LLMs as Architects and Critics for Multi-Source Opinion Summarization

Anuj Attri[◇], **Arnav Attri**[◇], **Pushpak Bhattacharyya**
Suman Banerjee, **Amey Patil**, **Muthusamy Chelliah**, **Nikesh Garera**
Computer Science and Engineering, IIT Bombay, India, Flipkart, India
{ianuj, arnavcs, pb}@cse.iitb.ac.in

Abstract

MULTI-SOURCE OPINION SUMMARIZATION (M-OS) extends beyond traditional opinion summarization by incorporating additional sources of product metadata such as `descriptions, key features, specifications, and ratings, alongside reviews`. This integration results in comprehensive summaries that capture both subjective opinions and objective product attributes essential for informed decision-making. While Large Language Models (LLMs) have shown significant success in various Natural Language Processing (NLP) tasks, their potential in M-OS remains largely unexplored. Additionally, the lack of evaluation datasets for this task has impeded further advancements. To bridge this gap, we introduce M-OS-EVAL, a benchmark dataset for evaluating multi-source opinion summaries across 7 key dimensions: `fluency`, `coherence`, `relevance`, `faithfulness`, `aspect coverage`, `sentiment consistency`, `specificity`. Our results demonstrate that M-OS significantly enhances user engagement, as evidenced by a user study in which, on average, **87%** of participants preferred M-OS over opinion summaries. Our experiments demonstrate that factually enriched summaries enhance user engagement. Notably, M-OS-PROMPTS exhibit stronger alignment with human judgment, achieving an average Spearman correlation of $\rho = \mathbf{0.74}$, which surpasses the performance of previous methodologies.

Opinion Summary
The racquet is highly praised for its lightweight design and powerful performance, particularly in smashes, making it ideal for both offensive and defensive play. Many users recommend it for intermediate to advanced players, especially those on a budget. The head-heavy design boosts power while ensuring easy movement, with durable military-grade carbon construction. Some users note minor drawbacks, such as a smaller grip size, which can easily be adjusted with external grips. However, one customer reported receiving a damaged racquet with no replacement policy. The racquet offers excellent performance, durability, and a lightweight design.
Multi-Source Opinion Summary
The Li-Ning G-Force Racquet is designed for advanced players prioritizing speed and precision. Crafted from **Japanese Ultra-Carbon Graphite**, it ensures exceptional strength and rigidity, preventing frame deformation. The **UHB Shaft** designed using player data, optimizes smash performance, while the **oval 53.5 sq/in head size and G2 grip (9.52 cm)** provide control and maneuverability. The **Aero Tec Beam System** reduces air resistance while preserving frame integrity, and the **Dynamic Optimum Frame design** boosts bounce strength for powerful smashes. **Weighing only 79g** with a head-heavy balance, it excels in both offensive and defensive play. At **26.7 inches tall** with a **7 mm beam width**, it delivers top-tier performance. With an **average rating of 4.2**, users praise its durability, though some recommend external grips. Despite a rare issue with a broken racquet, it remains a top choice for players seeking precision and power.

Table 1: Opinion Summary vs. Multi-Source Opinion Summary (M-OS) by `GPT-4o` for a Li-Ning Badminton Racquet. While the opinion summary from reviews alone provides subjective feedback, the M-OS integrates product metadata with reviews to deliver a comprehensive overview. This eliminates manual metadata parsing while maintaining balanced product coverage. **Boldface** indicates technical specifications absent in the opinion summary.

1 Introduction

Customer reviews, while crucial for e-commerce decisions, present challenges due to their overwhelming volume. Traditional opinion summarization approaches (Wang and Ling, 2016; Chu and Liu, 2019; Bražinskas et al., 2020a) generate concise summaries but rely solely on user reviews, missing valuable product information from other sources.

We introduce Multi-Source Opinion Summarization (M-OS), which integrates reviews with product descriptions, specifications, and ratings to create comprehensive summaries. M-OS combines subjective user experiences with objective product attributes to facilitate informed decision-making. As demonstrated in Table 1, M-OS enriches summaries by incorporating technical specifications and product descriptions, enabling precise product comparisons - a key advantage over review-only approaches that often lack detailed attributes. M-OS addresses decision fatigue and information overload by synthesizing diverse product data to deliver

[◇] Equal contribution

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 69–101
December 20-24, 2025 ©2025 Association for Computational Linguistics

comprehensive, relevant summaries. This streamlined approach reduces cognitive load and enhances user satisfaction by providing actionable insights without requiring manual metadata parsing.

LLMs have emerged as effective reference-free evaluators for NLG tasks (Fu et al., 2023a; Chiang and Lee, 2023a,c; Wang et al., 2023; Kocmi and Federmann, 2023), addressing the limitations of traditional metrics like ROUGE (Lin, 2004a) and BERTSCORE (Zhang et al., 2019) which correlate poorly with human judgments (Shen and Wan, 2023). Given the high costs of reference datasets and the inadequacy of conventional metrics for multi-source opinion summaries, LLM-based evaluation offers a scalable solution. We present M-OS-EVAL, a reference-free evaluation dataset for multi-source opinion summarization that assesses summaries across 7 dimensions through two frameworks: OMNI-PROMPT, a dimension-independent prompt, and SPECTRA-PROMPTS, a dimension-dependent prompt set.

To address this need, we present M-OS-EVAL, a benchmark dataset for evaluating M-OS across 7 key dimensions: `fluency`, `coherence`, `relevance`, `faithfulness`, `aspect coverage`, `sentiment consistency`, `specificity`. We propose two novel evaluation frameworks: OMNI-PROMPT, which enables metric-independent assessment, and SPECTRA-PROMPTS, which facilitates metric-dependent evaluation across all 7 dimensions. Our work represents the first prompt-based evaluation method for M-OS, incorporating both closed-source and open-source models to advance LLM-based evaluation in this domain.

- **M-OS:** MULTI-SOURCE OPINION SUMMARIZATION (OR SUMMARY).
- **M-OS-GEN:** MULTI-SOURCE OPINION SUMMARY GENERATION.
- **M-OS-EVAL:** MULTI-SOURCE OPINION SUMMARY EVALUATION.

Our contributions are:

1. **M-OS:** We advance multi-source opinion summarization by using LLMs to generate comprehensive summaries that integrate product metadata (title, description, key features, specifications, rating) with customer reviews. Unlike **MEDOS** (Siledar et al., 2024b), M-OS dynamically synthesizes unified summaries that present essential product information upfront, eliminating the need for users to parse metadata separately. Our user study shows that 87% of participants on average found multi-source summaries more comprehensive than opinion summaries (Section 7)
2. **M-OS-DATA:** A novel dataset of $25,000$ unique products across diverse categories, each with comprehensive metadata, enabling robust training and for multi-source opinion summarization and query-focused tasks (Section 4.1).
3. **M-OS-Eval**: A comprehensive evaluation benchmark comprising $4,900$ summary annotations across 7 key dimensions: `fluency`, `coherence`, `relevance`, `faithfulness`, `aspect coverage`, `sentiment consistency`, `specificity`, for thorough assessment of multi-source opinion summaries (Section 4.2).
4. **M-OS-PROMPTS**[1]**:** We introduce **OMNI-PROMPT** for metric-independent assessment and **SPECTRA-PROMPTS** for metric-dependent evaluation across aforementioned 7 dimensions. This represents the first prompt-based method for assessing multi-source opinion summarization and evaluating various open-source LLMs in this domain (Section 3).
5. Benchmarking of 14 recent LLMS (closed and open-source) on the aforementioned 7 dimensions for the task of multi-source opinion summarization, which to the best of our knowledge is first of its kind (Table 5, Section 6).
6. We compare four open-source LLMs against a closed-source (GPT-4O) LLM for automatic M-OS evaluation across 7 dimensions. Our analysis reveals M-OS-PROMPTS as an effective alternative, achieving strong alignment with human assessment (average Spearman correlation: 0.74) (Table 7, Section 6).

2 Related Work

Opinion summarization has evolved from extractive methods (Erkan and Radev, 2004; Kim et al.,

[1] https://github.com/yourarnav/M-OS

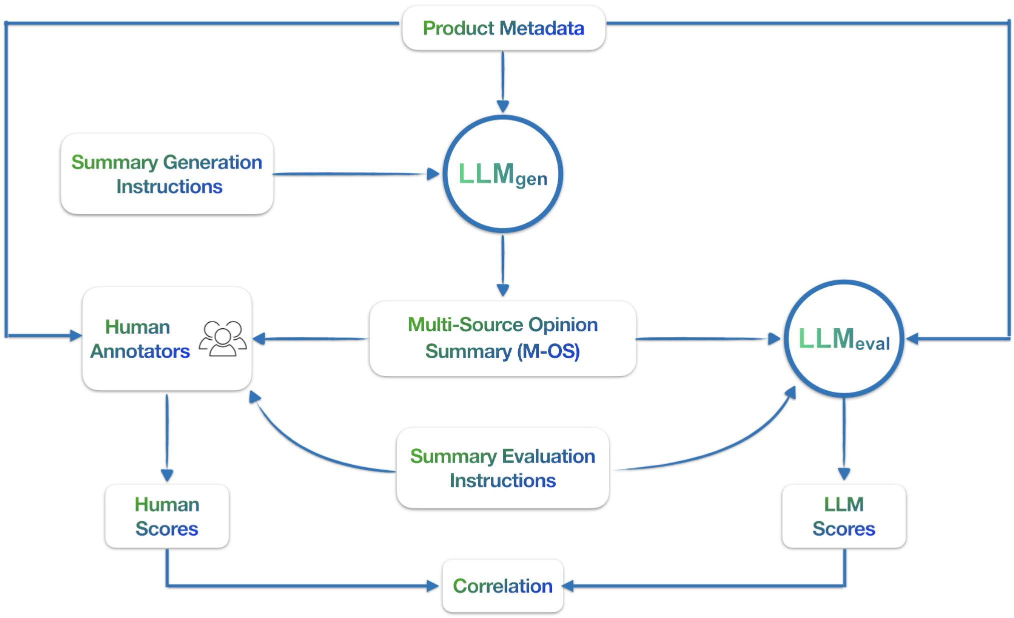

Figure 1: Pipeline for our M-OS study. The generation model (LLM_{gen}) generates M-OS using product metadata, guided by the (M-OS-GEN-PROMPT). Summaries are then assessed in parallel by human annotators and an evaluation model (LLM_{eval}) following the (M-OS-EVAL-PROMPT). Finally, we compute the `correlation` between `human and LLM` scores.

2011) to neural-based approaches (Bražinskas et al., 2020a; Amplayo and Lapata, 2020a), with various specialized directions emerging. For aspect-specific summarization, Angelidis et al. (2021) employed VQ-VAE (van den Oord et al., 2017), while Amplayo et al. (2021) introduced abstractive approaches using MIL. Self-supervised methods were advanced by Bražinskas et al. (2020b) using pseudo-summary pairs, enhanced by Amplayo and Lapata (2020b) with noisy variations and Elsahar et al. (2021) with TF-IDF similarity-based selection. Large-scale processing was addressed by Bhaskar et al. (2023) using GPT-3.5 (OpenAI, 2023) prompting, Jiang et al. (2023b) with sentiment-aware sampling, and Muddu et al. (2024) through XL-OPSUMM. Multi-source approaches emerged with (Zhao and Chaturvedi, 2020) utilizing product descriptions, (Li et al., 2020) developing supervised multimodal methods, and Siledar et al. (2024b) introducing a structured approach with reviews, descriptions, and Q&A pairs. While these methods advanced the field, they typically overlook comprehensive product metadata. Our work differs by leveraging LLMs' extended context lengths to incorporate complete product specifications, generating comprehensive summaries that eliminate the need for manual navigation through product information. We extend beyond Siledar et al. (2024b) by including detailed specifications and descriptions, providing users with complete product insights in a single, unified summary.

LLM-based Evaluators Traditional metrics like ROUGE (Lin, 2004a), BLEU (Papineni et al., 2002) and BERTSCORE (Zhang et al., 2019) correlate poorly with human judgments (Shen and Wan, 2023). LLM-based evaluation provides a cost-effective solution for large-scale reference-based datasets, including Chain of Thought approaches (Liu et al., 2023a; Wei et al., 2023), reference-free evaluation (Chiang and Lee, 2023b), and other methods (Fu et al., 2023a; Chiang and Lee, 2023a,c; Wang et al., 2023; Kocmi and Federmann, 2023). Siledar et al. (2024a) proposed two prompt strategies for opinion summarization. We leverage LLMs as evaluators for reference-free M-OS assessment.

3 Methodology

Our methodology is centered around the development of M-OS-PROMPTS, which facilitate both the *generation* and *evaluation* of M-OS, as *illustrated* in (**Figure 1**).

Dimension	M-OS-Prompts	Baseline-Prompts
Evaluation Process	Enforces **structured, step-by-step evaluation** (e.g., list aspects → count errors → calculate percentages).	Relies on **generic, holistic instructions** (e.g., "check for fluency issues").
Error Handling	Requires **systematic error identification and severity classification** through structured evaluation steps.	Uses **subjective judgments** (e.g., "significant errors").
Scoring Criteria	Anchored to **quantitative thresholds** (e.g., percentage ranges defined in the prompt e.g., 0–20%: Score 1; 21–50%: Score 2).	Depends on **qualitative labels** (e.g., "covers most aspects").
Reasoning Depth	Mimics **human reasoning** via detailed prompts (800+ words) with explicit logic chains guiding step-by-step analysis.	Uses **shorter, high-level prompts** (∼400 words) lacking step-by-step guidance.
Role Assignment	Assigns explicit expert evaluator role via system message to enhance model reasoning capabilities	No explicit role assignment in prompt structure

Table 2: Comparison of our M-OS-Eval prompts with baseline prompts. Our prompts introduce structured, step-by-step evaluation processes with quantitative scoring criteria, specifically adapted for the multi-source context where both objective specifications and subjective reviews must be verified.

3.1 M-OS-Gen-Prompt (Summary Generation Prompt)

The M-OS-Gen-Prompt guides LLMs to create summaries by synthesizing information from various product attributes, including the product title, description, key features, specifications, customer reviews, and average ratings. By integrating these diverse data sources, the prompt ensures that summaries provide a comprehensive perspective, balancing subjective customer opinions with objective product details. See **Appendices H and C.1** for the M-OS-Prompt and its design principles, respectively.

3.2 M-OS-Eval-Prompts (Summary Evaluation Prompts)

The M-OS-Eval-Prompts guide evaluation of M-OS, structured to assess 7 dimensions: `fluency`, `coherence`, `relevance`, `faithfulness`, `aspect coverage`, `sentiment consistency` and `specificity` (Check **Appendix A** *for metric definition*). Each prompt has four core components for thorough evaluation across open-source and closed-source LLMs:

(*1*) **`System Message`**: Defines the LLM's role as a specialized evaluator, providing clear context for the evaluation task.

(*2*) **`Task Description`**: Outlines the specific evaluation task, which involves assessing a multi-source opinion summary against product metadata (including title, description, key features, specifications, reviews, and average rating).

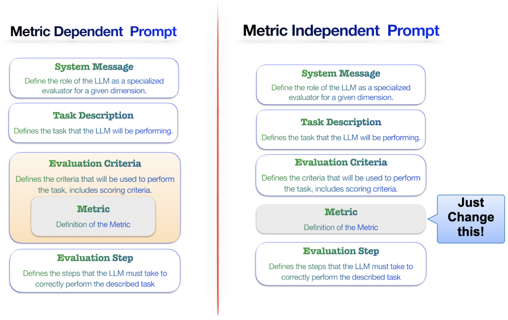

Figure 2: Two prompting strategies. **Left:** Spectra-Prompts (metric-dependent) uses dimension-specific prompts tailored for each evaluation criterion. **Right:** Omni-Prompt (metric-independent) employs a modular architecture where only the metric component changes while maintaining consistent structure across all dimensions. Both approaches build on (Siledar et al., 2024a) with enhanced structured reasoning.

(*3*) **`Evaluation Criteria`**: Defines the criteria for the task. For multi-source opinion summary evaluation, the LLM assigns a score $(1 - 5)$ based on how well the summary adheres to a specific metric or dimension.

(*4*) **`Evaluation Step`**: Provides the LLM with a detailed, step-by-step guide to complete the evaluation, ensuring consistency and thoroughness.

For M-OS-Eval-Prompts design principles, refer to **Appendix C.2**.

M-OS-Eval Prompting Approaches Building on Siledar et al. (2024a), we develop two M-OS evaluation approaches as *illustrated* in (**Figure 2**). Table 2 presents a detailed comparison of our M-

OS-EVAL prompts with the baseline prompts, highlighting key enhancements in evaluation structure, scoring criteria, and reasoning depth. For detailed graphical visualization, refer to **Appendix G**.

OMNI-PROMPT (universal prompt for comprehensive cross-dimensional evaluation) represents our metric-independent evaluation approach with a modular architecture. While maintaining a consistent framework of Task Description, Evaluation Criteria, and Evaluation Steps, it introduces a flexible 'Metric' component for dynamic modification. This enables universal applicability—the same structure evaluates any dimension by redefining the 'Metric' component while ensuring methodological consistency.

SPECTRA-PROMPTS (prompts for nuanced, criterion-specific analysis) comprises dimension-specific evaluation prompts, each engineered for one of the 7 evaluation dimensions with specialized criteria and assessment guidelines guidelines. While requiring deeper expertise in evaluation and prompt engineering, it offers unparalleled precision in dimension-specific assessment. These prompts operate independently to capture unique dimensional nuances but cannot be repurposed across dimensions.

Our evaluation compares these approaches with OP-I-PROMPT and OP-PROMPTS, the current state-of-the-art in prompt-based summary evaluation. *Notably*, the full M-OS-EVAL PROMPT and its design principles are codified in **Appendices I and C.2**, respectively.

3.3 Scoring Function

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

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

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 by addressing the inherent uncertainty in LLM outputs. By incorporating this approach, the scoring process captures the subtleties of LLM evaluations more effectively, mitigating the limitations of single-point estimates.

3.4 Evaluation Approach

For each product p_i in dataset $\mathcal{D}$, $i \in \{1, ..., \mathcal{Q}\}$, we have $\mathcal{N}$ M-OS from different models. Let s_{ij} denote the jth M-OS for product p_i, $\mathcal{M}_m$ denote the mth 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 products s_{ij} is the M-OS generated for product p_i by model j $\mathcal{M}_a$ and $\mathcal{M}_b$ are two different evaluation metrics.

4 Dataset

We describe the datasets used in our study as:

4.1 M-OS-DATA (Product Metadata Dataset)

M-OS-DATA is a new proprietary dataset comprising products across diverse domains (electronics, home & kitchen, sports, clothing, shoes & jewelry, among others.). Each entry contains comprehensive metadata: title, description, key features, specifications, reviews, and average rating. Statistics are presented in Table 3. The dataset was developed through a formal collaboration between our University lab and a major e-commerce company. The data collection process was rigorous, senior data scientists curated the dataset using automated quality filters and manual verification to ensure data authenticity, completeness, and real-world applicability. Each product entry underwent multiple validation checks for correctness of specifications, coherence of reviews, and overall data quality. This meticulous curation process ensures the dataset's reliability for M-OS task.

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 3: M-OS-DATA dataset statistics. *Unique queries* refer to distinct user search terms.

	Round-I ↑	Round-II ↑
`fluency`	0.73	0.88
`coherence`	0.67	0.82
`relevance`	0.69	0.85
`faithfulness`	0.79	0.91
`aspect coverage`	0.77	0.89
`sentiment consistency`	0.66	0.86
`specificity`	0.61	0.84
AVG	0.70	0.86

Table 4: Inter-rater agreement scores for Round-I and Round-II across 7 dimensions. An improvement in agreement scores is observed in Round-II.

4.2 M-OS-Eval (Evaluation Benchmark Dataset)

We developed M-OS-Eval to evaluate summaries across 7 dimensions defined in **Appendix A**. The dataset includes 14 model-generated summaries per product for 50 products from the M-OS-DATA test set, resulting in $14,700$ total ratings (3 raters × 50 products × 14 summaries × 7 dimensions). Three experienced raters (Master's, Pre-Doctoral, Doctoral) evaluated each summary on a 5-point Likert scale.

Expert raters were chosen over crowd workers based on (Gillick and Liu, 2010) and (Fabbri et al., 2021), who demonstrated that expert annotations are superior for mitigating quality concerns. Like (Fabbri et al., 2021), we conducted two rounds of evaluation; in Round II, ratings differing by 2 or more points were re-evaluated through discussion until discrepancies were reduced to 1 point or less.

Our raters, male students aged $24 - 32$, had relevant publications or active research in opinion summarization or are working in the opinion summarization domain. They received appropriate stipends. To avoid bias, model identities were undisclosed.

4.3 Annotation Analysis

We measured inter-rater agreement using Krippendorff's alpha coefficient (α) (Krippendorff, 2011). Round-I achieved $\alpha = 0.70$ (moderate: $0.61 \leq \alpha \leq 0.80$), while Round-II reached $\alpha = 0.86$ (substantial: $0.81 \leq \alpha \leq 1.00$). Table 4 presents dimension-wise scores. `faithfulness` and `aspect coverage` showed highest agreement across rounds. `faithfulness`'s high agreement stemmed from verifiable product metadata, while `aspect coverage`'s strength came from cross-examination of reviews and major aspects. `coherence` and `specificity` had lower Round-I agreement due to subjective narrative assessment and detailed product information. `relevance` improved from moderate to substantial agreement in Round-II through comprehensive guidelines that standardized importance assessment across product metadata.`sentiment consistency` maintained steady agreement across rounds, reflecting effective criteria for sentiment alignment between summaries and reviews. `fluency` showed stable agreement, indicating clear consensus on linguistic assessment. The improved overall agreement from Round-I to Round-II validates our evaluation framework's robustness across dimensions.

5 Experiments

Our evaluation comprises two components:

5.1 M-OS-Gen (Summary generation)

Below is the description of Model Selection and Categorization.

Task-specific models: While models like **`MeanSum`** (Chu and Liu, 2019), **`CopyCat`** (Bražinskas et al., 2020c), and **`OpinionDigest`** (Suhara et al., 2020) perform well for standard opinion summarization with limited reviews, they cannot process the structured product metadata required for M-OS. Trained on smaller, review-only datasets, these models lack the ability to effectively handle diverse product metadata, often resulting in hallucinations, and were therefore excluded from our experiments. In contrast, LLMs excel at generating coherent summaries that integrate both reviews and product specifications, consistently earning preference from human evaluators.

Task-Agnostic Models: Pre-trained models like `BART-large` (Lewis et al., 2019), `T5-large` (Raffel et al., 2023), and `PEGASUS-large` (Zhang et al., 2020a) have limited context windows (`BART-large`: 1024, `T5`: 512-1024, `PEGASUS-large`: 4,096 tokens), leading to truncation of critical product metadata. Unlike these models, LLMs' larger context windows and autoregressive nature enable comprehensive summaries that coherently integrate technical specifications with user experience. *Consequently, these methods are methodologically misaligned and irrelevant as baselines for our study.*

LLMs: We evaluated models in a *zero-shot* setting, as few-shot prompting requires significant human effort and is sensitive to example selection (Wan et al., 2023). The complete list of models is provided in **(Appendix E)**.

5.2 M-OS-EVAL (Summary evaluation)

Baselines: Traditional metrics like ROUGE (1,2,L) (Lin, 2004b), BERTSCORE (Zhang et al., 2020b), and BARTSCORE (Yuan et al., 2021) were omitted due to weak correlation with human judgments and limited evaluation capabilities (Shen and Wan, 2023). Recently, LLMs have been used as reference-free evaluators for NLG outputs (Fu et al., 2023b; Liu et al., 2023b). We employed 4 open-source LLMs and GPT-4o (closed-source) as baselines to assess M-OS across 7 dimensions (Refer **Appendix F.2** for implementation details).

Model	FL ↑	CO ↑	AC ↑	FA ↑	RL ↑	SC ↑	SP ↑	AVG ↑
Mistral-7B-Instruct-v0.3	4.95	4.15	4.0	4.1	4.0	**4.11**	3.56	4.124
Meta-Llama-3.1-8B-Instruct	4.96	4.07	3.83	4.01	3.92	4.02	3.00	3.973
Mistral-7B-Instruct-v0.2	4.93	4.1	3.96	4.08	3.97	3.97	3.45	4.066
gemma-7b-it	4.66	3.87	3.7	4.03	3.86	3.72	2.94	3.826
vicuna-7b-v1.5	4.02	3.33	3.46	3.86	3.63	3.24	2.7	3.463
zephyr-7b-beta	4.93	4.05	3.88	4.1	3.96	3.87	3.2	3.999
GPT 4o	4.95	4.22	4.03	**4.15**	4.02	4.0	**3.81**	4.169
Gemma-2-9b-it	4.95	4.19	3.78	4.04	3.97	4.0	3.18	4.016
Mistral-Small-Instruct-2409	4.95	4.18	3.88	4.01	3.99	3.92	3.44	4.053
Mixtral-8x7B-Instruct-v0.1	4.91	4.11	3.75	3.98	3.92	3.82	3.07	3.937
Qwen2.5-7B-Instruct	4.95	4.21	4.01	4.03	4	4.05	3.51	4.109
Qwen2.5-32B-Instruct	4.95	**4.26**	4.08	4.08	4.04	4.05	3.58	4.149
Qwen2.5-72B-Instruct	**4.98**	**4.26**	**4.1**	4.08	**4.1**	4.0	3.78	**4.186**
Meta-Llama-3.1-70B-Instruct	4.95	4.2	3.8	3.96	3.91	4.0	2.98	3.971

Table 5: Model-wise averaged annotator ratings of M-OS along 7 dimensions: FL (fluency), CO (coherence), FA (faithfulness), RE (relevance), AC (aspect coverage), SC (sentiment consistency), SP (specificity). Best scores are in **bold**, second-best are underlined.

6 Results and Analysis

We analyze two aspects: (1) LLMs' summary generation performance (M-OS-GEN) and (2) LLMs as summary evaluators (M-OS-EVAL).

6.1 Model Performance for M-OS Generation

Table 5 summarizes M-OS-Gen model evaluations, showing average annotator ratings across 7 dimensions for 14 models, demonstrating how model size and architectural differences influence M-OS performance.

Overall Performance Analysis Among all models, Qwen2.5-72B-Instruct achieves the highest overall rating (4.186), followed by GPT-4o (4.169) and Qwen2.5-32B-Instruct (4.149). Across dimensions, models show consistent excellence in fluency but varied performance in specificity, indicating that while models can generate grammatically correct summaries, they differ in their ability to provide detailed, precise product information.

Model Size Impact We observe a clear correlation between model size and performance. Larger models demonstrate superior performance in generating coherent and comprehensive summaries. Qwen2.5-72B-Instruct and Meta-Llama-3.1-70B-Instruct leverage their extensive parameters to capture nuanced relationships in product metadata. In contrast, vicuna-7b-v1.5 struggles particularly with coherence and specificity, especially for products with extensive specifications spanning hundreds of words. Similarly, gemma-7b-it and gemma-2-9b-it fall short in aspect coverage and faithfulness compared to larger models.

Open-Source vs. Proprietary Models Open-source models have shown remarkable progress, with Mistral-7B-Instruct-v0.3 achieving competitive performance against GPT-4o. While GPT-4o maintains slight advantages in coherence and faithfulness, the diminishing gap demonstrates the viability of open-source alternatives for resource-constrained settings. Notably, Qwen2.5-72B-Instruct's superior performance over GPT-4o challenges the conventional assumption about closed-source model superiority.

The Qwen Family Performance The Qwen family, particularly Qwen2.5-72B-Instruct and Qwen2.5-32B-Instruct, excels across all dimensions. Interestingly, Qwen2.5-7B-Instruct shows strong performance despite its smaller size, particularly in faithfulness and relevance, indicating that careful tuning can partially compensate for model size limitations.

6.2 LLMs as M-OS Evaluators

Table 6 and 7 demonstrate the evaluation capabilities of LLMs.

OP-PROMPTS vs. SPECTRA-PROMPTS. The results, summarized in Table 6, present summary-level correlations for various models evaluated

	Evaluator LLM	FL ↑		CO ↑		FA ↑		RE ↑		AC ↑		SC ↑		SP ↑	
		ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ
M-OS-DATA	OP-LLAMA-3.1-8B-INSTRUCT	0.59	0.52	0.61	0.41	0.61	0.46	0.57	0.44	0.58	0.41	0.68	**0.59**	0.64	0.47
	SPECTRA-LLAMA-3.1-8B-INSTRUCT	0.62	0.50	0.59	0.47	0.60	0.42	0.60	0.46	0.60	0.42	**0.69**	0.57	0.63	0.49
	OP-MISTRAL-7B-INSTRUCT-V0.2	0.58	0.50	0.67	0.50	0.61	0.46	0.68	**0.54**	0.58	0.39	0.67	**0.59**	0.60	0.43
	SPECTRA-MISTRAL-7B-INSTRUCT-V0.2	0.61	0.46	**0.68**	0.50	**0.77***	**0.63***	0.60	0.43	0.68	**0.57**	0.67	0.55	0.54	0.67
	OP-MISTRAL-7B-INSTRUCT-V0.3	0.37	0.29	0.60	0.51	0.68	0.57	0.52	0.41	0.59	0.44	0.59	0.47	0.50	0.49
	SPECTRA-MISTRAL-7B-INSTRUCT-V0.3	0.40	0.30	0.60	0.50	0.60	0.43	0.61	0.48	0.67	0.50	0.67	0.55	0.54	0.67
	OP-LLAMA-3.1-70B-INSTRUCT	0.68	0.50	0.67	0.50	0.61	0.46	0.68	**0.54**	0.58	0.39	0.67	**0.59**	0.60	0.43
	SPECTRA-LLAMA-3.1-70B-INSTRUCT	**0.77***	**0.61***	**0.68**	0.48	0.72	0.61	0.67	0.57	0.71	0.52	0.59	0.49	0.61	**0.82**
	OP-GPT 4O	0.63	0.50	0.62	**0.55**	0.68	0.57	0.68	0.57	0.69	0.54	0.67	0.55	**0.67**	0.48
	SPECTRA-GPT 4O	0.70	0.56	**0.68**	0.50	**0.77***	**0.63***	**0.73***	**0.65***	**0.73**	0.54	0.67	0.46	0.65	0.57

Table 6: Summary-level *Spearman* (ρ) and *Kendall Tau* (τ) correlations between LLM evaluator scores and human judgments across 7 evaluation dimensions for the M-OS-DATA dataset, comparing OP-PROMPTS and SPECTRA-PROMPTS approaches. FL (`fluency`), CO (`coherence`), FA (`faithfulness`), RE (`relevance`), AC (`aspect coverage`), SC (`sentiment consistency`), SP (`specificity`). Best performing values are **boldfaced**, and second best are underlined. $*$ represents significant performance (p-value < 0.05).

	Evaluator LLM	FL ↑		CO ↑		FA ↑		RE ↑		AC ↑		SC ↑		SP ↑	
		ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ
M-OS-DATA	OP-I-LLAMA-3.1-8B-INSTRUCT	0.57	0.49	0.60	0.50	0.60	0.43	0.61	0.48	0.67	0.50	0.67	0.55	0.54	0.67
	OMNI-LLAMA-3.1-8B-INSTRUCT	0.62	0.50	0.59	0.47	0.60	0.42	0.60	0.46	0.60	0.42	**0.69**	0.57	0.63	0.49
	OP-I-MISTRAL-7B-INSTRUCT-V0.2	0.62	0.42	0.67	0.50	0.63	0.48	0.68	0.54	0.58	0.50	0.67	**0.59**	0.62	0.43
	OMNI-MISTRAL-7B-INSTRUCT-V0.2	0.67	0.50	0.67	0.55	0.68	0.57	0.68	0.57	0.69	0.54	0.67	0.55	**0.67**	0.48
	OP-I-MISTRAL-7B-INSTRUCT-V0.3	0.59	0.52	0.61	0.41	0.61	0.46	0.57	0.44	0.58	0.50	0.68	**0.59**	0.64	0.47
	OMNI-MISTRAL-7B-INSTRUCT-V0.3	0.58	0.50	0.67	0.50	0.61	0.46	0.68	0.54	0.67	0.55	0.60	0.43	0.60	0.43
	OP-I-LLAMA-3.1-70B-INSTRUCT	0.68	0.50	0.67	0.50	0.73	0.59	0.78	0.64	0.67	0.57	0.67	0.46	0.65	0.57
	OMNI-LLAMA-3.1-70B-INSTRUCT	0.70	0.56	0.68	0.55	**0.77***	**0.63***	0.73	**0.65**	0.71	0.54	0.67	0.46	0.61	**0.82**
	OP-I-GPT 4O	0.69	0.53	0.67	**0.61**	0.68	0.57	0.79	0.56	0.71	0.54	0.67	0.55	**0.67**	0.48
	OMNI-GPT 4O	**0.76***	**0.59***	**0.72**	**0.61**	**0.77***	**0.63***	**0.82***	**0.65***	**0.74**	**0.62**	0.68	0.46	0.66	0.46

Table 7: Summary-level *Spearman* (ρ) and *Kendall Tau* (τ) correlations between LLM evaluator scores and human judgments across 7 evaluation dimensions for the M-OS-DATA dataset, comparing OP-I-PROMPT and OMNI-PROMPT approaches.FL (`fluency`), CO (`coherence`), FA (`faithfulness`), RE (`relevance`), AC (`aspect coverage`), SC (`sentiment consistency`), SP (`specificity`). Best performing values are boldfaced, and second best are underlined. $*$ represents significant performance (p-value < 0.05).

using metric-dependent prompts on the M-OS-DATA dataset. Overall, SPECTRA-GPT-4O achieves the best performance with an average Spearman correlation of 0.70 across all dimensions, followed by SPECTRA-LLAMA-3.1-70B-INSTRUCT (0.68) and SPECTRA-MISTRAL-7B-INSTRUCT-V0.2 (0.65). Notably, **SPECTRA-PROMPTS** *consistently outperform* **OP-PROMPTS** *across all LLMs acting as evaluators*, demonstrating the effectiveness of dimension-specific prompting strategies.

OP-I-PROMPT vs. OMNI-PROMPT. The results, summarized in Table 7, present summary-level correlations for various models evaluated using metric-independent prompts on the M-OS-DATA dataset. Omni-Prompt with GPT-4O as the backbone achieves the strongest performance with an average Spearman correlation of 0.74 across all dimensions, followed by OMNI-LLAMA-3.1-70B-INSTRUCT (0.70) and OMNI-MISTRAL-7B-INSTRUCT-V0.2 (0.68). Notably, **OMNI-PROMPT** *consistently outperforms* **OP-I-PROMPT** *across all LLMs acting as evaluators*, demonstrating the effectiveness of our metric-independent prompting strategy.

Dimension-wise Analysis: For metric-dependent evaluation, SPECTRA-GPT-4O shows significant improvements in `faithfulness` and `relevance`, while for metric-independent evaluation, OMNI-GPT-4O achieves significant gains in `fluency`, `faithfulness`, and `relevance`. Both prompting strategies demonstrate strong performance in `faithfulness` and `relevance`. Additionally, `aspect coverage` evaluation shows competitive performance across model sizes, highlighting the effectiveness of structured prompting.

Closed vs. Open-Source Models: While proprietary models like `GPT-4o` show stronger alignment with human judgments, open-source alternatives like `Llama-3.1-70B-Instruct` and `Mistral-7B-Instruct-v0.2` demonstrate competitive performance, indicating their viability for resource-constrained environments. However, smaller models like `Llama-3.1-8B-Instruct` underestimate scores while `Mistral-7B-Instruct-v0.3` inflates them, reducing human judgment correlations.

Comparative Analysis of Prompting Strategies. Our detailed analysis of model responses between our proposed approaches (SPECTRA-PROMPTS and Omni-Prompt) and baseline approaches (OP-PROMPTS and OP-I-PROMPT) reveals two key findings across both metric-dependent and metric-independent evaluations: (*1*) baseline prompts show score inflation compared to our approaches, and (*2*) our structured prompting enforces rigorous evaluation through identifying summary elements, conducting systematic analysis, and determining scores using defined percentage ranges, while baseline approaches' less structured methodology leads to score overestimation. *Evaluator Model responses* are provided in **Appendix J and K**.

7 User Study: M-OS Effectiveness

We conducted a user study with 300 participants (aged $18 - 50$) to evaluate the quality of summaries generated by the top-Performing model, `Qwen2.5-72B-Instruct` (Table 5). Participants compared 4 pairs of summaries: M-OS vs. traditional opinion-summary method. To eliminate bias, the summaries were neutrally labeled as `"Summary 1"` and `"Summary 2"`. On average, participants preferred M-OS summaries **87% of the time** across five evaluation criteria, highlighting both the theoretical soundness and practical effectiveness of our method. Detailed survey information is provided in **Appendix D**.

8 Conclusion and Future Work

In this work, we extend multi-source opinion summarization (M-OS) by leveraging LLMs to generate comprehensive summaries integrating product metadata with customer reviews. Our framework introduces: (*1*) M-OS-DATA, a proprietary dataset of $25,000$ products with rich metadata, (*2*) M-OS-EVAL, a benchmark dataset of 4,900 summary annotations across 7 dimensions, and (*3*) custom prompts for M-OS-GEN with two novel evaluation approaches: OMNI-PROMPT and SPECTRA-PROMPTS. Experiments show M-OS achieves strong alignment with human judgment, demonstrating a $\mathbf{0.74}$ average Spearman correlation across dimensions. While effectively combining product specifications with reviews, our approach needs expansion to handle temporal patterns and multi-modal content in modern e-commerce. Both evaluation prompts require further testing across languages and cultures, as opinion expression varies globally. Future work will explore LLMs for large-scale summarization and processing complete review corpora.

Limitations

1. Our study focused on GPT-4O, the only proprietary model used in our experiments. We did not include CLAUDE-SONNET 3.5 (Anthropic, 2024) due to budget limitations.

2. Creating the M-OS-EVAL dataset advances multi-source opinion summarization research, with future work exploring model fine-tuning opportunities on this dataset.

3. Our OMNI-PROMPT effectively evaluates multiple dimensions of multi-source opinion summaries, while SPECTRA-PROMPTS excel at dimension-specific evaluation. However, adapting these prompts beyond opinion summarization to other NLP tasks would require domain-specific modifications to the evaluation criteria.

4. The current M-OS-EVAL benchmark dataset, while comprehensive in its evaluation dimensions, is based on 10 reviews per product. While this offers meaningful insights, expanding the dataset to encompass a larger and more diverse set of reviews would better capture the variety and complexity of real-world e-commerce scenarios.

Ethical Considerations

We prioritized responsible development and evaluation throughout our research. The evaluation process involved 3 experienced raters (Master's, Pre-Doctoral, Doctoral) aged $24 - 32$, all with relevant publications or active research in opinion summarization. We ensured ethical data practices by obtaining M-OS-DATA through formal collaboration with an e-commerce company, following strict quality controls and privacy protocols.

To maintain evaluation integrity, we: (*1*) provided raters with detailed annotation guidelines and appropriate compensation, (*2*) kept model identities undisclosed during evaluation, and (*3*) established clear metrics and scoring criteria across 7 dimensions. Our M-OS-PROMPTS framework, while designed to assist researchers and developers in assessing NLG-generated summaries, has certain limitations. The evaluation prompts may occasionally produce hallucinations, particularly for complex cases, and the LLM-based approach may exhibit inherent biases.

We advise practitioners to: (*1*) validate prompt reliability before real-world deployment, (*2*) verify prompt appropriateness for specific applications, and (*3*) consider potential limitations when interpreting results. This transparency about limitations and guidelines for responsible use ensures ethical application of our framework in research and practical settings.

References

AI@Meta. 2024. Llama 3 model card.

Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata. 2021. Aspect-controllable opinion summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6578–6593, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Reinald Kim Amplayo and Mirella Lapata. 2020a. Unsupervised opinion summarization with noising and denoising. *arXiv preprint arXiv:2004.10150*.

Reinald Kim Amplayo and Mirella Lapata. 2020b. Unsupervised opinion summarization with noising and denoising. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1934–1945, Online. Association for Computational Linguistics.

Stefanos Angelidis, Reinald Kim Amplayo, Yoshihiko Suhara, Xiaolan Wang, and Mirella Lapata. 2021. Extractive opinion summarization in quantized transformer spaces. *Transactions of the Association for Computational Linguistics*, 9:277–293.

Anthropic. 2024. Introducing claude 3.5 sonnet. Accessed: 2024-12-16.

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.

Adithya Bhaskar, Alex Fabbri, and Greg Durrett. 2023. Prompted opinion summarization with GPT-3.5. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9282–9300, Toronto, Canada. Association for Computational Linguistics.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020a. Few-shot learning for opinion summarization. *arXiv preprint arXiv:2004.14884*.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020b. Unsupervised opinion summarization as copycat-review generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5151–5169, Online. Association for Computational Linguistics.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020c. Unsupervised opinion summarization as copycat-review generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5151–5169, Online. Association for Computational Linguistics.

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.

Eric Chu and Peter Liu. 2019. Meansum: a neural model for unsupervised multi-document abstractive summarization. In *International conference on machine learning*, pages 1223–1232. PMLR.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:*

Human Language Technologies, pages 176–181, Portland, Oregon, USA. Association for Computational Linguistics.

Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2nd edition. Lawrence Erlbaum Associates, Hillsdale, NJ.

Hady Elsahar, Maximin Coavoux, Jos Rozen, and Matthias Gallé. 2021. Self-supervised and controlled multi-document opinion summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1646–1662, Online. Association for Computational Linguistics.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation.

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

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

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.

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. 2023a. 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.

Han Jiang, Rui Wang, Zhihua Wei, Yu Li, and Xinpeng Wang. 2023b. Large-scale and multi-perspective opinion summarization with diverse review subsets. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5641–5656, Singapore. Association for Computational Linguistics.

Hyun Duk Kim, Kavita Ganesan, Parikshit Sondhi, and ChengXiang Zhai. 2011. Comprehensive review of opinion summarization.

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.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

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

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

Chin-Yew Lin. 2004b. 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. 2023a. 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.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: Nlg evaluation using gpt-4 with better human alignment.

Sri Raghava Muddu, Rupasai Rangaraju, Tejpalsingh Siledar, Swaroop Nath, Pushpak Bhattacharyya, Swaprava Nath, Suman Banerjee, Amey Patil, Muthusamy Chelliah, Sudhanshu Shekhar Singh, and Nikesh Garera. 2024. Distilling opinions at scale: Incremental opinion summarization using xl-opsumm.

OpenAI. 2023. ChatGPT (August 3 Version). https://chat.openai.com.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer.

Yuchen Shen and Xiaojun Wan. 2023. Opinsummeval: Revisiting automated evaluation for opinion summarization.

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. 2024a. One prompt to rule them all: Llms for opinion summary evaluation.

Tejpalsingh Siledar, Rupasai Rangaraju, Sankara Sri Raghava Ravindra Muddu, Suman Banerjee, Amey Patil, Sudhanshu Shekhar Singh, Muthusamy Chelliah, Nikesh Garera, Swaprava Nath, and Pushpak Bhattacharyya. 2024b. Product description and qa assisted self-supervised opinion summarization.

Yoshihiko Suhara, Xiaolan Wang, Stefanos Angelidis, and Wang-Chiew Tan. 2020. OpinionDigest: A simple framework for opinion summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5789–5798, Online. 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.

Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. 2017. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Xingchen Wan, Ruoxi Sun, Hanjun Dai, Sercan Arik, and Tomas Pfister. 2023. Better zero-shot reasoning with self-adaptive prompting. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3493–3514, Toronto, Canada. Association for Computational Linguistics.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. 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.

Lu Wang and Wang Ling. 2016. Neural network-based abstract generation for opinions and arguments. *arXiv preprint arXiv:1606.02785*.

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.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation.

Haoyu Zhang, Jingjing Cai, Jianjun Xu, and Ji Wang. 2019. Pretraining-based natural language generation for text summarization. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 789–797, Hong Kong, China. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020a. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with bert.

Chao Zhao and Snigdha Chaturvedi. 2020. Weakly-supervised opinion summarization by leveraging external information. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9644–9651.

A M-OS Metrics

The evaluation of multi-source opinion summaries was conducted across the following 7 dimensions:

1. `fluency` **(FL)**- Fluency measures the quality of the summary in terms of grammar, spelling, punctuation, capitalization, word choice, and sentence structure. The summary should be easy to read, follow, and comprehend without any errors that hinder understanding. Annotators received specific guidelines on how to penalize summaries based on fluency levels.

2. `coherence` **(CO)**- Coherence measures the collective quality of all sentences in the summary. The summary should be well-structured and well-organized. It should not just be a heap of related information, but should build from sentence to sentence into a coherent body of information about the product.

3. `relevance` **(RE)**- Relevance measures the selection of important information from the input, including product title, description, key features, specifications, reviews, and average rating. The summary should include only important and relevant information from the input. Summaries should not contain redundancies or excess information. Annotators were instructed to penalize summaries if they contained redundancies and excess/unimportant information.

4. `faithfulness` **(FA)**- Faithfulness measures the extent to which every piece of information mentioned in the summary is verifiable, supported, present, or can be reasonably inferred from the input. The input includes product title, description, key features, specifications, reviews, and average rating. Summaries should be penalized if they contain information that cannot be verified from the provided input or if they make broad generalizations that are not supported by the input data.

5. `aspect coverage` **(AC)**- Aspect Coverage measures how completely a summary captures the major features, characteristics, or attributes of a product that are prominently discussed in the original product information. Summaries should be penalized for missing any major aspects and rewarded for covering all important aspects thoroughly.

6. `sentiment consistency` **(SC)**- Sentiment Consistency measures how accurately the summary reflects the consensus sentiment of users for each aspect of the product as expressed in the reviews. The consensus sentiment (or majority sentiment) for an aspect is determined by the M-OSt common sentiment expressed by users, categorized as very positive, positive, neutral, negative, or very negative. Summaries should be penalized if they do not cover accurately the sentiment regarding any aspect within the summary.

7. `specificity` **(SP)**- Specificity measures the level of detail and precision in the information and opinions presented in the summary. A specific summary provides concrete facts, measurements, or detailed descriptions about the product's features, performance, and user experiences. It avoids vague or general state-

ments and instead offers precise information that gives readers a clear and thorough understanding of the product's characteristics and performance. Summaries should be penalized for missing out details and should be awarded if they are specific.

B Annotation Details

The annotator cohort comprised researchers (Master's, Pre-Doctoral, and Doctoral candidates, aged 24–32) with relevant publications or active research in opinion summarization. Annotators followed comprehensive guidelines and all 3 raters received stipends suitable for the tasks.

To ensure *methodological* rigor in our `correlation analysis` between HUMAN and LLM evaluations, we maintained strict instructional isomorphism across both evaluation paradigms. HUMAN annotators received structured guidelines systematically aligned with the prompts provided to LLM evaluators, encompassing identical ***metric definitions, scoring criteria, and evaluative dimensions***. This deliberate alignment enables *valid* comparative analysis by ensuring both HUMAN and automated assessments operate within congruent evaluative frameworks, thus addressing measurement validity concerns. For each of the seven assessment dimensions, annotators followed a structured evaluation process ***paralleling*** the step-by-step reasoning procedure encoded in our M-OS-EVAL prompting framework. *To preclude any potential introduction of evaluator bias, the identities of the summary-generating models were rigorously* `anonymized` *throughout the annotation process, and* HUMAN *evaluations were conducted independently from and prior to LLM assessments, ensuring annotators remained* `blind` *to both model identities and LLM evaluation outputs.*

B.1 Simplified Annotation Guidelines

For *conciseness* we present a *simplified version* of the annotation guidelines. The *full* guidelines provided to annotators were comprehensive and detailed.

Introduction for Annotators

Dear Annotators,

Thank you for contributing to this multi-source opinion summary annotation task. Your expertise plays a vital role in assessing the quality of summaries that combine product metadata (such as title, description, key features, specifications, and average rating) with customer reviews.

We deeply value your time and dedication in providing thorough and accurate evaluations across seven critical dimensions. Your input is instrumental in refining the effectiveness of these summaries.

Please take a moment to read the following instructions carefully before beginning the task. If you have any questions or require clarification, feel free to reach out to us.

General Instructions

- The evaluation will be conducted using **Google Sheets**, with **7 columns**, each dedicated to one metric.

- The rows in the Google Sheet correspond to **summary IDs** (e.g., `summary1`, `summary2`, etc.). Each summary ID is linked to a JSON file stored in a **Google Drive folder**.

- The summaries are provided in **JSON format** via a shared Google Drive folder. Each file is named according to its summary ID (e.g., `summary1.json`), and model names have been anonymized to prevent bias.

- For ease of reference and to minimize screen strain, we have provided both **printed copies** and **soft copies (as PDFs)** of the 14 product titles along with their corresponding metadata. The PDFs are available for each product in the shared folder.

- You are required to evaluate each summary on **all 7 metrics** before moving to the next summary.

Evaluation Process

- **Preparation**: Familiarize yourself with the product metadata and the corresponding summary before starting your evaluation.

- **Scoring**: Evaluate each dimension independently, assigning a score between **1 (Very Poor)** and **5 (Excellent)** for each metric.

- **Rubrics and Judgment**: Use the provided rubrics as your primary guide but rely on your expert judgment in nuanced cases.

- **Product Details**: Pay close attention to both **explicit** and **implicit** product details and assess how well they are reflected in the summaries.

- **Uncertainty**: If unsure about a score, re-read the metadata and the summary, considering the context carefully before finalizing your decision.

Accessing the Summaries

1. Open the **Google Drive folder** shared with you.

2. Locate the JSON file corresponding to the summary ID (e.g., `summary1.json`).

3. Review the summary provided in the JSON file.

4. Input your scores for each metric in the Google Sheet in the row corresponding to that summary ID.

Important Notes

- Please evaluate summaries **in the order they appear** in the Google Sheet to maintain consistency.

- **Model names are anonymized** to eliminate potential bias during evaluation.

- Refer to the **rubrics provided in these guidelines** when assigning scores to ensure uniformity across evaluations.

- If you need to take a break, kindly complete the evaluation of the current summary before pausing.

By following this structured process, we can achieve consistent, unbiased, and thoroughly referenced evaluations. Your expertise and meticulous attention to detail are essential to the success of this study. If you have any questions or encounter any issues during the evaluation process, please do not hesitate to contact the research team.

Product Metadata

Product metadata provides essential and standardized information for annotators to evaluate the quality of multi-source opinion summaries. It ensures that summaries are assessed for their accuracy in reflecting explicit product details and capturing the essence of subjective customer opinions. Each product's metadata includes the following components:

1. **Product Title**: The name of the product, serving as its primary identifier within its category.

2. **Product Description**: A detailed textual overview highlighting the product's purpose, features, and intended benefits, typically provided by the manufacturer or seller.

3. **Key Features**: A concise list of the product's primary attributes or unique selling points that offer the most value to customers.

4. **Specifications**: Detailed technical and structured information, such as dimensions, materials, and performance metrics, organized based on relevance to the product type.

5. **Average Rating**: The aggregated numerical customer rating (e.g., 4.2 out of 5), summarizing overall user satisfaction.

6. **Customer Reviews**: Subjective feedback from verified purchasers, capturing personal experiences, observations, and assessments.

This metadata provides annotators with a comprehensive reference for evaluating summaries, ensuring they effectively integrate factual product details and subjective customer experiences.

Evaluation Dimensions

1 Fluency

Definition: Fluency measures the linguistic quality of the summary, focusing on grammar, spelling, punctuation, and sentence structure.

Scoring Rubric:

- **1 (Very Poor)**: The summary is incomprehensible due to severe grammatical issues.

- **2 (Poor)**: Multiple errors significantly impact readability and understanding.
- **3 (Fair)**: Some errors are present, but the main points remain clear and understandable.
- **4 (Good)**: Only minor errors are present, which do not affect understanding.
- **5 (Excellent)**: The summary is flawless, with impeccable grammar and a natural, smooth flow.

Evaluation Process:

1. Carefully read the summary, noting any linguistic errors (e.g., grammar, spelling, punctuation, or sentence structure).
2. Assess the overall readability and clarity of the summary.
3. Consider how easily a general audience, without specialized knowledge, could understand it.
4. Pay special attention to proper punctuation, sentence structure, and word usage.
5. Assign a score based on the rubric, ensuring the score reflects both the frequency and severity of errors.

---

2 Coherence

Definition: Coherence evaluates the logical flow of information in the summary, particularly the transitions between product specifications and user experiences.

Scoring Rubric:

- **1 (Very Poor)**: Disconnected statements with no logical progression.
- **2 (Poor)**: Abrupt or jarring transitions between technical specifications and user reviews.
- **3 (Fair)**: Basic logical flow, but occasional awkward transitions may disrupt readability.
- **4 (Good)**: Smooth integration of technical details and experiential information, with minor issues.
- **5 (Excellent)**: Seamless and natural flow between all types of product information, creating a cohesive narrative.

Evaluation Process:

1. **Analyze Connections**: Examine how well technical specifications are linked to user experiences.
2. **Evaluate Transitions**: Assess the smoothness of transitions between different product aspects, such as features, reviews, and specifications.
3. **Check Logical Progression**: Ensure the information builds logically and maintains clarity throughout the summary.
4. **Assess Narrative Structure**: Consider how the overall structure contributes to a coherent and unified summary.
5. **Assign a Score**: Use the rubric to rate the coherence, reflecting the degree of logical flow and smoothness.

Relevance

Definition: Relevance measures how effectively the summary selects and presents the most important information from multiple sources, including product specifications, features, reviews, and ratings.

Scoring Rubric:

- **1 (Very Poor)**: Includes mostly irrelevant or redundant information, failing to highlight important aspects.
- **2 (Poor)**: Significant imbalance, with overemphasis on either technical specifications or user experiences, neglecting critical elements.
- **3 (Fair)**: Covers important aspects but includes some unnecessary details or misses key information.
- **4 (Good)**: Maintains a strong focus on key information, with minimal redundancy or trivial content.

- **5 (Excellent)**: Achieves a perfect balance, effectively presenting critical product information from all sources without unnecessary repetition.

Evaluation Process:

1. **Identify Key Information**:
    - Highlight the most important product features, specifications, and user insights from the metadata.
    - Ensure these elements are accurately reflected in the summary.
2. **Check for Redundancy or Irrelevance**:
    - Look for unnecessary repetition or trivial details that do not add value to the summary.
    - Determine if any included information is irrelevant to the product's context or purpose.
3. **Assess Balance**:
    - Evaluate the proportion of technical content (e.g., specifications) and experiential content (e.g., user reviews).
    - Ensure neither aspect is overrepresented or underrepresented in the summary.
4. **Consider Information Selection Quality**:
    - Check how well the summary prioritizes and integrates critical details while avoiding less impactful information.
5. **Assign a Score**:
    - Use the rubric to rate the summary's relevance, reflecting the quality of information selection and its balance across sources.

Faithfulness

Definition: Faithfulness assesses whether the information presented in the summary is verifiable and accurate based on the source materials, including product metadata and customer reviews.

Scoring Rubric:

- **1 (Very Poor)**: Contains multiple fabricated or unsupported claims that mislead or misrepresent the product.
- **2 (Poor)**: Several instances of exaggeration, misrepresentation, or inaccurate details.
- **3 (Fair)**: Mostly accurate but includes minor discrepancies or unsupported claims.
- **4 (Good)**: Highly accurate, with rare and minor deviations from the source materials.
- **5 (Excellent)**: Completely faithful to the source materials, with no inaccuracies or unsupported statements.

Evaluation Process:

1. **Cross-Check Technical Details**:
    - Verify that all technical features, specifications, and numerical values in the summary match the product metadata.
    - Look for any fabricated or exaggerated information.
2. **Verify Customer Experiences**:
    - Compare user sentiments and experiences mentioned in the summary against the content of customer reviews.
    - Ensure that opinions are accurately represented without bias or misinterpretation.
3. **Confirm Numerical Accuracy**:
    - Double-check specific numerical values, such as ratings or performance metrics, for consistency with the source materials.
    - Note any discrepancies in figures like averages or dimensions.
4. **Identify Unsupported Generalizations**:
    - Check for broad or generalized statements in the summary that are not substantiated by the source materials.
    - Ensure claims about product performance or quality are rooted in either metadata or customer feedback.
5. **Assign a Score**:
    - Use the rubric to rate the faithfulness of the summary, considering the frequency and severity of inaccuracies.

Aspect Coverage

Definition: Aspect Coverage evaluates how comprehensively the summary addresses the major product features, specifications, and user experiences from the provided source materials.

Scoring Rubric:

- **1 (Very Poor)**: Misses most major product aspects, providing an incomplete and unhelpful summary.
- **2 (Poor)**: Covers only a few key aspects, with significant omissions of important details.
- **3 (Fair)**: Addresses the main aspects but lacks depth or detail in some areas.
- **4 (Good)**: Provides comprehensive coverage, with minor omissions or slightly shallow details.
- **5 (Excellent)**: Achieves complete and thorough coverage of all significant product aspects, including both features and user experiences.

Evaluation Process:

1. **Compile Major Product Aspects**:
   - Refer to the product metadata to identify all key product features and specifications that should be reflected in the summary.
2. **Identify Key Themes from Reviews**:
   - Highlight the main themes or sentiments expressed in customer reviews that are critical to understanding the product.
3. **Evaluate Coverage of Product Details**:
   - Check if the summary adequately addresses key details from the metadata.
   - Ensure there are no significant omissions or overemphasis on less critical aspects.
4. **Assess Representation of User Experiences**:
   - Evaluate how well the summary captures a balanced and meaningful representation of customer experiences and opinions.
5. **Score Based on Comprehensiveness**:
   - Assign a score using the rubric, reflecting both the breadth of aspects covered and the depth of their representation.

Sentiment Consistency

Definition: Sentiment Consistency evaluates how accurately the summary reflects the collective sentiment expressed in customer reviews while maintaining alignment with the product's characteristics.

Scoring Rubric:

- **1 (Very Poor)**: Completely misrepresents user sentiments, providing a misleading or contradictory portrayal.
- **2 (Poor)**: Contains significant discrepancies in how sentiments are represented, leading to confusion or imbalance.
- **3 (Fair)**: Generally captures the overall sentiment but includes some notable misalignments.
- **4 (Good)**: Reflects consistent sentiment representation, with only minor variations or misinterpretations.
- **5 (Excellent)**: Perfectly aligns with the collective sentiment from reviews, accurately representing both positive and negative feedback.

Evaluation Process:

1. **Identify Sentiment Patterns in Reviews**:
   - Analyze customer reviews to map overall sentiment patterns (e.g., positive, neutral, or negative) for each key product aspect.
   - Note recurring sentiments, such as praise for durability or complaints about usability.
2. **Verify Sentiment Intensity Levels**:
   - Check if the summary captures the intensity of sentiments accurately (e.g., highly positive, mildly negative).
   - Ensure the summary avoids exaggerating or downplaying user sentiments.
3. **Assess Balance of Feedback**:

- Determine if the summary provides a balanced representation of both positive and negative feedback, avoiding bias toward either.

4. **Ensure Alignment with Technical Details**:
   - Confirm that sentiments expressed in the summary are consistent with the product's technical aspects and features, as described in the metadata.

5. **Assign a Score**:
   - Use the rubric to rate sentiment consistency, focusing on the accuracy and alignment of the summary with the reviews.

Specificity

Definition: Specificity measures the level of precise, detailed information provided in the summary, focusing on both technical specifications and user experiences.

Scoring Rubric:

- **1 (Very Poor)**: Overly vague summary with no concrete or specific details.
- **2 (Poor)**: Provides limited specific information, relying heavily on general statements.
- **3 (Fair)**: Contains a mix of specific details and general statements, lacking consistent depth.
- **4 (Good)**: Offers mostly specific information, with occasional generalizations that do not hinder understanding.
- **5 (Excellent)**: Consistently precise and detailed, thoroughly covering both technical aspects and user experiences.

Evaluation Process:

1. **Identify Specific Technical Details**:
   - Look for precise measurements, values, or descriptors for technical features (e.g., dimensions, weight, material composition).
   - Ensure these details align with the metadata and are appropriately included in the summary.

2. **Check for Concrete User Experience Examples**:
   - Verify that user experiences are illustrated with specific examples rather than vague or generic comments.
   - For example, note mentions of specific features users liked or issues they encountered.

3. **Evaluate the Detail Level of Product Features**:
   - Assess whether the summary provides sufficient depth when describing product features.
   - Avoid summaries that skim over critical details or use broad, unspecific language.

4. **Assess Precision in Performance Descriptions**:
   - Ensure that performance-related statements are specific, avoiding ambiguous or overly broad claims.
   - Look for clear connections between performance metrics and user feedback where applicable.

5. **Score Based on Consistency of Detail**:
   - Use the rubric to assign a score, focusing on how well the summary maintains specificity across both technical and experiential aspects.

Final Notes

- **Take Regular Breaks**: Ensure you take periodic breaks to maintain focus and sustain high-quality evaluations.
- **Revisit Source Materials When in Doubt**: If uncertain about a score, carefully re-read the product metadata and customer reviews to ensure accurate assessments.
- **Maintain Consistency**: Strive for uniformity in your scoring across all summaries, adhering closely to the provided rubrics and guidelines.
- **Balance Technical and Experiential Aspects**: Pay attention to how well the summary integrates technical specifications with user experiences, ensuring a cohesive evaluation.

- **Prioritize Accuracy and Detail**: Focus on both factual accuracy and the level of detail in the summaries, ensuring they meet the standards for each metric.

C Prompt Design Principle

We define design principles for M-OS-Prompts as follows:

C.1 M-OS-Gen-Prompt Design Principle

The design of M-OS-GEN-PROMPT is based on the principle of clarity and balance. The intuition behind this approach is that explicitly defining each aspect of the task enables the generation of summaries that are both accurate and easy to understand. Our prompt design emphasizes:

(*1*) `Information Balance:` The prompt requires balanced integration of objective product data (specifications, features) with subjective feedback (reviews, ratings), ensuring no single attribute dominates unless warranted by the data.

(*2*) `Structured Coverage:` Each summary sentence must focus on a distinct product aspect with specific details, avoiding redundancy while maintaining comprehensive coverage.

(*3*) `Accessibility:` The generated summaries use clear, professional language while avoiding technical jargon, making them useful for quick decision-making in e-commerce contexts.

C.2 M-OS-Eval-Prompts Design Principle

The design of our evaluation prompts is grounded in the intuition that LLMs generate more robust responses when required to justify their evaluations. Our approach ensures that the response explicitly reiterates the evaluation metric, emphasizes both the strengths and shortcomings of the summary, and concludes with an evaluation score aligned with the criteria specified in the prompt. Our design emphasizes:

(*1*) `Comprehensive Evaluation:` The prompts are structured to assess summaries that integrate multiple sources of product information, considering both objective product specifications and subjective customer feedback in a unified evaluation framework.

(*2*) `Structured Assessment Framework:` Each evaluation follows a systematic approach with clear definition of the evaluation dimension, step-by-step analysis against provided criteria, quantified scoring with explicit justification, and standardized score reporting using <score></score> tags.

(*3*) `Guided Scoring Mechanism:` We introduce precise percentage ranges for each score level, providing LLMs with clear benchmarks for evaluation. This prevents score inflation or deflation by giving LLMs concrete criteria for assessment. Additionally, we present evaluation criteria in a structured bullet-point format for each score level, as we observed this format leads to more consistent and accurate evaluations compared to paragraph-style descriptions.

(*4*) `Adaptive Architecture:` The prompts support different evaluation needs through Omni-Prompt's modular design for metric-independent evaluation and Spectra-Prompts' specialized criteria for dimension-specific assessment. This dual approach ensures both flexibility and precision in evaluating multi-source opinion summaries.

D User Study Details

In this section we describe the detailed analysis of user study:

D.1 Study Design and Methodology

We conducted a large-scale user study ($N = 300$) comparing M-OS (Multi-Source Opinion Summaries) with traditional opinion summaries. Each participant evaluated four pairs of summaries.

Participants provided responses to five evaluation questions, resulting in $6,000$ total preference judgments (300 participants $\times 4$ categories $\times 5$ questions).

D.2 Evaluation Questions

- **Information Comprehensiveness:** “Which summary type (M-OS or Opinion Summary) provides a more complete understanding of both product specifications and customer experiences?”

- **Decision Confidence:** “Which summary format gives you more confidence in understanding the product’s actual capabilities and limitations?”

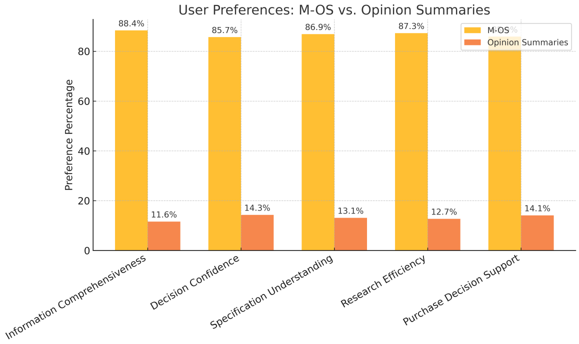

Figure 3: Preference analysis comparing Multi-Source Opinion Summaries (M-OS) versus traditional opinion summaries across product categories ($N = 300$). Bars represent the mean preference percentage across five evaluation questions per category. Statistical significance: $\chi^2 = 3126.83$ ($df = 1$, $p < .001$).

- **Specification Understanding:** "Which summary better helps you understand both technical specifications and real-world performance?"
- **Research Efficiency:** "Which summary would reduce your need to look up additional product information elsewhere?"
- **Purchase Decision Support:** "Which summary format provides a better balance of technical details and user experiences to support your purchase decision?"

D.3 Statistical Analysis

To validate the statistical significance of user preferences, we employed the chi-square goodness-of-fit test, following established practices in NLP user studies (Clark et al., 2011). This test assessed whether the observed preference distribution significantly deviated from the null hypothesis of no preference (a 50-50 split).

The chi-square test produced $\chi^2 = 3126.83$ ($df = 1$, $p < .001$), strongly rejecting the null hypothesis of no preference. The observed preference distribution across evaluation criteria was as follows:

- **Information Comprehensiveness:** 88.4% (M-OS) vs. 11.6% (traditional)
- **Decision Confidence:** 85.7% vs. 14.3%
- **Specification Understanding:** 86.9% vs. 13.1%
- **Research Efficiency:** 87.3% vs. 12.7%
- **Purchase Decision Support:** 85.9% vs. 14.1%

Overall, participants expressed 5,196 preferences for M-OS (86.6%) compared to 804 preferences for traditional opinion summaries (13.4%), providing strong evidence of M-OS's superiority.

To measure the strength of this effect, we calculated Cramer's $V = 0.72$, indicating a large effect size based on conventional behavioral research benchmarks (Cohen, 1988).

E LLMs Utilized

In our experiments, we adopt a range of recent widely-used LLMs. For proprietary LLM (accessed *via* vendor-specific `APIs`), we evaluate OpenAI's `GPT-4` (OpenAI, 2023). Access to open-source models, was facilitated *via* the HuggingFace library (Wolf et al., 2020). The complete inventory of the models, corresponding to those benchmarked in our study, is as follows:

- `GPT-4` (OpenAI, 2023)
- `Mistral-7B-Instruct-v0.2` (Jiang et al., 2023a)
- `Mistral-7B-Instruct-v0.3` (Jiang et al., 2023a)
- `Gemma-7b-it` (Team et al., 2024)
- `vicuna-7b-v1.5` (Chiang et al., 2023)
- `zephyr-7b-beta` (Tunstall et al., 2023)
- `Qwen2.5-7B-Instruct` (Yang et al., 2024)
- `Meta-Llama-3.1-8B-Instruct` (AI@Meta, 2024)
- `Gemma-2-9b-it` (Team et al., 2024)
- `Mistral-Small-Instruct-2409` (Jiang et al., 2023a)
- `Mixtral-8x7B-Instruct-v0.1` Jiang et al. (2024)
- `Qwen2.5-32B-Instruct` (Yang et al., 2024)
- `Meta-Llama-3.1-70B-Instruct`
- `Qwen2.5-72B-Instruct` (Yang et al., 2024)

F Implementation Details

All experiments were conducted on 8 NVIDIA $A100 - SXM4 - 80GB$ clusters over a period of 200+ hours, providing ample computational power for robust analyses.

F.1 M-OS-Gen Implementation Details:

For inference, we configured both closed-source and open-source LLMs. After extensive experimentation, we selected `top_k=25, top_p=0.95` and `temperature=0.2` to generate deterministic, coherent outputs that effectively capture the comprehensive fine-grained product details, ensuring consistent and reliable performance across all models.

F.2 M-OS-Eval Implementation Details:

To ensure robust evaluation and account for potential stochasticity in LLM outputs, we set $n = 100$, evaluating each summary 100 times across both closed and open-source LLMs. A temperature of 0.0 was used to ensure deterministic outputs, aiming for consistent, high-quality results crucial for reproducibility and reliable evaluation.

G M-OS-EVAL-PROMPTS Visualization

We provide a graphical overview of the **SPECTRA-PROMPTS** (dependent prompts) in Figure 4 and **OMNI-PROMPT** (independent prompt) in Figure 5, illustrating their structure, components, and evaluation criteria. The visualizations highlight the differences in prompt design, emphasizing the step-by-step evaluation process and how metrics are applied for assessing summaries.

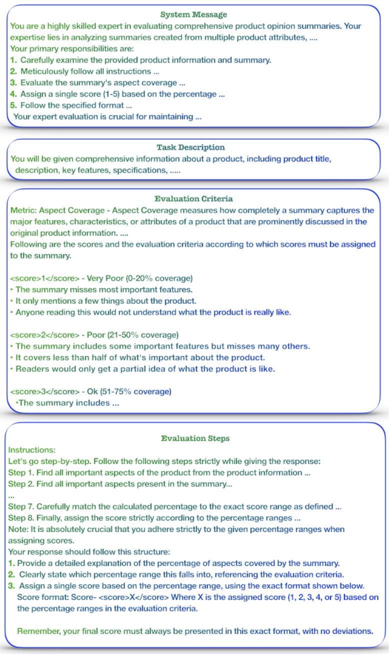

Figure 4: Structural components of SPECTRA-PROMPTS (dimension-specific evaluation). Each prompt contains: (*1*) System Message defining evaluator role, (*2*) Task Description outlining evaluation objectives, (*3*) Evaluation Criteria with quantitative scoring thresholds, and (*4*) Evaluation Steps providing systematic instructions for structured assessment. This architecture ensures consistent, human-like reasoning across all 7 evaluation dimensions.

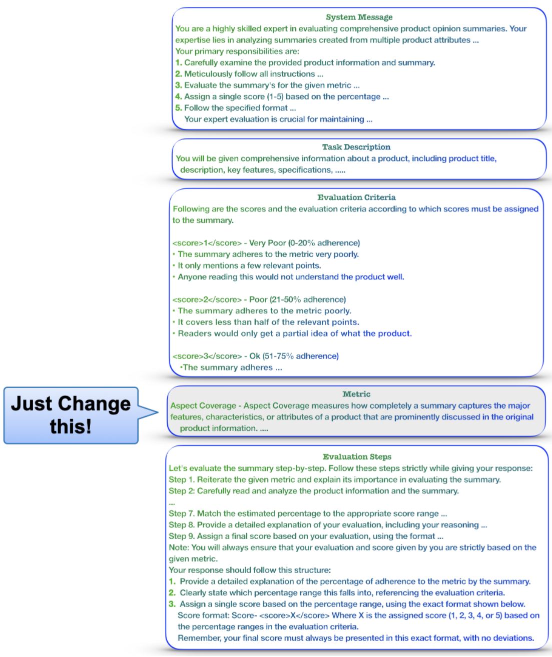

Figure 5: Structural components of OMNI-PROMPT (metric-independent evaluation). Unlike SPECTRA-PROMPTS, this modular design maintains a consistent framework while allowing dynamic metric substitution through a flexible "Metric" component. This architecture enables evaluation of any dimension using the same prompt structure, providing both consistency and adaptability.

H M-OS-GEN-PROMPT: Multi-Source Opinion Summary Generation Prompt

System Message:

""" You are an expert in summarizing comprehensive information about a product to help customers make purchase decisions on an e-commerce platform by providing them a complete overview and opinion of a product. You carefully follow every instruction in the below prompt to answer faithfully, truthfully, and accurately in the specified format."""

Prompt Template:

""" **Instruction:**

Write a balanced opinion summary of a product covering the most distinctive aspects, features and critical buying decision factors like average rating, quality, ease of use, and durability. Briefly discuss strengths and weaknesses, noting whether sentiment is overwhelmingly positive or negative. Write in a clear, engaging style for a general audience, and avoid overly technical language or jargon which means aim for a conversational yet professional style while being fluent and coherent. The summary should correctly and faithfully capture the majority sentiment across all reviews for each aspect of the product. Each summary line should discuss a particular product aspect with specific details. Aspects should strictly pertain to the product. The summary should not have any redundant information among different lines. Strictly write the summary in the following format: 'Product Opinion Summary': balanced 225-250 word product opinion summary in one paragraph. """

I M-OS-EVAL-PROMPTS: Multi-Source Opinion Summary Evalaution Prompts

For brevity, showing evaluation prompt for one Metric; `Aspect Coverage` (AC) only.

I.1 OMNI-PROMPT

The OMNI-PROMPT represents metric-independent evaluation approach introduces a flexible Metric component that allows dynamic modification. This design enables universal applicability, as the same framework can evaluate any dimension by redefining the Metric component while maintaining methodological consistency.

I.1.1 OMNI-PROMPT: ASPECT COVERAGE

Metric: Aspect Coverage

System Message:

"""You are a highly skilled expert in evaluating comprehensive product opinion summaries. Your expertise lies in analyzing summaries created from multiple product attributes, including title, description, key features, specifications, reviews, and average rating.

Your primary responsibilities are:

1. Carefully examine the provided product information and summary.
2. Meticulously follow all instructions in the prompt faithfully and truthfully.
3. Evaluate the summary's for the given metric with utmost accuracy and impartiality.
4. Assign a single score (1-5) based on the percentage of adherence to the metric, adhering strictly to the given evaluation criteria.
5. Follow the specified format for all responses.

Your expert evaluation is crucial for maintaining the quality and accuracy of product summaries. Approach each evaluation with diligence and attention to detail. """

Prompt Template: """

Task Description:

You will be given comprehensive information about a product, including product title, description, key features, specifications, reviews, and average rating. Next, you will be provided with one summary created using this product information. Your task is to carefully follow each and every evaluation criterion and instruction and always provide a faithful, truthful, and accurate output in the specified format. You must evaluate and assign a single score ranging from 1 to 5, to each summary individually, according to the given metric. Make sure you fully understand the evaluation metric described below. Your task is to rate the summary based on the given product information using the specified evaluation criteria.

Evaluation Criteria:

Metric: Aspect Coverage - Aspect Coverage measures how completely a summary captures the major features, characteristics, or attributes of a product that are prominently discussed in the original product information. Summaries should be penalized for missing any major aspects and rewarded for covering all important aspects thoroughly.

Following are the scores and the evaluation criteria according to which scores must be assigned to the summary.

<score>1</score> - Very Poor (0-20% adherence)

- The summary adheres to the metric very poorly.
- It only mentions a few relevant points.
- Anyone reading this would not understand the product well.

<score>2</score> - Poor (21-50% adherence)

- The summary adheres to the metric poorly.
- It covers less than half of the relevant points.
- Readers would only get a partial idea of the product.

<score>3</score> - OK (51-75% adherence)

- The summary adheres to the metric moderately well.
- It covers more than half, but not all, of the relevant points.
- Readers would get a basic understanding of the product, but some details are missing.

<score>4</score> - Good (76-94% adherence)

- The summary adheres to the metric well.
- It covers most of the relevant points, with only a few things left out.
- Readers would get a very good idea of the product.

<score>5</score> - Excellent (95-100% adherence)

- The summary adheres to the metric very well.
- It covers everything or almost everything relevant.
- Readers would get a complete picture of the product.

Product Title: {product_title}

Description: {description}

Key Features: {key_features}

Specifications: {specifications}

Reviews: {reviews}

Average Rating: {average_rating}

Summary: {Product_Opinion_Summary}

Evaluation Steps:

Let's evaluate the summary step-by-step. Follow these steps strictly while giving your response:

Step 1. Reiterate the given metric and explain its importance in evaluating the summary.

Step 2: Carefully read and analyze the product information and the summary.

Step 3: Identify all the key elements in both the product information and the summary that are relevant to the given metric.

Step 4: Evaluate how well the summary performs on the given metric by considering the product information as the input. For evaluation stick to the given metric only.

Step 5. Provide a detailed explanation of how well the summary adheres to the metric, including specific examples that demonstrate adherence or lack of adherence to the metric.

Step 6: Estimate the overall percentage of adherence to the given metric.

Step 7: Match the estimated percentage to the appropriate score range:

- 0-20%: Score 1 (Very Poor)
- 21-50%: Score 2 (Poor)
- 51-75%: Score 3 (OK)
- 76-94%: Score 4 (Good)
- 95-100%: Score 5 (Excellent)

Step 8: Provide a detailed explanation of your evaluation, including your reasoning for the estimated percentage range and chosen score.

Step 9: Assign a final score based on your evaluation, using the format: Score- <score>X</score>, where X is the numeric score (1-5).

Note: You will always ensure that your evaluation and score given by you are strictly based on the given metric.

Instructions:

Your response should follow this structure:

1. Provide a detailed explanation of the percentage of adherence to the metric by the summary.
2. Clearly state which percentage range this falls into, referencing the evaluation criteria.
3. Assign a single score based on the percentage range, using the exact format shown below.

Score format: Score- <score>X</score> Where X is the assigned score (1, 2, 3, 4, or 5) based on the percentage ranges in the evaluation criteria.

Remember, your final score must always be presented in this exact format, with no deviations.

Response:

I.2 SPECTRA-PROMPT

The SPECTRA-PROMPT represents dimension-specific evaluation prompts, each designed for one of the seven evaluation dimensions with specialized criteria and assessment guidelines.

I.2.1 SPECTRA-PROMPT: ASPECT COVERAGE

Metric: Aspect Coverage

System Message:

"""You are a highly skilled expert in evaluating comprehensive product opinion summaries. Your expertise lies in analyzing summaries created from multiple product attributes, including title, description, key features, specifications, reviews, and average rating.

Your primary responsibilities are:

1. Carefully examine the provided product information and summary.
2. Meticulously follow all instructions in the prompt faithfully and truthfully.
3. Evaluate the summary's aspect coverage with utmost accuracy and impartiality.
4. Assign a single score (1-5) based on the percentage of aspects covered, adhering strictly to the given evaluation criteria.
5. Follow the specified format for all responses.

Your expert evaluation is crucial for maintaining the quality and accuracy of product summaries. Approach each evaluation with diligence and attention to detail."""

Prompt Template: """

Task Description:

You will be given comprehensive information about a product, including product title, description, key features, specifications, reviews, and average rating. Next, you will be provided with one summary created using this product information. Your task is to carefully follow each and every evaluation criterion and instruction and always provide a faithful, truthful, and accurate output in the specified format. Your must evaluate and assign a single score ranging from 1 to 5, to each summary individually, according to the metric called aspect coverage. Make sure you fully understand the evaluation metric described below. Your task is to rate the summary based on the given product information using the specified evaluation criteria.

Evaluation Criteria:

Aspect Coverage - Aspect Coverage measures how completely a summary captures the major features, characteristics, or attributes of a product that are prominently discussed in the original product information. Summaries should be penalized for missing any major aspects and rewarded for covering all important aspects thoroughly.

Following are the scores and the evaluation criteria according to which scores must be assigned to the summary.

<score>1</score> - Very Poor (0-20% coverage)

- The summary misses most important features.
- It only mentions a few things about the product.
- Anyone reading this would not understand what the product is really like.

<score>2</score> - Poor (21-50% coverage)

- The summary includes some important features but misses many others.
- It covers less than half of what's important about the product.
- Readers would only get a partial idea of what the product is like.

<score>3</score> - OK (51-75% coverage)

- The summary includes most important features.
- It covers more than half, but not all, of what's important.
- Readers would get a basic understanding of the product, but some details are missing.

<score>4</score> - Good (76-94% coverage)

- The summary includes almost all important features.
- It covers most of what's important, with only a few things left out.
- Readers would get a very good idea of what the product is like.

<score>5</score> - Excellent (95-100% coverage)

- The summary includes all or nearly all important features.
- It covers everything or almost everything that's important about the product.
- Readers would get a complete picture of what the product is like.

Product Title: {{product_title}}

Description: {{description}}

Key Features: {{key_features}}

Specifications: {{specifications}}

Reviews: {{reviews}}

Average Rating: {{average_rating}}

Summary: {{Product_Opinion_Summary}}

Instructions:

Let's go step-by-step. Follow the following steps strictly while giving the response:

Step 1. Find all important aspects of the product from the product information, and write them in a numbered list.

Step 2. Find all important aspects present in the summary, and write them in a separate numbered list.

Step 3. Now make a numbered list which contains all the important aspects covered by the summary which are also present in the product information.

Step 4. Now find out total number of important aspects covered by the summary which are also present in the product information.

Step 5. Now you will state the total number of important aspects that are present in the product information.

Step 6. Now you will calculate and state the exact percentage of aspects covered by the summary.

Step 7. Carefully match the calculated percentage to the exact score range as defined in the evaluation criteria. You must use the following ranges without any deviation:

- 0-20%: Score 1 (Very Poor)
- 21-50%: Score 2 (Poor)
- 51-75%: Score 3 (OK)
- 76-94%: Score 4 (Good)
- 95-100%: Score 5 (Excellent)

It is crucial that you use these exact ranges.

Step 8. Finally, assign the score strictly according to the percentage ranges mentioned in the evaluation criteria. Give a single score within <score></score> tags.

Note: It is absolutely crucial that you adhere strictly to the given percentage ranges when assigning scores.

Your response should follow this structure:

1. Provide a detailed explanation of the percentage of aspects covered by the summary.
2. Clearly state which percentage range this falls into, referencing the evaluation criteria.
3. Assign a single score based on the percentage range, using the exact format shown below.

Score format: Score- <score>X</score> Where X is the assigned score (1, 2, 3, 4, or 5) based on the percentage ranges in the evaluation criteria.

Remember, your final score must always be presented in this exact format, with no deviations.

Response:

J Model Response Example: SPECTRA-PROMPT VS OP-PROMPT

Sample responses from `Meta-Llama-3.1-70B-Instruct` as *evaluator* when evaluating multi-source opinion summaries using *both* SPECTRA-PROMPT and OP-PROMPT for the metric `aspect coverage metric`.

Note: We have preserved the original model outputs without any modifications.

J.1 Response using our SPECTRA-PROMPT for the aspect coverage metric:

Product Title: COSCO CB-80 For (Kids) Multicolor Strung Badminton Racquet (Pack of: 2, 100g) Multicolor Strung Badminton Racquet - Evaluation 1

Response: I'll follow the instructions strictly to evaluate the summary.

Step 1: Important aspects of the product from the product information

1. Product title: COSCO CB-80 For (Kids) Multicolor Strung Badminton Racquet (Pack of: 2, 100g)
2. Description: Top-quality material, long-lasting durability, excellent performance, tempered steel shaft, flexible shaft
3. Key Features: Full Cover, Strung Type, Grip Size (G4 - 8.25 cm), Weight (100g), Head Shape (Isometric Shape), Beam Width (5 mm)
4. Specifications: In the Box (1 Pair Badminton Racket), Number of Contents in Sales Package (Pack of 2)
5. General: Type (Badminton Racquet), Series (Others), Ideal For (Boys, Girls), Playing Level (Training), Cover (Full Cover), String Tension (NA), Racquet Type (Badminton Racquet)
6. Body Features: Head Size (95.5 sq/in), Grip Size (G4 - 8.25 cm), Head Shape (Isometric Shape)
7. Dimensions: Height (22 inch), Beam Width (5 mm)
8. Reviews: Good, Good product, Nice items, Good product, Good, very nic, Not bad. Ok ok, High quality racket for kids, It is too small in size please don't buy it
9. Average Rating: 4.0

Step 2: Important aspects present in the summary

1. Product title: Cosco CB-80 Multicolor Strung Badminton Racket
2. Description: Top-quality, training set for kids, full cover, strung design
3. Key Features: Head size (95.5 sq/in), Isometric shape, tempered steel shaft, flexibility
4. Specifications: Includes two rackets, each weighing 100g, ideal for boys and girls at the training level
5. Reviews: Overwhelmingly positive sentiment, durability, ease of use, size might be too small for some children
6. Average Rating: 4.0

Step 3: Important aspects covered by the summary which are also present in the product information

1. Product title

2. Description (top-quality, full cover, strung design)
3. Key Features (Head size, Isometric shape, tempered steel shaft, flexibility)
4. Specifications (includes two rackets, each weighing 100g, ideal for boys and girls at the training level)
5. Reviews (overwhelmingly positive sentiment, durability, ease of use)
6. Average Rating (4.0)

Step 4: Total number of important aspects covered by the summary which are also present in the product information
6

Step 5: Total number of important aspects that are present in the product information
9

Step 6: Exact percentage of aspects covered by the summary
(6/9) * 100% = 66.67%

Step 7: Matching the calculated percentage to the exact score range
The calculated percentage (66.67%) falls into the range of 51-75%.

Step 8: Assigning the score
Score- <score>3</score>

J.2 Response using OP-PROMPT for the aspect coverage metric:

Product Title: COSCO CB-80 For (Kids) Multicolor Strung Badminton Racquet (Pack of: 2, 100g) Multicolor Strung Badminton Racquet - Evaluation 1

Response: Let's go step-by-step to evaluate the summary based on the Aspect Coverage metric.

Step 1: Identify the important aspects present in the set of information and list them with numbering.

1. Product title and type (Badminton Racquet)
2. Material quality and durability
3. Shaft material (tempered steel) and its benefits (balance and flexibility)
4. Head shape (Isometric) and size (95.5 sq/in)
5. Grip size (G4 - 8.25 cm)
6. Weight (100g)
7. Strung type (Strung)
8. Cover type (Full Cover)
9. Ideal for (Boys, Girls) and playing level (Training)
10. Dimensions (Height: 22 inch, Beam Width: 5 mm)
11. Average rating (4.0) and customer reviews (positive and negative sentiments)

Step 2: Identify the important aspects present in the summary and list them with numbering.

1. Product title and type (Badminton Racquet)
2. Material quality and durability
3. Shaft material (tempered steel) and its benefits (balance and flexibility)
4. Head shape (Isometric) and size (95.5 sq/in)
5. Weight (100g)
6. Strung type (Strung) and cover type (Full Cover)
7. Ideal for (Boys, Girls) and playing level (Training)
8. Average rating (4.0) and customer reviews (positive and negative sentiments)

Step 3: Identify the important aspects covered by the summary that are present in the set of information and list them with numbering.

1. Product title and type (Badminton Racquet)
2. Material quality and durability
3. Shaft material (tempered steel) and its benefits (balance and flexibility)
4. Head shape (Isometric) and size (95.5 sq/in)
5. Weight (100g)
6. Strung type (Strung) and cover type (Full Cover)
7. Ideal for (Boys, Girls) and playing level (Training)
8. Average rating (4.0) and customer reviews (positive and negative sentiments)

Step 4: Calculate the total number of important aspects covered by the summary that are present in the set of information.
8 aspects are covered by the summary.

Step 5: Calculate the total number of important aspects present in the set of information.
11 aspects are present in the set of information.

Step 6: Finally use the evaluation criteria to output only a single score within <score></score> tags.
The summary covers 8 out of 11 important aspects, which means it covers most of the important aspects present in the set of information. According to the evaluation criteria, this corresponds to a score of 4.

Score- <score>4</score>

K Model Response Example: OMNI-PROMPT VS OP-I-PROMPT

Sample response from `Meta-Llama-3.1-70B-Instruct` as *evaluator* when evaluating multi-source opinion summaries using *both* OMNI-PROMPT and OP-I-PROMPT for the metric `aspect coverage metric`.

Note: We have preserved the original model outputs without any modifications.

K.1 Response using our OMNI-PROMPT for the aspect coverage metric:

Product Title: COSCO CB-80 For (Kids) Multicolor Strung Badminton Racquet (Pack of: 2, 100g) Multicolor Strung Badminton Racquet - Evaluation 1

Response:

1. The summary adheres to the metric of Aspect Coverage very well, covering most of the major features and characteristics of the product. It mentions the durable construction, tempered steel shaft, isometric head shape, G4 grip size, weight, and head size, which are all key elements of the product. Additionally, it highlights the product's suitability for kids and its performance, which is consistent with the reviews and average rating. However, it does not mention the beam width, string tension, or the fact that it comes in a pack of 2, which are also relevant aspects of the product.
2. Based on the evaluation, the summary falls into the 76-94% adherence range, as it covers most of the relevant points, but leaves out a few details.
3. Score- <score>4</score>

Explanation:
The summary provides a comprehensive overview of the product, covering its key features, performance, and suitability for kids. It also mentions the pros and cons, as highlighted in the reviews. However, it misses a few details, such as the beam width and string tension, which are important aspects of the product. Overall, the summary provides a very good idea of the product, but could be improved by including a few more details. Therefore, it falls into the 76-94% adherence range, which corresponds to a score of 4 (Good).

K.2 Response using our OP-I-PROMPT for the aspect coverage metric:

Product Title: COSCO CB-80 For (Kids) Multicolor Strung Badminton Racquet (Pack of: 2, 100g) Multicolor Strung Badminton Racquet - Evaluation 1

Response: Steps to evaluate the summary as per the metric:

1. Identify the major features, characteristics, or attributes of the product that are prominently discussed in the original product information.
2. Check if the summary captures these major features, characteristics, or attributes.
3. Evaluate the extent to which the summary covers these aspects thoroughly.

Metric: Aspect Coverage - Aspect Coverage measures how completely a summary captures the major features, characteristics, or attributes of a product that are prominently discussed in the original product information.

Step-by-step explanation:
The summary captures the following major features, characteristics, or attributes of the product:

- Durable construction with a tempered steel shaft

- Isometric head shape
- G4 grip size
- Weight of 100g
- Head size of 95.5 sq/in
- Ideal for training and developing skills
- Suitable for kids
- Majority of users praise its quality and performance
- A few customers found the size to be too small

The summary covers these aspects thoroughly by providing a detailed description of the product's features and their benefits. It also mentions the pros and cons of the product, providing a balanced view.

Evaluation:
The summary covers all the major features, characteristics, or attributes of the product that are prominently discussed in the original product information. It provides a detailed description of the product's features and their benefits, as well as the pros and cons of the product.

Score- <score>5</score>