**4.81**	**4.53**	4.50	**4.61**	4.32	**4.55**
Qwen2-7B-instruct-DIA	4.30	4.33	3.81	4.18	3.89	4.10

Table 4: Model-wise averaged annotator ratings of QF-CES along 5 dimensions: clarity (CL), faithfulness (FA), informativeness (IF), format adherence (FoA) and query relevance (QR) Best scores are in **bold**, second-best are underlined. Qwen2-7B-instruct-DIA represents inputting raw data directly to LLM. Refer Figure 9 for graphical representation.

Model	FL ↑	CO ↑	AC ↑	FF ↑	RL ↑	SC ↑	SP ↑
Mistral-7B-Instruct-v0.3	4.73	4.64	4.09	4.08	4.05	4.24	4.06
LLaMA-3.1.8B-Instruct	4.90	4.63	3.94	4.05	4.01	4.16	3.63
Mistral-7B-Instruct-v0.2	4.70	4.52	3.94	4.05	4.01	4.13	3.99
Gemma-1.1-7b-it	4.30	4.35	3.82	4.04	3.99	3.97	3.25
vicuna-7b-v1.5	3.89	3.76	3.51	3.92	3.62	3.50	3.06
zephyr-7b-beta	4.79	4.56	3.86	4.14	4.06	4.09	3.79

Table 5: M-OS Model-wise performance across 7 dimensions: fluency (FL), coherence (CO), aspect coverage (AC), faithfulness (FF), relevance (RL), sentiment consistency (SC), and specificity (SP), evaluated by LLaMA-3.1.70B-Instruct over $n = 100$ generations. Refer Figure 7 for graphical representation.

and temperature=0.2 to produce deterministic outputs capturing product technicalities. For OpenAI's GPT-4 (OpenAI, 2023), we set decoding temperature=0 for increased determinism.

In **summary evaluation**, open-source LLMs use n=100, temperature=0.2 to account for stochasticity while maintaining consistency. GPT-4o (OpenAI, 2023) again uses decoding temperature=0.

All experiments run on 8 NVIDIA A100-SXM4-80GB clusters, ensuring robust computational capacity for our analyses.

6 Results and Analysis

Our obtained results are provided in Tables 3, 4 and 5.

M-OS Results: Table 5 presents averaged scores assigned by LLaMA-3.1.70B-Instruct (AI@Meta, 2024) across 7 dimensions for each model evaluated for M-OS generation, with Figure 7 providing a graphical representation of these results. Mistral-7B-Instruct-v0.3 (Jiang et al., 2023) and was selected to generate M-OS for use in QF-CES.

QF-CES Results: Table 4 (Refer to Figure 5 for a graphical view) presents averaged annotator ratings across 5 dimensions for each evaluated model, with Figure 9 providing a graphical representation. GPT-4o (OpenAI, 2023) excelled, leading in 4 dimensions and competing strongly in query relevance. Among open-source models, Qwen2-7B-instruct (Yang et al., 2024) achieved the highest average score, particularly in informativeness and query relevance. While larger models like GPT-4o and LLaMA-3.1.70B-Instruct (AI@Meta, 2024) generally performed better, smaller models like Qwen2-7B-instruct (Yang et al., 2024) showed competitive results in specific areas, highlighting the importance of task-specific model selection. Qwen2-7B-instruct's (Yang et al., 2024) performance using a DIRECT INPUT APPROACH was inferior, especially in informativeness and query relevance, likely due to data overload when generating QF-CES for top-3 products. These findings validate our M-OS approach as an effective intermediate step for high-quality QF-CES generation.

QF-CES Evaluation Results: Table 3 report the summary-level correlation scores on the CES-EVAL dataset, for 4 open-source models and one closed-source model. Overall, `LLaMA-3.1.70B-Instruct` ranks as the best evaluator achieving **0.74** average spearman correlation, outperforming OpenAI's `GPT-4o` (OpenAI, 2023) across all dimensions. Figure 8 provides a visual representation of the scores given by different LLMs acting as evaluators for `n=100` generations for the 5 dimensions of QF-CES across various models. This multi-model evaluation approach offers a comprehensive view of model performance, reinforcing the reliability of our findings.

Time Efficiency Results: To account for the stochasticity of LLMs, we generated 50 iterations of each QF-CES (`n=50`), which also helps mitigate instances where an LLM might loop until reaching the maximum token length, potentially skewing time measurements. We recorded generation times for both M-OS and DIA methods, excluding M-OS generation time to reflect real-world pre-computation scenarios. Analysis of average generation time per query (Figure 6) shows M-OS significantly accelerates inference by 40% compared to DIA. Queries 24 and 27, involving complex electronics specifications, demonstrated M-OS's efficiency in condensing large data volumes, outpacing DIA's raw data processing.

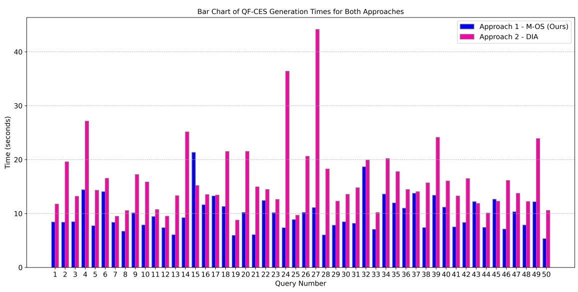

Figure 6: Comparison of inference times for QF-CES generation using M-OS and DIA approach. Each data point represents the average of 50 generations per query.

7 Conclusion & Future Work

This paper introduces Query-Focused Comparative Explainable Summarization QF-CES, a novel task that addresses the limitations of traditional opinion summarization in e-commerce recommendation systems. By leveraging LLMs and Multi-Source Opinion Summarization M-OS, we present a comprehensive approach to generate query-specific, comparative summaries of recommended products. The framework, validated through the MS-Q2P dataset and extensive evaluations, showed `GPT-4` superior performance, with `Qwen2-7B-instruct` as a strong open-source contender for QF-CES. The evaluation using QF-CES-PROMPT with `LLaMA-3.1.70B-Instruct` yielded an average *Spearman correlation* of **0.74** with human judgments, highlighting its reliability. Future work will focus on improving LLM performance on complex products with extensive specifications, reducing inference latency while maintaining high summary quality, refining prompting strategies to better select relevant attributes based on user queries, and exploring the applicability of QF-CES-PROMPT to other domains beyond e-commerce to assess its generalizability and necessary adaptations.

Limitations

1. We evaluate using `GPT-4o` for QF-CES, omitting `GPT-4` due to cost constraints. Our primary goal was to design prompts applicable to open-source and closed-source models for generating and evaluating M-OS and QF-CES.

2. Our QF-CES-PROMPT targets a specific dimension of comparative summarization. Its broader applicability requires further study and potential prompt adjustments.

3. The M2-Q2P dataset's limitation of only 10 reviews highlights a broader issue in opinion summarization. Future benchmarks should incorporate datasets with more reviews. ‘

4. During QF-CES generation, LLMs occasionally struggled with products having extensive specifications, resulting in incomplete or stalled summaries.

5. QF-CES with M-OS consistently outperformed DIA in inference latency across 50 queries. However, a larger query set is necessary to solidify and generalize these findings.

Ethical Considerations

We engaged 3 raters with diverse academic backgrounds: a Master's student, a Pre-Doctoral researcher, and a Doctoral candidate. All were male, aged 24-32, with publications or active research in opinion summarization. Raters were compensated appropriately for their contributions.

QF-CES-PROMPTS generate and evaluate QF-CES across 5 dimensions, aiding the assessment of NLG-produced comparative summaries. While insightful, these prompts may occasionally produce hallucinations, especially in complex cases. We urge judicious use and validation of reliability for specific applications. Researchers should verify prompt appropriateness before integration into evaluation processes, ensuring careful application in real-world scenarios.

References

AI@Meta. 2024. Llama 3 model card.

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9347–9359, Online. Association for Computational Linguistics.

Chong Chen, Min Zhang, Yiqun Liu, and Shaoping Ma. 2018. Neural attentional rating regression with review-level explanations. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, page 1583–1592, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

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

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

Cheng-Han Chiang and Hung-yi Lee. 2023c. A closer look into using large language models for automatic evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8928–8942, Singapore. Association for Computational Linguistics.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Anthony Colas, Jun Araki, Zhengyu Zhou, Bingqing Wang, and Zhe Feng. 2023. Knowledge-grounded natural language recommendation explanation. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 1–15, Singapore. Association for Computational Linguistics.

Jessica Maria Echterhoff, An Yan, and Julian McAuley. 2023. Comparing apples to apples: Generating aspect-aware comparative sentences from user reviews. *ArXiv*, abs/2307.03691.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire.

Shen Gao, Yifan Wang, Jiabao Fang, Lisi Chen, Peng Han, and Shuo Shang. 2024. Dre: Generating recommendation explanations by aligning large language models at data-level.

Dan Gillick and Yang Liu. 2010. Non-expert evaluation of summarization systems is risky. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 148–151, Los Angeles. Association for Computational Linguistics.

Jinbae Im, Moonki Kim, Hoyeop Lee, Hyunsouk Cho, and Sehee Chung. 2021. Self-supervised multimodal opinion summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 388–403, Online. Association for 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)*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris

Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Trung-Hoang Le and Hady W. Lauw. 2021. Explainable recommendation with comparative constraints on product aspects. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, WSDM '21, page 967–975, New York, NY, USA. Association for Computing Machinery.

Haoran Li, Peng Yuan, Song Xu, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020a. Aspect-aware multimodal summarization for chinese e-commerce products. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8188–8195.

Lei Li, Yongfeng Zhang, and Li Chen. 2020b. Generate neural template explanations for recommendation. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 755–764, New York, NY, USA. Association for Computing Machinery.

Lei Li, Yongfeng Zhang, and Li Chen. 2020c. Generate neural template explanations for recommendation. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 755–764, New York, NY, USA. Association for Computing Machinery.

Lei Li, Yongfeng Zhang, and Li Chen. 2021. Personalized transformer for explainable recommendation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4947–4957, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. 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 188–197, Hong Kong, China. Association for Computational Linguistics.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Yicui Peng, Hao Chen, Chingsheng Lin, Guo Huang, Jinrong Hu, Hui Guo, Bin Kong, Shu Hu, Xi Wu, and Xin Wang. 2024. Uncertainty-aware explainable recommendation with large language models.

Tejpalsingh Siledar, Jigar Makwana, and Pushpak Bhattacharyya. 2023. Aspect-sentiment-based opinion summarization using multiple information sources. In *Proceedings of the 6th Joint International Conference on Data Science & Management of Data (10th ACM IKDD CODS and 28th COMAD)*, CODS-COMAD '23, page 55–61, New York, NY, USA. Association for Computing Machinery.

Tejpalsingh Siledar, Swaroop Nath, Sankara Sri Raghava Ravindra Muddu, Rupasai Rangaraju, Swaprava Nath, Pushpak Bhattacharyya, Suman Banerjee, Amey Patil, Sudhanshu Shekhar Singh, Muthusamy Chelliah, and Nikesh Garera. 2024. One prompt to rule them all: Llms for opinion summary evaluation.

Juntao Tan, Shuyuan Xu, Yingqiang Ge, Yunqi Li, Xu Chen, and Yongfeng Zhang. 2021. Counterfactual explainable recommendation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 1784–1793, New York, NY, USA. Association for Computing Machinery.

Wenyi Tay. 2019. Not all reviews are equal: Towards addressing reviewer biases for opinion summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 34–42, Florence, Italy. Association for Computational Linguistics.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam

Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of lm alignment.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023a. Is ChatGPT a good NLG evaluator? a preliminary study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11, Singapore. Association for Computational Linguistics.

Nan Wang, Hongning Wang, Yiling Jia, and Yue Yin. 2018. Explainable recommendation via multi-task learning in opinionated text data. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, page 165–174, New York, NY, USA. Association for Computing Machinery.

Yequan Wang, Hengran Zhang, Aixin Sun, and Xuying Meng. 2023b. Gcre-gpt: A generative model for comparative relation extraction. *ArXiv*, abs/2303.08601.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report.

Aobo Yang, Nan Wang, Renqin Cai, Hongbo Deng, and Hongning Wang. 2022. Comparative explanations of recommendations. In *Proceedings of the ACM Web Conference 2022*. ACM.

Aobo Yang, Nan Wang, Hongbo Deng, and Hongning Wang. 2021. Explanation as a defense of recommendation. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, WSDM '21, page 1029–1037, New York, NY, USA. Association for Computing Machinery.

A QF-CES Dimensions

We define QF-CES evaluation dimensions as follows:

1. `clarity` **(CL)**- Clarity measures the degree to which the information in the Comparative Summary is clearly presented, avoiding ambiguity and ensuring that comparisons are easy to understand. The summary should be clear, concise, and easy to comprehend, using simple language and avoiding technical jargon whenever possible. It should be well-structured and well-organized, presenting comparison of the three products in a straightforward manner. The metric evaluates

the readability of the entire summary, ensuring it is free from grammatical errors and has a logical flow between different sections and points. Additionally, the clarity of the tabular data is assessed to ensure it clearly conveys the comparisons between three products.

2. `faithfulness` **(FL)**- Faithfulness measures the degree to which the information presented in the Comparative Summary is accurate, verifiable, and directly supported by the input data. The Comparative Summary must faithfully represent the content provided, ensuring that all details, including the query and attributes of each product are correct and inferred directly from the input. Comparative Summary will be penalized for any information that cannot be verified from the input data or if they make broad generalizations that are not supported by the input data.

3. `informativeness` **(IF)**- Informativeness evaluates the extent to which the Comparative Summary comprehensively covers all relevant aspects and attributes of the products being compared. This metric assesses the presence and completeness of essential attributes and features in the comparison, including the product title, base price, final price, key attributes dynamically selected from the product opinion summaries, pros, cons, and average rating. The summary should ensure that all majorly discussed aspects are covered and any missing values are properly marked as "N/A". Summaries should be penalized for missing significant aspects and rewarded for thorough coverage of the aspects from the provided information.

4. `format adherence` **(FoA)**- This metric evaluates the extent to which the Comparative Summary follows the prescribed format. The Comparative Summary should consist of two main parts: (*1*) A tabular comparison of the three products. (*2*) A final verdict summary.

   The tabular comparison should list products in columns and attributes in rows,including dynamically selected attributes based on the user query and essential attributes such as Base Price, Final Price, Average Rating, Pros, and Cons. It verifies that dynamically selected attributes are appropriately named and not using placeholders. The final verdict summary should provide a concise overview of the comparison among three products. The metric assesses the presence, completeness, and proper formatting of both these components (the tabular comparison along with the final verdict), as well as the overall organization and consistency of the entire summary.

5. `query relevance` **(QR)**- This metric evaluates how well the Comparative Summary addresses the user's query. It assesses two main components: (*1*) *The tabular comparison:* Ensures that only the most relevant information and dynamic attributes are present, directly addressing the user query without including irrelevant details. (*2*) *The final verdict summary:* Verifies that the user query is explicitly addressed, providing clear suggestions that enable the user to make an informed buying decision.

   The metric measures the overall relevance and usefulness of the Comparative Summary in helping the user make an informed decision based on their specific query.

B M-OS Evaluation

Figure 7 represents the model-wise performance across 7 dimensions: `fluency` (FL), `coherence` (CO), `aspect coverage` (AC), `faithfulness` (FF), `relevance` (RL), `sentiment consistency` (SC), and `specificity` (SP). The scores are given by `LLaMA-3.1.70B-Instruct` as evaluator, for `n=100` generations.

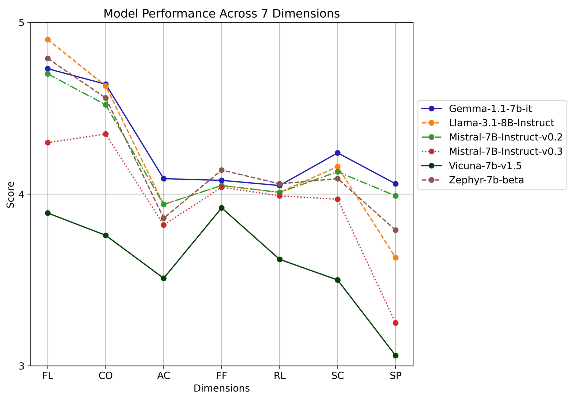

Figure 7: Various open-source models performance across 7 dimensions by `LLaMA-3.1.70B-Instruct` as evaluator

C QF-CES Evaluation

Figure 8 represents model-wise averaged score given by various LLM as evaluators of QF-CES along 5 dimensions: `clarity` (CL), `faithfulness` (FA), `informativeness` (IF), `format adherence` (FoA) and `query relevance` (QR)

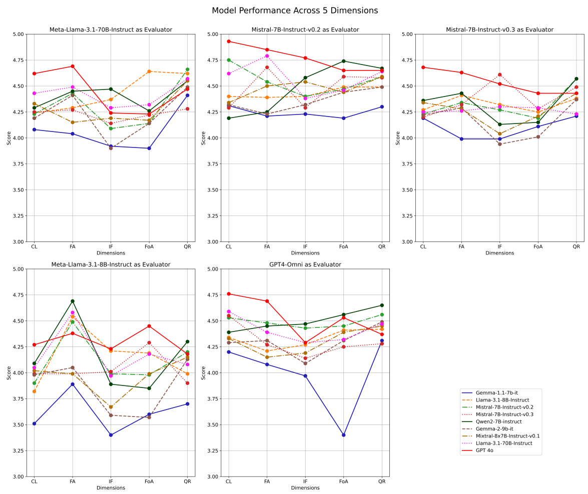

Figure 8: Performance of different models as rated by human annotators (Round-II). We observe that `GPT-4` performs the best followed by `Qwen2-7B-instruct`.

D LLM as Evaluators

Figure 9 represents model-wise averaged annotator ratings of QF-CES along 5 dimensions: `clarity` (CL), `faithfulness` (FA), `informativeness` (IF), `format adherence` (FoA) and `query relevance` (QR)

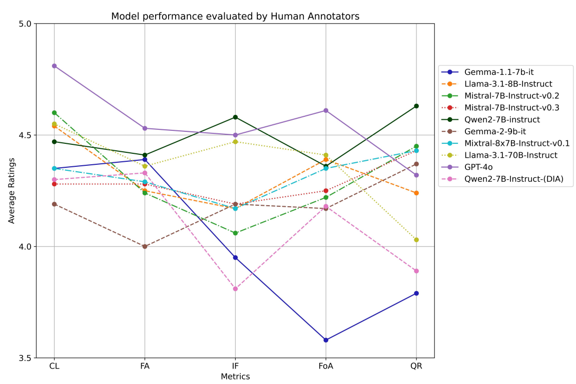

Figure 9: Performance of different models as rated by human annotators (Round-II). We observe that `GPT-4` performs the best followed by `Qwen2-7B-instruct`.

E Rater Agreement

Table 6 reports the pairwise correlations between raters as well as the correlation between each rater and average ratings for both Round-I and Round-II.

		CL ↑		FA ↑		IF ↑		FoA ↑		QR ↑	
		ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ
Round-I	*Pairwise correlation among raters*										
	A1-A2	0.59	0.59	0.59	0.57	0.49	0.47	0.53	0.51	0.39	0.37
	A2-A3	0.58	0.57	0.58	0.57	0.33	0.31	0.46	0.45	0.32	0.29
	A1-A3	0.60	0.60	0.57	0.56	0.38	0.36	0.58	0.56	0.47	0.44
	AVG-I	0.59	0.59	0.58	0.57	0.40	0.38	0.52	0.51	0.39	0.37
	Pairwise correlation between raters and the overall average ratings										
	A-A1	0.82	0.78	0.86	0.80	0.74	0.68	0.83	0.77	0.79	0.72
	A-A2	0.84	0.79	0.79	0.74	0.78	0.72	0.79	0.73	0.72	0.64
	A-A3	0.84	0.80	0.83	0.78	0.73	0.67	0.81	0.75	0.78	0.71
	AVG-II	0.83	0.79	0.83	0.77	0.75	0.69	0.81	0.75	0.76	0.69
Round-II	*Pairwise correlation among raters*										
	A1-A2	0.80	0.80	0.80	0.79	0.77	0.76	0.81	0.80	0.85	0.83
	A2-A3	0.78	0.78	0.79	0.79	0.77	0.75	0.78	0.77	0.72	0.70
	A1-A3	0.78	0.78	0.82	0.81	0.75	0.73	0.81	0.80	0.70	0.68
	AVG-I	0.79	0.78	0.80	0.80	0.76	0.75	0.80	0.79	0.75	0.74
	Pairwise correlation between raters and the overall average ratings										
	A-A1	0.91	0.87	0.92	0.88	0.86	0.82	0.92	0.88	0.90	0.85
	A-A2	0.92	0.87	0.91	0.87	0.88	0.84	0.92	0.88	0.92	0.87
	A-A3	0.92	0.87	0.94	0.89	0.91	0.87	0.92	0.88	0.88	0.83
	AVG-II	0.92	0.87	0.92	0.88	0.89	0.85	0.92	0.88	0.90	0.85

Table 6: Rater Correlations: Pairwise *Spearman* (ρ) and *Kendall Tau* (τ) correlations at summary-level for 3 raters A1, A2, and A3 along with the average of their ratings.