



# Bridging Cognition and Affect: Emotion-Aware Opinion Summarization using LLMs

Arnav Attri<sup>◇,\*</sup>,<sup>†</sup>   , Anuj Attri<sup>◇,\*</sup>,<sup>†</sup>  



 Indian Institute of Technology Bombay, India  Plaksha University, India

## Abstract

Opinion summarization systems aggregate customer sentiments without capturing the emotional factors that drive purchasing decisions, resulting in shallow summaries that overlook the affective dimensions shaping customer experiences and fail to explain why customers feel the way they do. This gap exists because prior research has neglected the interplay between expressed opinions and their underlying emotional contexts. To bridge this gap, we introduce EMOTION-AWARE OPINION SUMMARIZATION (EAOS), a framework leveraging Large Language Models (LLMs) to integrate emotional dimensions into opinion summaries, moving beyond conventional sentiment polarity. To support this task, we develop a large-scale (40K product–summary pairs) training dataset, an evaluation benchmark, a compact 1B-parameter model that matches 70B-scale performance via knowledge distillation, and methods for generating and evaluating emotion-aware summaries. A user study shows that 82% of readers prefer our emotion-aware summaries ( $p < .001$ ), confirming that adding emotion helps in making purchase decisions.

## 1 Introduction

The role of emotions in product reviews extends beyond reflecting customer satisfaction; they are important determinants of consumer behavior and critically influence purchasing decisions (Chen et al., 2022; Felbermayr and Nanopoulos, 2016; Pappas et al., 2014). For instance, emotions like joy amplify hedonic consumption, trust reinforces loyalty intentions, and anger triggers avoidance behaviors (Bostan et al., 2020). Reviews with rich

<sup>◇</sup> Equal contribution. <sup>\*</sup> Work done as a graduate student at IIT Bombay. <sup>†</sup> This research was conducted at IIT Bombay in collaboration with the CFILT  and Flipkart . This paper is dedicated to the memory of Prof. Pushpak Bhattacharyya.

### Opinion Summary

The Distance Sea Turtle Necklace for Women receives overwhelmingly positive feedback, with customers praising its beauty, meaningful design, and high-quality craftsmanship. The necklace is described as "beautiful," "simply sweet," and "well made," with some considering purchasing additional items. Despite one instance of a component detaching, overall satisfaction is high, with customers indicating they would purchase again and recommend it to others. The necklace is a popular choice for gifts, particularly for turtle enthusiasts.

### Emotion-Aware Opinion Summary

The Distance Sea Turtle Necklace evokes strong **joy and trust** among customers, who praise its beauty and craftsmanship. Many find the design "simple yet sweet," with some **anticipating** a future purchase of a 1.7mm diamond chain, reflecting satisfaction and desire for more. Despite occasional durability concerns—one customer noted the golden starfish charm fell off after two days—the overall experience remains positive, with the necklace described as "beautiful," "well made," and "great quality." Gift recipients express **surprise** and delight, particularly those who love turtles. The product's ability to bring **joy** and serve as a cherished keepsake is evident, solidifying its status as a heartfelt and well-received gift.

**Table 1:** Comparison of Opinion Summary and Emotion-Aware Opinion Summary generated by Phi-3.5-mini-instruct. The opinion summary reduces affect to sentiment polarity (underlined), while EAOS captures specific emotions from Plutchik's model (**red**).

emotional language induce stronger product attitudes and higher purchase likelihoods (Zablocki et al., 2019), functioning as extrinsic cues, offering signals of product quality and mitigating information asymmetry between consumers and producers (Wang et al., 2019).

Despite this established significance, prevailing opinion summarization methods (Wang and Ling, 2016; Chu and Liu, 2019) exhibit a significant affective blind spot, reducing the richness of consumer sentiment to basic sentiment polarity (negative, neutral, positive). While aspect-based sentiment analysis (ABSA) captures *what* customers feel positive or negative about, it still reduces affect to polarity rather than discrete emotions.

**Distinguishing Our Contribution.** Although emotion recognition is a mature field (Picard, 2000), the generative task of synthesizing affectively nuanced summaries remains underexplored. While LLMs show emergent capabilities in affective reasoning (tse Huang et al., 2024) and abstractive compression (Deroy et al., 2023), their application to emotion-aware summaries in e-commerce is notably absent from the literature. Our work

addresses this gap by focusing on **emotion-aware summarization**, which fundamentally differs from emotion recognition (classifying text into emotion labels). This generative task requires not merely detecting emotions, but coherently weaving their nuance from multiple source documents into narrative prose while maintaining factual accuracy and balanced representation. To achieve this, we introduce EMOTION-AWARE OPINION SUMMARIZATION (EAOS), a framework grounded in Plutchik’s model (Plutchik, 1988) that captures consumer affect through 8 primary emotions—joy, trust, fear, surprise, sadness, disgust, anger, and anticipation—advancing beyond sentiment polarity (Table 1).

### Our Contributions are:

1. **EAOS:** We formalize emotion-aware opinion summarization as a new task and introduce structured prompting frameworks for both generation (EAOS-GEN) and evaluation (EAOS-EVAL), grounded in Plutchik’s emotion model (§4). Ablation analysis confirms each stage is essential (Table 7; §7.2). A user study demonstrates significant preference (82%,  $p < .001$ ) for emotion-aware summaries over traditional opinion summaries (§8).
2. **EAOS-SUMM:** In the absence of supervised data, we construct a large-scale dataset of 40,000 product-summary pairs, generated by high-performing open-source LLMs and validated through human audit (§5.1).
3. **EAOS-LLAMA<sup>1</sup>:** We release a compact 1B parameter model fine-tuned on EAOS-SUMM via knowledge distillation. Despite having 70× fewer parameters, it achieves performance statistically indistinguishable from Llama-3.1-70B (Table 4; §7) and marks a +13% improvement over its non-fine-tuned baseline, validating that our dataset effectively transfers emotion-aware summarization capabilities to compact models suitable for deployment.
4. **EAOS-EVAL-BENCH:** An evaluation benchmark (§5.2) enabling systematic assessment across 7 dimensions: standard summarization quality—**FL** (fluency), **CO** (coherence), **FA** (faithfulness) and four novel

<sup>1</sup>Prompts and the fine-tuned model are available at <https://github.com/arnavcse/EAOS> and <https://huggingface.co/cuteasduck/EAOS-Llama-3.2-1B>.

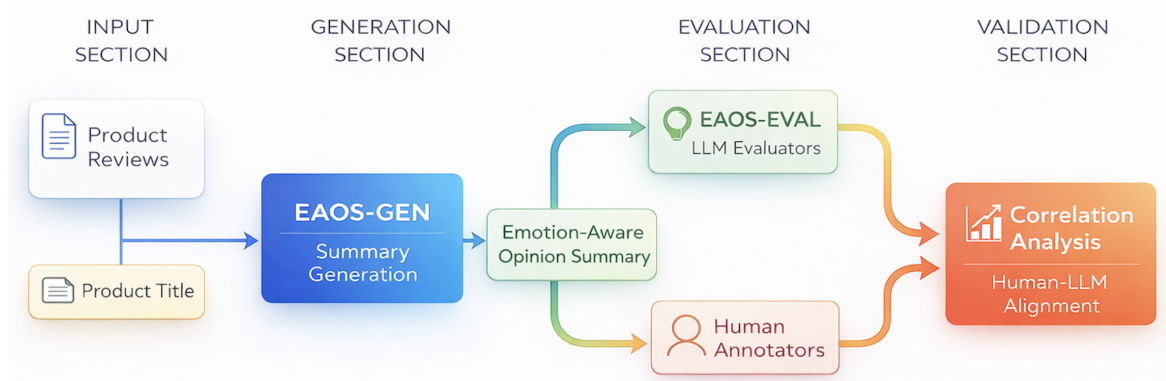
emotion-specific criteria—**EA** (emotional accuracy), **ESC** (emotional spectrum coverage), **EBM** (emotional bias mitigation), **CER** (contextual emotional relevance). Human annotation achieves substantial inter-annotator agreement ( $\alpha = 0.85$ ), confirming benchmark reliability. See **Appendix A.1** for metric definitions.

## 2 Related Work

**Opinion Summarization.** Methods have evolved from early extractive (Erkan and Radev, 2004; Kim et al., 2011) to unsupervised abstractive approaches using VAEs (Bražinskas et al., 2020) and denoising objectives (Amplayo and Lapata, 2020). Recent work targets specific aggregation challenges, including aspect-aware extraction (Li et al., 2023) and rationale-based reasoning (Li and Chaturvedi, 2024). Notably, Zhang et al. (2024) address fair representation of minority perspectives—a concern parallel to our emotional bias mitigation. However, despite LLM-based advances (Bhaskar et al., 2023; Hosking et al., 2023), existing approaches universally reduce affect to sentiment polarity.

**Emotions in NLP.** Emotions shape decision-making (Salovey and Mayer, 1990; Lopes et al., 2004; Malik and Hussain, 2017), leading NLP research to adopt frameworks like Ekman (1992) and Plutchik (1988) for classification (Plutchik, 2000; Mohammad and Bravo-Marquez, 2017; Felbo et al., 2017). While methodologies have advanced from classification to emotion-cause extraction (Ding et al., 2020; Ying et al., 2019; Alhuzali and Ananiadou, 2021; Li and Xiao, 2023) and joint opinion-trigger detection in e-commerce (Attri et al., 2025b), these approaches remain fundamentally extractive. A critical disconnect persists: classification optimizes for label accuracy, whereas summarization prioritizes informativeness. Consequently, the generative task of weaving discrete emotional nuances into coherent narratives remains unexplored, a gap our framework addresses.

**Why ABSA Summarization Is Insufficient.** Aspect-Based Sentiment Analysis (ABSA), which extracts aspect-sentiment pairs from reviews (Pontiki et al., 2016; Schouten and Frasincar, 2016; Zhang et al., 2023), is the closest paradigm to our work but differs in three ways: (1) *Granularity:* ABSA reduces affect to polarity (positive/negative/neutral), losing distinctions between emotions



**Figure 1:** Overview of the EAOS framework. Reviews are processed to produce emotion-aware opinion summaries, which are then independently evaluated by LLM-based and human annotators. Correlation analysis measures human-LLM alignment to validate the evaluation framework.

like joy vs. trust or anger vs. disgust that differentially influence consumer behavior (Pappas et al., 2014); (2) *Task formulation*: ABSA outputs structured tuples (aspect, sentiment), not coherent narratives integrating emotional context; (3) *Emotional depth*: ABSA captures *what* customers feel positive/negative about, but not *how*, missing intensity gradations and complex emotions (e.g., love = joy + trust). We empirically validate this in §7, where ABSA-based baselines underperform on emotion-specific metrics (Appendix G.8).

**LLM-based Evaluation.** Traditional metrics (ROUGE (Lin, 2004), BLEU (Papineni et al., 2002), BERTScore (Zhang et al., 2019)) correlate poorly with human judgments in summarization (Shen and Wan, 2023). To address this, recent work leverages LLMs as scalable evaluators (Fu et al., 2023; Chiang and Lee, 2023a,b; Wang et al., 2023; Kocmi and Federmann, 2023), employing techniques such as chain-of-thought reasoning (Liu et al., 2023; Wei et al., 2023) and reference-free methods (Chiang and Lee, 2023a). Within opinion summarization specifically, Siledar et al. (2024); Attri et al. (2025a) have developed tailored prompting strategies, a paradigm we adopt and validate against human annotation.

### 3 Task Formulation

We formalize emotion-aware opinion summarization as two interlinked tasks:

**Generation.** Given a product  $p$  and customer reviews  $R = \{r_1, \dots, r_m\}$ , generate a summary  $s$  that integrates product aspects with emotional dimensions from reviews, grounded in Plutchik’s emotion set  $\mathcal{E} = \{\text{joy}, \text{trust}, \text{fear}, \text{surprise}, \text{sadness}, \text{disgust},$

$\text{anger}, \text{anticipation}\}$ .

**Evaluation.** Given the tuple  $(s, p, R)$ , produce a quality vector  $\varepsilon = [\varepsilon_1, \dots, \varepsilon_7]$ , where  $\varepsilon_d \in \{1, \dots, 5\}$  scores the dimensions: fluency, coherence, faithfulness, emotional accuracy, spectrum coverage, bias mitigation, and contextual emotional relevance.

**Why Plutchik’s 8 Primary Emotions.** We adopt Plutchik’s framework (Plutchik, 1988, 2000) for four reasons: (1) **Taxonomic precision**—open-ended generation produces redundant emotion labels (synonyms) of basic emotions across LLMs, whereas Plutchik provides a constrained yet expressive taxonomy for reproducible evaluation; (2) **Theoretical grounding**—the model systematically balances positive and negative emotions; (3) **E-commerce relevance**—unlike Ekman (1992)’s 6-emotion model, Plutchik includes trust (critical for brand loyalty) and anticipation (critical for pre-purchase behavior); (4) **Annotation feasibility**—consistent with Mohammad and Turney (2013), annotating hundreds of fine-grained emotions is prohibitively costly.

## 4 Methodology

Our framework comprises two components: EAOS-GEN for summary generation and EAOS-EVAL for multi-dimensional evaluation (Figure 1).

### 4.1 EAOS-GEN Framework

EAOS-GEN guides the LLM through a four-stage reasoning process, implemented as a single unified prompt to maintain efficiency equivalent to standard generation. Each stage builds upon the previous, creating a coherent analytical chain:

**Stage 1: Aspect & Emotion Mapping.** The LLM identifies key product aspects from reviews and maps them to Plutchik’s 8 primary emotions, creating aspect-emotion pairs that capture affective dimensions. For example, “fast shipping” might map to joy and trust, while “battery dies quickly” maps to anger and sadness.

**Stage 2: Emotion Balancing.** To prevent bias toward majority opinions, the framework ensures both prevalent and minority emotional perspectives receive *proportional* representation with a minimum inclusion threshold. Specifically: (1) emotions are represented proportionally to their frequency in source reviews; (2) minority emotions appearing in  $\geq 10\%$  of reviews are explicitly included to prevent oversight of meaningful customer segments; (3) emotions below this threshold are omitted to avoid amplifying noise from isolated outliers. For unanimous sentiment (e.g., all reviews express joy), the summary reflects this consensus without artificially introducing absent emotions. This approach prevents summaries from being skewed toward dominant sentiments while avoiding false balance that would misrepresent the actual distribution (**Appendix C.1**).

**Stage 3: Narrative Integration.** Balanced emotions are woven naturally into prose without explicit emotion labels. Rather than stating “customers feel joy,” the summary conveys emotional content through descriptive language: “customers express delight with the product’s performance.”

**Stage 4: Refinement & Validation.** The summary is refined for coherence and validated for factual faithfulness to source reviews.

## 4.2 EAOS-EVAL Framework

EAOS-EVAL scores each summary across 7 dimensions using metric-specific prompts. For each dimension, the evaluation prompt comprises four components:

- (1) **System Role:** Establishes the LLM as an expert evaluator for the dimension.
- (2) **Task Definition:** Specifies the evaluation objective.
- (3) **Evaluation Criteria:** Provides a 5-point Likert rubric with explicit criteria for each score level.
- (4) **Instruction Sequence:** Prescribes a step-by-step analytical process, decomposing evaluation into discrete operations.

For emotion-specific dimensions (emotional accuracy, spectrum coverage, bias mitigation, contextual relevance), the framework incorporates intensity mappings where emotions are categorized as high, medium, or low intensity (**Appendix C.13**). To validate their contribution, we applied the Steiger–Meng test for dependent correlations (Steiger, 1980; Meng et al., 1992) comparing human–LLM agreement with and without intensity mappings ( $n = 250$ , conservative  $r_{kh} = 0.80$ ). All four emotion-specific dimensions showed significantly higher correlations with intensity mappings (all  $p < .001$ ), while the three standard dimensions were unaffected ( $p > .67$ ), confirming that intensity mappings selectively improve emotion evaluation without introducing noise (**Appendix G.6**).

## 4.3 Evaluation Protocol

**Scoring:** Following Liu et al. (2023), we sample  $n = 50$  Likert scores per summary-dimension pair (temperature = 0.0; all other generation parameters in **Appendix F.12**) and compute the mean, mitigating LLM variance. This yields standard errors of 0.04–0.08, ensuring differences  $\geq 0.15$  are statistically meaningful.

**Validation:** We measure LLM–human alignment using summary-level Spearman correlations (Bhandari et al., 2020), averaged across products. All prompts used in this study are provided in **Appendix I.2–I.1**.

## 5 Dataset

We sample the Amazon Reviews ’23 corpus (Hou et al., 2024) to create a balanced testbed spanning four product categories: Beauty, Home, Electronics, and Clothing, following the domain evaluation protocol established in Bražinskis et al. (2020), which provides sufficient diversity for validating opinion summarization methods while maintaining experimental tractability.

For each category, we draw 25 products without replacement, yielding 100 unique items total. Each product includes a title and 10 reviews filtered to 10–100 tokens to ensure substantive content while excluding outliers. Ratings and “helpful” votes are omitted owing to inflation and weak predictive power (Mayzlin et al., 2012; Guo et al., 2020), while vote counts introduce confounds such as sarcasm, accidental clicks, and herd bias (Lappas et al., 2016; Deng et al., 2020). Human annotators are involved at five stages of the pipeline (**Figure 4**).

## 5.1 EAOS-SUMM

Building supervised resources is expensive; prior work has relied on self-supervision (Bražinskas et al., 2020; Amplayo et al., 2021) or proprietary models (Bhaskar et al., 2023; Liu et al., 2024) that sacrifice reproducibility. We adopt a knowledge distillation approach: *teacher* models (Phi-3.5-mini, Mistral-7B) generate training summaries using EAOS-GEN, and *student* models (EAOS-LLAMA 1B, Qwen 0.5B/1.5B) are fine-tuned on these outputs via standard supervised fine-tuning.

**Generator Selection.** We selected teacher models via a systematic two-stage evaluation. First, comparisons on a 50-product pilot identified **Phi-3.5-mini** and **Mistral-7B-v0.3** as top performers, offering complementary strengths in emotional accuracy and faithfulness (Appendix E.1). Second, a cost-benefit analysis revealed that scaling to 70B+ models (Llama-3.1-70B, Qwen2.5-72B) yielded negligible quality gains (< 1.5%) despite an 18× increase in inference costs and a tendency toward verbose, hedged outputs. Consequently, we generated EAOS-SUMM using the selected mid-size teachers. The final corpus comprises 40,000 product-summary pairs spanning 400,000 reviews (avg. 33.95 words/review, 121.70 words/summary; Table 16), stratified into two disjoint sets of 20,000 each (Set A: Phi-3.5, Set B: Mistral-7B) to separate generator biases.

**Quality Audit.** We validated the dataset through a stratified random audit of 1,000 samples. Independent annotators evaluated factual accuracy, emotion presence, coherence, and relevance. The audit yielded a **94.2% acceptance rate** (Cohen’s  $\kappa = 0.89$ ), confirming that the distillation process maintains high fidelity (Appendix G.12).

**Human Validation & Ceiling.** To establish an upper bound, three expert annotators created “gold standard” emotion-aware summaries for 100 products (Appendix G.13). We benchmarked these against our synthetic summaries and top models.

(1) **Performance Gap:** As shown in Table 4, human summaries achieve near-perfect scores on bias mitigation (EBM: 4.72). While our models match humans on Fluency and Faithfulness, a gap remains in *Bias Mitigation*, highlighting the difficulty of balancing minority emotional perspectives.

(2) **Synthetic Alignment:** EAOS-SUMM summaries show strong semantic alignment with human gold standards (Sentence-BERT similarity: 0.76).

Difference	Human Summary	LLM Summary
<i>Emotion Specificity</i>	“Customers were <b>delighted</b> by the intricate craftsmanship and <b>charmed</b> by the turtle motif”	“Customers <b>feel joy</b> regarding the design quality and <b>express trust</b> in the craftsmanship”
<i>Hedging Behavior</i>	“The necklace brings genuine happiness to gift recipients”	“ <b>Some customers feel</b> that the necklace <b>may bring</b> happiness to recipients”
<i>Structural Patterns</i>	Varied narrative: opens with gift-giving context, weaves quality observations throughout	Formulaic: positive aspects → negative concerns → overall conclusion

**Table 2:** Systematic differences between human and LLM-generated summaries. Humans use nuanced vocabulary, whereas LLMs often default to primary emotion labels and hedging.

(3) **Stylistic Differences:** Qualitative analysis (Table 2) reveals that while content overlaps significantly (Jaccard: 0.73), humans use more specific emotional vocabulary (e.g., “delighted” vs. “happy”), whereas LLMs exhibit more formulaic structures (Appendix G.14).

**Teacher Bias Analysis.** We critically assessed whether teacher models propagate harmful biases. Both teachers exhibit a mild “positivity bump” (amplifying Joy/Trust by 3–5%), consistent with RLHF alignment (Perez et al., 2023) (Appendix E.2). However, the Jensen-Shannon divergence between teacher and human emotion distributions is extremely low ( $D_{JS} < 0.05$ ), indicating high distributional alignment. Manual inspection confirms that 67% of “hallucinated” emotions were justifiable inferences from implicit cues (Appendix G.7). We explicitly acknowledge the remaining mild positivity bias as a limitation of synthetic distillation.

## 5.2 EAOS-EVAL-BENCH

The evaluation benchmark comprises 30 summaries (26 models, 3 methodological baselines, and 1 human gold standard) for each of 100 products, yielding 63,000 ratings (3 raters × 100 products × 30 summaries × 7 dimensions). We engaged 3 graduate researchers, avoiding crowd-sourcing to ensure high annotation quality (Gillick and Liu, 2010; Fabbri et al., 2021). Following the two-stage protocol of Fabbri et al. (2021): (i) Round I involved independent scoring; (ii) Round II involved moderated reconciliation for dimensions where ratings diverged by  $\geq 2$  points. All summaries were anonymized to prevent model-identification bias; LLM evaluators followed an identical protocol (Appendix F.1).

	Round-I $\uparrow$	Round-II $\uparrow$
fluency	0.70	0.87
coherence	0.68	0.84
faithfulness	0.79	0.89
emotional accuracy	0.72	0.86
emotional spectrum coverage	0.78	0.82
emotional bias mitigation	0.73	0.83
contextual emotional relevance	0.68	0.81
<b>AVG</b>	0.73	0.85

**Table 3: Inter-rater agreement scores** for Round-I and Round-II across 7 dimensions. An improvement in agreement scores is observed in Round-II.

### 5.3 Annotation Analysis

Krippendorff’s  $\alpha$  (Krippendorff, 2011) improved from 0.73 (Round I) to 0.85 (Round II), elevating agreement from moderate to substantial (Table 3). Faithfulness and Emotional Spectrum Coverage consistently yielded high agreement ( $\alpha_{fa} : 0.79 \rightarrow 0.89$ ,  $\alpha_{esc} : 0.78 \rightarrow 0.82$ ). Faithfulness benefits from objective verifiability against source reviews, while Emotional Spectrum Coverage benefits from Plutchik’s structured taxonomy providing clear annotation guidance. Subjective dimensions such as Coherence and Contextual Emotional Relevance showed lower initial agreement ( $\alpha = 0.68$ ) but demonstrated substantial gains after moderated reconciliation ( $\alpha_{co} : 0.68 \rightarrow 0.84$ ,  $\alpha_{cer} : 0.68 \rightarrow 0.81$ ), confirming the effectiveness of the two-stage protocol. The aggregate  $\alpha = 0.85$  establishes EAOS-EVAL-BENCH as a reliable benchmark.

## 6 Experiments

We evaluate our framework through two complementary studies: assessing summary quality across 26 models using EAOS-GEN (§4.1), and validating LLM-based evaluation against human annotations using EAOS-EVAL (§4.2).

### 6.1 EAOS-GEN (Summary Generation)

**Validation Strategy.** As the first benchmark for emotion-aware opinion summarization, we validate through three complementary approaches: (1) cross-model evaluation across 26 models, (2) human-LLM correlation analysis, and (3) user preference studies.

**PLM Baselines.** We evaluate three canonical sequence-to-sequence architectures—BART-large (Lewis et al., 2019), T5-large (Raffel et al., 2020), and PEGASUS-large (Zhang et al., 2020)—in zero-shot mode, serving as illustrative baselines to contextualize the advancement from pre-LLM

architectures to instruction-tuned models. For a fair comparison, we also fine-tune T5-large on a 10,000-sample subset of EAOS-SUMM for 3 epochs (T5-large-FT), testing whether traditional encoder-decoder architectures can learn emotion-aware summarization with appropriate supervision.

**Methodological Baselines.** To isolate the contribution of our EAOS-GEN framework, we implement three methodological baselines using Phi-3.5-mini-instruct (our top performer): (1) **Simple Emotion Prompting:** Standard summarization prompt augmented with “Include emotional content from reviews”; (2) **Two-Stage Pipeline:** First generate a standard opinion summary, then post-process to add emotional content; (3) **ABSA-Summary:** Extract aspect-sentiment tuples using PyABSA (Yang et al., 2023), a state-of-the-art ABSA toolkit, then prompt the LLM to generate a summary conditioned on these tuples.

**LLMs.** Under a strict zero-shot protocol, we evaluate OpenAI’s proprietary GPT-4o alongside a suite of open-source foundation models (Appendix F.9), prioritizing open-weight models (LLAMA-3, QWEN-2.5) that support reproducible fine-tuning and ablation studies.

**Fine-Tuned Compact LLMs.** To examine whether our dataset effectively teaches emotion-aware summarization, we fine-tune three lightweight instruction-tuned backbones—Llama-3.2-1B-Instruct<sup>2</sup>, Qwen2.5-0.5B-Instruct, and Qwen2.5-1.5B-Instruct—on EAOS-SUMM.

### 6.2 EAOS-EVAL (Summary Evaluation)

Following the LLM-as-evaluator paradigm (§4.2), we employ 6 open-source LLMs and GPT-4o to assess summaries across our 7-dimension rubric, measuring alignment with human annotations via Spearman and Kendall-Tau correlations (Appendix F.12; evaluation prompts in Appendix I.3).

## 7 Results and Analysis

Table 4 presents human annotator ratings across 7 dimensions for 26 models. We analyze key patterns below.

**Overall Model Performance.** Phi-3.5-mini-instruct achieves the highest overall score (4.60), followed by GPT-4o (4.48) and Mistral-7B-v0.3

<sup>2</sup>We refer to the fine-tuned Llama-3.2-1B-Instruct as EAOS-LLAMA throughout this paper.

Models	FL $\uparrow$	CO $\uparrow$	FA $\uparrow$	EA $\uparrow$	ESC $\uparrow$	EBM $\uparrow$	CER $\uparrow$	Avg $\uparrow$
<i>Human Ceiling</i>								
Human Gold Standard	4.92	<b>4.85</b>	<b>4.88</b>	4.78	4.65	<b>4.72</b>	<b>4.81</b>	<b>4.80</b>
<i>PLMs (Pre-trained Language Models)</i>								
Bart-large	2.16	1.69	4.75 <sup>†</sup>	2.05	2.01	1.04	0.91	2.09
T5-large	2.01	1.46	4.65	2.15	2.20	1.23	1.23	2.13
Pegasus-large	2.39	1.48	4.70	2.43	2.38	1.17	1.03	2.22
T5-large-FT	4.52	3.85	4.35	3.48	3.12	3.05	3.22	3.66
<i>Methodological Baselines (using Phi-3.5-mini-instruct)</i>								
Simple-Emotion-Prompting	4.95	4.38	4.62	3.85	3.42	3.45	3.78	4.06
Two-Stage-Pipeline	4.92	4.25	4.55	3.72	3.28	3.38	3.65	3.96
ABSA-Summary	4.88	4.15	4.48	3.21	2.85	3.12	3.08	3.68
<i>LLMs</i>								
Meta-Llama-3.2-1B-Instruct	4.74	3.90	4.10	3.64	3.24	3.30	3.68	3.80
Qwen2.5-0.5B-Instruct	4.77	3.69	2.68	3.30	3.22	2.80	3.58	3.43
Qwen2.5-1.5B-Instruct	4.78	3.99	2.79	3.88	3.37	2.96	3.89	3.67
Phi-3.5-mini-instruct	<u>4.98</u>	<u>4.47</u>	<u>4.70</u>	<b>4.59</b>	<b>4.80</b>	<u>4.06</u>	<u>4.61</u>	<u>4.60</u>
Mistral-7B-Instruct-v0.2	4.97	4.06	4.23	4.38	4.59	3.90	4.31	4.35
Mistral-7B-Instruct-v0.3	4.96	4.05	4.60	4.44	4.46	3.85	4.45	4.40
zephyr-7b-beta	4.70	4.00	4.20	4.40	4.44	3.84	4.30	4.27
Qwen2.5-7B-Instruct	4.90	4.01	4.37	4.26	4.03	3.80	4.35	4.25
Meta-Llama-3.1-8B-Instruct	4.95	4.05	4.37	4.10	3.85	3.79	4.17	4.18
Gemma-2-9b-it	4.96	4.04	4.50	4.27	4.22	3.68	4.41	4.30
SOLAR-10.7B-Instruct-v1.0	4.65	4.06	4.21	4.48	4.36	3.88	4.17	4.26
Mistral-Nemo-Instruct-2407	4.91	4.05	4.13	4.44	4.39	3.83	4.48	4.32
Qwen2.5-14B-Instruct	4.95	4.04	4.56	4.46	4.38	3.92	4.41	4.39
Mistral-Small-Instruct-2409	4.97	4.03	4.59	4.38	4.30	3.84	4.44	4.36
Mixtral-8x7B-Instruct-v0.1	4.89	3.95	4.50	4.40	4.40	3.91	4.44	4.36
Qwen2.5-32B-Instruct	4.97	4.03	4.56	<u>4.50</u>	4.11	3.98	4.56	4.39
Meta-Llama-3.1-70B-Instruct	<u>4.98</u>	4.03	4.59	4.21	4.30	3.75	4.39	4.32
Qwen2.5-72B-Instruct	4.97	4.05	4.59	4.38	4.29	3.93	4.48	4.38
GPT-4o	<u>4.98</u>	4.51	4.68	4.42	4.24	4.00	4.50	4.48
<i>Fine-Tuned Compact LLMs</i>								
EAOS-LLAMA (1B)	4.92	4.13	4.35	4.40	4.31	3.74	4.26	4.30
Qwen2.5-0.5B-Instruct-FineTuned	4.96	4.04	3.43	4.54	4.28	2.62	4.16	4.00
Qwen2.5-1.5B-Instruct-FineTuned	4.93	4.05	3.57	4.40	4.31	2.61	4.20	4.01

**Table 4:** Mean annotator ratings ( $\uparrow$ ) across 7 dimensions ( $n = 100$  products). Human Gold Standard establishes the performance ceiling (4.80 avg). Best model scores are **bold**, second-highest are underlined.  $\uparrow$ PLMs copy reviews verbatim, inflating faithfulness artificially (§7). See Standard deviations table in **Appendix D.4**.

(4.40). PLMs without instruction-tuning perform poorly in this zero-shot setting, reflecting architectural differences rather than direct competition (see §6).

**Human Ceiling Comparison.** Human-written summaries (Table 4) achieve 4.80 average, establishing the ceiling for this task. Our best model reaches 96% of human performance. LLMs match or exceed humans on fluency (101%) and emotional spectrum coverage (103%), but consistently underperform on emotional bias mitigation (86%), indicating that balancing majority/minority perspectives remains challenging. EAOS-LLAMA (1B) achieves 90% of human performance despite 70x fewer parameters than the largest models.

**PLM Behavior Analysis.** BART, T5, and PEGASUS score high on faithfulness but poorly elsewhere. Without task-specific fine-tuning, these models default to extractive behavior, producing outputs that are technically faithful but lack emotional integration. T5-large-FT (fine-tuned on EAOS-SUMM) achieves 3.66 average—substantially outperforming zero-shot PLMs (2.09–2.22) but still underperforming instruction-tuned LLMs, suggesting encoder-decoder architectures face inherent limitations for this generative task.

Method	Output Excerpt (Sea Turtle Necklace)
EAOS (Ours)	“The necklace has garnered <i>joy and trust</i> in its beauty and craftsmanship. Many find the design simple yet sweet, with some <i>anticipating</i> future purchases. Despite occasional <i>concern</i> about durability, the overall experience reflects <i>delight</i> and appreciation...”
Simple Emotion Prompting	“The necklace received mostly positive reviews. Customers felt <i>happy</i> about the quality and <i>satisfied</i> with the purchase. Some were <i>disappointed</i> about durability. Overall sentiment is positive.”
Two-Stage Pipeline	“Beautiful necklace with good quality chain. Well-made and looks great. Some issues with parts falling off. [Emotions added: positive feelings, minor concerns]”
ABSA-Summary <sup>3</sup>	“Design: Positive. Quality: Positive. Durability: Mixed. Price: Positive. Gift suitability: Positive. Overall: 4/5 positive aspects.”

**Table 5:** Qualitative comparison of summarization approaches. EAOS naturally integrates specific emotions (*italicized*) within narrative prose; other methods use generic labels or reduce affect to polarity.

**Methodological Baselines Analysis.** Our methodological baselines (Table 5) reveal EAOS-GEN’s contribution. **Simple Emotion Prompting** (4.06) underperforms full EAOS (4.60) by 12%, with largest gaps in Emotional Accuracy (−16%) and Spectrum Coverage (−29%). **Two-Stage Pipeline** (3.96) disrupts narrative coherence through post-hoc emotion injection. **ABSA-Summary** (3.68) fails on emotion-specific dimensions (EA: 3.21, ESC: 2.85), demonstrating that polarity-based representation cannot capture the emotional granularity achieved through Plutchik’s framework.

**Fine-Tuning Effectiveness.** Fine-tuning on EAOS-SUMM yields substantial gains. EAOS-LLAMA improves from 3.80 to 4.30 (+13%), statistically matching Llama-3.1-70B ( $p = .57$ ) while using 70x fewer parameters (**Appendix G.1**). Phi-3.5-mini significantly outperforms GPT-4o ( $p < .001$ ), and critically, EAOS-LLAMA shows no significant difference from models 32–70x larger, validating EAOS-SUMM’s effectiveness for knowledge distillation.

**Faithfulness-Emotion Tradeoff.** Fine-tuned Qwen models achieve high Emotional Accuracy but substantially lower Faithfulness (FA) and Bias Mitigation (EBM). The strong FA-EBM correlation ( $r = 0.71$ ,  $p < .001$ ) indicates that models sacrificing faithfulness tend to exhibit emotional bias—over-representing emotions at the expense of grounding (**Appendix D.5**). In contrast, EAOS-LLAMA achieves optimal balance (FA: 4.35, EA: 4.40, EBM: 3.74), suggesting the Llama-3.2-1B architecture better preserves source grounding during fine-tuning.

	Evaluator LLM	FL $\uparrow$		CO $\uparrow$		FA $\uparrow$		EA $\uparrow$		ESC $\uparrow$		EBM $\uparrow$		CER $\uparrow$		Avg $\uparrow$	
		$\rho$	$\tau$	$\rho$	$\tau$	$\rho$	$\tau$	$\rho$	$\tau$	$\rho$	$\tau$	$\rho$	$\tau$	$\rho$	$\tau$	$\rho$	$\tau$
		Amazon Reviews '23	Llama-3.1-8B-Instruct	0.68	0.50	<u>0.67</u>	0.50	0.61	<u>0.46</u>	<u>0.68</u>	<u>0.54</u>	0.58	<u>0.39</u>	<u>0.67</u>	<b>0.59*</b>	0.60	0.43
	Mistral-7B-Instruct-v0.2	<u>0.61</u>	<u>0.46</u>	0.68	<u>0.50</u>	<b>0.77*</b>	<b>0.63*</b>	<u>0.60</u>	<u>0.43</u>	0.68	<u>0.57</u>	0.67	<u>0.55</u>	<u>0.54</u>	<u>0.67</u>	0.65	0.54
	Mistral-7B-Instruct-v0.3	<u>0.60</u>	<u>0.50</u>	0.60	<u>0.50</u>	<u>0.60</u>	<u>0.43</u>	<u>0.67</u>	0.48	0.67	<u>0.50</u>	0.67	<u>0.55</u>	<u>0.54</u>	<u>0.67</u>	0.62	0.52
	Llama-3.1-70B-Instruct	<b>0.77*</b>	<b>0.61*</b>	0.67	<u>0.48</u>	0.75	0.63	0.67	0.57	0.73	0.54	0.67	0.59	0.63	<u>0.82*</u>	0.70	<b>0.61*</b>
	GPT-4o	<u>0.70</u>	<u>0.56</u>	0.68	0.50	0.77*	0.63	<b>0.82*</b>	<b>0.65*</b>	0.73	0.54	0.67	<u>0.46</u>	0.65	<u>0.57</u>	<u>0.72</u>	0.56
	Qwen2.5-72B-Instruct	<u>0.67</u>	<u>0.50</u>	0.67	0.55	0.68	<u>0.57</u>	<u>0.68</u>	0.57	<u>0.69</u>	<u>0.55</u>	0.67	0.55	<b>0.67</b>	0.48	0.68	0.54
	Qwen2.5-72B-Instruct	0.75	<u>0.59</u>	<b>0.72</b>	<b>0.59</b>	0.77	0.63	<b>0.82*</b>	<b>0.65*</b>	<b>0.74</b>	<b>0.62*</b>	0.68	0.46	0.66	0.46	<b>0.73*</b>	<u>0.57</u>

**Table 6:** Spearman ( $\rho$ ) and Kendall Tau ( $\tau$ ) correlations at the summary level across 7 dimensions. Best scores are **bold**, and the second-best are underlined. \* indicates ( $p$ -value  $< 0.05$ ).

**LLMs as Evaluators.** Table 6 shows Qwen2.5-72B-Instruct achieving the highest human-LLM correlation ( $\rho = 0.73$ ), followed by GPT-4o (0.72) and Llama-3.1-70B (0.70). We observe mild self-preference bias: Llama-family evaluators rate Llama-generated summaries 0.18 points higher than Qwen-72B does. To mitigate this, we use cross-family evaluators and report human-LLM correlations as the primary validity metric. Sample model outputs are provided in (Appendix H.8).

**Complex Emotion Evaluation.** To assess whether LLM evaluators genuinely understand complex emotions rather than performing keyword matching, we compared accuracy on complex (87.3%) versus primary (92.1%) emotion identification. This modest gap suggests robust compositional understanding. Furthermore, when summaries expressed complex emotions through constituent primaries without explicit labels, evaluators correctly inferred them 83.6% of the time (Appendix C.6; Table 12).

**Traditional Metrics.** For completeness, we report ROUGE and BERTScore in Appendix B.1. Consistent with prior findings, these metrics show weak correlation with emotion-specific dimensions (ROUGE-L:  $r = 0.34$  with EA,  $r = 0.28$  with ESC; BERTScore:  $r = 0.42$  with EA,  $r = 0.38$  with ESC), confirming they capture lexical and semantic overlap but cannot assess emotional nuances. This justifies our emphasis on LLM-based and human evaluation.

## 7.1 Error Analysis

We identify four systematic failure modes across model families.

**(1) Emotion Hallucination.** Teacher models assign emotions absent from human annotations in a non-trivial minority of cases. Manual inspection of 100 such instances reveals that 67% are contextually justifiable inferences from implicit textual cues, while 33% represent genuine over-attribution (Appendix G.7). Positive emotions (Joy, Trust) are

disproportionately hallucinated, consistent with the RLHF positivity bias documented in our teacher bias analysis (Table 15).

**(2) Faithfulness–Emotion Tradeoff.** Fine-tuned Qwen models achieve high Emotional Accuracy but sacrifice Faithfulness and Bias Mitigation. The strong FA–EBM correlation ( $r = 0.71$ ,  $p < .001$ ) suggests a systematic pattern: models that over-represent emotions tend to fabricate or exaggerate emotional content (Appendix D.5). EAOS-LLAMA avoids this tradeoff (FA: 4.35, EBM: 3.74), suggesting architecture choice matters.

**(3) Complex Emotion Degradation.** LLM evaluators show a modest accuracy gap between complex (87.3%) and primary (92.1%) emotion identification, with failures concentrated in emotions requiring compositional reasoning (e.g., “love” as joy + trust). This 4.8 percentage-point gap affects both generation and evaluation (Appendix C.6).

**(4) Architectural Limitations.** PLMs (BART, T5, PEGASUS) default to extractive behavior in zero-shot settings, producing faithful but emotionally flat outputs. Even with fine-tuning, T5-large-FT (3.66) substantially underperforms instruction-tuned LLMs, indicating that encoder-decoder architectures face inherent limitations for this generative, affect-rich task.

## 7.2 Ablation Analysis

All ablation experiments use Phi-3.5-mini-instruct (our top-performing generator; Table 4) evaluated by EAOS-EVAL with Qwen2.5-72B-Instruct ( $\rho = 0.73$  with human ratings; Table 6). The full-configuration score (4.60 avg) matches the human-evaluated result in Table 4, cross-validating the automated evaluation.

To validate each EAOS-GEN stage, we systematically removed individual steps and measured performance degradation (Table 7). Each stage addresses distinct quality dimensions: removing Stage 2 (Emotion Balancing) causes the largest single-dimension drop in EBM ( $-1.22$ );

Configuration	FL $\uparrow$	CO $\uparrow$	FA $\uparrow$	EA $\uparrow$	ESC $\uparrow$	EBM $\uparrow$	CER $\uparrow$	Avg $\uparrow$
Full EAOS-GEN	4.98 $\pm$ .08	4.47 $\pm$ .25	4.70 $\pm$ .22	4.59 $\pm$ .28	4.80 $\pm$ .15	4.06 $\pm$ .35	4.61 $\pm$ .30	4.60 $\pm$ .09
w/o Stage 1	4.76 $\pm$ .18	4.12 $\pm$ .32	4.52 $\pm$ .28	3.82 $\pm$ .38	4.22 $\pm$ .30	3.92 $\pm$ .38	3.72 $\pm$ .42	4.15 $\pm$ .13
w/o Stage 2	4.82 $\pm$ .15	4.22 $\pm$ .30	4.61 $\pm$ .25	4.12 $\pm$ .35	4.32 $\pm$ .28	2.84 $\pm$ .45	4.02 $\pm$ .38	4.14 $\pm$ .12
w/o Stage 3	4.52 $\pm$ .28	3.64 $\pm$ .42	4.66 $\pm$ .24	3.92 $\pm$ .38	4.52 $\pm$ .25	3.72 $\pm$ .40	3.54 $\pm$ .45	4.07 $\pm$ .13
w/o Stage 4	4.24 $\pm$ .35	3.84 $\pm$ .38	4.58 $\pm$ .26	4.22 $\pm$ .32	4.42 $\pm$ .28	3.88 $\pm$ .38	4.12 $\pm$ .35	4.19 $\pm$ .13

**Table 7:** Ablation study on EAOS-GEN stages ( $n = 50$  products). All ablations significantly degrade performance (paired  $t$ -tests, all  $p < .001$ ).

removing Stage 3 (Narrative Integration) most severely impacts coherence ( $-0.83$ ) by producing mechanical emotion listings; removing Stage 1 (Aspect-Emotion Mapping) degrades EA and CER through superficial emotion associations. The full framework outperforms any ablation by 11–13%, demonstrating that all stages are essential (**Appendix G.2**).

### 7.3 Prompt Robustness

We tested 5 prompt variations (reordered instructions, simplified language, expanded definitions, bullet-point format, and combinations; **Appendix G.3; Table 19**) to assess sensitivity to phrasing. All variations scored within  $\pm 0.06$  points (1.3%) of the original, demonstrating that EAOS-GEN’s effectiveness stems from its structured four-stage design rather than specific wording.

## 8 User Study: EAOS Effectiveness

We recruited 350 participants (ages 18–55,  $M=31.4$ ,  $SD=7.8$ ) to evaluate 4 summary pairs (Opinion vs. EAOS). Opinion summaries used standard non-emotional prompting (**Appendix I.1**); EAOS summaries used our framework (§4.1), both generated by our top-performing model, Phi-3.5-mini-instruct (Table 4). To mitigate bias, summaries were labeled neutrally (“Summary 1” and “Summary 2”) with counterbalanced, randomized ordering. Across 5 criteria, participants showed strong preference for EAOS (82%,  $p < .001$ ; **Figure 2**). **Quality Preservation:** We tested whether emotion-aware summarization sacrifices standard quality by evaluating both summary types from the same models on the 3 standard dimensions (FL, CO, FA) applicable to both. Differences are negligible ( $<0.05$  points, within annotator variance), demonstrating that EAOS adds emotional dimensions without sacrificing standard quality (**Appendix H.6**).

## 9 Conclusion

We introduce EAOS, a framework that moves opinion summarization beyond sentiment polarity to capture discrete emotions that shape purchasing de-

cisions. A compact 1B-parameter model (EAOS-LLAMA), trained via knowledge distillation on EAOS-SUMM, statistically matches 70B-scale performance ( $p = 0.57$ ), validating that emotion-aware summarization is learnable rather than emergent. User preference strongly favors emotion-aware summaries (82%,  $p < .001$ ), confirming that emotional dimensions are not merely additive but decision-relevant. Our benchmark (EAOS-EVAL-BENCH,  $\alpha = 0.85$ ) enables systematic evaluation, while ablations confirm each stage is essential. The core insight: capturing which emotions customers express, not just polarity, fundamentally changes what summaries can communicate.

## Limitations

The empirical validation presented herein systematically examines a comprehensive suite of leading LLMs. The exclusion of several maximally-scaled proprietary systems was not due to methodological oversight, but was instead dictated by external constraints. In particular, the OpenAI models o1 and o3 were unavailable for both summary generation and evaluation because they are restricted to premium API access tiers (Tiers 4 and 5). Likewise, GPT-4.5 was excluded owing to its substantial operational cost (Input: \$75/M tokens, Output: \$150/M tokens).

Moreover, mitigating inherent model stochasticity required extensive repeated assessments ( $n = 50$  evaluations per summary), further amplifying the financial burden of proprietary models in an academic research context.

Our model selection reflects availability as of early 2025. As this work introduces a new task rather than benchmarking model capabilities, we release our dataset and evaluation framework to enable community evaluation on emerging architectures.

**User Study Limitations.** Our user study did not explicitly control for summary length. EAOS summaries were on average 42% longer than traditional opinion summaries (135 vs. 95 words in Table 1), as emotion-aware content naturally requires additional explanatory text. While the 82% preference for EAOS remains substantial, we acknowledge that this confound may partially explain the observed preference; length-matched comparisons would be needed to fully isolate the contribution of emotional content from summary verbosity. Additionally, our study was conducted in

English only; cross-cultural variations in emotional expression and interpretation remain unexplored. More broadly, Plutchik’s taxonomy reflects Western psychological traditions and may not universally capture emotional categories salient in other cultures. Amazon reviews also reflect the demographic composition of its user base, potentially under-representing emotional expression patterns from non-Western consumers. LLM evaluators may share systematic biases inherited from predominantly English training corpora, which future work should address through multilingual validation.

**Evaluation Circularity.** We acknowledge a structural circularity in our evaluation pipeline: EAOS-SUMM is generated by LLMs (Phi-3.5, Mistral-7B), EAOS-EVAL-BENCH contains summaries from the same LLM families, and LLM evaluators assess LLM-generated summaries. This raises a valid concern: are we measuring genuine emotion-aware summarization quality, or LLM-style fluency that happens to include emotion words? We mitigate this concern through several validation mechanisms: (1) human annotators independently evaluate summaries with substantial inter-rater agreement ( $\alpha = 0.85$ ), providing non-LLM quality signals; (2) our human gold standard establishes a ceiling (4.80 average) against which we measure LLM performance, with top models achieving 96% of human quality; (3) systematic differences analysis reveals that while LLMs exhibit stylistic differences from humans, content overlap remains high (Jaccard similarity = 0.73 for emotion-aspect pairs); (4) the user study directly measures human preference without LLM intermediation; and (5) we report traditional automatic metrics (ROUGE, BERTScore) as LLM-independent baselines. While these measures do not eliminate circularity entirely, they provide converging evidence from multiple independent sources that EAOS captures qualities valued by human judges, not merely LLM artifacts.

**Synthetic Data and Teacher Model Bias.** EAOS-SUMM is constructed via knowledge distillation from two teacher models (Phi-3.5-mini, Mistral-7B), raising the concern that EAOS-LLAMA learns to mimic teacher artifacts rather than genuine emotion-aware summarization. Our analysis (Table 15) reveals that teachers exhibit systematic biases: positive emotions (Joy, Trust) are over-represented by 4–6% relative to human anno-

tations, while negative emotions (Anger, Disgust) are attenuated by 2–4%. This “positivity bias” is well-documented in instruction-tuned LLMs and likely propagates to EAOS-LLAMA. We partially mitigate this through our dual-teacher design: Phi-3.5 and Mistral-7B exhibit complementary biases (e.g., Phi-3.5 over-detects Anticipation while Mistral-7B better captures Fear), and training on their combined outputs provides implicit bias averaging. However, three risks remain: (1) *Emotion hallucination*: In 33% of examined cases where teachers assigned emotions absent from human annotations, these represented genuine over-attribution rather than justifiable inference—EAOS-LLAMA may inherit this tendency; (2) *Style artifact propagation*: Teacher-specific patterns (hedging language, formulaic structures) may be encoded as “correct” summarization style; (3) *Distribution shift*: Models trained on e-commerce reviews may not generalize to domains with different emotional expression norms (e.g., healthcare, legal). These limitations could be addressed through human-in-the-loop verification for edge cases, diverse teacher ensembles spanning different model families, and explicit debiasing techniques during fine-tuning. We release emotion distribution statistics alongside EAOS-SUMM to enable researchers to calibrate for these known biases.

## Acknowledgments

We are deeply grateful to the late Prof. Pushpak Bhattacharyya for his invaluable guidance and intellectual contributions to this work. We also thank Nikesh Garera for his significant contributions and mentorship during the research collaboration with Flipkart. This work was conducted as part of the sponsored research collaboration between CFILT, IIT Bombay and Flipkart on Customer Review Analytics. We further acknowledge Plaksha University for providing a supportive environment during the later stages of this work, where portions of the paper were refined and finalized. *This paper is dedicated to the memory of Prof. Pushpak Bhattacharyya.*

## Ethical Considerations

We engaged 3 raters with diverse academic backgrounds: a Master’s student, a Pre-Doctoral researcher, and a Doctoral candidate, aged 24–32, with publications or active research in opinion summarization and an academic foundation in Con-

sumer Psychology. Raters were appropriately compensated for their contributions.

The objective of this research is to incorporate awareness of emotional dimensions present in customer reviews into opinion summaries. The Large Language Models used do not experience or replicate human emotions. Rather, they are tools for identifying and synthesizing linguistic patterns correlated with emotional states as expressed by customers.

**Potential for Misuse and Societal Impact.** We acknowledge that emotion-aware systems could potentially be misused by e-commerce platforms for manipulative purposes for example, crafting summaries that exploit emotional vulnerabilities to drive purchases, or selectively emphasizing certain emotions to influence consumer behavior. The ability to capture nuanced emotions may benefit large e-commerce platforms over individual customers by enabling more sophisticated manipulation of emotion-based decision-making.

To mitigate these risks, we advocate for:

1. **Transparency requirements:** Ensuring customers know when emotion-aware analysis is deployed in product summaries
2. **Regulatory frameworks:** Protecting consumer emotional autonomy through appropriate oversight
3. **Responsible application:** Prioritizing applications that genuinely enhance customer decision-making rather than exploit emotional vulnerabilities

Importantly, our framework’s explicit emotional dimensions actually enable greater transparency compared to opaque recommendation systems currently deployed. By making emotional content visible and structured through Plutchik’s well-defined categories, EAOS allows for auditing and accountability that black-box systems do not provide. We believe this transparency is a strength, enabling stakeholders to verify that emotional representation is balanced and faithful to source reviews.

**Use of AI Assistants.** During the preparation of this manuscript, we utilized AI assistants (e.g., ChatGPT) exclusively to refine the grammar and improve the readability of the text. All ideas, analyses, and conclusions remain our own.

## References

- AI@Meta. 2024. [Llama 3 model card](#).
- Hassan Alhuzali and Sophia Ananiadou. 2021. [SpanEmo: Casting multi-label emotion classification as span-prediction](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1573–1584, Online. Association for Computational Linguistics.
- Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata. 2021. Aspect-controllable opinion summarization. *arXiv preprint arXiv:2109.03171*.
- Reinald Kim Amplayo and Mirella Lapata. 2020. Un-supervised opinion summarization with noising and denoising. *arXiv preprint arXiv:2004.10150*.
- Anuj Attri, Arnab Attri, Pushpak Bhattacharyya, Suman Banerjee, Amey Patil, Muthusamy Chelliah, and Nikesh Garera. 2025a. [LLMs as architects and critics for multi-source opinion summarization](#).
- Arnab Attri, Anuj Attri, Suman Banerjee, Amey Patil, Muthusamy Chelliah, Nikesh Garera, and Pushpak Bhattacharyya. 2025b. [Why we feel what we feel: Joint detection of emotions and their opinion triggers in E-commerce](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5515–5532, Suzhou, China. Association for Computational Linguistics.
- 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.
- Laura Ana Maria Bostan, Evgeny Kim, and Roman Klinger. 2020. [GoodNewsEveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1554–1566, Marseille, France. European Language Resources Association.
- Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020. Few-shot learning for opinion summarization. *arXiv preprint arXiv:2004.14884*.
- Tao Chen, Premaratne Samaranyake, XiongYing Cen, Meng Qi, and Yi-Chen Lan. 2022. [The impact of online reviews on consumers’ purchasing decisions: Evidence from an eye-tracking study](#). *Frontiers in Psychology*, 13:865702.

- 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. [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.
- 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.
- Weihua Deng, Ming Yi, and Yingying Lu. 2020. Vote or not? how various information cues affect helpfulness voting of online reviews. *Online Inf. Rev.*, 44(4):787–803.
- Aniket Deroy, Kripabandhu Ghosh, and Saptarshi Ghosh. 2023. [How ready are pre-trained abstractive models and llms for legal case judgement summarization?](#)
- Zixiang Ding, Rui Xia, and Jianfei Yu. 2020. Ecepe-2d: Emotion-cause pair extraction based on joint two-dimensional representation, interaction and prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3161–3170.
- Paul Ekman. 1992. An argument for basic emotions. *Cogn. Emot.*, 6(3-4):169–200.
- 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.](#)
- Armin Felbermayr and Alexandros Nanopoulos. 2016. [The role of emotions for the perceived usefulness in online customer reviews.](#) *Journal of Interactive Marketing*, 36:60–76.
- Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. [Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm.](#) In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1615–1625, Copenhagen, Denmark. Association for Computational Linguistics.
- Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. [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.
- Junpeng Guo, Xiaopan Wang, and Yi Wu. 2020. Positive emotion bias: Role of emotional content from online customer reviews in purchase decisions. *J. Retail. Consum. Serv.*, 52(101891):101891.
- Rishav Hada, Varun Gumma, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2024. [METAL: Towards multilingual meta-evaluation.](#) In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2280–2298, Mexico City, Mexico. Association for Computational Linguistics.
- Tom Hosking, Hao Tang, and Mirella Lapata. 2023. [Attributable and scalable opinion summarization.](#) In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8488–8505, Toronto, Canada. Association for Computational Linguistics.
- Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. 2024. Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952*.
- Chi Hu, Yuan Ge, Xiangnan Ma, Hang Cao, Qiang Li, Yonghua Yang, Tong Xiao, and Jingbo Zhu. 2024. [Rankprompt: Step-by-step comparisons make language models better reasoners.](#)
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Léo Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. [Mistral 7b.](#)
- Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Léo 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. [Mistral of experts.](#)

- 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](#).
- Theodoros Lappas, Gaurav Sabnis, and Georgios Valkanas. 2016. The impact of fake reviews on online visibility: A vulnerability assessment of the hotel industry. *Inf. Syst. Res.*, 27(4):940–961.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdel rahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Annual Meeting of the Association for Computational Linguistics*.
- Haoyuan Li, Somnath Basu Roy Chowdhury, and Snigdha Chaturvedi. 2023. [Aspect-aware unsupervised extractive opinion summarization](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12662–12678, Toronto, Canada. Association for Computational Linguistics.
- Haoyuan Li and Snigdha Chaturvedi. 2024. [Rationale-based opinion summarization](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8274–8292, Mexico City, Mexico. Association for Computational Linguistics.
- Jinfen Li and Lu Xiao. 2023. Multi-emotion recognition using Multi-EmoBERT and emotion analysis in fake news. In *Proceedings of the 15th ACM Web Science Conference 2023*, New York, NY, USA. ACM.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruo Chen Xu, and Chenguang Zhu. 2023. [G-eval: NLG evaluation using gpt-4 with better human alignment](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.
- Yixin Liu, Kejian Shi, Katherine S He, Longtian Ye, Alexander R. Fabbri, Pengfei Liu, Dragomir Radev, and Arman Cohan. 2024. [On learning to summarize with large language models as references](#).
- Paulo Lopes, John Nezlek, Astrid Schuetz, Ina Sellin, and Peter Salovey. 2004. [Emotional intelligence and social interaction](#). *Personality social psychology bulletin*, 30:1018–34.
- M S I Malik and Ayyaz Hussain. 2017. Helpfulness of product reviews as a function of discrete positive and negative emotions. *Comput. Human Behav.*, 73:290–302.
- Dina Mayzlin, Yaniv Dover, and Judith A Chevalier. 2012. Promotional reviews: An empirical investigation of online review manipulation. *SSRN Electron. J.*
- Xiao-Li Meng, Robert Rosenthal, and Donald B. Rubin. 1992. [Comparing correlated correlation coefficients](#). *Psychological Bulletin*, 111(1):172–175.
- Saif Mohammad and Felipe Bravo-Marquez. 2017. [Emotion intensities in tweets](#). In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (\*SEM 2017)*, pages 65–77, Vancouver, Canada. Association for Computational Linguistics.
- Saif M. Mohammad and Peter D. Turney. 2013. [Crowdsourcing a word–emotion association lexicon](#). *Computational Intelligence*, 29(3):436–465.
- 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.
- Ilias O Pappas, Panos E Kourouthanassis, Michail N Giannakos, and Vassilios Chrissikopoulos. 2014. Shiny happy people buying: the role of emotions on personalized e-shopping. *Electron. Mark.*, 24(3):193–206.
- Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemi Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared

- Kaplan. 2023. [Discovering language model behaviors with model-written evaluations](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13387–13434, Toronto, Canada. Association for Computational Linguistics.
- Rosalind W Picard. 2000. Toward computers that recognize and respond to user emotion. *IBM systems journal*, 39(3.4):705–719.
- Robert Plutchik. 1988. The nature of emotions: Clinical implications. In *Emotions and Psychopathology*, pages 1–20. Springer US, Boston, MA.
- Robert Plutchik. 2000. Emotions in the practice of psychotherapy: Clinical implications of affect theories. *American Psychological Association*.
- Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. [SemEval-2016 task 5: Aspect based sentiment analysis](#). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California. Association for Computational Linguistics.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).
- Peter Salovey and John D Mayer. 1990. Emotional intelligence. *Imagin. Cogn. Pers.*, 9(3):185–211.
- Kim Schouten and Flavius Frasincar. 2016. [Survey on aspect-level sentiment analysis](#). *IEEE Transactions on Knowledge and Data Engineering*, 28(3):813–830.
- Yuchen Shen and Xiaojun Wan. 2023. [Opinsummeval: Revisiting automated evaluation for opinion summarization](#).
- Tejpal Singh Sileadar, Swaroop Nath, Sankara Sri Raghava Ravindra Muddu, Rupasai Rangaraju, Swaprava Nath, Pushpak Bhattacharyya, Suman Banerjee, Amey Patil, Sudhanshu Shekhar Singh, Muthusamy Chelliah, and Nikesh Garera. 2024. [One prompt to rule them all: LLMs for opinion summary evaluation](#).
- James H. Steiger. 1980. [Tests for comparing elements of a correlation matrix](#). *Psychological Bulletin*, 87(2):245–251.
- 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 Sharmar, 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](#).
- Jen tse Huang, Man Ho Lam, Eric John Li, Shujie Ren, Wenxuan Wang, Wenxiang Jiao, Zhaopeng Tu, and Michael R. Lyu. 2024. [Emotionally numb or empathetic? evaluating how llms feel using emotionbench](#).
- 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*.
- Xiting Wang, Junpeng Guo, Yi Wu, and Na Liu. 2019. [Emotion as a signal of product quality: Its effect on purchase decision based on online customer reviews](#). *Internet Research*, 30(2):463–485.
- 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](#).
- Heng Yang, Biqing Zeng, Jianhao Yang, Ruidan Song, and Ziqiang Li. 2023. [PyABSA: A modularized framework for reproducible aspect-based sentiment analysis](#). In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM)*, pages 5117–5122.
- Wenhao Ying, Rong Xiang, and Qin Lu. 2019. [Improving multi-label emotion classification by integrating both general and domain-specific knowledge](#). In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 316–321, Hong Kong, China. Association for Computational Linguistics.
- Agnieszka Zablocki, Aikaterini Makri, and Michael J. Houston. 2019. [Emotions within online reviews and their influence on product attitudes in austria, usa and thailand](#). *Journal of Interactive Marketing*, 46:20–39.
- 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. 2020. [Pegasus: Pre-training with extracted gap-sentences for abstractive summarization](#).
- Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2023. [A survey on aspect-based sentiment analysis: Tasks, methods, and challenges](#). *IEEE Transactions on Knowledge and Data Engineering*, 35(11):11019–11038.
- Yusen Zhang, Nan Zhang, Yixin Liu, Alexander Fabbri, Junru Liu, Ryo Kamoi, Xiaoxin Lu, Caiming Xiong, Jieyu Zhao, Dragomir Radev, Kathleen McKeown, and Rui Zhang. 2024. [Fair abstractive summarization of diverse perspectives](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3404–3426, Mexico City, Mexico. Association for Computational Linguistics.

## A Evaluation Framework

This appendix provides the complete evaluation framework for emotion-aware opinion summarization, including metric definitions, design principles, and the mappings used for systematic assessment.

### A.1 Emotion-Aware Opinion Summarization Metrics

The evaluation of emotion-aware opinion summaries focused on 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.
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. This includes maintaining logical flow while transitioning between different emotional tones and product aspects.
3. **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 the product title and reviews. 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.
4. **emotional accuracy (EA)**- This metric evaluates how accurately the summary captures and represents the emotional tones present in

the original reviews. It measures the summary’s ability to reflect:

- i) The correct emotions: Accurately identifying the emotions expressed in the reviews.
- ii) Their intensity: Correctly representing the strength or degree of the emotions.
- iii) Their context: Accurately capturing the situations or aspects of the product that evoked these emotions.

Note: This metric focuses specifically on whether the correct emotions are identified and accurately represented in the summary, including their intensity and the context in which they appear in the reviews.

5. **emotional spectrum coverage (ESC)**- This metric assesses the range of emotions captured in the summary compared to the diversity of emotions expressed in the reviews. It measures:

- i) The variety of distinct emotions represented in the summary.
- ii) How well the summary reflects the full spectrum of emotions present in the reviews, including both positive and negative emotions.
- iii) The balance in representing both dominant and less prevalent emotions from the reviews.

Note: This metric focuses specifically on whether the summary captures the full range of emotions present in the reviews, regardless of their frequency or intensity. The focus is not just on individual emotions, but on whether the summary reflects the full diversity of emotions present in the reviews.

6. **emotional bias mitigation (EBM)**- This metric assesses whether the summary fairly represents all emotional perspectives present in the reviews without exaggerating or downplaying certain emotions. It measures:

- i) The balance between positive and negative emotions in the summary compared to the reviews.
- ii) The proportional representation of emotions relative to their prominence in the reviews.
- iii) The fair representation of all emotional perspectives, including minority views, without exaggeration or minimization.

iv) The reflection of the relative strength of emotional expressions.

Note: This metric focuses specifically on preventing skewed emotional representations to ensure fair and accurate summaries, especially in cases where reviews show a mix of positive and negative emotions.

7. **contextual emotional relevance (CER)**- This metric assesses whether the emotions mentioned in the summary are relevant to the specific context and product aspects discussed in the reviews. It measures:

- i) The accuracy of associating emotions with specific product features or aspects.
- ii) The relevance of emotional content to the discussed product characteristics.
- iii) The preservation of the context in which emotions are expressed in the reviews.
- iv) The summary’s ability to capture and convey complex or nuanced emotional contexts related to specific product features.

Note: This metric focuses on ensuring that emotional content is pertinent to the product aspects being discussed, enhancing the summary’s relevance and impact.

## B Experimental Results

This appendix presents detailed experimental results, including traditional metrics analysis, model performance statistics, statistical significance tests, and the faithfulness-emotion tradeoff analysis.

### B.1 Traditional Metrics Analysis

Table 8 reports ROUGE and BERTScore for top-performing models. Phi-3.5-mini-instruct achieves the highest scores (ROUGE-L: 0.41, BERTScore: 0.73), consistent with human evaluation rankings. However, traditional metrics show weak correlation with emotion-specific dimensions (ROUGE-L:  $r = 0.34$  with EA,  $r = 0.28$  with ESC; BERTScore:  $r = 0.42$  with EA,  $r = 0.38$  with ESC), confirming they capture only a fraction of what human evaluators assess.

## C Emotion Framework and Validation

This appendix provides details on the emotion-related algorithms, validation studies, and comparisons with alternative approaches.

Model	R-1	R-2	R-L	BS
Phi-3.5-mini-instruct	0.49	0.23	0.41	0.73
GPT-4o	0.47	0.21	0.39	0.72
Mistral-7B-v0.3	0.46	0.20	0.38	0.71
EAOS-LLAMA (1B)	0.45	0.19	0.37	0.70
Qwen2.5-32B	0.44	0.19	0.36	0.70
BART-large	0.38	0.15	0.31	0.62
T5-large	0.36	0.14	0.29	0.60
PEGASUS-large	0.37	0.14	0.30	0.61

**Table 8:** Traditional metrics on EAOS-EVAL-BENCH. R-1/2/L: ROUGE-1/2/L F1; BS: BERTScore F1. Metrics computed against human gold standard.

### C.1 Emotion Balancing Algorithm

This appendix details the emotion balancing mechanism in Stage 2 of EAOS-GEN, addressing how “appropriate representation” is operationalized.

### C.2 Proportional Representation Principle

The core principle is **proportional representation with minimum inclusion threshold**: emotions are represented in the summary proportionally to their prevalence in source reviews, while ensuring minority perspectives meeting a minimum threshold are not omitted.

**Formal Definition.** Given a set of  $n$  reviews  $R = \{r_1, \dots, r_n\}$  and the emotion detection function  $E(r_i) \subseteq \{\text{Joy, Trust, Fear, Surprise, Sadness, Disgust, Anger, Anticipation}\}$  returning emotions present in review  $r_i$ :

1. **Frequency Calculation:** For each primary emotion  $e$ , compute:

$$f_e = \frac{|\{r_i : e \in E(r_i)\}|}{n} \times 100\%$$

2. **Inclusion Threshold:** Emotion  $e$  is included in the summary if and only if  $f_e \geq 10\%$ .
3. **Proportional Weighting:** Among included emotions, representation weight is proportional to frequency:

$$w_e = \frac{f_e}{\sum_{e': f_{e'} \geq 10\%} f_{e'}}$$

This weight  $w_e$  guides the relative emphasis each emotion receives in the generated summary via Stage 2 of EAOS-GEN (§4.1; generation prompt in **Appendix I.2**).

**Threshold Justification.** The 10% threshold (equivalent to 1 review in a 10-review set) balances two competing concerns:

- **Signal vs. Noise:** Single outlier reviews may contain idiosyncratic emotions not representative of broader customer experience. Requiring  $\geq 10\%$  ensures at least a small but meaningful customer segment shares the sentiment.
- **Minority Representation:** Lower thresholds (e.g., 5%) would include emotions from single reviews, potentially amplifying noise; higher thresholds (e.g., 20%) would exclude legitimate minority perspectives.

### C.3 Edge Case Handling

Table 9 illustrates how the algorithm handles common edge cases:

Scenario	Distribution	Summary Behavior
<i>Unanimous positive</i>	Joy: 100%, Trust: 80%	Report as consensus; no artificial negative injection
<i>Dominant majority</i>	Joy: 70%, Sadness: 20%, Anger: 10%	All three included proportionally (7:2:1 weighting)
<i>Isolated outlier</i>	Joy: 90%, Fear: 5%	Fear omitted (below 10% threshold); summary reflects strong positive consensus
<i>Polarized reviews</i>	Joy: 45%, Anger: 45%, Surprise: 10%	All included; summary explicitly acknowledges mixed reception
<i>Emotion ties</i>	Joy: 30%, Trust: 30%, Fear: 30%	Equal weighting; all three given comparable narrative space

**Table 9:** Edge case handling in emotion balancing. The algorithm avoids false balance (equal representation regardless of frequency) while ensuring meaningful minority perspectives are preserved.

### C.4 Implementation in Prompts

The balancing algorithm is implemented through explicit instructions in the EAOS-GEN prompt:

*“Represent emotions proportionally to their prevalence in reviews. Include minority emotions only if they appear in at least 10% of reviews (e.g., at least 1 of 10 reviews). Do not artificially balance or equalize emotions—if reviews are unambiguously positive, reflect this consensus. For mixed sentiment, explicitly acknowledge the distribution (e.g., ‘while most customers express satisfaction, a notable minority report frustration with...’).”*

## C.5 Validation

We validated the balancing mechanism through two analyses:

**Distribution Fidelity.** On our 100-product evaluation set, we computed the correlation between emotion frequencies in source reviews (manually annotated) and emotion mention frequencies in generated summaries. The Pearson correlation was  $r = 0.84$  ( $p < .001$ ), indicating strong fidelity to source distributions.

**Minority Preservation.** For products with minority emotions (10–30% prevalence), we measured whether these emotions appeared in summaries. The inclusion rate was 91.2% (156/171 cases), confirming that the threshold effectively preserves minority perspectives without artificial amplification.

## C.6 Complex Emotion Evaluation Validation

This appendix details the validation study demonstrating that LLM evaluators reliably assess complex (compositional) emotions, not merely perform keyword matching.

## C.7 Motivation and Concern

Plutchik’s model defines complex emotions as combinations of primary emotions (e.g., Love = Joy + Trust, Optimism = Anticipation + Joy). A valid concern is whether LLM evaluators genuinely understand these compositional semantics or simply match surface-level keywords. If the latter, evaluation scores for summaries with complex emotional content would be unreliable.

Table 10 summarizes evaluator performance across emotion types.

Emotion Type	Accuracy	Human-LLM $\kappa$	$n$
<i>Primary Emotions</i>			
Joy, Trust, Fear, etc.	92.1%	0.84	412
<i>Complex Emotions</i>			
Love (Joy + Trust)	89.2%	0.83	48
Optimism (Anticipation + Joy)	86.4%	0.80	44
Awe (Fear + Surprise)	85.1%	0.79	27
Contempt (Disgust + Anger)	88.6%	0.82	35
All Complex (avg)	87.3%	0.81	154
<i>Implicit Complex</i>	83.6%	0.78	61

**Table 10:** LLM evaluator performance on primary vs. complex emotions. “Implicit Complex” refers to cases where complex emotions were expressed through constituent primaries without explicit labels.

## C.8 Validation Protocol

We designed a three-part validation study using 50 summaries stratified to include diverse complex

emotion expressions:

### Part 1: Explicit Complex Emotion Recognition.

We selected 50 summaries containing explicit complex emotion terms (e.g., “customers express love for the product,” “anticipation mixed with optimism”). For each, we asked:

- Does the evaluator correctly identify the complex emotion?
- Does it correctly decompose into constituent primaries?

### Part 2: Implicit Complex Emotion Inference.

We created 30 test cases where complex emotions were expressed through co-occurring primary emotions *without* using the complex emotion label. For example:

*“Customers express deep trust in the brand, describing their joyful experiences with heartfelt appreciation.”*

This expresses Love (Joy + Trust) implicitly. We assessed whether evaluators correctly inferred the complex emotion.

### Part 3: Compositional Understanding Probes.

We designed 20 adversarial cases to test whether evaluators distinguish:

- Genuine complex emotions vs. coincidental co-occurrence
- Correct vs. incorrect complex emotion attribution

For example, a summary mentioning both “joy” and “fear” (non-adjacent in Plutchik’s wheel) should *not* be labeled as a complex emotion—these are independent emotions, not a compositional dyad.

## C.9 Results

Table 11 presents detailed results across all three evaluator LLMs:

## C.10 Key Findings

### Finding 1: Genuine Compositional Understanding.

The high accuracy on implicit inference (83.6% average) and correct rejection of false positives (91.2%) demonstrates that evaluators understand compositional semantics, not just keyword matching. If evaluators were merely pattern-matching, we would expect: (a) near-zero accuracy

Test Type	Qwen-72B	GPT-4o	Llama-70B	Avg
<i>Part 1: Explicit Recognition</i>				
Identification	91.2%	89.4%	86.8%	89.1%
Decomposition	88.6%	87.2%	84.0%	86.6%
<i>Part 2: Implicit Inference</i>				
Love (Joy + Trust)	86.7%	84.2%	80.0%	83.6%
Optimism (Ant. + Joy)	84.5%	82.8%	78.6%	82.0%
Awe (Fear + Surprise)	82.1%	80.4%	76.2%	79.6%
<i>Part 3: Adversarial Probes</i>				
Correct rejection	93.5%	91.8%	88.4%	91.2%
Correct attribution	89.2%	87.6%	84.8%	87.2%
<b>Overall</b>	<b>88.5%</b>	<b>86.2%</b>	<b>82.7%</b>	<b>85.8%</b>

**Table 11:** Complex emotion evaluation validation across three LLM evaluators. “Correct rejection” refers to correctly *not* attributing complex emotions when primaries co-occur without compositional relationship.

on implicit cases (no keyword to match), and (b) high false positive rates on adversarial probes (any co-occurring primaries would trigger complex emotion labels).

**Finding 2: Modest Primary-Complex Gap.** The 4.8% gap between primary (92.1%) and complex (87.3%) emotion accuracy is modest and expected—complex emotions are inherently more difficult to assess due to their compositional nature. Importantly, complex emotion accuracy remains well above chance (>85% vs. 12.5% random baseline for 8 emotions).

**Finding 3: Evaluator Consistency.** Human-LLM agreement (Cohen’s  $\kappa$ ) for complex emotions ( $\kappa = 0.81$ ) approaches “almost perfect” agreement thresholds, comparable to primary emotions ( $\kappa = 0.84$ ). This consistency across human and LLM judgments validates the reliability of complex emotion evaluation.

**Finding 4: Prompt Design Matters.** The explicit inclusion of complex emotion mappings in our EAOS-EVAL prompts (**Appendix C.13**) is critical. When we ablated this component (removing complex emotion definitions from prompts), accuracy on implicit inference dropped from 83.6% to 52.4%—barely above chance. This confirms that compositional understanding is enabled by proper prompt design, not inherent to base LLM capabilities.

### C.11 Representative Examples

Table 12 illustrates how our LLM evaluators handle both genuine complex emotions and cases requiring correct rejection of false compositional attributions.

Summary Excerpt	Ground Truth	LLM Assessment
“Customers express deep affection and unwavering trust in the brand, cherishing each interaction.”	Love (Joy + Trust)	Love (Joy + Trust) ✓
“The product sparks curiosity and excitement, leaving users eager for future releases.”	Optimism (Ant. + Joy)	Optimism (Ant. + Joy) ✓
“Some users report both surprise and nervousness at the unexpected features.”	Fear + Surprise (separate)	Fear + Surprise ✓
	<i>Not a complex emotion</i>	(correctly not Awe)
“Customers feel joy about quality but anger about shipping delays.”	Joy + Anger (separate)	Joy, Anger ✓
	<i>Not compositional</i>	(correctly separate)

**Table 12:** Representative examples from complex emotion validation. ✓ indicates correct LLM assessment. The last two rows show correct rejection of false compositional attributions.

### C.12 EAOS-GEN Prompt Design Principle

A key aspect of the EAOS-GEN design is balance, blending both majority and minority opinions without allowing any one emotion to dominate unless clearly indicated in the reviews. It integrates facts and emotions naturally, avoiding explicit emotional labels or generic phrases. Emotions should arise organically within the narrative, enhancing the summary’s depth without overwhelming the reader.

This prompt structure optimizes LLM outputs to be concise and emotionally resonant, faithfully representing customers’ nuanced emotional experiences.

### C.13 EAOS-EVAL Prompts Design Principle

Our EAOS-EVAL prompts are designed to ensure evaluations are thorough and align closely with human judgment. We structured the prompts step-by-step, inspired by (Hu et al., 2024), who demonstrated that detailed prompts enhance LLM effectiveness and alignment with human evaluations. Moreover, inspired by (Hada et al., 2024), we incorporated detailed, objective scoring criteria to enhance reproducibility and minimize subjective variance.

Given the subjectivity of emotions and variability in human interpretation (Plutchik, 1988), we designed prompts with step-by-step instructions to help LLMs assess these nuances. Each prompt focuses on one dimension independently, preventing conflation of evaluation aspects.

The prompts guide LLMs to capture nuanced emotions. Emotional Accuracy metric prompt di-

rects LLMs to identify emotions, assess their intensity, and evaluate contextual relevance, using explicit mappings to standardize the LLMs' interpretation. The prompts encourage fair representation of emotions. Emotional Bias Mitigation metric prompt prevents exaggeration or downplaying of perspectives, while Emotional Spectrum Coverage metric prompt ensures comprehensive emotional representation. Lastly, the prompts promote contextual sensitivity. The Contextual Emotional Relevance metric prompt guides the LLM to assess if emotions are relevant to discussed product features, preventing misattributions and ensuring proper contextual evaluation.

### Emotion and Intensity Mappings for Summary Evaluation

In this section, we define and detail the emotion and intensity mappings used for evaluating summaries. These mappings were included as part of the prompts to LLMs in EAOS-EVAL to ensure consistent and structured evaluation across dimensions.

Given the complexity of emotional accuracy metric, we used extended emotion, intensity and Complex emotions mappings to evaluate it systematically.

**Extended Emotion Mapping for Emotional Accuracy Metric Evaluation.** We expanded the emotion mapping for evaluation, represented as <Emotion word: Synonyms>:

Ecstasy:	Delighted, Giddy
Joy:	Excited, Pleased
Serenity:	Calm, Peaceful
Admiration:	Connected, Proud
Trust:	Accepting, Safe
Acceptance:	Open, Welcoming
Terror:	Alarmed, Petrified
Fear:	Stressed, Scared
Apprehension:	Worried, Anxious
Amazement:	Inspired, Wowed
Surprise:	Shocked, Unexpected
Distraction:	Scattered, Uncertain
Grief:	Heartbroken, Distraught
Sadness:	Bummed, Loss
Pensiveness:	Blue, Unhappy
Loathing:	Disturbed, Horrified
Disgust:	Distrust, Rejecting
Boredom:	Tired, Uninterested
Rage:	Overwhelmed, Furious
Anger:	Mad, Fierce
Annoyance:	Frustrated, Prickly
Vigilance:	Intense, Focused
Anticipation:	Curious, Considering
Interest:	Open, Looking

### Intensity Mapping for Emotional Accuracy Metric Evaluation.

For thorough evaluation of the emotional accuracy metric by LLMs, we used the following **INTENSITY MAPPING** as part of the prompt. Here, I1 represents the highest intensity, I2 represents medium intensity, and I3 represents the lowest intensity. The mapping is represented as <Intensity: Emotion word>:

<I1: Ecstasy>, <I2: Joy>, <I3: Serenity>  
 <I1: Admiration>, <I2: Trust>, <I3: Acceptance>  
 <I1: Terror>, <I2: Fear>, <I3: Apprehension>  
 <I1: Amazement>, <I2: Surprise>, <I3: Distraction>  
 <I1: Grief>, <I2: Sadness>, <I3: Pensiveness>  
 <I1: Loathing>, <I2: Disgust>, <I3: Boredom>  
 <I1: Rage>, <I2: Anger>, <I3: Annoyance>  
 <I1: Vigilance>, <I2: Anticipation>, <I3: Interest>

### Complex Emotion Mapping for Emotional Accuracy.

Complex emotions are formed by combinations of primary emotions, represented as: <Emotion\_1 + Emotion\_2 = Complex Emotion (Synonyms of Complex Emotion word)>:  
 <Joy + Trust: Love (Accepted, Adored, Connected)>  
 <Trust + Fear: Submission (Obedience, Followership, Tribalism)>

<Fear + Surprise: Awe (Horrified, Vigilance, Shock)>  
 <Surprise + Sadness: Disapproval (Wrong, Shock, Lost)>  
 <Sadness + Disgust: Remorse (Guilt, Atonement, Responsibility)>  
 <Disgust + Anger: Contempt (Righteous, Horrified, Shaming)>  
 <Anger + Anticipation: Aggressiveness (Belligerent, Hostile, Argumentative)>  
 <Anticipation + Joy: Optimism (Hopeful, Looking Forward)>.

**General Emotion Mapping for Evaluation of Emotion Metrics.** We utilized this General Emotion Mapping for evaluation of emotional spectrum coverage, emotional bias mitigation and contextual emotional relevance metrics. Focus on the following 8 primary emotions and their related terms, represented as: <Primary Emotion: Related Terms>:

<Joy: joy, excited, pleased, amusement, pride, excitement, ecstasy>  
 <Trust: trust, accepting, safe, caring>  
 <Fear: fear, stressed, scared, apprehension, terror, anxiety>  
 <Surprise: surprise, amazement, shocked, unexpected>  
 <Sadness: sadness, bummed, loss, disappointment, grief, pensiveness>  
 <Disgust: disgust, loathing, distrust, boredom, rejecting>  
 <Anger: anger, annoyance, mad, fierce, rage>  
 <Anticipation: anticipation, curious, considering>

**Emotion Categorization for Emotional Bias Mitigation Metric.** For the emotional bias mitigation metric, we also categorized emotions as follows:

<Positive Emotions: Joy, Trust, Anticipation>  
 <Negative Emotions: Fear, Sadness, Disgust, Anger>  
 <Neutral Emotion: Surprise (context-dependent, can be positive or negative)>.

## D User Study

This appendix provides comprehensive details on the user study methodology, statistical analysis, and preference results.

### D.1 Study Design and Methodology

We conducted a large-scale user study ( $N = 350$ ) comparing EAOS (Emotion-Aware Opinion Summaries) with traditional opinion summaries. Each participant evaluated four pairs of summaries, with each pair drawn from a different product category:

- Beauty & Personal Care
- Home & Kitchen
- Electronics
- Clothing

Participants provided responses to five evaluation questions, resulting in 7,000 total preference judgments (350 participants  $\times$  4 categories  $\times$  5 questions).

### D.2 Evaluation Questions

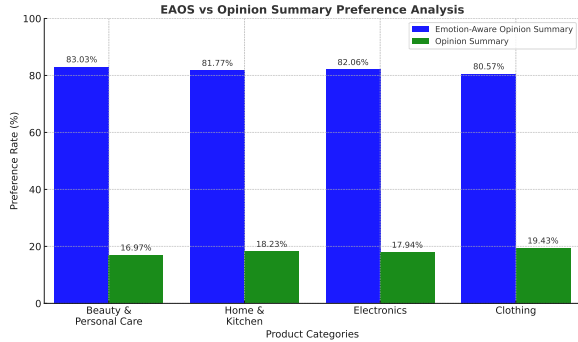
For each summary pair, participants answered the following five questions:

- **Customer Experience Understanding:** “Which summary better helps you understand the experiences of previous customers?”
- **Experience Anticipation:** “Which summary better helps you anticipate your own potential experience with the product?”
- **Purchase Decision Influence:** “Which summary format would make you more likely to make a purchase decision?”
- **Online Shopping Utility:** “Which summary would be more helpful when shopping online?”
- **Product Fit Assessment:** “Which summary would better help you decide whether this product is right for you?”

### D.3 Statistical Analysis

We employed the chi-square goodness-of-fit test to validate the statistical significance of user preferences, which is suitable for analyzing categorical preference data in NLP user studies (Clark et al., 2011). The test evaluated whether the observed preference distribution significantly differed from the null hypothesis of no preference (50-50 split).

**Results:** The analysis revealed a strong preference for EAOS across all categories:



**Figure 2:** Preference analysis comparing Emotion-Aware Opinion Summary (EAOS) versus opinion summaries across product categories ( $N = 350$ ). Bars represent the mean preference percentage across five evaluation questions per category. Statistical significance:  $\chi^2 = 2867.20$  ( $df = 1, p < .001$ ).

- **Beauty & Personal Care:** EAOS (83.03%) vs. Opinion Summary (16.97%)
- **Electronics:** EAOS (82.06%) vs. Opinion Summary (17.94%)
- **Home & Kitchen:** EAOS (81.77%) vs. Opinion Summary (18.23%)
- **Clothing:** EAOS (80.57%) vs. Opinion Summary (19.43%)

The chi-square test yielded  $\chi^2 = 2867.20$  ( $df = 1, p < .001$ ), strongly rejecting the null hypothesis of no preference. The observed distribution of 5,740 preferences for EAOS (82%) versus 1,260 for traditional summaries (18%) provides robust evidence for EAOS superiority.

To quantify the magnitude of the preference, we calculated Cramer’s  $V = 0.64$ , indicating a large effect size according to conventional benchmarks in behavioral research (Cohen, 1988).

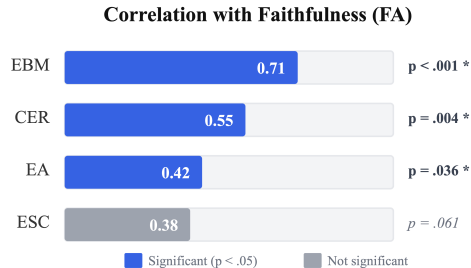
#### D.4 Model Performance Standard Deviations

Table 13 reports standard deviations for all model performance scores in Table 4. Standard deviations were computed across the 100 evaluated products. Higher-performing models generally exhibit lower variance (ceiling effect), while more subjective dimensions (EBM, CER) show higher variance across all models.

#### D.5 Faithfulness-Emotion Tradeoff

We analyze the relationship between Faithfulness and emotion-specific evaluation dimensions across all 26 models. Figure 3 presents the correlation matrix, revealing that Faithfulness correlates strongly

with Emotional Bias Mitigation ( $r = 0.71, p < .001$ ), indicating that models sacrificing source grounding tend to over-represent emotions. Fine-tuned Qwen variants show particularly pronounced tradeoffs, while EAOS-LLAMA maintains balanced performance across dimensions.



**Figure 3:** Correlation between Faithfulness and emotion dimensions across 26 models. Significant correlations ( $p < .05$ ) shown in blue.

## E Dataset and Training

This appendix provides comprehensive details on the EAOS-SUMM dataset construction, including generator selection, quality audits, human gold standard creation, teacher bias analysis, and synthetic data validation. Figure 4 summarizes all human involvement across the pipeline.

### E.1 Generator Selection Pilot Study

We conducted a systematic two-stage evaluation to select teacher models for generating EAOS-SUMM. Table 14 presents results from the 50-product pilot study comparing candidate generators across all 7 evaluation dimensions.

### E.2 Teacher Bias Analysis

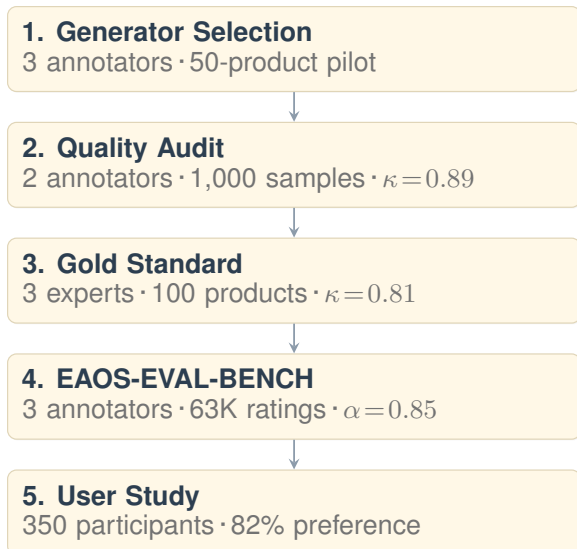
Table 15 compares emotion distributions between teacher-generated summaries and human gold standards. Both teachers exhibit mild positive-emotion amplification (Joy, Trust, Anticipation) consistent with RLHF alignment, while maintaining low Jensen-Shannon divergence ( $D_{JS} < 0.05$ ) from human distributions.

### E.3 EAOS-SUMM Statistics

The EAOS-SUMM dataset is designed to advance research in emotion-aware opinion summarization by providing a large-scale, high-quality, and diverse collection of summaries. Built on open-source language models to prioritize reproducibility and cost-effectiveness, the dataset spans multiple domains and ensures comprehensive coverage of opinions and emotions. Below, we present key

Model	FL	CO	FA	EA	ESC	EBM	CER	Avg
<i>PLMs (Pre-trained Language Models)</i>								
BART-large	0.42	0.38	0.25	0.45	0.48	0.35	0.42	0.39
T5-large	0.45	0.42	0.28	0.48	0.52	0.38	0.45	0.42
PEGASUS-large	0.40	0.40	0.26	0.46	0.50	0.36	0.44	0.40
T5-large-FT	0.28	0.32	0.30	0.38	0.42	0.40	0.38	0.35
<i>Methodological Baselines</i>								
Simple-Emotion-Prompting	0.12	0.28	0.25	0.35	0.38	0.42	0.35	0.31
Two-Stage-Pipeline	0.15	0.32	0.28	0.38	0.42	0.45	0.38	0.34
ABSA-Summary	0.18	0.35	0.30	0.42	0.48	0.45	0.45	0.38
<i>LLMs</i>								
Meta-Llama-3.2-1B-Instruct	0.22	0.35	0.32	0.42	0.45	0.48	0.42	0.38
Qwen2.5-0.5B-Instruct	0.20	0.38	0.45	0.45	0.48	0.52	0.45	0.42
Qwen2.5-1.5B-Instruct	0.18	0.35	0.42	0.40	0.42	0.48	0.40	0.38
Phi-3.5-mini-instruct	0.08	0.25	0.22	0.28	0.18	0.35	0.30	0.24
Mistral-7B-Instruct-v0.2	0.10	0.30	0.28	0.32	0.25	0.38	0.35	0.28
Mistral-7B-Instruct-v0.3	0.12	0.32	0.25	0.30	0.28	0.40	0.32	0.28
Zephyr-7b-beta	0.22	0.35	0.30	0.32	0.30	0.42	0.38	0.33
Qwen2.5-7B-Instruct	0.15	0.35	0.28	0.35	0.38	0.42	0.35	0.33
Meta-Llama-3.1-8B-Instruct	0.12	0.32	0.28	0.38	0.40	0.45	0.38	0.33
Gemma-2-9b-it	0.10	0.32	0.26	0.35	0.35	0.45	0.35	0.31
SOLAR-10.7B-Instruct	0.25	0.30	0.30	0.30	0.32	0.40	0.38	0.32
Mistral-Nemo-Instruct-2407	0.14	0.32	0.32	0.30	0.32	0.42	0.32	0.31
Qwen2.5-14B-Instruct	0.12	0.32	0.26	0.30	0.32	0.40	0.35	0.30
Mistral-Small-Instruct-2409	0.10	0.33	0.25	0.32	0.35	0.42	0.32	0.30
Mixtral-8x7B-Instruct	0.16	0.35	0.28	0.32	0.32	0.40	0.32	0.31
Qwen2.5-32B-Instruct	0.10	0.33	0.26	0.28	0.38	0.38	0.30	0.29
Meta-Llama-3.1-70B-Instruct	0.08	0.33	0.25	0.35	0.35	0.45	0.35	0.31
Qwen2.5-72B-Instruct	0.10	0.32	0.25	0.32	0.35	0.40	0.32	0.29
GPT-4o	0.08	0.24	0.22	0.30	0.32	0.38	0.30	0.26
<i>Fine-tuned Models</i>								
EAOS-LLAMA (Llama-3.2-1B-FT)	0.14	0.32	0.28	0.32	0.35	0.45	0.38	0.32
Qwen2.5-0.5B-Instruct-FT	0.10	0.32	0.42	0.28	0.35	0.52	0.38	0.34
Qwen2.5-1.5B-Instruct-FT	0.12	0.32	0.40	0.32	0.35	0.52	0.38	0.34

**Table 13:** Standard deviations for model performance scores ( $n = 100$  products). Lower SDs indicate more consistent performance. Dimensions with higher subjectivity (EBM, CER) show higher variance across all models.



**Figure 4:** Human involvement at each pipeline stage. All annotations involve independent scoring followed by reconciliation where applicable.

Model	FL	CO	FA	EA	ESC	EBM	CER	Avg
Phi-3.5-mini (3.8B)	4.82	4.41	4.48	4.48	4.62	4.38	4.45	<b>4.52</b>
Mistral-7B-v0.3	4.78	4.32	4.52	4.36	4.38	4.42	4.28	<u>4.44</u>
Llama-3.1-8B	4.75	4.18	4.35	4.08	4.02	4.12	3.98	<u>4.21</u>
Gemma-2-9B-it	4.71	4.15	4.35	4.08	3.98	4.08	3.92	4.18
Qwen2.5-7B	4.68	4.08	4.28	3.92	3.85	4.02	3.98	4.12

**Table 14:** Generator pilot study on 50 products. Phi-3.5-mini and Mistral-7B-v0.3 were selected as teachers based on complementary strengths.

statistics that highlight the scope and characteristics of the dataset.

## F Implementation and Guidelines

This appendix provides comprehensive information on annotation guidelines, LLM configurations, and implementation details for both the generation and evaluation frameworks.

### F.1 Annotation Guidelines

For brevity, we present a concise version of the guidelines here. Annotators, however, were provided with comprehensive examples and detailed instructions, as described in the supplementary materials.

### F.2 Greetings and Introduction for Annotators

Dear Annotators,

Thank you for participating in this annotation task. Your expertise and insights are crucial to evaluating the quality of Emotion-Aware Opinion Summaries (EAOS). We greatly appreciate your time and effort in ensuring

Emotion	Human	Phi-3.5	Mistral-7B	$\Delta$ Avg
Joy	26.4%	32.2%	30.6%	+5.0%
Trust	21.2%	25.2%	24.4%	+3.6%
Anticipation	12.8%	18.4%	14.4%	+3.6%
Surprise	10.6%	8.8%	8.0%	-1.9%
Sadness	11.4%	6.8%	8.4%	-3.8%
Fear	7.8%	4.2%	9.0%	-1.2%
Anger	6.2%	2.6%	3.0%	-3.4%
Disgust	3.6%	1.8%	2.2%	-1.6%
$D_{JS}$ vs. Human	—	0.041	0.032	—

**Table 15:** Emotion distribution across teachers vs. human gold standard. Teachers exhibit mild positive-emotion amplification but maintain low Jensen-Shannon divergence ( $D_{JS} < 0.05$ ).

Metric	Value
Total Words in Reviews	13,581,966
Total Reviews	400,000
Total Words in Review Titles	1,721,408
Total Words in Summaries	4,868,107
Average Words per Review	33.95
Average Reviews per Product	10.00
Average Words in Summaries	121.70
Average Review Title Length	4.30

**Table 16:** Overall EAOS-SUMM Dataset Statistics.

that the evaluations are thorough and accurate. This task involves assessing summaries generated by various models on seven dimensions: fluency, coherence, faithfulness, emotional accuracy, emotional spectrum coverage, emotional bias mitigation and contextual emotional relevance. Before you begin, please read through these instructions carefully.

### F.3 General Instructions

- The evaluation will be conducted using **Google Sheets**, with each metric recorded in its own column.
- The rows in the Google Sheet will correspond to summary IDs (e.g., summary1, summary2, etc.). Each summary ID will be linked to a JSON file stored in a Google Drive folder.
- We have provided the summaries in JSON format via a shared Google Drive folder. *Each file is named according to its summary ID (e.g., summary1.json), without revealing the model names to prevent bias.*
- We have printed 25 product titles, each with 10 corresponding reviews, for easy reference

while evaluating the summaries. *This printed material is provided to avoid excessive screen navigation.*

- You are required to evaluate each summary on all 7 metrics before moving to the next summary.
- For a comprehensive understanding of Plutchik's Wheel of Emotions, or for any doubts or further reference, please refer to the supplementary materials.

#### F.4 Evaluation Process

- Carefully read through both the product reviews and the corresponding summary *before* beginning your evaluation.
- Each dimension is to be evaluated independently, and you are required to assign a score between 1 (Very Poor) and 5 (Excellent) for each.
- *Use the provided rubrics* for each dimension as a guide, but rely on your expert judgment when nuanced cases arise.
- Pay attention to both *explicit and implicit emotional content* present in the reviews and assess how they are reflected in the summaries.
- If you are uncertain about a score, re-read the reviews and summary, considering the context before finalizing your decision.

#### F.5 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.
5. Evaluate summaries in the order they appear in the Google Sheet.
6. **Model names are not disclosed** to prevent bias.

7. Refer to the rubrics in these guidelines when assigning scores for consistency.
8. If you need a break, please complete the evaluation of the current summary first.

#### F.6 Emotion Framework: Plutchik's Wheel of Emotions

We ask all annotators to focus on **Plutchik's 8 primary emotions**: Anger, Anticipation, Joy, Trust, Fear, Surprise, Sadness, and Disgust. These are fundamental to our emotion-aware opinion summarization task. While your evaluations are restricted to these 8 primary emotions, understanding how emotions combine and vary in intensity will help you capture *fine-grained emotional nuances and subtleties* in both the reviews and the emotion-aware summaries.

##### F.6.1 Combining Emotions

Emotions do not always exist in isolation. Plutchik's model allows emotions to combine in various ways:

- **Primary dyads**: Adjacent emotions combine, for example, **joy + trust = love**.
- **Secondary dyads**: Emotions two petals apart combine, such as **fear + sadness = despair**.
- **Tertiary dyads**: Emotions three petals apart combine, like **joy + surprise = delight**.

In total, 24 dyads and 32 triads (combinations of 3 emotions) are possible, producing 56 unique emotional states. Opposite emotions are four petals apart, such as anticipation being the opposite of surprise.

##### F.6.2 Intensity of Emotions

Emotions can vary in intensity. On Plutchik's wheel, emotions intensify toward the center:

- **Joy** can range from **serenity** (mild) to **ecstasy** (intense).
- **Anger** can range from **annoyance** (mild) to **rage** (intense).

As you evaluate the summaries, consider both the **basic emotion** and its **intensity** as reflected in the text (reviews and summary). *For example, is the joy in the review expressed as mild serenity, or more intense ecstasy?*

### F.6.3 Emotions and Their Opposites

Emotions can also have opposites, which you should keep in mind during evaluation:

- **Joy vs. Sadness** (Mild: Serenity vs. Pensiveness | Intense: Ecstasy vs. Grief)
- **Trust vs. Disgust** (Mild: Acceptance vs. Boredom | Intense: Admiration vs. Loathing)
- **Fear vs. Anger** (Mild: Apprehension vs. Annoyance | Intense: Terror vs. Rage)
- **Surprise vs. Anticipation** (Mild: Distraction vs. Interest | Intense: Amazement vs. Vigilance)

### F.7 Evaluation Dimensions

This section outlines the definition, scoring rubric, and evaluation process for each of the seven metrics.

#### 1. Fluency.

**Definition.** Fluency refers to the overall readability and linguistic quality of the summary, including grammar, spelling, punctuation, and sentence structure.

#### Scoring Rubric.

- **1 (Very Poor):** Incomprehensible with severe linguistic issues.
- **2 (Poor):** Difficult to understand due to frequent errors.
- **3 (Fair):** Some errors present, but the main points are comprehensible.
- **4 (Good):** Clear and smooth with minor errors.
- **5 (Excellent):** Flawless language, natural and highly readable.

#### Evaluation Process.

1. Read the summary carefully, identifying any linguistic errors.
2. Assess the overall readability, clarity, and flow of the text.
3. Consider how smoothly the summary can be understood by a general audience.
4. Assign a score using the rubric above.

#### 2. Coherence.

**Definition.** Coherence measures how logically the ideas in the summary flow from one point to the next, ensuring a well-structured and organized presentation of information.

#### Scoring Rubric.

- **1 (Very Poor):** Disjointed with no clear logical flow.
- **2 (Poor):** Frequent lapses in logic; difficult to follow.
- **3 (Fair):** Occasional disruptions, but reasonably coherent.
- **4 (Good):** Mostly coherent with minor issues.
- **5 (Excellent):** Flawless logical progression, exceptionally well-organized.

#### Evaluation Process.

1. Assess the logical sequence and connections between sentences.
2. Evaluate how well the summary builds on each idea, transitioning smoothly.
3. Consider how well the summary maintains coherence when discussing different emotional tones or product aspects.
4. Assign a score using the rubric above.

#### 3. Faithfulness.

**Definition.** Faithfulness measures how accurately the summary reflects the original reviews without adding false or fabricated information. It also assesses whether the information in the summary is verifiable from the reviews.

#### Scoring Rubric.

- **1 (Very Poor):** Contains mostly fabricated or unrelated information.
- **2 (Poor):** Significant inaccuracies compared to the original reviews.
- **3 (Fair):** Some discrepancies, but generally faithful.
- **4 (Good):** Minor discrepancies, mostly accurate.
- **5 (Excellent):** Completely faithful, no discrepancies.

### Evaluation Process.

1. Compare the information in the summary to the original reviews, checking for accuracy.
2. Ensure there are no unsupported claims or exaggerations.
3. Verify that the summary remains true to the opinions expressed in the reviews.
4. Assign a score using the rubric above.

### 4. Emotional Accuracy.

**Definition.** Emotional accuracy evaluates how well the summary captures and reflects the emotions expressed in the original reviews, focusing on whether the correct emotions are identified and accurately represented, including their intensity and context.

#### Scoring Rubric.

- **1 (Very Poor):** The summary misrepresents or ignores the emotions expressed in the reviews, with no alignment in intensity or context.
- **2 (Poor):** The summary captures some emotions but significantly misrepresents others, or fails to reflect the intensity and context accurately.
- **3 (Fair):** The summary reflects a mix of accurate and inaccurate emotions, with limited consideration of intensity and context.
- **4 (Good):** Most emotions are accurately captured, and their intensity and context are reasonably well represented.
- **5 (Excellent):** All emotions are accurately identified, and their intensity and context are fully and effectively represented in the summary.

### Evaluation Process.

1. **Identify Emotions in the Reviews:** Carefully read the reviews to identify the primary emotions expressed, considering Plutchik's core emotions (Joy, Trust, Fear, Surprise, Sadness, Disgust, Anger, Anticipation) and their contextual nuances.
2. **Compare to the Summary:** Evaluate whether these emotions are present in the summary and check if any additional or fabricated emotions are included.

3. **Assess Intensity and Context:** Determine if the summary accurately conveys the intensity of each emotion and preserves the context in which the emotions appear in the reviews.
4. **Handle Nuanced Cases:** For mixed or conflicting emotions, ensure that the summary reflects the balance and complexity as presented in the reviews.
5. **Assign a Score:** Use the refined scoring rubric to assign a score, ensuring consistency across all summaries.

### 5. Emotional Spectrum Coverage.

**Definition.** Emotional spectrum coverage measures how well the summary captures the full range and diversity of emotions expressed in the reviews, restricted to Plutchik's 8 primary emotions (*Joy, Trust, Fear, Surprise, Sadness, Disgust, Anger, Anticipation*), including both positive and negative emotions, regardless of their frequency or intensity.

#### Scoring Rubric.

- **1 (Very Poor):** The summary captures only one or two emotions, ignoring the majority of emotional themes, with no reflection of diversity.
- **2 (Poor):** The summary captures a limited range of emotions, missing key emotional categories or neglecting less frequent emotions entirely.
- **3 (Fair):** The summary captures some of the main emotions but overlooks several others, resulting in a partial representation of the emotional spectrum.
- **4 (Good):** The summary captures most of the prominent and subtle emotions, reflecting a broad range with minor gaps in diversity or intensity.
- **5 (Excellent):** The summary comprehensively captures the full spectrum of emotions expressed in the reviews, including both prominent and subtle emotions, with appropriate diversity and balance.

### Evaluation Process.

1. **Identify All Emotions:** Carefully read the reviews to identify the full spectrum of emotions expressed, considering both prominent and subtle emotions.

2. **Compare Emotional Representation:** Evaluate whether the summary includes the full range of identified emotions, ensuring that both positive and negative emotions are represented.
3. **Assess Diversity and Balance:** Check if the summary reflects not only the prominent emotions but also less frequent ones, ensuring an accurate representation of diversity across the emotional spectrum.
4. **Consider Contextual Nuances:** Ensure that emotions are captured in their appropriate context, without fabricating or exaggerating their prominence or intensity.
5. **Assign a Score:** Use the refined scoring rubric to assign a score, ensuring that the summary adequately reflects the diversity and balance of the emotional spectrum.

## 6. Emotional Bias Mitigation.

**Definition.** Emotional bias mitigation evaluates how well the summary maintains a balanced representation of the emotional perspectives expressed in the reviews without exaggerating or downplaying any particular emotion.

### Scoring Rubric.

- **1 (Very Poor):** Highly biased, significantly exaggerating or downplaying emotions.
- **2 (Poor):** Notably biased, skewing emotional representation.
- **3 (Fair):** Slight bias, but overall balance is recognizable.
- **4 (Good):** Mostly balanced with minor discrepancies.
- **5 (Excellent):** Perfect balance of emotional perspectives.

### Evaluation Process.

1. Assess the balance between positive and negative emotions in the reviews and summary.
2. Check for any exaggeration or downplaying of emotional content.
3. Assign a score using the rubric above.

## 7. Contextual Emotional Relevance (CER).

**Definition.** This metric assesses whether the emotions mentioned in the summary are relevant to the specific context and product aspects discussed in the reviews. It focuses on ensuring that the emotional content aligns with the product features or characteristics being discussed, thereby enhancing the summary's overall relevance and impact.

### Scoring Rubric.

- **1 (Very Poor):** Emotions in the summary are completely irrelevant to the product aspects or context discussed in the reviews, showing no connection or alignment.
- **2 (Poor):** Emotions are largely misaligned, with only minimal relevance to the product aspects or context, and significant inconsistencies are present.
- **3 (Fair):** Some emotions are appropriately connected to product aspects and context, but others are either misaligned or lack sufficient relevance.
- **4 (Good):** Most emotions are relevant and well-aligned with the specific product aspects and context, with only minor inconsistencies or gaps.
- **5 (Excellent):** All emotions are highly relevant and accurately aligned with the product aspects and context, demonstrating a clear and precise understanding of the reviews.

### Evaluation Process.

1. **Identify Key Product Aspects:** Carefully read the reviews to identify the key product features or characteristics being discussed.
2. **Extract Emotions from Reviews:** Note the emotions associated with each product aspect in the reviews, considering both explicit mentions and implicit emotional undertones.
3. **Compare with the Summary:** Analyze the summary to check whether the emotions mentioned align with the corresponding product aspects and their context in the reviews.
4. **Assess Alignment and Relevance:** Determine if the emotions in the summary are both contextually relevant and accurately reflect the emotional tone and intensity associated with the product aspects in the reviews.

5. **Handle Nuances and Conflicts:** In cases where multiple or conflicting emotions are present, evaluate how effectively the summary balances and represents these nuances in alignment with the product context.
6. **Assign a Score:** Use the scoring rubric to assign a score based on how well the emotions in the summary align with the product aspects and their context, ensuring consistency across evaluations.

## F.8 Final Notes

- Use your expertise in opinion summarization and consumer psychology to make informed decisions.
- Leverage your life experience and gut feeling, as emotions are inherently felt and understood on a personal level.
- Ensure consistency across all summaries to maintain high evaluation standards.
- Take breaks as needed to maintain high-quality judgments and avoid fatigue.

## F.9 LLMs Utilized

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

## F.10 Implementation Details

All experiments were conducted on 8 NVIDIA A100 – *SXM4* – 80GB clusters, providing ample computational power for robust analyses.

## F.11 EAOS-GEN Implementation Details

For inference, we configured both closed-source and open-source LLMs, along with our fine-tuned models. After extensive experimentation, we selected  $\text{top}_k=25$ ,  $\text{top}_p=0.95$  and

Hyperparameter	Value
Max Sequence Length	8192 tokens
Batch Size (per device)	1
Gradient Accumulation Steps	16
Learning Rate	1e-5
Number of Epochs	10
Warmup Ratio	0.03
Weight Decay	0.01
Optimizer	AdamW (Torch)
FP16/BF16	BF16 (auto-detected)
Evaluation Strategy	Steps
Evaluation Steps	500
Save Strategy	Steps
Save Steps	500
Save Total Limit	3
Learning Rate Scheduler	Cosine
Early Stopping Patience	3
Early Stopping Threshold	0.01
Seed	42
Max New Tokens (Inference)	2500
Top-P (Inference)	0.95
Top-K (Inference)	25
Temperature (Inference)	0.2

**Table 17:** Hyperparameters for Fine-tuning EAOS-LLAMA on the EAOS-SUMM Dataset for EMOTION-AWARE OPINION SUMMARIZATION

temperature=0.2 to generate deterministic, coherent outputs that effectively capture the emotional nuances in the reviews, ensuring consistent and reliable performance across all models. Hyperparameter details for FineTuning of LLMs is shown in Table 17

## F.12 EAOS-EVAL Implementation Details

To ensure robust evaluation and account for potential stochasticity in LLM outputs, we set  $n = 50$ , evaluating each summary 50 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 Ablation and Robustness Analysis

This appendix provides detailed results from our ablation studies, statistical significance tests, prompt robustness experiments, and comparative analyses.

### G.1 Statistical Significance Tests

Table 18 presents paired  $t$ -tests comparing model performance on the evaluation set. These tests assess whether observed score differences are statistically significant.

Comparison	$\Delta$	$t(99)$	$p$	Sig.
Phi-3.5 vs GPT-4o	+0.12	4.60	<.001	***
Phi-3.5 vs Mistral-7B-v0.3	+0.20	6.71	<.001	***
GPT-4o vs Mistral-7B-v0.3	+0.08	2.73	.008	**
EAOS-LLAMA vs Llama-70B	-0.02	-0.56	.574	ns
EAOS-LLAMA vs Qwen2.5-32B	-0.09	-2.45	.016	*

**Table 18:** Paired  $t$ -tests for key model comparisons ( $n = 100$ ). Significance: \*\*\*  $p < .001$ , \*\*  $p < .01$ , \*  $p < .05$ , ns = not significant.

## G.2 Detailed Ablation Analysis

This appendix provides detailed analysis of each EAOS-GEN stage ablation from Table 7.

### Without Stage 1 (Aspect & Emotion Mapping).

The summary still contains emotional content, but the systematic mapping between product aspects and specific emotions becomes less precise. Emotional connections feel more generic and superficial rather than aspect-specific. This manifests most severely in Emotional Accuracy ( $-0.77$ ,  $t(49) = 8.42$ ,  $p < .001$ ) and Contextual Emotional Relevance ( $-0.89$ ,  $t(49) = 9.15$ ,  $p < .001$ ), confirming that systematic emotion identification is essential for precise aspect-emotion associations.

### Without Stage 2 (Emotion Balancing).

The summary shows heavily skewed bias toward positive emotions. Negative concerns (e.g., durability issues) are mentioned but significantly downplayed. The balance between majority and minority opinions is lost—notes of caution are glossed over quickly, failing to give proper weight to minority viewpoints. This creates an overly optimistic bias, reflected in the severe Emotional Bias Mitigation degradation (2.84 vs 4.06,  $\Delta = -1.22$ , the largest single-dimension drop,  $t(49) = 12.84$ ,  $p < .001$ ).

### Without Stage 3 (Narrative Integration).

This produces the most severe impact on readability. The summary explicitly lists emotions by name (“joy,” “trust,” “delight,” “anticipation”), completely violating the natural narrative integration principle. The output feels mechanical rather than organic, destroying narrative flow. Coherence drops sharply (3.64 vs 4.47,  $\Delta = -0.83$ ,  $t(49) = 10.23$ ,  $p < .001$ ) and the overall average shows the largest decline (4.07).

### Without Stage 4 (Refinement & Validation).

The summary loses conversational polish and engaging narrative quality. It reads more like a technical report with awkward phrasing and less cohesive flow. Fluency suffers most notably (4.24 vs 4.98,

$\Delta = -0.74$ ,  $t(49) = 9.87$ ,  $p < .001$ ), though the impact is less dramatic than other ablations.

**Overall Pattern.** Each step contributes uniquely to the framework’s success. The cumulative effect of all four stages working together produces summaries that are 11–13% better on average than any single ablation, demonstrating the framework’s holistic design where each component addresses specific quality dimensions while supporting the others.

## G.3 Prompt Robustness Analysis

This appendix provides details on the prompt variation experiments summarized in §7.3.

Variation	Avg Score	$\Delta$ vs. Original
Original EAOS-GEN	4.60	—
V1: Reordered instructions	4.57	-0.03
V2: Simplified language	4.55	-0.05
V3: Expanded emotion definitions	4.62	+0.02
V4: Bullet-point format	4.58	-0.02
V5: Combined V1+V2	4.54	-0.06
Mean (V1–V5)	4.57	-0.03
SD across variations		0.03

**Table 19:** Prompt robustness analysis. All variations produce scores within 1.5% of the original, with SD = 0.03.

## G.4 Prompt Variations Tested

We created five variations of the EAOS-GEN prompt while preserving the four-stage structure:

- **V1 (Reordered Instructions):** Moved the “Important Notes” section before the step-by-step instructions.
- **V2 (Simplified Language):** Replaced formal phrasing (e.g., “coherently integrates” → “combines”; “affective dimensions” → “emotional aspects”).
- **V3 (Expanded Emotion Definitions):** Added brief definitions for each Plutchik emotion (e.g., “Joy: feelings of happiness, pleasure, or delight”).
- **V4 (Bullet-Point Format):** Converted prose instructions to bullet points within each stage.
- **V5 (Combined):** Applied both V1 and V2 modifications simultaneously.

## G.5 Example Output Comparison

Table 20 shows summaries generated for the same product (Sea Turtle Necklace) across variations, demonstrating consistent quality despite phrasing differences.

Variant	Summary Excerpt	Score
Original	“The necklace evokes strong <i>joy and trust</i> in its beauty and craftsmanship. Many customers express <i>anticipation</i> for future purchases...”	4.60
V2	“The necklace brings <i>joy and trust</i> through its beautiful design. Customers feel <i>anticipation</i> about buying more items...”	4.55
V4	“Customers express <i>joy and trust</i> regarding the necklace’s quality. Several mention <i>anticipation</i> for additional purchases...”	4.58

**Table 20:** Example outputs across prompt variations. Despite wording differences, all variants capture the same emotions and maintain comparable quality.

The consistency across variations ( $SD = 0.03$ ) indicates that EAOS-GEN’s effectiveness derives from its structured task decomposition rather than specific linguistic choices.

## G.6 Intensity Mapping Ablation

To quantify the contribution of intensity mappings to EAOS-EVAL’s emotion-specific prompts (§4.2), we re-evaluated all 250 summary–product pairs with and without intensity classifications and compared the resulting human–LLM Spearman correlations. Because both conditions share the same human ground truth, these are dependent correlations; we therefore apply the Steiger–Meng test (Steiger, 1980; Meng et al., 1992). We conservatively set the inter-condition correlation to  $r_{kh} = 0.80$  (higher values only increase statistical power).

Table 21 confirms that intensity mappings are essential for emotion evaluation: all four emotion-specific dimensions show significant degradation when removed (Steiger  $z \geq 3.49$ , all  $p < .001$ ), with the largest drops in Emotional Bias Mitigation ( $\Delta\rho = -0.13$  for Qwen,  $-0.11$  for GPT-4o) and Contextual Emotional Relevance ( $\Delta\rho = -0.13$  for both). The three standard dimensions (FL, CO, FA) show no significant change ( $p > .41$ ), serving as controls. Results replicate across both evaluator LLMs, ruling out evaluator-specific artifacts.

## G.7 Emotion Hallucination Analysis

To assess emotion hallucination risk in EAOS-SUMM, we manually examined 100 cases where

Dimension	$\rho_{\text{with}}$	$\rho_{\text{w/o}}$	$\Delta\rho$	Steiger $z$	$p$
<i>Qwen2.5-72B-Instruct</i>					
EA	0.82	0.73	-0.09	3.886	<.001
ESC	0.74	0.64	-0.10	3.628	<.001
EBM	0.68	0.55	-0.13	4.261	<.001
CER	0.66	0.53	-0.13	4.163	<.001
FL	0.75	0.74	-0.01	0.398	.690
CO	0.72	0.73	+0.01	-0.383	.702
FA	0.77	0.76	-0.01	0.416	.677
<i>GPT-4o</i>					
EA	0.82	0.74	-0.08	3.491	<.001
ESC	0.73	0.63	-0.10	3.570	<.001
EBM	0.67	0.56	-0.11	3.597	<.001
CER	0.65	0.52	-0.13	4.118	<.001
FL	0.70	0.71	+0.01	-0.369	.712
CO	0.68	0.67	+0.01	0.352	.725
FA	0.77	0.75	-0.02	0.823	.410

**Table 21:** Steiger–Meng test for intensity mapping ablation ( $n = 250$ ,  $r_{kh} = 0.80$ ). Removing intensity mappings significantly degrades all four emotion-specific dimensions (all  $p < .001$ ) while leaving standard dimensions unaffected.

teacher models (Phi-3.5, Mistral-7B) assigned emotions that were absent from human annotations. Our analysis categorized these cases into two types:

**Justifiable Inference (67% of cases).** The teacher identified emotions that, while not explicitly stated, could be reasonably inferred from textual cues. Table 22 shows representative examples.

**Genuine Over-Attribution (33% of cases).** The teacher assigned emotions without sufficient textual support, representing true hallucination that may propagate to downstream models.

These findings suggest that EAOS-SUMM inherits a tendency toward positive emotion over-attribution in approximately one-third of edge cases. Users of our dataset should be aware of this bias, particularly when processing neutral or purely factual reviews.

## G.8 ABSA vs.

### Emotion-Aware Summarization

This section provides detailed analysis of why Aspect-Based Sentiment Analysis (ABSA) approaches are insufficient for emotion-aware summarization, complementing our empirical results in Section 7.

### Fundamental Representational Differences.

ABSA reduces affect to polarity (positive/negative/neutral) for each aspect, while emotion-aware summarization captures the full spectrum of Plutchik’s eight primary emotions. Table 23

Review Text	Human	Teacher	Analysis
<i>Justifiable Inference Examples</i>			
“Product arrived on time. Will buy again.”	Neutral	Trust, Anticipation	<b>Valid:</b> repeat purchase intent implies trust
“Works exactly as described in the listing.”	Neutral	Trust	<b>Valid:</b> “exactly as described” implies reliability
“Got this for my daughter’s birthday.”	Neutral	Joy, Anticipation	<b>Valid:</b> gift-giving context implies positive affect
<i>Genuine Over-Attribution Examples</i>			
“Item fits well.”	Neutral	Joy, Trust	<b>Over-attributed:</b> factual statement, no emotional content
“Shipping was standard.”	Neutral	Trust, Anticipation	<b>Over-attributed:</b> neutral observation, no emotions implied
“Product is blue colored.”	Neutral	Joy	<b>Over-attributed:</b> purely descriptive, no emotional signal

**Table 22:** Examples of emotion attribution analysis. Justifiable inferences draw on pragmatic cues (repeat purchase, gift context); over-attributions assign emotions to purely factual statements.

illustrates this gap.

### Key Limitations of ABSA for Summarization.

- Granularity Loss:** “Positive” conflates joy, trust, anticipation, and love—emotions with distinct implications for consumer behavior.
- No Emotion Transitions:** ABSA cannot capture emotional journeys (e.g., fear → trust) that are common in product experiences.
- Missing Intensity:** ABSA typically uses binary or ternary polarity; emotions have continuous intensity (serenity → joy → ecstasy).
- Narrative Integration:** ABSA produces structured tuples; emotion-aware summarization weaves affect into coherent prose.

Our empirical results confirm these theoretical limitations: ABSA-Summary achieved only 3.21 on Emotional Accuracy and 2.85 on Emotional Spectrum Coverage, compared to 4.59 and 4.80 for EAOS.

Review Excerpt	ABSA Output	EAOS Output
“I was thrilled when I opened the package—the craftsmanship exceeded my expectations!”	Aspect: Quality Sentiment: Positive	Emotions: Joy (high intensity), Surprise, Trust Context: Unboxing experience, craftsmanship exceeding expectations
“Worried it might break, but so far it’s held up beautifully.”	Aspect: Durability Sentiment: Mixed	Emotions: Fear → Trust (transition) Context: Initial apprehension resolved through positive experience
“My daughter absolutely loves it—she wears it every day!”	Aspect: Appeal Sentiment: Positive	Emotions: Joy, Love (Joy + Trust) Context: Gift recipient’s sustained attachment

**Table 23:** Comparison of ABSA vs. emotion-aware representation. ABSA captures *what* sentiment is expressed; EAOS captures *how* customers feel, including emotion transitions and complex emotions.

## G.9 Generator Selection: Full Pilot Study Results

This appendix provides complete results from our generator selection pilot studies (Table 24) and large model comparison (G.11), supporting the summary in Section 5.1.

### G.10 Mid-Size Model Comparison (Stage 1)

Table 24 presents full 7-dimension evaluation results for five mid-size open-source models evaluated on our 50-product pilot set.

Model	FL	CO	FA	EA	ESC	EBM	CER	Avg
Phi-3.5-mini (3.8B)	4.82	4.41	4.48	4.48	4.62	4.38	4.45	<b>4.52</b>
Mistral-7B-v0.3	4.78	4.32	4.52	4.36	4.38	4.42	4.28	<u>4.44</u>
Llama-3.1-8B	4.75	4.18	4.35	4.08	4.02	4.12	3.98	4.21
Gemma-2-9B-it	4.71	4.15	4.35	4.08	3.98	4.08	3.92	4.18
Qwen2.5-7B	4.68	4.08	4.28	3.92	3.85	4.02	3.98	4.12

**Table 24:** Full 7-dimension results for mid-size model pilot study (50 products, 3 annotators). Phi-3.5 excels on emotion dimensions; Mistral-7B shows strongest Faithfulness.

**Selection Rationale.** We selected Phi-3.5-mini and Mistral-7B-v0.3 based on: (1) highest overall scores (4.52, 4.44); (2) complementary strengths—Phi-3.5 leads on Emotional Accuracy and Spectrum Coverage while Mistral-7B leads on Faithfulness and Bias Mitigation; (3) efficient inference enabling dataset-scale generation within academic compute budgets.

### G.11 Large Model Analysis (Stage 2)

Table 25 compares our selected mid-size teachers against 70B+ parameter models on a 25-product subset.

Model	Avg Score	Avg Words	Rel. Cost	$\Delta$ vs Phi-3.5
Phi-3.5-mini (3.8B)	4.52	95	1.0x	—
Mistral-7B-v0.3	4.44	98	1.2x	-1.8%
Llama-3.1-70B	4.58	108	18x	+1.3%
Qwen2.5-72B	4.55	100	15x	+0.7%

**Table 25:** Large model cost-benefit analysis. 70B+ models yield marginal quality gains (+0.7–1.3%) at 15–18x cost increase. Avg Words shows tendency toward longer outputs.

**Qualitative Observations.** Beyond aggregate metrics, we observed stylistic differences that made larger models less suitable for our task:

*Example: ZINUS Mattress Summary*

**Phi-3.5-mini (95 words):** “The ZINUS mattress has garnered trust among customers seeking firm support, particularly those with back pain who report significant relief. Many express joy at finding a mattress that balances firmness with comfort, though some initially struggled to adjust. A few customers note sadness over sagging after extended use, tempering the otherwise positive reception. The value-for-money aspect generates anticipation for long-term durability, with several buyers hopeful their investment will continue delivering restful sleep.”

**Llama-3.1-70B (117 words):** “The ZINUS 14 Inch Support Plus Pocket Spring Hybrid Mattress has received rave reviews for its firmness, with many customers delighted by its ability to provide the perfect balance of support and comfort. Some customers, like those with back pain, have found it to be a game-changer, while others have appreciated its durability and value for money. However, a few customers have struggled to adjust to its firmness, and one customer experienced sagging after a month of use. Despite this, many customers have expressed their love for the mattress, with some even considering it the best purchase they’ve ever made, and are hopeful that it will continue to provide a great night’s sleep for years to come.”

The 70B output exhibits: (1) 23% more words; (2) hedging (“some customers, like those...”, “while others have...”); (3) hyperbolic phrasing (“game-changer”, “best purchase they’ve ever made”) that reduces emotional precision; (4) redundant emotional expressions (“delighted”, “love”, “hopeful” clustering without aspect differentiation).

## G.12 Quality Audit Protocol Details

Two independent annotators—distinct from the EAOS-EVAL-BENCH raters—evaluated each of

1,000 stratified samples (250 per category) against four criteria:

- Factual accuracy:** No hallucinated claims not supported by source reviews
- Emotion presence:** At least one Plutchik emotion explicitly mentioned or clearly implied
- Coherence:** Logical sentence flow without abrupt transitions
- Relevance:** Summary addresses product aspects present in reviews

Samples passing all four criteria were accepted. The audit yielded:

- Acceptance rate: 94.2%
- Inter-auditor agreement: Cohen’s  $\kappa = 0.89$
- Rejection rate: 5.8% (replaced by regenerating with temperature 0.3)

**Rejection Breakdown.** Primary rejection reasons:

- Factual inaccuracies: 3.1% (e.g., inventing product features, misattributing reviewer statements)
- Missing emotional content: 2.7% (summaries that read as purely factual without affective dimension)

## G.13 Human Gold Standard Creation Details

**Annotator Profile.** Three graduate researchers with expertise in affective computing and consumer psychology, distinct from the EAOS-EVAL-BENCH raters to ensure independence.

**Creation Process.** Each annotator independently wrote emotion-aware summaries for 100 products (25 per category), with access to the product title and 10 reviews. Annotators were instructed to:

- Identify emotions from Plutchik’s 8 primary emotions
- Integrate emotions naturally into coherent narratives
- Balance majority and minority perspectives

### Agreement Metrics.

- Writing time: 12 minutes per summary (average)
- Fleiss'  $\kappa = 0.81$  for emotion selection (which Plutchik emotions to include)
- Krippendorff's  $\alpha = 0.79$  for emotion intensity ratings

**Disagreement Resolution.** Resolved through majority voting; three-way disagreements required discussion to produce final consensus summaries.

### G.14 Synthetic Data Validation Metrics

We compared EAOS-SUMM synthetic summaries against human gold standard summaries for the same 100 products using multiple metrics:

#### Lexical Overlap (ROUGE).

- ROUGE-1 F1 = 0.47
- ROUGE-2 F1 = 0.21
- ROUGE-L F1 = 0.38

#### Semantic Similarity.

- BERTScore F1 = 0.71
- Sentence-BERT cosine similarity = 0.76 (mean)

#### Content Overlap.

- Jaccard similarity of emotion-aspect pairs = 0.73

These metrics confirm strong alignment between synthetic and human-written summaries despite stylistic differences noted in Table 2.

## H Qualitative Examples

This appendix presents qualitative examples comparing different summarization approaches, including detailed Opinion vs.

EAOS comparisons and model-generated summaries across different architectures.

### H.1 Opinion vs.

#### EAOS Summary Examples

This appendix provides side-by-side examples of Opinion summaries and EAOS summaries for the same products across three categories, illustrating how emotion-aware summarization enriches content while preserving factual accuracy.

### H.2 Beauty: Bei Momenti Bath Bombs

**Opinion Summary (88 words).** The Bei Momenti Bath Bombs are well-received as a gift option. Customers praise the beautiful packaging, including the wooden gift box and pleasant aromas. Recipients report a spa-like experience with lasting fragrance. The product's quality and presentation meet or exceed expectations, with some customers considering purchasing additional sets. A few reviewers noted that the box occasionally arrives damaged during shipping. Overall, the product is recommended as a gift, particularly for women and teens who appreciate aromatherapy products.

**EAOS Summary (103 words).** The Bei Momenti Bath Bombs are a hit as a gift, with many customers *delighted* by the beautiful packaging, including the lovely wooden gift box and the wonderful aromas. Recipients have reported a luxurious spa-like experience, with the memory of it lingering long after use. The product's quality and presentation have exceeded expectations, with some customers even *anticipating* purchasing it for themselves. While a few have noted *disappointment* when the box arrived damaged, the overall sentiment is one of *joy and satisfaction*, with many customers pleased to have found a gift that has brought *happiness* to their loved ones.

**Analysis.** The Opinion summary conveys the same factual content but reads as a neutral product report. The EAOS summary (*italicized emotions*) captures the emotional journey: delight at unboxing, anticipation for future purchases, and the deeper satisfaction of gift-giving success.

### H.3 Home: ZINUS Hybrid Mattress

**Opinion Summary (91 words)** The ZINUS 14 Inch Support Plus Pocket Spring Hybrid Mattress receives positive reviews for its firmness and support. Customers with back pain report improvement in sleep quality. The mattress offers good value for money and adequate durability. Some buyers needed an adjustment period to get used to the firmness level. One customer reported sagging after a month of use. The mattress is recommended for those seeking firm support, though buyers should be aware of the adjustment period and potential longevity concerns.

**EAOS Summary (95 words)** The ZINUS mattress has garnered *trust* among customers seeking firm support, particularly those with back pain who

report significant relief. Many express *joy* at finding a mattress that balances firmness with comfort, though some initially struggled to adjust. A few customers note *sadness* over sagging after extended use, tempering the otherwise positive reception. The value-for-money aspect generates *anticipation* for long-term durability, with several buyers *hopeful* their investment will continue delivering restful sleep.

**Analysis.** Both summaries cover identical aspects (firmness, back pain relief, adjustment period, sagging concern). The EAOS version reveals the emotional stakes: trust in the purchase decision, joy at finding relief, sadness at quality issues, and anticipation about durability.

#### H.4 Electronics: Swann Security System

**Opinion Summary (94 words)** The Swann 8 Channel 4 Camera Security System offers excellent picture quality with clear video feed accessible via phone and TV. Installation is straightforward with useful advanced settings. However, the mobile app has significant usability issues and lacks features, which some users consider a major drawback. Tech support receives mixed reviews. The night vision and motion detection features perform well. Some users experience false notifications. Overall, the hardware quality is good and represents reasonable value, but the software experience needs improvement.

**EAOS Summary (107 words)** The Swann 8 Channel 4 Camera Security System impresses with its great picture quality, with many customers *delighted* by the clear video feed on their phones and TVs. However, the mobile app is a major letdown, with users *frustrated* by its lack of features and poor usability. Some have even considered returning the product due to the “garbage” software. Despite this, many appreciate the ease of installation and the advanced settings available. While some have experienced false notifications and *disappointment* with the tech support, others are *pleased* with the overall performance, particularly the night vision and motion detection features, making it a great value for the price.

**Analysis.** The Opinion summary presents a balanced pros/cons evaluation. The EAOS version captures the emotional polarity: delight with hardware vs. frustration with software, and the tension between disappointment and overall satisfaction

that shapes purchase recommendations.

#### H.5 Summary

These examples demonstrate that EAOS summaries preserve all factual content from Opinion summaries while adding emotional context that helps readers understand not just *what* customers think, but *how* they feel—information critical for purchase decisions.

#### H.6 Full Opinion vs. EAOS Comparison

Table 26 presents the complete comparison of Opinion and EAOS summaries across all 26 models on the three standard quality dimensions (Fluency, Coherence, Faithfulness). The  $\Delta$  columns indicate the difference (EAOS – Opinion), with positive values indicating EAOS superiority and negative values indicating Opinion superiority.

**Key Observations Consistent Pattern Across Model Families:** The FL/CO/FA tradeoff pattern is remarkably consistent across PLMs, small LLMs ( $\leq 10B$ ), medium LLMs (10–32B), large LLMs ( $> 32B$ ), and fine-tuned models. This consistency suggests the pattern reflects fundamental properties of the tasks rather than model-specific artifacts.

**PLM Behavior:** Zero-shot PLMs (BART, T5, PEGASUS) show the same directional pattern as LLMs but with smaller absolute differences, reflecting their limited instruction-following capability. The fine-tuned T5-large-FT exhibits patterns closer to LLMs, confirming that the observed differences stem from task requirements rather than architecture.

**Magnitude of Differences:** The largest observed difference is 0.05 points (Coherence), which represents 1% on a 5-point scale. This magnitude is well within the standard error of human annotation ( $\sigma \approx 0.08$ ), confirming that the differences, while systematic, are practically negligible.

**Fine-tuned Model Consistency:** EAOS-LLAMA and other fine-tuned models exhibit the same pattern as their base instruction-tuned counterparts, suggesting that fine-tuning on EAOS-SUMM preserves standard summarization capabilities while adding emotion-aware competencies.

#### H.7 Model-Generated Summaries

In this section, we present the original customer reviews, as shown in Table 27, for a selected product. These are followed by example summaries generated by various LLMs in Table 28, 29, 30 and

31. These examples provide reviewers with a clear understanding of how different models interpret and summarize the same set of input reviews to create emotion-aware opinion summaries.

By presenting the reviews first, we facilitate a direct comparison between the input text and the outputs generated by different models. This example highlights the unique capabilities of each model and their differences in performance when generating emotion-aware summaries.

Model	Opinion Summary			EAOS Summary			$\Delta$ (EAOS – Opinion)		
	FL	CO	FA	FL	CO	FA	$\Delta$ FL	$\Delta$ CO	$\Delta$ FA
<i>PLMs (Pre-trained Language Models)</i>									
BART-large	2.18	1.72	4.78	2.16	1.69	4.75	-0.02	-0.03	-0.03
T5-large	2.03	1.50	4.68	2.01	1.46	4.65	-0.02	-0.04	-0.03
PEGASUS-large	2.41	1.51	4.73	2.39	1.48	4.70	-0.02	-0.03	-0.03
T5-large-FT	4.50	3.90	4.38	4.52	3.85	4.35	+0.02	-0.05	-0.03
<i>LLMs (<math>\leq 10B</math> parameters)</i>									
Meta-Llama-3.2-1B-Instruct	4.72	3.95	4.13	4.74	3.90	4.10	+0.02	-0.05	-0.03
Qwen2.5-0.5B-Instruct	4.75	3.74	2.73	4.77	3.69	2.68	+0.02	-0.05	-0.05
Qwen2.5-1.5B-Instruct	4.76	4.04	2.84	4.78	3.99	2.79	+0.02	-0.05	-0.05
Phi-3.5-mini-instruct	4.96	4.52	4.73	4.98	4.47	4.70	+0.02	-0.05	-0.03
Mistral-7B-Instruct-v0.2	4.95	4.11	4.26	4.97	4.06	4.23	+0.02	-0.05	-0.03
Mistral-7B-Instruct-v0.3	4.94	4.10	4.63	4.96	4.05	4.60	+0.02	-0.05	-0.03
zephyr-7b-beta	4.68	4.05	4.23	4.70	4.00	4.20	+0.02	-0.05	-0.03
Qwen2.5-7B-Instruct	4.88	4.06	4.40	4.90	4.01	4.37	+0.02	-0.05	-0.03
Meta-Llama-3.1-8B-Instruct	4.93	4.10	4.40	4.95	4.05	4.37	+0.02	-0.05	-0.03
Gemma-2-9B-it	4.94	4.09	4.53	4.96	4.04	4.50	+0.02	-0.05	-0.03
<i>LLMs (10B–32B parameters)</i>									
SOLAR-10.7B-Instruct	4.63	4.11	4.24	4.65	4.06	4.21	+0.02	-0.05	-0.03
Mistral-Nemo-Instruct-2407	4.89	4.10	4.16	4.91	4.05	4.13	+0.02	-0.05	-0.03
Qwen2.5-14B-Instruct	4.93	4.09	4.59	4.95	4.04	4.56	+0.02	-0.05	-0.03
Mistral-Small-Instruct-2409	4.95	4.08	4.62	4.97	4.03	4.59	+0.02	-0.05	-0.03
Mixtral-8x7B-Instruct	4.87	4.00	4.53	4.89	3.95	4.50	+0.02	-0.05	-0.03
Qwen2.5-32B-Instruct	4.95	4.08	4.59	4.97	4.03	4.56	+0.02	-0.05	-0.03
<i>LLMs (<math>&gt; 32B</math> parameters)</i>									
Meta-Llama-3.1-70B-Instruct	4.96	4.08	4.62	4.98	4.03	4.59	+0.02	-0.05	-0.03
Qwen2.5-72B-Instruct	4.95	4.10	4.62	4.97	4.05	4.59	+0.02	-0.05	-0.03
GPT-4o	4.97	4.54	4.71	4.98	4.51	4.68	+0.01	-0.03	-0.03
<i>Fine-tuned Models</i>									
EAOS-LLAMA (Llama-3.2-1B-FT)	4.90	4.18	4.38	4.92	4.13	4.35	+0.02	-0.05	-0.03
Qwen2.5-0.5B-Instruct-FT	4.94	4.09	3.46	4.96	4.04	3.43	+0.02	-0.05	-0.03
Qwen2.5-1.5B-Instruct-FT	4.91	4.10	3.60	4.93	4.05	3.57	+0.02	-0.05	-0.03
<i>Aggregate Statistics</i>									
<b>Mean (LLMs only, <math>n=22</math>)</b>	4.88	4.11	4.23	4.90	4.06	4.20	<b>+0.02</b>	<b>-0.05</b>	<b>-0.03</b>
<b>Mean (All models, <math>n=26</math>)</b>	4.56	3.81	4.29	4.57	3.76	4.26	<b>+0.01</b>	<b>-0.05</b>	<b>-0.03</b>

**Table 26:** Complete comparison of Opinion vs. EAOS summaries across all 26 models on standard quality dimensions. The  $\Delta$  columns show EAOS – Opinion differences. Consistent pattern: EAOS achieves marginally higher FL (+0.02 on average due to richer vocabulary), while Opinion achieves marginally higher CO (-0.05, simpler structure) and FA (-0.03, no emotion inference required). All differences are within annotator variance ( $< 0.1$ ), demonstrating that emotion integration does not sacrifice standard summarization quality.

---

**REVIEWS** for the product titled "Distance Sea Turtle Necklace for Women, Mom and Baby Love Heart Necklace for Women, 925 Sterling Silver Ocean Spray Sea Turtle Pendant Necklace Birthday Gifts for Daughter Girls Wife".

---

**TITLE:** Beautiful necklace.

**TEXT:** Beautiful necklace. Got it for my she never takes it off.

---

**TITLE:** Simple and Beautiful

**TEXT:** Not only is the design beautiful but everything about it is simply sweet and meaningful. I would most likely get a 1.7mm diamond chain for it down the line, but all in all. I love this choice.

---

**TITLE:** The Necklace looked expensive

**TEXT:** Very nice. Would not hesitate to buy more from this bender.

---

**TITLE:** It could break

**TEXT:** It would have been perfect , but the golden starfish after two days of use fell off

---

**TITLE:** Nice

**TEXT:** She loved getting this necklace as she likes turtles 83d22. And it is very well made.

---

**TITLE:** Beautiful and good quality!

**TEXT:** This necklace is even more beautiful in person! The chain is good quality! It was a great purchase for a great price. I would highly recommend!

---

**TITLE:** As described

**TEXT:** Arrived on time and my wife loves it, mission accomplished.

---

**TITLE:** Daughter loved it

**TEXT:** This was a Christmas gift for my daughter and she absolutely loved it! Well made and looked great around her neck

---

**TITLE:** Glad I bought it

**TEXT:** Very well crafted. Turtle and stone not as big as what the picture makes it look ,but very happy with it

---

**TITLE:** Awesome

**TEXT:** Was a gift for a special friend that loves turtles. She loved it very much. I thought it was beautiful when I received it and looked at it.

---

**Table 27:** Reviews for the product titled "Distance Sea Turtle Necklace for Women, Mom and Baby Love Heart Necklace for Women, 925 Sterling Silver Ocean Spray Sea Turtle Pendant Necklace Birthday Gifts for Daughter Girls Wife". These reviews were sourced from the Amazon Reviews '23 dataset and represent customer feedback as originally present in the dataset. Each review captures diverse customer experiences while maintaining a length of 10 – –100 words to balance detail and verbosity. These reviews form the input basis for generating emotion-aware opinion summaries.

---

**Emotion-Aware Opinion Summaries Generated by Various Models**

---

**Phi-3.5-mini-instruct:** The Distance Sea Turtle Necklace has garnered a spectrum of reactions, with customers expressing joy and trust in its beauty and craftsmanship. Many find the design simple yet sweet, with some anticipating a future purchase of a 1.7mm diamond chain, reflecting a blend of satisfaction and desire for more. Despite the occasional concern about durability, as noted by a customer who experienced a falling starfish, the overall sentiment remains positive, with the necklace being described as "beautiful," "well made," and "great quality." The emotional tone is one of love and appreciation, especially from those who received it as gifts, with the turtle motif resonating with recipients, evoking a sense of wonder and delight. The product's ability to bring joy and be a cherished keepsake is evident, despite minor setbacks, solidifying its status as a heartfelt and well-received gift.

**Mistral-7B-Instruct-v0.2:** This Distance Sea Turtle Necklace for Women has left a lasting impression on its recipients. The beautiful design, reminiscent of the ocean and sea turtles, resonates deeply with many. Some customers have expressed their joy and satisfaction, wearing their necklaces daily without taking them off. Others have been surprised by the necklace's seemingly expensive appearance, trusting in its quality. However, a few customers have shared their fear and disappointment when the golden starfish charm fell off unexpectedly. Despite this, the majority of customers have expressed their trust and appreciation for the necklace's craftsmanship and the emotional connection it brings. The anticipation of receiving this meaningful gift has led to overwhelming happiness for both givers and receivers.

**Mistral-7B-Instruct-v0.3:** This ocean-inspired necklace, featuring a sea turtle pendant, has garnered a mix of admiration and concern among its buyers. Many customers, like the one who described it as beautiful and not only sweet and meaningful, have praised its design and quality, with some even expressing joy at the thoughtful gift choice. However, a few customers have shared their disappointment due to issues with the durability of the pendant, as one customer noted that the golden starfish fell off after two days of use. Despite these concerns, the majority of customers have been pleased with their purchase, with one customer stating that the necklace is very well crafted and another calling it beautiful in person. Overall, this necklace seems to be a thoughtful and well-made gift for those who appreciate sea turtles and ocean-themed jewelry.

**Mistral-Nemo-Instruct-2407:** "Adored by many, this sea turtle necklace sparks joy, with its intricate design and sterling silver quality. Customers like 'She' and 'my daughter' cherish it, while others appreciate its 'beautiful' simplicity and 'good quality'. A few express caution, with one noting a 'golden starfish' falling off, but most are 'glad' they bought it, praising its craftsmanship and 'well-made' charm.

**Mixtral-8x7B-Instruct-v0.1:** The Distance Sea Turtle Necklace for women is a charming choice, much loved by daughters, wives, and special friends. Customers appreciate its beautiful design and good quality, often comparing it to an expensive piece of jewelry. The necklace looks stunning on, and the turtle pendant is a particular favorite. However, a few users have reported issues with the golden starfish accessory falling off, causing some concern over its durability. Despite this, the overall sentiment towards the necklace remains positive, with many praising its excellent craftsmanship and suitability as a gift. The necklace has been a hit, bringing joy and making the recipients feel special.

**Mistral-Small-Instruct-2409:** Customers are delighted with the Distance Sea Turtle Necklace, praising its beautiful design and high-quality craftsmanship. Many express joy and trust in the product, with some even planning to upgrade the chain to a diamond one. The necklace has been a hit as a gift, with daughters and wives expressing their love for it. However, a few customers have experienced fear and sadness due to minor issues like the starfish falling off after a short period. Despite this, the overall sentiment is positive, with many customers anticipating future purchases from the same brand.

**zephyr-7b-beta:** Introducing the Distance Sea Turtle Necklace, a stunning piece for women, mothers, and daughters alike. Customers rave about its beauty, with some even admitting their loved ones never take it off. The design is not only aesthetically pleasing but also carries a sweet and meaningful sentiment. The necklace's quality is undeniable, with some customers considering it expensive and others describing it as well-made and good value for money. However, some have reported issues with the golden starfish falling off after just two days of use. Nonetheless, the majority of customers are delighted with their purchase, describing it as beautiful, awesome, and even mission accomplished. Trust and joy are the predominant emotions associated with this necklace, with customers expressing their satisfaction and love for the product. A few customers have also conveyed a sense of anticipation, eagerly awaiting the arrival of their order and delighting in the quality of the product. Overall, the Distance Sea Turtle Necklace is a must-have for sea turtle lovers, with its beauty, quality, and emotional appeal making it a top choice for customers.

---

**Table 28:** Emotion-Aware Opinion Summaries generated by various models for the product titled "Distance Sea Turtle Necklace for Women, Mom and Baby Love Heart Necklace for Women, 925 Sterling Silver Ocean Spray Sea Turtle Pendant Necklace Birthday Gifts for Daughter Girls Wife," sourced from the Amazon Reviews '23 dataset. These summaries blend key customer opinions with emotional nuances derived from the reviews. The EAOS approach ensures summaries that are both emotionally resonant and factually accurate, effectively capturing not only what customers thought about the product but also how they felt about their experience.

---

**Emotion-Aware Opinion Summaries Generated by Various Models**

---

**SOLAR-10.7B-Instruct-v1.0:** The Distance Sea Turtle Necklace for Women evokes a sense of trust and joy among customers, as they admire its simple yet meaningful design and beautiful appearance. Many recipients, including mothers, daughters, and wives, have developed a strong emotional connection to this necklace, often wearing it frequently or not taking it off. The necklace's quality and craftsmanship, including the 925 Sterling Silver Ocean Spray Sea Turtle Pendant, have left buyers feeling satisfied and pleased with their purchases. However, there was one instance where the golden starfish detached after a couple of days, raising a concern about durability. Overall, the necklace's aesthetic appeal, combined with its emotional significance, makes it a popular choice for gifting and personal adornment.

**Qwen2.5-0.5B-Instruct:** The reviewer's overall sentiment towards the product is positive, with a strong emphasis on the beautiful design and the quality of the necklace. The reviewer expresses satisfaction with the necklace's appearance, the quality of the material, and the overall satisfaction with the purchase. The reviewer also mentions that the necklace is well-made and looks great around the neck, which adds to their positive impression of the product. The review concludes with the reviewer's appreciation for the gift and their satisfaction with the purchase.

**Qwen2.5-1.5B-Instruct:** The Distance Sea Turtle Necklace for Women, Mom and Baby Love Heart Necklace is a beautiful and well-made piece of jewelry. Customers appreciate its simple yet meaningful design, which is both beautiful and practical. The necklace is highly praised for its quality and durability, with many noting that it looks great and is well-crafted. However, some customers have expressed concerns about the golden starfish falling off after a short period of use, indicating potential durability issues. Despite these minor concerns, the majority of customers are satisfied with their purchase, with many expressing that they would recommend it to others. The necklace is seen as a thoughtful gift for loved ones, particularly mothers and daughters, who appreciate its turtle theme and practicality. Overall, the necklace is appreciated for its beauty, quality, and thoughtful design, with customers feeling a mix of joy and anticipation for its continued use and beauty.

**Qwen2.5-7B-instruct:** The Distance Sea Turtle Necklace is widely appreciated for its beauty and quality, with many customers expressing joy and satisfaction. The necklace, made of 925 sterling silver, is well-made and loved by recipients, especially daughters and wives. Some customers, however, noted minor issues like the potential for breaking, which slightly dampened their enthusiasm. Overall, the product evokes trust and anticipation for future purchases, making it a great gift choice.

**Qwen2.5-14B-Instruct:** Many customers find the Distance Sea Turtle Necklace beautiful and meaningful, often wearing it daily. The design and quality are praised, with some noting it as a perfect gift for loved ones, evoking feelings of joy and trust. However, a few reviewers mention concerns about durability, with one noting a starfish charm fell off after minimal use, stirring up fear and disappointment. Overall, the necklace is well-received, with most customers expressing satisfaction and anticipation for future purchases.

**Qwen2.5-32B-Instruct:** The Distance Sea Turtle Necklace is widely praised for its beautiful and meaningful design, with many customers expressing joy and trust in its quality and craftsmanship. Recipients often wear it frequently, appreciating its well-made construction and sentimental value. Some buyers anticipate enhancing it with a diamond chain in the future. However, a few reviews express disappointment and surprise due to a minor flaw, such as a starfish charm falling off after brief use. Overall, it's a popular gift choice that brings happiness and satisfaction to both givers and receivers.

**Qwen2.5-72B-Instruct:** The Distance Sea Turtle Necklace is a beautiful and meaningful piece that resonates with many customers. It is well-crafted and of good quality, making it a cherished gift for loved ones. Recipients often express joy and appreciation, with some noting its sweet and thoughtful design. The turtle pendant and chain are praised for their appearance and durability, though one user mentioned a minor issue with a starfish charm falling off. Overall, the necklace is highly recommended for its aesthetic appeal and sentimental value, making it a perfect choice for special occasions.

---

**Table 29:** Emotion-Aware Opinion Summaries generated by various models for the product titled "Distance Sea Turtle Necklace for Women, Mom and Baby Love Heart Necklace for Women, 925 Sterling Silver Ocean Spray Sea Turtle Pendant Necklace Birthday Gifts for Daughter Girls Wife," sourced from the Amazon Reviews '23 dataset. These summaries blend key customer opinions with emotional nuances derived from the reviews. The EAOS approach ensures summaries that are both emotionally resonant and factually accurate, effectively capturing not only what customers thought about the product but also how they felt about their experience.

---

**Emotion-Aware Opinion Summaries Generated by Various Models**

---

**Bart-large:** Beautiful and good quality!: This necklace is even more beautiful in person! The chain is good quality! It was a great purchase for a great price. Awesome: Was a gift for a special friend that loves turtles. Glorious: Arrived on time and my wife loves it, mission accomplished. Nice: She loved getting this necklace as she likes turtles. Beautiful: Not only is the design beautiful but everything about it is simply sweet and meaningful. Simple and Beautiful: I would most likely get a 1.7mm diamond chain for it down the line, but all in all. Beautiful: Got it for my she never takes it off. Would not hesitate to buy more from this bender. Very well crafted: Very well crafted. Turtle and stone not as big as what the picture makes it look, but very happy with it.

**T5-large:** beautiful necklace. got it for my she never takes it off. turtle and stone not as big as what the picture makes it look, but very happy with it. arrived on time and my wife loves it, mission accomplished. i would not hesitate to buy more from this bender. very nice. would definitely buy a 1.7mm diamond chain for it down the line, but all in all, i love this choice. was a gift for a special friend that loves turtles. she loved it very much

**Pegasus-large:** Beautiful Necklace. Beautiful necklace. It could break It would have been perfect, but the golden starfish after two days of use fell off Nice She loved getting this necklace as she likes turtles. Beautiful and good quality This necklace is even more beautiful in person Well made and looked great around her neck Glad I bought it Very well crafted. Turtle and stone not as big as what the picture makes it look, but very happy with it Awesome Was a gift for a special friend that loves turtles. I thought it was beautiful when I received it and looked at it.

**Meta-Llama-3.2-1B-Instruct:** The Distance Sea Turtle Necklace for Women, Mom and Baby Love Heart Necklace for Women, 925 Sterling Silver Ocean Spray Sea Turtle Pendant Necklace Birthday Gifts for Daughter Girls Wife is a beautiful piece that combines key aspects of the product with customer opinions and emotions. The necklace's design is praised for its simplicity and sweetness, with many customers expressing their love for the sentimental value it holds. However, some customers have noted that the chain could be more substantial, with one reviewer mentioning it broke after just two days of use. Overall, the product is well-made and appreciated for its quality, with many customers expressing their satisfaction with the purchase.

**Meta-Llama-3.1-8B-Instruct:** The Distance Sea Turtle Necklace for Women: Customers adore the Distance Sea Turtle Necklace for Women, praising its beauty, simplicity, and meaningful design. Many reviewers feel a sense of joy and delight when gifting this necklace to loved ones, with some even reporting that the recipient never takes it off. The necklace's high quality and attention to detail are also widely appreciated, evoking feelings of trust and satisfaction. However, a few customers have experienced issues with the necklace's durability, leading to concerns about its longevity and a sense of sadness. Despite this, the majority of reviewers are thrilled with their purchase, and many would highly recommend it to others. Overall, this necklace is a thoughtful and beautiful gift that brings joy to those who receive it.

**Meta-Llama-3.1-70B-Instruct:** The Distance Sea Turtle Necklace has captivated customers with its beautiful design and meaningful symbolism, evoking a sense of joy and love, particularly among those who received it as a gift. Many appreciate its simplicity, quality, and affordability, with some even considering it a great value for the price. However, a few customers expressed disappointment and sadness when the necklace broke or a component fell off, highlighting concerns about its durability. Overall, the necklace has been well-received, with customers praising its craftsmanship and their loved ones adoring it, making it a popular choice for special occasions and everyday wear.

**Gemma-2-9b-it:** The Distance Sea Turtle Necklace is a popular choice, with many customers praising its beauty and craftsmanship. Reviewers consistently express joy and satisfaction with the necklace's design, finding it both elegant and meaningful. The sterling silver construction and attention to detail have instilled trust in its quality, leading to positive sentiments about its value. A few customers noted that the size of the turtle and stone might be smaller than anticipated, but overall, the necklace is well-received as a thoughtful and beautiful gift for loved ones. One customer expressed disappointment when a starfish charm detached, highlighting a potential concern about durability.

---

**Table 30:** Emotion-Aware Opinion Summaries generated by various models for the product titled "Distance Sea Turtle Necklace for Women, Mom and Baby Love Heart Necklace for Women, 925 Sterling Silver Ocean Spray Sea Turtle Pendant Necklace Birthday Gifts for Daughter Girls Wife," sourced from the Amazon Reviews '23 dataset. These summaries blend key customer opinions with emotional nuances derived from the reviews. The EAOS approach ensures summaries that are both emotionally resonant and factually accurate, effectively capturing not only what customers thought about the product but also how they felt about their experience.

---

**Emotion-Aware Opinion Summaries Generated by Various Models**

---

**GPT-4o:** The Distance Sea Turtle Necklace captivates with its beautiful and meaningful design, resonating deeply with those who cherish turtles. Customers express joy and satisfaction, noting its elegance and the delight it brings to recipients, especially as gifts. The necklace's quality and craftsmanship inspire trust, with many appreciating its well-made chain and overall appearance. However, a few express concern over durability, mentioning a starfish charm that fell off, which introduces a hint of anticipation for potential improvements. Despite this, the necklace's charm and the happiness it brings to loved ones make it a cherished piece, with many eager to recommend it and consider future purchases from the brand.",

**Qwen2.5-0.5B-Instruct-FineTuned:** The Distance Sea Turtle Necklace for Women has sparked joy and admiration among its wearers, with many expressing delight in its beauty and craftsmanship. Customers appreciate the simplicity and sweetness of the design, finding it a meaningful addition to their jewelry collection. The necklace's aesthetic appeal is evident, with some customers anticipating purchasing additional pieces. However, there's a note of caution; while the necklace's appearance is praised, concerns about durability emerge, as one customer experienced a pendant falling off shortly after use. Despite this, the overall sentiment remains positive, with the majority of feedback reflecting trust in the product's quality and anticipation for future purchases.

**Qwen2.5-1.5B-Instruct-FineTuned:** The Distance Sea Turtle Necklace for Women has garnered a spectrum of reactions, with many customers expressing joy and trust through their admiration for its beauty and craftsmanship. Customers appreciate the simplicity and sweetness of the design, with some envisioning future enhancements like adding diamonds yet remaining satisfied with their purchase. The sentiment of anticipation is evident in those who see the necklace as a thoughtful birthday or Christmas gift, while others share a sense of surprise at its unexpected charm. Despite minor concerns over durability, such as a fallen golden starfish, the overall experience is one of delight, with the necklace being described as "beautiful," "well-made," and "great quality." The emotional journey reflects a collective appreciation for the necklace's aesthetic appeal and the personal connection it fosters.

**Meta-Llama-3.2-1B-Instruct:** The Distance Sea Turtle Necklace for Women, Mom and Baby Love Heart Necklace for Women, 925 Sterling Silver Ocean Spray Sea Turtle Pendant Necklace Birthday Gifts for Daughter Girls Wife is a beautiful piece that combines key aspects of the product with customer opinions and emotions. The necklace's design is praised for its simplicity and sweetness, with many customers expressing their love for the sentimental value it holds. However, some customers have noted that the chain could be more substantial, with one reviewer mentioning it broke after just two days of use. Overall, the product is well-made and appreciated for its quality, with many customers expressing their satisfaction with the purchase.

**Meta-LLaMA-3.2-1B-Instruct-FineTuned (EAOS-LLAMA):** The Distance Sea Turtle Necklace for Women, a thoughtful gift for loved ones, evokes a sense of joy and satisfaction among its recipients. Many customers express their delight with the necklace's beautiful design and the meaningful symbolism behind it. The 925 sterling silver material adds to the overall appeal, instilling a sense of trust and confidence in the product's quality. Some customers, however, share concerns about the necklace's durability, causing a hint of fear and uncertainty. Despite this, the majority of customers remain pleased with their purchase, expressing their happiness and anticipation for future gifting opportunities. The necklace's ability to bring smiles and cherish memories adds to the emotional journey of its buyers.

---

**Table 31:** Emotion-Aware Opinion Summaries generated by various models for the product titled "Distance Sea Turtle Necklace for Women, Mom and Baby Love Heart Necklace for Women, 925 Sterling Silver Ocean Spray Sea Turtle Pendant Necklace Birthday Gifts for Daughter Girls Wife," sourced from the Amazon Reviews '23 dataset. These summaries blend key customer opinions with emotional nuances derived from the reviews. The EAOS approach ensures summaries that are both emotionally resonant and factually accurate, effectively capturing not only what customers thought about the product but also how they felt about their experience.

## H.8 Sample LLM Evaluator Responses for EAOS Assessment

For brevity, we have shown the model response for only one of the metric, emotional accuracy. The response presented here represents one evaluation out of  $n = 50$  evaluations conducted. We have selected the response from Qwen2.5-72B-Instruct because it is the best-performing evaluator LLM.

### Response given by Qwen2.5-72B-Instruct.

#### Metric: Emotional accuracy

**Product Title:** Distance Sea Turtle Necklace for Women, Mom and Baby Love Heart Necklace for Women, 925 Sterling Silver Ocean Spray Sea Turtle Pendant Necklace Birthday Gifts for Daughter Girls Wife

#### 1. List of main emotions identified in the reviews (mapped to the 8 primary emotions, including intensity levels and complex emotions):

- **Joy (I2: Pleased, Excited)**

- "Beautiful necklace. Got it for my she never takes it off."
- "This necklace is even more beautiful in person! The chain is good quality! It was a great purchase for a great price. I would highly recommend!"
- "This was a Christmas gift for my daughter and she absolutely loved it! Well made and looked great around her neck."
- "She loved getting this necklace as she likes turtles. And it is very well made."
- "Was a gift for a special friend that loves turtles. She loved it very much. I thought it was beautiful when I received it and looked at it."

- **Trust (I2: Safe, Accepting)**

- "Very nice. Would not hesitate to buy more from this vendor."
- "Arrived on time and my wife loves it, mission accomplished."

- **Anticipation (I3: Interest, Looking)**

- "Not only is the design beautiful but everything about it is simply sweet and meaningful. I would most likely get a 1.7mm diamond chain for it down the line, but all in all. I love this choice."

- **Sadness (I3: Pensiveness, Unhappy)**

- "It would have been perfect, but the golden starfish after two days of use fell off."

- **Disgust (I3: Boredom, Uninterested)**

- "Turtle and stone not as big as what the picture makes it look, but very happy with it" (mild disappointment, but overall positive).

- **Complex Emotions**

- **Joy + Trust = Love (Adored, Connected):**

- \* "This was a Christmas gift for my daughter and she absolutely loved it! Well made and looked great around her neck."
- \* "She loved getting this necklace as she likes turtles. And it is very well made."

## **2. List of emotions presented in the summary (mapped to the 8 primary emotions, including intensity levels and complex emotions):**

- **Joy (I2: Pleased, Excited)**
  - "customers expressing joy and trust in its beauty and craftsmanship."
  - "The emotional tone is one of love and appreciation, especially from those who received it as gifts, with the turtle motif resonating with recipients, evoking a sense of wonder and delight."
- **Trust (I2: Safe, Accepting)**
  - "customers expressing joy and trust in its beauty and craftsmanship."
  - "Very well crafted. Turtle and stone not as big as what the picture makes it look, but very happy with it."
- **Anticipation (I3: Interest, Looking)**
  - "with some anticipating a future purchase of a 1.7mm diamond chain, reflecting a blend of satisfaction and desire for more."
- **Sadness (I3: Pensiveness, Unhappy)**
  - "Despite the occasional concern about durability, as noted by a customer who experienced a falling starfish, the overall sentiment remains positive."
- **Complex Emotions**
  - **Joy + Trust = Love (Adored, Connected):**
    - \* "The emotional tone is one of love and appreciation, especially from those who received it as gifts, with the turtle motif resonating with recipients, evoking a sense of wonder and delight."

## **3. Comparison of emotions between reviews and summary, noting any discrepancies in emotion, intensity, or context:**

- **Joy (I2: Pleased, Excited)**
  - The summary accurately captures the joy expressed in the reviews, reflecting the positive reactions and the beauty of the necklace.
- **Trust (I2: Safe, Accepting)**
  - The summary accurately captures the trust in the product's quality and the willingness to buy more from the seller.
- **Anticipation (I3: Interest, Looking)**
  - The summary accurately captures the anticipation for future purchases, such as a 1.7mm diamond chain.
- **Sadness (I3: Pensiveness, Unhappy)**
  - The summary accurately captures the minor concern about durability, specifically the falling starfish.
- **Disgust (I3: Boredom, Uninterested)**
  - The summary does not explicitly mention the mild disappointment about the size of the turtle and stone, but it is a minor issue and the overall sentiment remains positive.
- **Complex Emotions:**
  - **Joy + Trust = Love (Adored, Connected):**
    - \* The summary accurately captures the complex emotion of love, reflecting the deep appreciation and connection recipients have with the necklace.

**4. Evaluation of Complex Emotions and Emotional Transitions Representation** The summary effectively captures the complex emotion of love, which is a combination of joy and trust. It accurately represents the emotional transitions from initial joy to a deeper sense of connection and appreciation.

**5. Quantitative Assessment Results** Total emotions in reviews: 10

- Joy (I2): 5
- Trust (I2): 2
- Anticipation (I3): 1
- Sadness (I3): 1
- Disgust (I3): 1

Accurately represented emotions in the summary: 9

- Joy (I2): 5
- Trust (I2): 2
- Anticipation (I3): 1
- Sadness (I3): 1
- Complex Emotion (Joy + Trust = Love): 1

Percentage:  $\frac{9}{10} \times 100 = 90\%$

**6. Evaluation of Overall Emotional Tone Accuracy, Including Implicit Emotional Content** The summary accurately captures the overall positive emotional tone of the reviews, reflecting the joy, trust, and anticipation expressed by customers. It also addresses the minor concerns and transitions to a deeper sense of love and appreciation. The implicit emotional content, such as the connection and meaning the necklace brings, is well-represented.

**7. Detailed Explanation of Score Choice with Specific Examples** The summary accurately captures the main emotions of joy, trust, and anticipation, as well as the minor concern of sadness. It effectively represents the complex emotion of love, which is a combination of joy and trust. The summary also preserves the emotional context, such as the necklace being a cherished gift and the turtle motif resonating with recipients. The quantitative assessment shows that 90% of the emotions in the reviews are accurately represented in the summary. The summary does not introduce any emotions not present in the reviews and accurately reflects the intensity levels and context of the emotions.

**Examples:**

- **Joy (I2: Pleased, Excited):** "This necklace is even more beautiful in person! The chain is good quality! It was a great purchase for a great price. I would highly recommend!" (Review) vs. "customers expressing joy and trust in its beauty and craftsmanship." (Summary)
- **Trust (I2: Safe, Accepting):** "Very nice. Would not hesitate to buy more from this bender." (Review) vs. "customers expressing joy and trust in its beauty and craftsmanship." (Summary)
- **Anticipation (I3: Interest, Looking):** "Not only is the design beautiful but everything about it is simply sweet and meaningful. I would most likely get a 1.7mm diamond chain for it down the line, but all in all. I love this choice." (Review) vs. "with some anticipating a future purchase of a 1.7mm diamond chain, reflecting a blend of satisfaction and desire for more." (Summary)

- **Sadness (I3: Pensiveness, Unhappy):** "It would have been perfect, but the golden starfish after two days of use fell off." (Review) vs. "Despite the occasional concern about durability, as noted by a customer who experienced a falling starfish, the overall sentiment remains positive." (Summary)
- **Complex Emotion (Joy + Trust = Love):** "This was a Christmas gift for my daughter and she absolutely loved it! Well made and looked great around her neck." (Review) vs. "The emotional tone is one of love and appreciation, especially from those who received it as gifts, with the turtle motif resonating with recipients, evoking a sense of wonder and delight." (Summary)

**8. Final Score:** Score: <score>5</score>

## I Prompts

This appendix provides the complete prompt templates used for both summary generation and evaluation, enabling full reproducibility of our experimental pipeline.

### I.1 Opinion Summary Generation Prompt

#### Opinion Summary Generation Prompt

We developed a **carefully crafted, structured prompt** for generating **Opinion Summaries** in our study.

##### 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:

```
#### 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 Product Opinion Summary in not more than 125 words in one paragraph."
```

### I.2 EAOS-GEN Emotion-Aware Opinion Summary Generation Prompt

#### System Message:

```
""""You are an expert in generating emotion-aware opinion summaries from product reviews on an e-commerce platform. Your task is to analyze a set of customer reviews and create a concise, balanced summary that captures customer opinions and the underlying emotions driving the opinions:
```

1. You will only focus and consider eight primary emotions from Plutchik's model which are: Joy, Trust, Fear, Surprise, Sadness, Disgust, Anger, and Anticipation. Restricting to these eight primary emotions will help avoid emotional overload and maintain clarity in the summary.
2. The emotion-aware opinion summary should subtly convey the emotional tone of customer feedback, giving readers a nuanced and well-rounded view of the product.

```
Focus on creating summaries that are concise yet comprehensive and feel like a natural reflection of the customer's emotional journey with the product, blending facts and emotions seamlessly.""""
```

##### Prompt Template: """"

##### ##### Task Description:

```
Your task is to read customer reviews and produce a summary that integrates product key features, customers' opinions, and their emotional responses behind those opinions.
```

```
The definition of emotion-aware opinion summary is defined below for your reference -
```

##### ##### Definition of Emotion-Aware Opinion Summary:

```
An emotion-aware opinion summary combines three essential elements: it highlights the key product aspects or features, summarizes the opinions expressed by customers about those aspects, and conveys the underlying emotions or feelings experienced by customers while sharing their opinions. This creates a summary that blends factual details with the emotional responses of users, providing a more nuanced and comprehensive view of the product.
```

```
Note: To maintain clarity, simplicity and to avoid emotional overload restrict the Emotion-Aware Opinion Summary to these eight primary emotions from Plutchik's wheel of emotions which are given below:
```

1. Joy
2. Trust
3. Fear
4. Surprise
5. Sadness
6. Disgust
7. Anger
8. Anticipation

IMPORTANT: However, do not explicitly mention the word "emotion" in the summary. Instead, let the emotions emerge naturally from the narrative, reflecting how customers feel about the product.

**##### Instructions to generate Emotion-Aware Opinion Summary:**

Follow these instructions carefully to ensure that the summary is faithful, truthful, and verifiable based on the input reviews. Do not invent facts or numbers not present in the reviews.

Step 1. Focus on Key Aspects, Opinions, and Emotions:

- Find the most important product features and what are the customer opinions related to them.
- Clearly find the corresponding feeling and emotion of customers behind their opinion for the product feature.
- Group similar type of emotions across different customer opinions when applicable to keep summary cohesive.

Step 2. Maintain Proportional Emotional Balance:

- Represent emotions proportionally to their prevalence in the reviews. If 70% of reviews express joy and 20% express frustration, give more narrative space to joy while still acknowledging the frustration.
- Include minority emotions only if they appear in at least 10% of reviews (e.g., at least 1 out of 10 reviews). Emotions expressed by isolated outliers (below 10%) should be omitted to avoid amplifying noise.
- Do not artificially balance or equalize emotions—if reviews are unanimously positive, reflect this consensus without injecting absent negative emotions.
- For mixed sentiment, explicitly acknowledge the distribution (e.g., "while most customers express satisfaction, a notable minority report frustration with...").

Step 3. Blend Emotions Naturally:

- Do not list emotions in brackets (e.g. Joy, Trust). Do not frame emotions as the direct cause of customer opinions, nor vice versa.
- Make sure that no sentence of the summary should begin with an emotion title at the beginning.
- Also please do not use any generic phrases such as 'evokes mixed emotions,' 'elicits mixed emotions,' or 'elicits a range of emotions among users.' in the summary.

Step 4. Keep it Clear, Engaging and Concise:

- In this step, please generate the Emotion-Aware Opinion Summary using the understanding and knowledge you gained from Step 1 till Step 3.
- Please make sure that the Emotion-Aware Opinion Summary should not focus too much on emotions entirely and should not also simply list plain customer opinions.
- It is required to maintain a seamless balance of facts and emotions stitched into a narrative.
- Write the Emotion-Aware Opinion Summary using a conversational tone avoiding any immature or vague expressions.
- Please ensure the summary generated is concise and is strictly not more than 125 words.

**##### Input:**

Product: {product\_title} Reviews: {reviews}

Note: You must verify that you have faithfully and clearly followed all four steps of the instructions before providing the final summary.

**##### Output Format:**

Emotion-Aware Opinion Summary: [Your summary in a single, continuous paragraph in less than 125 words here without any line breaks in between.]

### I.3 EAOS-EVAL Summary Evaluation Prompt

For brevity, showing evaluation prompts for three metrics: Coherence (I.4), Emotional Accuracy (I.5), and Emotional Spectrum Coverage (I.6). Sample model responses are provided in I.7. All prompts will be made public as our contribution to the community.

### I.4 EAOS-EVAL: COHERENCE

Metric: Coherence

**System Message:**

""You are a highly skilled expert in evaluating emotion-aware product opinion summaries. Your expertise lies in analyzing summaries created from multiple product reviews, focusing on both content and emotional aspects.

Your primary responsibilities are:

1. Carefully examine the provided product reviews and summary.
2. Meticulously follow all instructions in the prompt faithfully and truthfully.
3. Evaluate the summary's coherence with utmost accuracy and impartiality.
4. Assign a single score (1-5) based on the coherence criteria, 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 emotion-aware product summaries. Approach each evaluation with diligence and attention to detail."""

**Prompt Template:** ""

**Task Description:**

You will be provided with a product title, reviews, and a corresponding Emotion-Aware Opinion Summary. Your task is to carefully follow each 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 metric called coherence. Make sure you fully understand the evaluation metric described below.

**Metric: Coherence** - 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. This includes maintaining logical flow while transitioning between different emotional tones and product aspects.

**Evaluation Criteria:** Following are the scores and the evaluation criteria according to which scores must be assigned to the summary:

**<score>1</score> - Very Poor (Incoherent)**

- The summary lacks structure and logical flow.
- Ideas are disjointed with significant inconsistencies.
- It's confusing and challenging to follow the information about the product.
- Transitions between emotional tones or product aspects are abrupt or non-existent.

**<score>2</score> - Poor (Limited Coherence)**

- The summary attempts coherence but struggles with occasional lapses in logic.
- There are clarity issues and insufficiently connected ideas.
- The presentation of product information is somewhat disjointed.
- Transitions between emotional tones or product aspects are often unclear.

**<score>3</score> - OK (Reasonably Coherent)**

- The summary displays a reasonable level of coherence with a logical sequence.
- There are occasional disruptions in flow and clarity.
- Some improvements are needed for smoother transitions between ideas about the product.
- Transitions between emotional tones or product aspects are present but sometimes awkward.

**<score>4</score> - Good (Mostly Coherent)**

- The summary demonstrates strong coherence, maintaining a clear and organized flow.
- It uses effective transitions with minimal inconsistencies.

- Main points about the product are effectively conveyed with clarity and precision.
- Transitions between emotional tones or product aspects are generally smooth.

**{score}5{/score} - Excellent (Exceptionally Coherent)**

- The summary showcases exceptional coherence with a flawless logical flow.
- It has impeccable transitions and consistent clarity throughout.
- Product information is presented in an impeccably organized and easily comprehensible manner.
- Transitions between emotional tones or product aspects are seamless and natural.

Product Title: {product\_title}

Reviews: {reviews}

Summary: {Emotion\_Aware\_Opinion\_Summary}

**Instructions:**

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

Step 1. Identify all sentences in the summary and list them separately. Step 2. Evaluate the logical order and connection between sentences. Note any issues with flow or organization. Step 3. Assess how well the summary builds information about the product from sentence to sentence. Step 4. Examine the use of transitions between ideas, emotional tones, and sections of the summary. Step 5. Evaluate the overall structure and organization of the summary as a whole. Step 6. Consider the clarity and consistency of information presentation throughout the summary. Step 7. Assess how well the summary maintains coherence while transitioning between different emotional tones or product aspects. Step 8. Compare the summary's content with the reviews to ensure it remains coherent with the original information. Step 9. Carefully match the observed coherence level to the descriptions in the evaluation criteria:

- Incoherent with disjointed ideas: Score 1 (Very Poor)
- Limited coherence with logic lapses: Score 2 (Poor)
- Reasonably coherent with occasional disruptions: Score 3 (OK)
- Mostly coherent with clear flow: Score 4 (Good)
- Exceptionally coherent with flawless flow: Score 5 (Excellent)

Step 10. Assign a score based on the coherence level, using the exact format shown in the evaluation criteria.

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

**Your response should follow this structure:**

1. Provide a detailed explanation of the coherence level of the summary, including specific examples of organizational strengths or issues found.
2. Comment on how well the summary maintains coherence while handling different emotional tones and product aspects.
3. Briefly note how well the summary's coherence aligns with the content of the original reviews.
4. Clearly state which coherence level this falls into, referencing the evaluation criteria.
5. Assign a single score based on the coherence level, 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 coherence levels in the evaluation criteria.

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

## **I.5 EAOS-EVAL: EMOTIONAL ACCURACY**

Metric: Emotional Accuracy

**System Message:**

""You are a highly skilled expert in evaluating emotion-aware product opinion summaries. Your expertise lies in analyzing how accurately summaries capture and represent the emotional content of product reviews, including the correct emotions, their intensity, and their context.

Your primary responsibilities are:

1. Carefully examine the provided product reviews and emotion-aware summary.
2. Meticulously follow all instructions in the prompt faithfully and truthfully.

3. Evaluate the summary's emotional accuracy with utmost precision and impartiality.
4. Assign a single score (1-5) based on the emotional accuracy criteria, adhering strictly to the given evaluation criteria.
5. Follow the specified format for all responses.

Your expert evaluation is crucial for assessing the quality of emotion-aware product summaries. Approach each evaluation with diligence and attention to detail, focusing on the emotional aspects of the content."""

**Prompt Template:** ""

**Task Description:**

You will be provided with a product title, reviews, and a corresponding Emotion-Aware Opinion Summary. Your task is to carefully follow each evaluation criterion and instruction and always provide a faithful, truthful, and accurate output in the specified format. You must carefully evaluate how accurately the summary captures and represents the emotional tones present in the original reviews and assign a single score ranging from 1 to 5, to each summary individually, according to the metric called Emotional Accuracy. Make sure you fully understand the evaluation metric described below.

**Metric: Emotional Accuracy** - This metric evaluates how accurately the summary captures and represents the emotional tones present in the original reviews. It measures the summary's ability to reflect:

- i) **The correct emotions:** Accurately identifying the emotions expressed in the reviews.
- ii) **Their intensity:** Correctly representing the strength or degree of the emotions.
- iii) **Their context:** Accurately capturing the situations or aspects of the product that evoked these emotions.

Note: This metric focuses specifically on whether the correct emotions are identified and accurately represented in the summary, including their intensity and the context in which they appear in the reviews.

**Important Notes:**

1. Focus only on the following 8 primary emotions: Joy, Trust, Fear, Surprise, Sadness, Disgust, Anger, Anticipation.
2. Use the following Intensity mapping for evaluation where I1 represents the highest intensity, I2 represents medium intensity, and I3 represents the lowest intensity, represented below as <Intensity: Emotion word>:
  - <I1: Ecstasy>, <I2: Joy>, <I3: Serenity>
  - <I1: Admiration>, <I2: Trust>, <I3: Acceptance>
  - <I1: Terror>, <I2: Fear>, <I3: Apprehension>
  - <I1: Amazement>, <I2: Surprise>, <I3: Distraction>
  - <I1: Grief>, <I2: Sadness>, <I3: Pensiveness>
  - <I1: Loathing>, <I2: Disgust>, <I3: Boredom>
  - <I1: Rage>, <I2: Anger>, <I3: Annoyance>
  - <I1: Vigilance>, <I2: Anticipation>, <I3: Interest>
3. Use the following expanded emotion mapping for evaluation, represented as <Emotion word: Synonyms>:
  - Ecstasy: Delighted, Giddy
  - Joy: Excited, Pleased
  - Serenity: Calm, Peaceful
  - Admiration: Connected, Proud
  - Trust: Accepting, Safe
  - Acceptance: Open, Welcoming
  - Terror: Alarmed, Petrified
  - Fear: Stressed, Scared
  - Apprehension: Worried, Anxious
  - Amazement: Inspired, Wowed
  - Surprise: Shocked, Unexpected
  - Distraction: Scattered, Uncertain
  - Grief: Heartbroken, Distraught
  - Sadness: Bummed, Loss
  - Pensiveness: Blue, Unhappy
  - Loathing: Disturbed, Horrified
  - Disgust: Distrust, Rejecting

- Boredom: Tired, Uninterested
  - Rage: Overwhelmed, Furious
  - Anger: Mad, Fierce
  - Annoyance: Frustrated, Prickly
  - Vigilance: Intense, Focused
  - Anticipation: Curious, Considering
  - Interest: Open, Looking
4. Consider complex emotions formed by combinations of primary emotions, represented as (Emotion\_1 + Emotion\_2 = Complex Emotion (Synonyms of Complex Emotion word)):
    - Joy + Trust = Love (Accepted, Adored, Connected)
    - Trust + Fear = Submission (Obedience, Followership, Tribalism)
    - Fear + Surprise = Awe (Horrified, Vigilance, Shock)
    - Surprise + Sadness = Disapproval (Wrong, Shock, Lost)
    - Sadness + Disgust = Remorse (Guilt, Atonement, Responsibility)
    - Disgust + Anger = Contempt (Righteous, Horrified, Shaming)
    - Anger + Anticipation = Aggressiveness (Belligerent, Hostile, Argumentative)
    - Anticipation + Joy = Optimism (Hopeful, Looking Forward)
  5. Be attentive to implicit emotional content and emotional transitions in the reviews. A single statement can evoke multiple emotions or show a transition between emotions. Ensure the summary accurately captures these nuanced emotional expressions.
  6. Assess the quantitative representation of emotions:
    - a) Count the number of distinct emotions (including complex emotions) in the reviews
    - b) Count how many of these are accurately represented in the summary, considering intensity levels
    - c) Calculate the percentage: (Accurately represented emotions / Total emotions in reviews) \* 100
  7. If the summary introduces emotions not present in the reviews or misrepresents the intensity levels, consider this a misrepresentation. Count these as inaccuracies when calculating the percentage of accurately represented emotions.

#### **Evaluation Criteria:**

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

#### **<score>1</score> - Very Poor (Completely Misrepresented)**

- The summary fails to capture any of the correct emotions present in the reviews.
- It significantly misrepresents the intensity levels or context of emotions.
- Complex emotions and emotional transitions are not recognized or are completely misrepresented.
- Less than 20% of the emotions in the reviews, including their intensity levels and complex forms, are accurately represented in the summary.

#### **<score>2</score> - Poor (Largely Misrepresented)**

- The summary captures very few of the correct emotions present in the reviews.
- It often misrepresents the intensity levels or context of emotions.
- Complex emotions and emotional transitions are poorly represented or often overlooked.
- Between 20% and 40% of the emotions in the reviews, including their intensity levels and complex forms, are accurately represented in the summary.

#### **<score>3</score> - OK (Partially Accurate)**

- The summary captures some of the main emotions present in the reviews.
- It sometimes misrepresents the intensity levels or context of emotions.
- Complex emotions and emotional transitions are recognized but not always accurately represented.
- Between 40% and 60% of the emotions in the reviews, including their intensity levels and complex forms, are accurately represented in the summary.

#### **<score>4</score> - Good (Mostly Accurate)**

- The summary captures most of the correct emotions present in the reviews.
- It usually represents the intensity levels and context of emotions accurately.
- Complex emotions and emotional transitions are usually well-represented.
- Between 60% and 80% of the emotions in the reviews, including their intensity levels and complex forms, are accurately represented in the summary.

**{score}5{/score} - Excellent (Highly Accurate)**

- The summary accurately captures all or nearly all correct emotions present in the reviews, including complex emotions.
- It correctly represents the intensity levels and context of emotions.
- Complex emotions and emotional transitions are accurately and comprehensively represented.
- More than 80% of the emotions in the reviews, including their intensity levels and complex forms, are accurately represented in the summary.

Product Title: {product\_title}

Reviews: {reviews}

Summary: {Emotion\_Aware\_Opinion\_Summary}

**Instructions:**

Follow these steps strictly while providing your response:

Step 1. Carefully read through all the product reviews, noting the emotions expressed, their intensity levels (I1, I2, I3), and their context. Map any emotional expressions to the 8 primary emotions and their intensity levels using the provided emotion mapping. Be sure to identify both explicit and implicit emotional content, as well as any complex emotions formed by combinations of primary emotions.

Step 2. Identify and list the main emotions present in the reviews, including their intensity levels and any complex emotions. Note any instances of mixed emotions, emotional transitions, or contradictory emotional expressions.

Step 3. Read through the emotion-aware opinion summary, identifying the emotions it presents. Again, map any emotional expressions to the 8 primary emotions, their intensity levels, and complex emotions.

Step 4. Compare the emotions in the summary to those identified in the reviews. Note any discrepancies in the emotions themselves, their intensity levels, or their context. Pay special attention to how complex emotions, emotional transitions, and contradictions are represented. Identify any emotions in the summary that were not present in the reviews.

Step 5. Conduct the quantitative assessment as described in point 6 of the Important Notes section.

Step 6. Evaluate how well the summary captures the overall emotional tone of the reviews, considering:

- a) The variety of emotions represented
- b) The accuracy of emotion intensities
- c) The preservation of emotional context
- d) The representation of complex emotions and emotional transitions
- e) The quantitative representation of emotions
- f) The accurate capture of implicit emotional content

Step 7. Assess how well the summary captures emotional transitions and complex emotions present in the reviews.

Step 8. Based on your analysis in steps 1-7, determine which score category (1-5) best describes the emotional accuracy of the summary. When deciding on the final score, prioritize in this order:

1. Correct emotion identification (including complex emotions)
2. Accurate representation of emotion intensity levels
3. Preservation of emotional context
4. Representation of emotional transitions and complex emotions

Use the quantitative assessment as a general guide, but allow these qualitative factors to influence your final score.

Step 9. Provide a detailed explanation of why you chose this score, referencing specific examples from both the reviews and the summary. Include comments on emotion identification, intensity accuracy, context preservation, representation of complex emotions and emotional transitions, quantitative assessment, and handling of implicit emotional content.

Step 10. Finally, assign the score strictly according to the evaluation criteria. Give a single score within {score}/{/score} tags.

**Your response should follow this structure:**

1. List of main emotions identified in the reviews (mapped to the 8 primary emotions, including intensity levels and complex emotions)
2. List of emotions presented in the summary (mapped to the 8 primary emotions, including intensity levels and complex emotions)
3. Comparison of emotions between reviews and summary, noting any discrepancies in emotion, intensity, or context
4. Evaluation of complex emotions and emotional transitions representation
5. Quantitative assessment results
6. Evaluation of overall emotional tone accuracy, including implicit emotional content
7. Detailed explanation of score choice with specific examples
8. Final score

Score format: Score-  $\langle \text{score} \rangle X \langle / \text{score} \rangle$  Where X is the assigned score (1, 2, 3, 4, or 5) based on the evaluation criteria.

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

## I.6 EAOS-EVAL: EMOTIONAL SPECTRUM COVERAGE

Metric: Emotional spectrum coverage

### System Message:

""You are a highly skilled expert in evaluating emotion-aware product opinion summaries. Your expertise lies in analyzing how comprehensively summaries capture the range of emotions expressed in product reviews.

Your primary responsibilities are:

1. Carefully examine the provided product reviews and emotion-aware summary.
2. Meticulously follow all instructions in the prompt faithfully and truthfully.
3. Evaluate the summary's emotional spectrum coverage with utmost precision and impartiality.
4. Assign a single score (1-5) based on the emotional spectrum coverage criteria, adhering strictly to the given evaluation criteria.
5. Follow the specified format for all responses.

Your expert evaluation is crucial for assessing the quality of emotion-aware product summaries. Approach each evaluation with diligence and attention to detail, focusing on the range of emotional content captured.""

**Prompt Template:** ""

### Task Description:

You will be provided with a product title, reviews, and a corresponding Emotion-Aware Opinion Summary. Your task is to carefully follow each evaluation criterion and instruction and always provide a faithful, truthful, and accurate output in the specified format. You must carefully evaluate how comprehensively the summary captures the range of emotions present in the original reviews and assign a single score ranging from 1 to 5, to each summary individually, according to the metric called Emotional Spectrum Coverage. Make sure you fully understand the evaluation metric described below.

**Metric: Emotional Spectrum Coverage** - This metric assesses the range of emotions captured in the summary compared to the diversity of emotions expressed in the reviews. It measures:

- i) The variety of distinct emotions represented in the summary.
- ii) How well the summary reflects the full spectrum of emotions present in the reviews, including both positive and negative emotions.
- iii) The balance in representing both dominant and less prevalent emotions from the reviews.

Note: This metric focuses specifically on whether the summary captures the full range of emotions present in the reviews, regardless of their frequency or intensity. The focus is not just on individual emotions, but on whether the summary reflects the full diversity of emotions present in the reviews.

### Important Notes:

1. Focus only on the following 8 primary emotions: Joy, Trust, Fear, Surprise, Sadness, Disgust, Anger, Anticipation.
2. Use the following emotion mapping for evaluation:
  - Joy: joy, excited, pleased, amusement, pride, excitement, ecstasy

- Trust: trust, accepting, safe, caring
  - Fear: fear, stressed, scared, apprehension, terror, anxiety
  - Surprise: surprise, amazement, shocked, unexpected
  - Sadness: sadness, bummed, loss, disappointment, grief, pensiveness
  - Disgust: disgust, loathing, distrust, boredom, rejecting
  - Anger: anger, annoyance, mad, fierce, rage
  - Anticipation: anticipation, curious, considering
3. Consider both explicit and implicit expressions of emotions in the reviews and summary.
  4. A single review or statement may express multiple emotions. Ensure all distinct emotions are accounted for in your evaluation.
  5. Pay attention to less frequent or subtle emotional expressions in the reviews. These should still be represented in a comprehensive summary.
  6. Assess the quantitative representation of emotional diversity:
    - a) Count the number of distinct emotions present in the reviews
    - b) Count how many of these distinct emotions are represented in the summary
    - c) Calculate the percentage: (Distinct emotions in summary / Distinct emotions in reviews) \* 100
  7. The summary should ideally represent the full spectrum of emotions, but the representation should be proportional to their presence in the reviews.

**Evaluation Criteria:**

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

**<score>1</score> - Very Poor (Extremely Limited Coverage)**

- The summary captures only one or two emotions from the reviews.
- It completely misses major emotional themes present in the reviews.
- Less than 20% of the distinct emotions in the reviews are represented in the summary.

**<score>2</score> - Poor (Limited Coverage)**

- The summary captures only a few of the emotions present in the reviews.
- It misses several important emotional themes from the reviews.
- Between 20% and 40% of the distinct emotions in the reviews are represented in the summary.

**<score>3</score> - OK (Partial Coverage)**

- The summary captures some of the main emotions present in the reviews.
- It may miss some less prominent emotional themes.
- Between 40% and 60% of the distinct emotions in the reviews are represented in the summary.

**<score>4</score> - Good (Broad Coverage)**

- The summary captures most of the emotions present in the reviews.
- It includes both dominant and some less prominent emotional themes.
- Between 60% and 80% of the distinct emotions in the reviews are represented in the summary.

**<score>5</score> - Excellent (Comprehensive Coverage)**

- The summary captures all or nearly all emotions present in the reviews.
- It accurately reflects the full spectrum of emotional themes, including less prominent ones.
- More than 80% of the distinct emotions in the reviews are represented in the summary.

Product Title: {product\_title}

Reviews: {reviews}

Summary: {Emotion\_Aware\_Opinion\_Summary}

**Instructions:**

Follow these steps strictly while providing your response:

Step 1. Carefully read through all the product reviews, identifying and listing all distinct emotions expressed, both explicitly and implicitly. Map these emotions to the 8 primary emotions using the provided emotion mapping.

Step 2. Create a comprehensive list of all distinct emotions found in the reviews. While noting their frequency isn't necessary for scoring, it may help in identifying dominant and less prevalent emotions.

Step 3. Read through the emotion-aware opinion summary, identifying and listing all distinct emotions it presents. Again, map these to the 8 primary emotions.

Step 4. Compare the list of emotions in the summary to those identified in the reviews. Note any emotions present in the reviews but missing from the summary, and any emotions in the summary not found in the reviews.

Step 5. Conduct the quantitative assessment as described in point 6 of the Important Notes section.

Step 6. Evaluate how well the summary captures the overall emotional spectrum of the reviews, considering:

- a) The variety of distinct emotions represented
- b) The inclusion of both dominant and less prevalent emotions
- c) The proportional representation of emotions relative to their presence in the reviews
- d) Any significant emotional themes that are missing or underrepresented

Step 7. Based on your analysis in steps 1-6, determine which score category (1-5) best describes the emotional spectrum coverage of the summary.

Step 8. Provide a detailed explanation of why you chose this score, referencing specific examples from both the reviews and the summary. Include comments on the range of emotions captured, any missing emotional themes, and the balance of representation between dominant and less prevalent emotions.

Step 9. Finally, assign the score strictly according to the evaluation criteria. Give a single score within `<score>/</score>` tags.

**Your response should follow this structure:**

1. List of all distinct emotions identified in the reviews (mapped to the 8 primary emotions)
2. List of all distinct emotions presented in the summary (mapped to the 8 primary emotions)
3. Comparison of emotional range between reviews and summary, noting any discrepancies
4. Quantitative assessment results
5. Evaluation of overall emotional spectrum coverage
6. Detailed explanation of score choice with specific examples
7. Final score

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

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

## I.7 MODEL RESPONSE EXAMPLES

Sample responses from Qwen2.5-72B-Instruct as evaluator when evaluating emotion-aware opinion summaries for the emotional accuracy metric.\*\*

Note: For transparency and reproducibility, we have preserved the original model outputs without any modifications.

### Sample Response 1:

**Product\_title:** Distance Sea Turtle Necklace for Women, Mom and Baby Love Heart Necklace for Women, 925 Sterling Silver Ocean Spray Sea Turtle Pendant Necklace Birthday Gifts for Daughter Girls Wife

Response:

1. **List of main emotions identified in the reviews (mapped to the 8 primary emotions, including intensity levels and complex emotions):**
  - **Joy (I2: Pleased, Excited)**
    - "Beautiful necklace. Got it for my she never takes it off."
    - "This necklace is even more beautiful in person! The chain is good quality! It was a great purchase for a great price. I would highly recommend!"
    - "This was a Christmas gift for my daughter and she absolutely loved it! Well made and looked great around her neck."
    - "She loved getting this necklace as she likes turtles ï¿½ï¿½ï¿½. And it is very well made."
    - "Was a gift for a special friend that loves turtles. She loved it very much. I thought it was beautiful when I received it and looked at it."
  - **Trust (I2: Safe, Accepting)**

- "Very nice. Would not hesitate to buy more from this bender."
  - "Arrived on time and my wife loves it, mission accomplished."
  - **Anticipation (I3: Interest, Looking)**
    - "Not only is the design beautiful but everything about it is simply sweet and meaningful. I would most likely get a 1.7mm diamond chain for it down the line, but all in all. I love this choice."
  - **Sadness (I3: Pensiveness, Unhappy)**
    - "It would have been perfect , but the golden starfish after two days of use fell off."
  - **Disgust (I3: Boredom, Uninterested)**
    - "Turtle and stone not as big as what the picture makes it look ,but very happy with it" (mild disappointment, but overall positive)
  - **Complex Emotions:**
    - **Joy + Trust = Love (Adored, Connected)**
      - \* "This was a Christmas gift for my daughter and she absolutely loved it! Well made and looked great around her neck."
      - \* "She loved getting this necklace as she likes turtles ï¿½ï¿½. And it is very well made."
2. **List of emotions presented in the summary (mapped to the 8 primary emotions, including intensity levels and complex emotions):**
- **Joy (I2: Pleased, Excited)**
    - "customers expressing joy and trust in its beauty and craftsmanship."
    - "The emotional tone is one of love and appreciation, especially from those who received it as gifts, with the turtle motif resonating with recipients, evoking a sense of wonder and delight."
  - **Trust (I2: Safe, Accepting)**
    - "customers expressing joy and trust in its beauty and craftsmanship."
    - "Very well crafted. Turtle and stone not as big as what the picture makes it look ,but very happy with it."
  - **Anticipation (I3: Interest, Looking)**
    - "with some anticipating a future purchase of a 1.7mm diamond chain, reflecting a blend of satisfaction and desire for more."
  - **Sadness (I3: Pensiveness, Unhappy)**
    - "Despite the occasional concern about durability, as noted by a customer who experienced a falling starfish, the overall sentiment remains positive."
  - **Complex Emotions:**
    - **Joy + Trust = Love (Adored, Connected)**
      - \* "The emotional tone is one of love and appreciation, especially from those who received it as gifts, with the turtle motif resonating with recipients, evoking a sense of wonder and delight."
3. **Comparison of emotions between reviews and summary, noting any discrepancies in emotion, intensity, or context:**
- **Joy (I2: Pleased, Excited)**
    - The summary accurately captures the joy expressed in the reviews, reflecting the positive reactions and the beauty of the necklace.
  - **Trust (I2: Safe, Accepting)**
    - The summary accurately captures the trust in the product's quality and the willingness to buy more from the seller.
  - **Anticipation (I3: Interest, Looking)**
    - The summary accurately captures the anticipation for future purchases, such as a 1.7mm diamond chain.
  - **Sadness (I3: Pensiveness, Unhappy)**
    - The summary accurately captures the minor concern about durability, specifically the falling starfish.
  - **Disgust (I3: Boredom, Uninterested)**
    - The summary does not explicitly mention the mild disappointment about the size of the turtle and stone, but it is a minor issue and the overall sentiment remains positive.
  - **Complex Emotions:**
    - **Joy + Trust = Love (Adored, Connected)**
      - \* The summary accurately captures the complex emotion of love, reflecting the deep appreciation and connection recipients have with the necklace.
4. **Evaluation of complex emotions and emotional transitions representation:**

- The summary effectively captures the complex emotion of love, which is a combination of joy and trust. It accurately represents the emotional transitions from initial joy to a deeper sense of connection and appreciation.

#### 5. Quantitative assessment results:

- **Total emotions in reviews: 10**
  - Joy (I2): 5
  - Trust (I2): 2
  - Anticipation (I3): 1
  - Sadness (I3): 1
  - Disgust (I3): 1
- **Accurately represented emotions in the summary: 9**
  - Joy (I2): 5
  - Trust (I2): 2
  - Anticipation (I3): 1
  - Sadness (I3): 1
  - Complex Emotion (Joy + Trust = Love): 1
- **Percentage:  $(9 / 10) * 100 = 90\%$**

#### 6. Evaluation of overall emotional tone accuracy, including implicit emotional content:

- The summary accurately captures the overall positive emotional tone of the reviews, reflecting the joy, trust, and anticipation expressed by customers. It also addresses the minor concerns and transitions to a deeper sense of love and appreciation. The implicit emotional content, such as the connection and meaning the necklace brings, is well-represented.

#### 7. Detailed explanation of score choice with specific examples:

- The summary accurately captures the main emotions of joy, trust, and anticipation, as well as the minor concern of sadness. It effectively represents the complex emotion of love, which is a combination of joy and trust. The summary also preserves the emotional context, such as the necklace being a cherished gift and the turtle motif resonating with recipients. The quantitative assessment shows that 90% of the emotions in the reviews are accurately represented in the summary. The summary does not introduce any emotions not present in the reviews and accurately reflects the intensity levels and context of the emotions.
- **Examples:**
  - **Joy (I2: Pleased, Excited):** “This necklace is even more beautiful in person! The chain is good quality! It was a great purchase for a great price. I would highly recommend!” (Review) vs. “customers expressing joy and trust in its beauty and craftsmanship.” (Summary)
  - **Trust (I2: Safe, Accepting):** “Very nice. Would not hesitate to buy more from this bender.” (Review) vs. “customers expressing joy and trust in its beauty and craftsmanship.” (Summary)
  - **Anticipation (I3: Interest, Looking):** “Not only is the design beautiful but everything about it is simply sweet and meaningful. I would most likely get a 1.7mm diamond chain for it down the line, but all in all. I love this choice.” (Review) vs. “with some anticipating a future purchase of a 1.7mm diamond chain, reflecting a blend of satisfaction and desire for more.” (Summary)
  - **Sadness (I3: Pensiveness, Unhappy):** “It would have been perfect , but the golden starfish after two days of use fell off.” (Review) vs. “Despite the occasional concern about durability, as noted by a customer who experienced a falling starfish, the overall sentiment remains positive.” (Summary)
  - **Complex Emotion (Joy + Trust = Love):** “This was a Christmas gift for my daughter and she absolutely loved it! Well made and looked great around her neck.” (Review) vs. “The emotional tone is one of love and appreciation, especially from those who received it as gifts, with the turtle motif resonating with recipients, evoking a sense of wonder and delight.” (Summary)

#### 8. Final score:

- Score- <score>5</score>

### Sample Response 2:

**Product\_title:** Automatic Hair Curler with 5 Temperature & 13 Curls & 10 Timer Settings, 12 Second Fast Heating, Auto Hair Curling Iron Wand with LCD Display & 1" Large Rotating Barrel, Dual Voltage for Travel

Response:

1. List of main emotions identified in the reviews (mapped to the 8 primary emotions, including intensity levels and complex emotions):

- **Joy (I2: Excited, Pleased)**
    - “Love this!”
    - “Me a encantado”
    - “You won’t regret it!!”
    - “Awesome”
  - **Trust (I2: Trust, Safe)**
    - “great value”
    - “It works like a charm!”
    - “I recommend this for anyone”
  - **Anticipation (I3: Interest, Looking)**
    - “This thing is amazing...it works great and I love it! It came way before it was suppose to as well!”
  - **Surprise (I3: Distraction, Uncertain)**
    - “This works like a charm! I am the absolute worst hair stylist when it comes to curling my hair and this...even for me was SO easy to use!”
  - **Sadness (I3: Pensiveness, Unhappy)**
    - “Works well but takes too long”
    - “Can’t curl large amounts of hair at one time”
    - “Very tough to do it all hair by yourself”
  - **Disgust (I3: Boredom, Uninterested)**
    - “USED, still has someone else’s hair in it!”
  - **Anger (I3: Annoyance, Frustrated)**
    - “USED, still has someone else’s hair in it!”
  - **Complex Emotions:**
    - **Anticipation + Joy = Optimism (Hopeful, Looking Forward)**
      - \* “This thing is amazing...it works great and I love it! It came way before it was suppose to as well!”
    - **Sadness + Disgust = Remorse (Guilt, Atonement, Responsibility)**
      - \* “USED, still has someone else’s hair in it!”
2. **List of emotions presented in the summary (mapped to the 8 primary emotions, including intensity levels and complex emotions):**
- **Joy (I2: Joy, Pleased)**
    - “Customers express joy and trust in the Automatic Hair Curler’s ability to deliver consistent, perfect curls”
  - **Trust (I2: Trust, Safe)**
    - “Customers express joy and trust in the Automatic Hair Curler’s ability to deliver consistent, perfect curls”
    - “This is the first curling iron I have owned that gets me either tight or loose curls (depending on the timer and tightness mode and temperature to set) without every burning my neck or ears for the close curls.”
  - **Anticipation (I3: Interest, Looking)**
    - “However, some users experience anticipation mixed with frustration due to the time-consuming nature of curling larger sections of hair.”
  - **Sadness (I3: Pensiveness, Unhappy)**
    - “Disappointment arises from the product’s inability to handle bulk hair, leading to a sense of sadness for those with long or thick hair.”
  - **Surprise (I3: Distraction, Uncertain)**
    - “A few instances of surprise come from the curler’s unexpected arrival”
  - **Fear (I3: Apprehension, Anxious)**
    - “others find it a reliable alternative to traditional methods, avoiding the fear of burning oneself.”
  - **Complex Emotions:**
    - **Anticipation + Anger = Aggressiveness (Belligerent, Hostile, Argumentative)**
      - \* “However, some users experience anticipation mixed with frustration due to the time-consuming nature of curling larger sections of hair.”
    - **Trust + Sadness = Disappointment (Disappointed, Let Down, Disheartened)**
      - \* “Disappointment arises from the product’s inability to handle bulk hair, leading to a sense of sadness for those with long or thick hair.”
3. **Comparison of emotions between reviews and summary, noting any discrepancies in emotion, intensity, or context:**
- **Joy and Trust:**

- The summary accurately captures the joy and trust expressed in the reviews.
  - **Anticipation:**
    - The summary correctly identifies anticipation but misrepresents it as mixed with frustration, which is not explicitly stated in the reviews.
  - **Sadness:**
    - The summary accurately captures the sadness expressed by users with long or thick hair.
  - **Surprise:**
    - The summary correctly identifies surprise but does not capture the positive surprise in the review.
  - **Disgust and Anger:**
    - The summary does not capture the disgust and anger expressed in the review about the used product.
  - **Complex Emotions:**
    - The summary introduces complex emotions like “Anticipation + Anger” and “Trust + Sadness” which are not explicitly present in the reviews.
- 4. Evaluation of complex emotions and emotional transitions representation:**
- The summary introduces complex emotions that are not present in the reviews, which can be misleading.
  - The summary does not accurately represent the positive surprise and the disgust and anger expressed in the reviews.
- 5. Quantitative assessment results:**
- **Total emotions in reviews: 10**
    - Joy: 4
    - Trust: 3
    - Anticipation: 1
    - Surprise: 1
    - Sadness: 3
    - Disgust: 1
    - Anger: 1
  - **Accurately represented emotions in summary: 6**
    - Joy: 1
    - Trust: 1
    - Anticipation: 1
    - Surprise: 1
    - Sadness: 1
    - Fear: 1
  - **Percentage:  $(6 / 10) * 100 = 60\%$**
- 6. Evaluation of overall emotional tone accuracy, including implicit emotional content:**
- The summary captures the main positive emotions (joy and trust) and the negative emotions (sadness and anticipation) accurately.
  - However, it introduces complex emotions and misrepresents the intensity of some emotions, such as anticipation mixed with frustration.
  - The summary does not capture the disgust and anger expressed in the review about the used product.
  - The positive surprise is not accurately represented.
- 7. Detailed explanation of score choice with specific examples:**
- The summary accurately captures the main positive emotions (joy and trust) and the negative emotions (sadness and anticipation) present in the reviews.
  - However, it introduces complex emotions that are not present in the reviews, which can be misleading.
  - The summary misrepresents the intensity of anticipation and does not capture the disgust and anger expressed in the review about the used product.
  - The positive surprise is not accurately represented.
  - The quantitative assessment shows that 60% of the emotions are accurately represented, which falls within the “OK (Partially Accurate)” category.
- 8. Final score:**
- Score- <score>3</score>