# Why We Feel What We Feel: Joint Detection of Emotions and Their Opinion Triggers in E-commerce


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



## Abstract

Customer reviews on e-commerce platforms capture critical affective signals that drive purchasing decisions. However, *no* existing research has explored the joint task of emotion detection and explanatory span identification in e-commerce reviews - a crucial gap in understanding what triggers customer emotional responses. To bridge this gap, we propose a novel joint task unifying **Emotion** detection and **Opinion Trigger** extraction (**EOT**), which explicitly models the relationship between causal text spans (opinion triggers) and affective dimensions (emotion categories) grounded in Plutchik's theory of 8 primary emotions. In the absence of labeled data, we introduce **EOT-X**, a human-annotated collection of **2,400** reviews with fine-grained emotions and opinion triggers. We evaluate **23** Large Language Models (LLMs) and present **EOT-DETECT**, a structured prompting framework with systematic reasoning and self-reflection. Our framework surpasses zero-shot and chain-of-thought techniques, across e-commerce domains.


## 1 Introduction

Emotional content in customer reviews fundamentally drives consumer behavior (Chen et al., 2022) demonstrates their direct influence on purchasing decisions, while (Rosario et al., 2016) establishes their impact on post-purchase satisfaction. These emotional expressions powerfully shape how future consumers perceive and engage with products (Greifeneder et al., 2007; Pham, 2007; Kim et al., 2019; Wang et al., 2022; Bian and Yan, 2022).

Despite these clear benefits, *no* existing research in literature has explored the joint task of emotion detection and explanatory span identification in e-commerce reviews, leaving a significant gap in

**Review:** *I've been buying these for my mom for 2 years now.* She works in food service, on her feet 10+ hours at a time and *these are the only shoes she will wear to work!* A touch wider than normal shoes (like most skechers) so *she can wear thicker socks in winter* and *in summer she has room in case her feet swell a little. Non slip, great arch support; my mom replaces about every 6mo* as she works 6 days a week.

**Emotions:** Joy, Trust, Anticipation

**Figure 1:** EOT-X dataset example: Product review with emotion triggers highlighted by color to indicate corresponding emotions. Each trigger is a contiguous span showing the emotion's cause.

understanding *what* customers feel and *why* they feel it.

For instance, merely detecting "*disgust*" in a review provides limited value compared to identifying that this emotion stems from specific triggers explicitly stated by customers, such as *finding hair strands in makeup products or discovering expired skincare items.*

We address this gap by introducing Emotion detection and Opinion Trigger extraction (EOT), a joint task that directly models the bidirectional relationship between emotions and their corresponding opinion triggers, grounded in Plutchik's theory (Plutchik, 1980).

**Why Plutchik's 8 Primary Emotions** Several factors guided our choice of Plutchik's 8 primary emotions:

(*1*) `LLM Behavior Control`: In our preliminary experiments, unconstrained LLMs generated redundant emotion labels (synonyms and variants) of basic emotions, creating an *"emotion explosion"* that hindered systematic evaluation and comparison. Plutchik's framework enforces taxonomic constraints without compromising expressivity.

[◇] Equal contribution




*Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5515–5532
November 4-9, 2025 ©2025 Association for Computational Linguistics

(*2*) **`Theoretical Strength`**: Plutchik's model balances positive and negative emotions, backed by psychological and empirical research. Prior work (Strapparava and Mihalcea, 2007) often limited analysis to 6 emotions or focused mainly on negative ones.

(*3*) **`Balanced Scope`**: (Ekman, 1992) identifies 6 basic emotions, but Plutchik's model adds granularity by including `trust` and `anticipation`, which are crucial for analyzing customer feedback.

(*4*) **`Practical Implementation`**: Consistent with (Mohammad and Turney, 2013), our observations show that annotating hundreds of emotions is costly and cognitively demanding.

**OPINION TRIGGER** is a *verbatim text* segment from a review that directly shows what caused a customer's emotional response, allowing verification of the connection between emotions and their specific causes.

Our contributions are:

1. **EMOTION-OPINION TRIGGER (EOT):** The pioneering joint task that leverages LLMs to unify fine-grained emotion detection and opinion trigger extraction from *customer reviews*, enabling interpretable and actionable analysis of user feedback (Section 3).

2. **EOT-X[1]:** The first human-annotated benchmark dataset (Section 5.2) of **2, 400** reviews curated from Amazon, Yelp, and TripAdvisor to support cross-domain generalization. Each review is labeled with fine-grained emotions from Plutchik's 8 primary emotions and extractive opinion triggers identifying the emotion sources, enabling model fine-tuning.

3. **EOT-LLaMA:** An edge-deployable LLM (fine-tuned from `LLaMA-3.2.1B-INSTRUCT`) designed to detect Plutchik's 8 emotions and extract opinion triggers from customer reviews. No existing model addresses this dual task. It outperforms models up to 7x larger counterparts (Table 4) while remaining deployable on *consumer-tier hardware*.

4. **EOT-DETECT:** A structured prompting framework that leverages systematic reasoning and self-reflection steps (Section 4.3), achieving superior performance over zero-shot and zero-shot chain-of-thought (Section 7).

5. **COMPREHENSIVE BENCHMARKING:** We conduct the first large-scale evaluation of 23 LLMs (closed- and open-source) for joint emotion detection and opinion trigger extraction on customer reviews (Table 2, 3).

*To the best of our knowledge*, we present the first unified framework for joint emotion-opinion analysis, conducting model comparisons at an unprecedented scale. We will publicly release our curated benchmark dataset, and fine-tuned model as *foundational contributions* to the research community.

## 2 Related Work

**EMOTION ANALYSIS** is crucial in natural language processing, demonstrating how emotions shape online discourse, decision-making, and user behavior, particularly in e-commerce where customer reviews influence purchasing decisions (Mohammad and Turney, 2013; Malik and Bilal, 2024). While early approaches focused on sentiment polarity *(negative, neutral, positive)*, researchers recognized the need for nuanced emotion taxonomies to capture human emotional expression complexity (Russell, 1980; Ekman, 1992; Plutchik, 2001).

**EMOTION-TRIGGER ANALYSIS** remains an under-explored dimension of emotion understanding. Early research utilized rule-based approaches ((Neviarouskaya et al., 2009; Lee et al., 2010) and statistical methods (Gui et al., 2016; Xia and Ding, 2019) to identify emotion causes in text. Recent studies examined graph-based models and attention mechanisms for joint emotion-cause extraction (Wei et al., 2020; Fan et al., 2021; Singh et al., 2021) and context-aware trigger identification (Li et al., 2019). Studies demonstrate the complexity of linking emotional expressions to triggers. (Singh et al., 2024) presented EMOTRIGGER dataset, testing LLMs on social media data, showing their limitations in trigger identification despite strong emotion recognition. *Emotion-trigger analysis remains unexplored in e-commerce.*

**DATASET DEVELOPMENT** has driven emotion analysis research advancement. Key datasets include SemEval (Strapparava and Mihalcea, 2007),

[1]https://github.com/yourarnav/EOT-X

GoEmotions (Demszky et al., 2020), and ISEAR (Scherer and Wallbott, 1994). Domain-specific datasets like CancerEmo (Sosea and Caragea, 2020) and EmoCause (Gui et al., 2016) emerged. *However,* ***no*** *existing dataset provides emotion labels with trigger annotations for e-commerce platforms.*

**LLMs** have shown exceptional capabilities in understanding and generating emotionally nuanced text (Brown et al., 2020; Ouyang et al., 2022). Research has explored their potential for emotion analysis (Acheampong et al., 2023; Huang and Rust, 2024), demonstrating success in zero-shot and few-shot settings. *The application of LLMs to combined emotion detection and trigger identification in e-commerce remains unexplored.*

While prior research has significantly advanced emotion analysis and trigger detection, a critical gap remains in analyzing customer feedback within e-commerce contexts—*a gap that our work seeks to bridge.*

## 3 Task Formulation

We define the *key* components as:

### 3.1 Review and Emotion Representation

Let $\mathcal{R} = \{R_i\}_{i=1}^{N}$ denote a review consisting of N tokens. We focus on the set of eight primary emotions from Plutchik's model: $\mathcal{E} = \{\texttt{Joy}, \texttt{Trust}, \texttt{Fear}, \texttt{Surprise}, \texttt{Sadness}, \texttt{Disgust}, \texttt{Anger}, \texttt{Anticipation}\}$, with an additional `Neutral` label for cases in which no discernible emotion is expressed. Thus, our extended set of possible emotions is: $\widehat{\mathcal{E}} = \mathcal{E} \cup \{\texttt{Neutral}\}$.

### 3.2 Emotion Detection and Trigger Extraction

A review can have multiple emotions and a single emotion can have many opinion triggers. We *seek* to detect emotions in $\widehat{\mathcal{E}}$ present in $\mathcal{R}$ and extract their corresponding OPINION TRIGGERS characterized as $\mathcal{T}_e = \{(i,j) \mid (w_i, \ldots, w_j)$ is a substring explaining emotion $e\}$. If no emotions in $\mathcal{E}$ are detected, then we assign: $\mathcal{O}(\mathcal{R}) = \{(\texttt{Neutral}, \varnothing)\}$. $\mathcal{T}_e$ extractive design enables direct verification against source reviews.

### 3.3 Extractive Constraints

Let $\mathcal{M}$ be an LLM that, given $\mathcal{R}$ and a structured prompt $\mathcal{P}$, solves:

**1. EMOTION IDENTIFICATION**: A subset $\mathcal{E}'$ of the 8 primary emotions $\mathcal{E}$ expressed in $\mathcal{R}$ is selected by $f_{\text{emo}}(\mathcal{R}) \rightarrow \mathcal{E}' \subseteq \mathcal{E}$. If $\mathcal{E}' = \varnothing$, we label the review as `Neutral`.

**2. OPINION TRIGGER EXTRACTION**: For each identified emotion $e_j \in \mathcal{E}'$, the model extracts a set of triggers: $f_{\text{trig}}(\mathcal{R}, e_j) \rightarrow \mathcal{T}_j \subseteq \mathcal{S}(\mathcal{R})$, where $\mathcal{S}(\mathcal{R})$ is the set of all possible contiguous substrings in $\mathcal{R}$. For evaluation and verification, each candidate trigger $t \in \mathcal{T}_j$ must satisfy the extractive constraint: $\forall t \in \mathcal{T}_j,\ \text{span}(t) \subseteq \mathcal{R}$.

### 3.4 Output Representation

The final outcome: $\mathcal{O}(\mathcal{R}) = \{(e, \mathcal{T}_e) \mid e \in \widehat{\mathcal{E}}, \mathcal{T}_e \neq \varnothing\}$ if at least one emotion from $\mathcal{E}$ is present, and $\{(\texttt{Neutral}, \varnothing)\}$ otherwise.

## 4 Methodology

Our methodology evaluates three prompting strategies of incrementally increasing complexity, as illustrated in Figure 2. We begin with a Zero-Shot (ZS) prompt that establishes a baseline with a basic task description. We then build on this with a Zero-Shot Chain-of-Thought (ZS-COT) prompt, which encourages the model to generate its own reasoning steps. Finally, we introduce our proposed EOT-DETECT framework, which provides a complete, multi-step reasoning and self-reflection sequence. This incremental design allows us to systematically evaluate the benefits of structured prompting through a clear, strategy-level comparison.

### 4.1 Zero-Shot Prompting (ZS)

In zero-shot prompting (Brown et al., 2020), we instruct the LLM to identify emotions and extract opinion triggers from a given review without any task-specific examples.

### 4.2 Zero-Shot Chain-of-Thought (ZS-CoT)

Following (Kojima et al., 2022), this approach explicitly prompts the LLM to <think step-by-step> before generating its final output, encouraging structured reasoning. A key distinction is that the <reasoning> block is *generated* by the model, not provided as input, ensuring the model articulates its analysis before identifying emotion-trigger pairs.

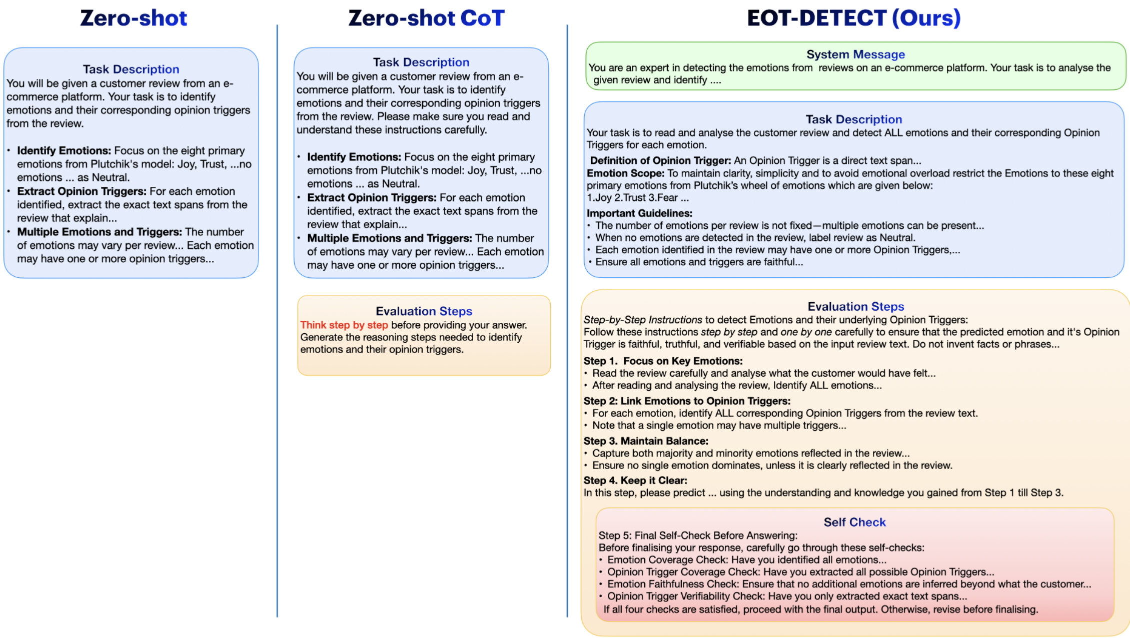


**Figure 2:** This figure illustrates three prompting strategies for joint emotion detection and opinion trigger extraction from e-commerce reviews: Zero-Shot (ZS), Zero-Shot Chain-of-Thought (ZS-CoT), and our proposed EOT-DETECT. ZS provides a basic task description. ZS-CoT adds a "think step-by-step" instruction to encourage reasoning. EOT-DETECT structures the prompt with a System Message (defining the LLM's role), a detailed Task Description (specifying constraints), and a 5-step Instruction sequence (guiding the reasoning process). Critically, EOT-DETECT incorporates a final Self-Check step to ensure comprehensive emotion coverage, accurate trigger extraction, and verifiable results.

### 4.3 EOT-DETECT Framework

We introduce **EOT-DETECT**, a structured prompting framework that guides LLMs to perform emotion-opinion trigger analysis on reviews with improved *instruction compliance*. Inspired by insights from *human* cognitive processes (Fiske and Taylor, 1991) , our framework interleaves *human-like reasoning steps* and *self-reflection mechanisms* (or "self-checks") to ensure extractive trigger identification and comprehensive emotion coverage.

Following our task formulation, we design a structured prompt $\mathcal{P}$ to enforce explicit constraints on both emotion detection and trigger extraction. Concretely, EOT-DETECT is defined as a 4-tuple: $\mathcal{P} = \langle \mathtt{S}, \mathtt{T}, \mathtt{I}, \mathtt{R} \rangle$

(*1*) **`System Message (S)`**: Defines the model's expert persona and sets the global context for how the model should interpret subsequent instructions.

(*2*) **`Task Description (T)`**: A structured component that delineates the scope and purpose of the prompt, composed of: $T = \langle \text{Task Definition}, \text{Opinion Trigger Definition}, \text{Emotion Scope}, \text{Guidelines} \rangle$

- `Task Definition`: Explains that the model should detect emotions and extract Opinion Triggers directly from the text.
- `Opinion Trigger Definition`: Clarifies that triggers must be exact substrings from $R$.
- `Emotion Scope`: Restricts focus to the 8 Plutchik emotions plus `Neutral`.
- `Guidelines`: Reinforces constraints such as "no modifications" and "multiple emotions possible," ensuring faithfulness to the review text.

(*3*) **`Instructions (I)`**: A sequence of **5** carefully designed *reasoning steps* that progressively guide the LLM through the analysis.

Formally: $I = \langle \mathtt{I}_1, \mathtt{I}_2, \mathtt{I}_3, \mathtt{I}_4, \mathtt{I}_5 \rangle$ where,

- **$\mathtt{I}_1$:** `Focus on Key Emotions` – Directs the model to read the review carefully and identify all relevant emotions (or label as `Neutral` if none are found).

- **$\mathtt{I}_2$: `Link Emotions to Opinion Triggers`** – For each emotion, prompts the LLM to extract one or more text spans from $R$ explaining why the emotion is present.
- **$\mathtt{I}_3$: `Maintain Balance`** – Ensures that both majority and minority emotions are captured, so no single emotion is over- or underrepresented.
- **$\mathtt{I}_4$: `Keep it Clear`** – Instructs the model to finalize all detected emotions and triggers, adhering to **extractive** requirements.
- **$\mathtt{I}_5$: `Final Self-Check`** – A critical self-reflection step, prompting the LLM to verify **Emotion Coverage**, **Trigger Coverage**, **Emotion Faithfulness**, and **Opinion Trigger Verifiability** before producing the final output.

(*4*) **`Input Review (R)`**: The text from which the model must **extract** triggers and identify emotions.

**Prompt Execution** When the structured prompt $\mathcal{P}$ is fed into an LLM (denoted by $\mathcal{M}$, the model sequentially processes the system message **`(S)`**, the task description **`(T)`**, the stepwise instructions **`(I)`**, and finally the review **`(R)`**. We denote this process as: $M(\mathcal{P}) \longmapsto \mathcal{O}(\mathcal{R})$, where $\mathcal{O}(\mathcal{R})$ is the emotion-trigger assignment defined in Task Formulation.

**Self-Reflection Mechanism** A standout aspect of our framework is the $\mathtt{I}_5$ `Final Self-Check`, which asks the model to confirm 4 critical conditions before providing its final answer: (*1*) **`Emotion Coverage Check`**: All expressed emotions (including minority ones) must be captured. (*2*) **`Opinion Trigger Coverage Check`**: All relevant triggers for each emotion are extracted, allowing multiple triggers for a single emotion. (*3*) **`Emotion Faithfulness Check`**: No extra (unwarranted) emotions are added. (*4*) **`Opinion Trigger Verifiability Check`**: Every trigger is exactly a sub-string of $\mathcal{R}$.

This step counters the tendency of LLMs to *overgeneralize or hallucinate* content by explicitly requiring a final verification loop.

# 5 Dataset

We utilized multi-domain dataset $\mathcal{D} = \{\mathtt{Amazon}, \mathtt{TripAdvisor}, \mathtt{Yelp}\}$, we employed Simple Random Sampling Without Replacement (SRSWOR) (Cochran, 1977). The Amazon subset (from Amazon Reviews '23; (Hou et al., 2024) ) spans subdomains $\mathcal{S} = \{\mathtt{Beauty}, \mathtt{Home}, \mathtt{Electronics}, \mathtt{Clothing}\}$. Let $P_s$ denote products in subdomain $s \in \mathcal{S}$. We sample $n_s = 40$ products per subdomain using SRSWOR: $p_i \sim \text{SRSWOR}(P_s, n_s), \quad \forall s \in \mathcal{S}$, yielding $\sum_{s\in\mathcal{S}} n_s = 160$ products. For TripAdvisor (Li et al., 2014) and Yelp (Yelp, 2025) , we sampled $n_t = 40$ items each using SRSWOR.

For each product $p$, we define a length-filtered review set:$\mathcal{R}_p = \{r \in \mathcal{R}'_p \mid 10 \leq \text{len}(r) \leq 100\}$ to control for verbosity (Kim and Sullivan, 2019; Herrando and Constantinides, 2021; Xie et al., 2022). We then sample $m = 10$ reviews using temporal stratification: $\mathcal{R}^*_p \sim \text{SRSWOR}(\mathcal{R}_p, m) \quad \text{s.t.} \quad \forall t \in \mathcal{T} : |\mathcal{R}^*_p \cap \mathcal{R}_t| \propto |\mathcal{R}_t|$. This ensures: (*1*) Uniform coverage $\pi = \frac{n_s}{|P_s|}$, (*2*) Temporal representation via stratification (error reduction by $\sqrt{1-f}$), (*3*) Cross-domain independence. Total sample size $N = 2,400$ achieves power $\beta > 0.8$ for medium effects ($d \geq 0.5$) at $\alpha = 0.05$ with Bonferroni correction (Bonferroni, 1936). We excluded ratings (Mayzlin et al., 2012; de Langhe et al., 2015; Guo et al., 2020) and helpful votes (Yin et al., 2014; Lappas et al., 2016; Deng et al., 2020) due to documented biases.

## 5.1 EOT-X

We introduce **EOT-X**, the first human-annotated benchmark dataset for emotion-opinion analysis in e-commerce. We constructed EOT-X by collecting annotations for $2,400$ reviews. To ensure annotation quality, each review was independently annotated by three expert raters to identify emotions (based on Plutchik's 8 primary emotions) and their corresponding opinion triggers. This process yielded a total of $\mathbf{7,200}$ annotations (3 raters × $2,400$ reviews).

## 5.2 Annotation

**Annotation Quality** Expert raters were selected over crowd workers due to concerns regarding annotation quality, inter-annotator agreement, and domain expertise. (Mohammad and Turney, 2013) identify key challenges in crowdsourcing, such as unqualified annotators, malicious inputs, and inconsistent judgments, emphasizing the need for rigorous quality control. Given our study's focus on nuanced emotion analysis, we employed expert raters with relevant academic training to ensure higher reliability and consistency in annotations. They followed detailed guidelines (Appendix D) and received appropriate stipends.

| Domain | A1/A2 | A2/A3 | A1/A3 | Average |
|---|---|---|---|---|
| *Emotion Agreement* | | | | |
| Beauty | 0.88 | 0.91 | 0.88 | 0.89 |
| Clothing | 0.88 | 0.90 | 0.86 | 0.88 |
| Home | 0.87 | 0.89 | 0.87 | 0.88 |
| Electronics | 0.87 | 0.88 | 0.87 | 0.87 |
| TripAdvisor | 0.85 | 0.89 | 0.87 | 0.87 |
| Yelp | 0.88 | 0.91 | 0.89 | 0.89 |
| *Overall Average* | 0.87 | 0.90 | 0.87 | 0.88 |
| *Trigger Agreement* | | | | |
| Beauty | 0.82 | 0.85 | 0.84 | 0.84 |
| Clothing | 0.81 | 0.84 | 0.81 | 0.82 |
| Home | 0.81 | 0.86 | 0.82 | 0.83 |
| Electronics | 0.82 | 0.85 | 0.83 | 0.83 |
| TripAdvisor | 0.83 | 0.86 | 0.84 | 0.85 |
| Yelp | 0.83 | 0.87 | 0.84 | 0.84 |
| *Overall Average* | 0.82 | 0.85 | 0.83 | 0.84 |

**Table 1:** Domain-wise annotator agreement scores for Emotion and Trigger identification.

**Inter-Annotator Agreement** We compute inter-annotator agreement using Fleiss' Kappa (Fleiss, 1971) ($\kappa$) for emotions and triggers. For emotions, pairwise scores range from $\mathbf{0.85}$ to $\mathbf{0.91}$ across domains (Table 1), with an average of $\mathbf{0.88}$, indicating *almost perfect agreement* ($0.81 \leq \kappa \leq 1.00$).

For triggers, we employ a **token-level analysis** approach to better accommodate the inherent variations in span annotation. By tokenizing trigger spans, we capture partial agreements where annotators identify the same triggers with different boundaries (e.g., `"very happy"` vs. `"happy"`). Pairwise scores range from $\mathbf{0.81}$ to $\mathbf{0.87}$ across domains (Table 1), with an average of $\mathbf{0.84}$, indicating *almost perfect agreement*. This strong token-level agreement demonstrates consistent identification of the same key trigger elements despite span variations.

This indicates humans can reliably identify emotions and their triggers in customer reviews. Examples of EOT-X are shown in Figure 1.

**Handling Annotation Complexity** The high agreement scores for this complex task reflect our systematic approach to addressing e-commerce review challenges. Our annotation guidelines specifically addressed linguistic complexity: annotators were instructed to identify *intended emotions* rather than literal meanings when encountering sarcasm, to use the entire review context when emotional cues were subtle or indirect, and to apply detailed protocols for edge cases such as nested spans, discontinuous triggers, and comparative expressions. Expert annotators performed systematic self-checks for trigger verifiability, emotion coverage, and annotation completeness before finalizing judgments. For overlapping trigger spans, we preserved all annotator-identified spans and applied a longest-span heuristic to maintain maximal context. This combination of expert domain knowledge, comprehensive linguistic guidelines, and systematic quality control procedures accounts for the robust agreement despite the inherent complexity of emotion-trigger annotation in informal e-commerce text.

## 6 Experiments

**Task-Specific Models:** Emotion-cause pair extraction (ECPE) systems (Xia and Ding, 2019; Bao et al., 2020; Ding et al., 2020; Yuan et al., 2020) were designed for two-step, clause-level pair extraction in newswire texts and primarily focus on linking a single emotion clause to one cause clause. ***In stark contrast***, our work addresses a fundamentally different task—jointly detecting multiple emotions and their associated opinion triggers in e-*commerce reviews*—where each emotion may be linked to several distinct textual spans. *Consequently, these methods are methodologically misaligned and irrelevant as baselines for our study*.

**Language Models:** We conducted a comprehensive evaluation of 23 language models, comprising 3 proprietary models and 20 open-source models. The full list of models is provided in **(Appendix A)**, with their corresponding inference configurations detailed in **(Appendix B)** for brevity[2].

**Fine-Tuned LLMs:** We fine-tuned `LLaMA-3.2-1B-Instruct`, `Qwen2.5-0.5B-Instruct`, and `DeepSeek-R1-Distill-LLaMA-8B` on the **EOT-X** dataset. Our objective was to assess whether compact models can match the performance of larger counterparts while reducing reliance on costly GPU infrastructure. Implementation details are provided in Appendix C.

**Rationale for Omitting PLMs for Fine-Tuning:** We chose the aforementioned LLMs over

[2]All experiments were conducted on H100 GPUs over a period of 120+ hours.

PLMs due to their superior capabilities at similar parameter counts. They support extended context (up to 128K tokens), generate longer outputs (up to 8K tokens), exhibit enhanced reasoning, and enable efficient fine-tuning with PEFT (Hu et al., 2021).

## 7 Results

**Evaluation on Aggregated Gold Standard Dataset** We evaluate our models on **EOT-X**. We create an *aggregated gold standard* by combining annotations from three expert annotators. For emotion detection, a majority vote (at least two out of three) determines inclusion, ensuring both reliability and nuance. For trigger identification, all unique triggers are preserved, and in cases of overlap, the longest span is retained to capture complete context. Preprocessing steps include standardizing product titles and review texts, exact span matching, and deduplication. For statistics of the aggregated gold standard, see Figure 3 and Appendix Table 5.

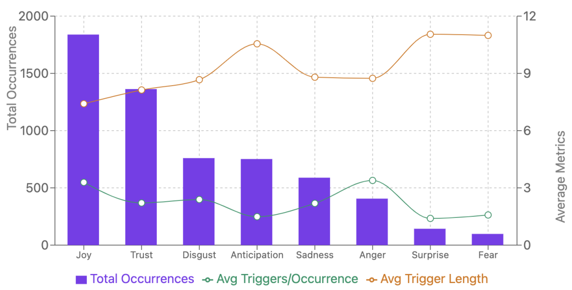


**Figure 3:** Distribution of emotions and triggers across domains in the EOT-X aggregated gold standard dataset.

**Model Family Behavior** Different model families exhibit distinct characteristics. The LLAMA series shows moderate performance, while the MISTRAL family—especially the `Mistral-7B-Instruct-v0.2` variants—demonstrate significant improvements with our enhanced EOT-DETECT framework. Similar trends are observed across the PHI and QWEN families, indicating that intrinsic family characteristics and pre-training data greatly influence performance.

| Models | Emotion | | | Opinion Trigger | | | |
|---|---|---|---|---|---|---|---|
| | P | R | F1 | EM | PM | R1 | RL |
| LLaMA-3.2-1B-Instruct_ZS | 0.13 | 0.20 | 0.10 | 0.03 | 0.10 | 0.01 | 0.01 |
| LLaMA-3.2-1B-Instruct_ZS CoT | 0.16 | 0.31 | 0.11 | 0.09 | 0.09 | 0.10 | 0.10 |
| LLaMA-3.2-1B-Instruct_EOT | 0.23 | 0.57 | 0.33 | 0.11 | 0.13 | 0.33 | 0.28 |
| LLaMA-3.2-3B-Instruct_ZS | 0.17 | 0.07 | 0.10 | 0.04 | 0.04 | 0.05 | 0.05 |
| LLaMA-3.2-3B-Instruct_ZS CoT | 0.21 | 0.19 | 0.16 | 0.11 | 0.09 | 0.019 | 0.17 |
| LLaMA-3.2-3B-Instruct_EOT | 0.49 | 0.39 | 0.44 | 0.26 | 0.31 | 0.42 | 0.37 |
| LLaMA-3.1-8B-Instruct_ZS | 0.59 | 0.52 | 0.51 | 0.26 | 0.37 | 0.53 | 0.47 |
| LLaMA-3.1-8B-Instruct_ZS CoT | 0.10 | 0.08 | 0.06 | 0.10 | 0.10 | 0.09 | 0.05 |
| LLaMA-3.1-8B-Instruct_EOT | 0.18 | 0.16 | 0.17 | 0.10 | 0.10 | 0.17 | 0.14 |
| LLaMA-3.3-70B-Instruct_ZS | 0.19 | 0.16 | 0.18 | 0.06 | 0.13 | 0.17 | 0.14 |
| LLaMA-3.3-70B-Instruct_ZS CoT | 0.14 | 0.11 | 0.08 | 0.03 | 0.03 | 0.07 | 0.04 |
| LLaMA-3.3-70B-Instruct_EOT | 0.18 | 0.21 | 0.19 | 0.06 | 0.12 | 0.18 | 0.15 |
| Mistral-7B-Instruct-v0.2_ZS | 0.69 | 0.45 | 0.54 | 0.41 | 0.33 | 0.59 | 0.50 |
| Mistral-7B-Instruct-v0.2_ZS CoT | 0.71 | 0.63 | 0.69 | 0.41 | 0.33 | 0.39 | 0.48 |
| Mistral-7B-Instruct-v0.2_EOT | **0.86** | **0.87** | **0.86** | **0.55** | 0.34 | **0.82** | **0.68** |
| Mistral-7B-Instruct-v0.3_ZS | 0.34 | 0.18 | 0.23 | 0.22 | 0.13 | 0.25 | 0.22 |
| Mistral-7B-Instruct-v0.3_ZS CoT | 0.38 | 0.30 | 0.34 | 0.16 | 0.41 | 0.58 | 0.50 |
| Mistral-7B-Instruct-v0.3_EOT | 0.59 | 0.54 | 0.57 | 0.30 | 0.32 | 0.60 | 0.50 |
| Mistral-Small-24B-Instruct-2501_ZS | 0.38 | 0.37 | 0.37 | 0.15 | 0.27 | 0.38 | 0.30 |
| Mistral-Small-24B-Instruct-2501_ZS CoT | 0.33 | 0.36 | 0.34 | 0.10 | 0.26 | 0.35 | 0.28 |
| Mistral-Small-24B-Instruct-2501_EOT | 0.37 | 0.38 | 0.38 | 0.13 | 0.26 | 0.38 | 0.30 |
| Zephyr-7b-beta_ZS | 0.52 | 0.22 | 0.31 | 0.37 | 0.17 | 0.40 | 0.35 |
| Zephyr-7b-beta_ZS CoT | 0.59 | 0.54 | 0.57 | 0.30 | 0.32 | 0.60 | 0.50 |
| Zephyr-7b-beta_EOT | 0.66 | 0.30 | 0.41 | 0.26 | 0.17 | 0.40 | 0.33 |
| Phi-3.5-mini-instruct_ZS | 0.63 | 0.53 | 0.58 | 0.51 | 0.30 | 0.68 | 0.58 |
| Phi-3.5-mini-instruct_ZS CoT | 0.71 | 0.59 | 0.65 | 0.51 | 0.29 | 0.71 | 0.60 |
| Phi-3.5-mini-instruct_EOT | 0.74 | 0.66 | 0.70 | 0.51 | 0.29 | 0.72 | 0.61 |
| Phi-4_ZS | 0.81 | 0.68 | 0.74 | 0.49 | 0.38 | 0.78 | 0.65 |
| Phi-4_ZS CoT | 0.70 | 0.69 | 0.70 | 0.40 | 0.41 | 0.73 | 0.60 |
| Phi-4_EOT | **0.86** | **0.87** | **0.86** | 0.54 | 0.35 | 0.82 | 0.67 |
| Gemma-2-2b-it_ZS | 0.72 | 0.44 | 0.55 | 0.38 | 0.39 | 0.44 | 0.40 |
| Gemma-2-2b-it_ZS CoT | 0.70 | 0.69 | 0.70 | 0.51 | 0.29 | 0.71 | 0.60 |
| Gemma-2-2b-it_EOT | 0.75 | 0.49 | 0.59 | 0.27 | 0.42 | 0.52 | 0.46 |
| Gemma-2-9b-it_ZS | 0.72 | 0.54 | 0.61 | 0.32 | 0.47 | 0.64 | 0.56 |
| Gemma-2-9b-it_ZS CoT | 0.34 | 0.18 | 0.23 | 0.22 | 0.13 | 0.25 | 0.22 |
| Gemma-2-9b-it_EOT | 0.56 | 0.39 | 0.44 | 0.21 | 0.36 | 0.47 | 0.41 |
| Gemma-2-27b-it_ZS | 0.49 | 0.38 | 0.43 | 0.31 | 0.28 | 0.45 | 0.38 |
| Gemma-2-27b-it_ZS CoT | 0.33 | 0.36 | 0.34 | 0.10 | 0.26 | 0.35 | 0.28 |
| Gemma-2-27b-it_EOT | 0.34 | 0.29 | 0.31 | 0.18 | 0.21 | 0.31 | 0.26 |
| Qwen2.5-0.5B-Instruct_ZS | 0.38 | 0.30 | 0.34 | 0.16 | 0.41 | 0.58 | 0.50 |
| Qwen2.5-0.5B-Instruct_ZS CoT | 0.35 | 0.37 | 0.36 | 0.12 | 0.36 | 0.55 | 0.48 |
| Qwen2.5-0.5B-Instruct_EOT | 0.30 | 0.46 | 0.36 | 0.12 | 0.31 | 0.52 | 0.45 |
| Qwen2.5-3B-Instruct_ZS | 0.63 | 0.38 | 0.48 | 0.26 | 0.30 | 0.35 | 0.32 |
| Qwen2.5-3B-Instruct_ZS CoT | 0.56 | 0.67 | 0.61 | 0.31 | 0.30 | 0.56 | 0.48 |
| Qwen2.5-3B-Instruct_EOT | 0.63 | 0.43 | 0.51 | 0.29 | 0.35 | 0.48 | 0.42 |
| Qwen2.5-7B-Instruct_ZS | 0.75 | 0.44 | 0.56 | 0.26 | 0.48 | 0.54 | 0.47 |
| Qwen2.5-7B-Instruct_ZS CoT | 0.61 | 0.62 | 0.61 | 0.28 | 0.37 | 0.58 | 0.49 |
| Qwen2.5-7B-Instruct_EOT | 0.67 | 0.46 | 0.55 | 0.26 | 0.44 | 0.54 | 0.47 |
| Qwen2.5-32B-Instruct_ZS | 0.78 | 0.54 | 0.63 | 0.33 | **0.51** | 0.72 | 0.60 |
| Qwen2.5-32B-Instruct_ZS CoT | 0.71 | 0.66 | 0.69 | 0.31 | 0.48 | 0.71 | 0.59 |
| Qwen2.5-32B-Instruct_EOT | **0.80** | 0.80 | 0.80 | 0.42 | 0.43 | 0.78 | 0.64 |
| Qwen2.5-72B-Instruct_ZS | 0.85 | 0.61 | 0.71 | 0.33 | 0.51 | 0.72 | 0.60 |
| Qwen2.5-72B-Instruct_ZS CoT | 0.71 | 0.71 | 0.71 | 0.38 | 0.41 | 0.74 | 0.61 |
| Qwen2.5-72B-Instruct_EOT | 0.76 | 0.73 | 0.75 | 0.30 | 0.51 | 0.76 | 0.62 |
| QwQ-32B-Preview_ZS | 0.63 | 0.61 | 0.61 | 0.35 | 0.35 | 0.69 | 0.56 |
| QwQ-32B-Preview_ZS CoT | 0.62 | 0.59 | 0.61 | 0.28 | 0.38 | 0.66 | 0.55 |
| QwQ-32B-Preview_EOT | 0.64 | 0.73 | 0.68 | 0.32 | 0.40 | 0.72 | 0.59 |
| Deepseek-R1_ZS | 0.78 | 0.66 | 0.71 | 0.32 | 0.49 | 0.71 | 0.60 |
| Deepseek-R1_ZS CoT | 0.75 | 0.70 | 0.72 | 0.27 | 0.51 | 0.70 | 0.59 |
| Deepseek-R1_EOT | 0.75 | 0.73 | 0.74 | 0.28 | 0.45 | 0.72 | 0.60 |

**Table 2:** Performance comparison of `open-source` models. Best scores are in **bold**, and second-best scores are underlined. **Emotion Metrics**: P (Precision), R (Recall), F1 (F1-score). **Opinion Trigger Metrics**: EM (Exact Match), PM (Partial Match), R1 (ROUGE-1), RL (ROUGE-L). Prompting strategies: ZS (Zero-Shot), ZS-CoT (Zero-Shot Chain-of-Thought), EOT-DETECT (our framework).

**Technique Comparison** Across both open- and closed-source settings, our EOT-DETECT framework consistently outperforms standard zero-shot and chain-of-thought approaches. For example, Mistral-7B-Instruct-v0.2 with EOT-DETECT achieves the highest emotion metrics (P: 0.86, R: 0.87, F1: 0.86), and among closed-source models, CLAUDE 3.5 SONNET with EOT-DETECT

| Models | Emotion | | | Opinion Trigger | | | |
|---|---|---|---|---|---|---|---|
| | P | R | F1 | EM | PM | R1 | RL |
| GPT-4o_ZS | 0.79 | 0.71 | 0.75 | 0.24 | 0.54 | 0.73 | 0.62 |
| GPT-4o_ZS CoT | 0.59 | 0.52 | 0.51 | 0.26 | 0.37 | 0.53 | 0.47 |
| GPT-4o_EOT | 0.69 | 0.61 | 0.53 | 0.28 | 0.38 | 0.57 | 0.49 |
| o1-mini_ZS | 0.63 | 0.53 | 0.58 | **0.51** | 0.30 | 0.68 | 0.58 |
| o1-mini_ZS CoT | 0.71 | 0.59 | 0.65 | **0.51** | 0.29 | 0.71 | 0.60 |
| o1-mini_EOT | 0.74 | 0.66 | 0.70 | **0.51** | 0.29 | 0.72 | 0.61 |
| Claude 3.5 Sonnet_ZS | 0.81 | 0.64 | 0.71 | 0.18 | 0.52 | 0.65 | 0.56 |
| Claude 3.5 Sonnet_ZS CoT | 0.69 | 0.63 | 0.66 | 0.20 | **0.57** | 0.70 | 0.59 |
| Claude 3.5 Sonnet_EOT | **0.94** | **0.72** | **0.81** | 0.35 | 0.52 | **0.83** | **0.68** |

**Table 3:** Performance comparison of `closed-source` models. Best scores are in **bold**, and second-best scores are underlined. **Emotion Metrics**: P (Precision), R (Recall), F1 (F1-score). **Opinion Trigger Metrics**: EM (Exact Match), PM (Partial Match), R1 (ROUGE-1), RL (ROUGE-L). Prompting strategies: ZS (Zero-Shot), ZS-CoT (Zero-Shot Chain-of-Thought), EOT-DETECT (our framework).

records top scores (P: 0.94, R: 0.72, F1: 0.81). This confirms that our structured prompting framework is more effective in eliciting nuanced and complete responses from models. Example responses from a range of models using our EOT-DETECT framework are provided in **Appendix E**.

**Impact of Parameter Size** Our results indicate a clear trend: larger models generally yield better performance. Within the QWEN family, increasing parameters from 0.5B to 72B leads to substantial gains in both emotion detection and trigger identification. However, improvements are also dependent on model architecture and training strategies, suggesting that model design is as crucial as scale. It is worth noting that pre-LLM approaches performed so poorly on these tasks—largely due to their limited training on token data—that we omit their results for clarity.

**Closed-Source vs. Open-Source** Closed-source models (e.g., GPT-4o, CLAUDE 3.5 SONNET) achieve higher absolute performance across metrics. Nonetheless, the best open-source models (e.g., `Mistral-7B-Instruct-v0.2 and Phi-4` with EOT-DETECT) perform competitively, demonstrating that state-of-the-art open-source approaches are rapidly closing the gap with commercial systems.

**Fine-Tuned Models:** Additionally, fine-tuned variants (e.g., LLaMA-3.2-1B-Instruct-FT (EOT-LLaMA)) generally show improved consistency over non-fine-tuned counterparts. While fine-tuning provides task-specific gains, our prompting strategies—particularly EOT—remain robust and effective across diverse domains.

| Models | Emotion | | | Opinion Trigger | | | |
|---|---|---|---|---|---|---|---|
| | P | R | F1 | EM | PM | R1 | RL |
| Qwen2.5-0.5B-Instruct-FT | 0.33 | 0.32 | 0.32 | 0.01 | 0.05 | 0.20 | 0.15 |
| LLaMA-3.2-1B-Instruct-FT (EOT-LLaMA) | **0.45** | **0.52** | **0.49** | **0.20** | **0.33** | **0.43** | **0.37** |
| DeepSeek-R1-Distill-LLaMA-8B | 0.36 | 0.33 | 0.35 | 0.02 | 0.04 | 0.20 | 0.16 |

**Table 4:** Performance comparison of `fine-tuned` models. Best scores are in **bold**, and second-best scores are underlined. **Emotion Metrics**: P (Precision), R (Recall), F1 (F1-score). **Opinion Trigger Metrics**: EM (Exact Match), PM (Partial Match), R1 (ROUGE-1), RL (ROUGE-L).

**Error Analysis** Through detailed examination of model outputs, we identified several common failure patterns:

- **Trigger Boundary Detection:** Models often struggle with precise trigger span boundaries, particularly for complex emotional expressions. For instance, in cases where emotions are expressed through multiple connected clauses, models sometimes over-extend or under-extend the trigger spans.
- **Emotion Conflation:** We observed that models occasionally conflate similar emotions, particularly within Plutchik's adjacent categories. For example, "Trust" and "Joy" are frequently confused when the review expresses satisfaction with product reliability.
- **Context Sensitivity:** Performance degrades noticeably when emotions are expressed through implicit or contextual cues rather than explicit statements. This is particularly evident in reviews containing sarcasm or subtle emotional undertones.
- **Trigger Multiplicity:** Models show inconsistent performance when handling multiple triggers for a single emotion, often missing secondary or tertiary triggers that human annotators readily identify.

**Summary** Our comprehensive evaluation—based on a rigorously aggregated gold standard—shows that: (*1*) Model families differ markedly in their baseline and enhanced performance. (*2*) The EOT-DETECT framework delivers consistent and significant gains across a wide range of models compared to standard methods. (*3* Larger parameter sizes generally translate to improved performance, with architectural design playing a critical role. (*4*) Although closed-source models currently hold a performance edge, open-source models are rapidly catching up.

These findings underscore the importance of structured prompt engineering, model scale, and aggregation methodologies in advancing the state of emotion and trigger detection in natural language processing.

## 8 Conclusion & Future Work

This work introduced **EOT** (EMOTION-OPINION TRIGGER), the first joint task that unifies emotion detection and opinion trigger extraction in e-commerce reviews. We demonstrated why a focused set of Plutchik's 8 primary emotions provides the optimal framework for analyzing customer feedback, supported by empirical evidence of strong inter-annotator agreement (emotion $\kappa = 0.88$, trigger $\kappa = 0.84$). Some key takeaways are: (*a*) Our work presents the **EOT-X** benchmark—a rigorously annotated dataset. (*b*) **EOT-DETECT**, a structured prompting framework that achieves superior performance over conventional zero-shot and chain-of-thought methods. (*c*) **`EOT-LLaMA`**, an edge-deployable 1B-parameter model that outperforms larger models while maintaining efficiency.

Our work aims to establish a foundation for understanding not just *what* emotions customers express, but *why* they feel them, enabling more nuanced analysis of user feedback. We hope our work catalyzes further advances in emotion-opinion analysis systems for e-commerce applications.

## Limitations

We could not test state-of-the-art OpenAI's reasoning models `o1`, `o3-mini` and `o3-mini-high` as these required higher-tier API access beyond our research budget. We observed that when prompted, OpenAI models `GPT-4o` and `o1-mini` prevented access to their internal chain-of-thought (CoT) reasoning steps. This opacity limited our ability to analyze how these models interacted with our framework and hinders further improvements. Note that this is due to the proprietary, closed-source nature of these AI models, rather than a limitation of our framework. While we believe the development of the **EOT-X** dataset offers new research opportunities in emotion-opinion analysis, its moderate size of $\mathbf{2,400}$ samples stems from the complexity of annotating emotions and opinion triggers at the phrase level, which is both labor-intensive and financially demanding in academic research.

## Ethical Considerations

We engaged three raters with diverse academic backgrounds: a Master's student, a Pre-Doctoral researcher, and a Doctoral candidate. All were male, aged 24–32, with publications or active research in emotion analysis and coursework in Consumer Psychology. Raters were compensated appropriately for their contributions.

**EOT-DETECT** framework and the `EOT-LLaMA` model, a 1B-parameter model fine-tuned for emotion-trigger analysis in customer reviews—may occasionally produce hallucinations or unintended outputs, particularly in complex cases.

In this work, we define emotion detection and opinion trigger extraction within the context of analyzing customer reviews to understand the interplay between expressed emotions and their underlying causes. We do not claim that LLMs can experience or replicate emotions. Researchers should verify framework and model appropriateness before integration into evaluation processes, ensuring careful application in real-world e-commerce scenarios.

## References


Francisca Adoma Acheampong, Henry Nunoo-Mensah, and Wei Chen. 2023. Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*, 5(4):e12549.

AI@Meta. 2024. Llama 3 model card.

Anthropic. 2024. Claude 3.5 sonnet. https://www.anthropic.com/news/claude-3-5-sonnet.

Yujie Bao, Min Zhang, and Yang Liu. 2020. Emotion-cause pair extraction with graph convolutional networks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1010–1020, Online. Association for Computational Linguistics.

Weijun Bian and Gong Yan. 2022. Analyzing intention to purchase brand extension via brand attribute associations: The mediating and moderating role of emotional consumer-brand relationship and brand commitment. *Frontiers in Psychology*, 13:884673.

Corrado Bonferroni. 1936. Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8:3–62.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Tao Chen, Premaratne Samaranayake, 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.

William G. Cochran. 1977. *Sampling Techniques*, 3rd edition. Wiley.

Bart de Langhe, Philip M. Fernbach, and Donald R. Lichtenstein. 2015. Navigating by the Stars: Investigating the Actual and Perceived Validity of Online User Ratings. *Journal of Consumer Research*, 42(6):817–833.

DeepSeek AI. 2025. Deepseek-r1: 67b-parameter moe reasoning llm. https://huggingface.co/deepseek-ai/DeepSeek-R1-6B.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

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.

Zixiang Ding, Rui Xia, and Jianfei Yu. 2020. Ecpe-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. *Cognition and Emotion*, 6(3-4):169–200.

Wenqian Fan, Pengfei Li, Wei Li, Rui Yan, and Lide Wu. 2021. Joint constrained learning with boundary-adjusting for emotion-cause pair extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 1570–1580, Online. Association for Computational Linguistics.

Susan T. Fiske and Shelley E. Taylor. 1991. *Social Cognition*, 2nd edition. McGraw-Hill, New York.

Jerome L. Fleiss. 1971. *Measuring Nominal Scale Agreement Among Many Raters*, volume 76. American Psychological Association.

Rainer Greifeneder, Herbert Bless, and Thorsten Kuschmann. 2007. Extending the brand image on new products: The facilitative effect of happy mood states. *Journal of Consumer Behaviour: An International Research Review*, 6(1):19–31.

Lin Gui, Dongyin Wu, Ruifeng Xu, Qin Lu, and Yu Zhou. 2016. Event-driven emotion cause extraction with corpus construction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1639–1649, Austin, Texas. 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.

Carolina Herrando and Efthymios Constantinides. 2021. Emotional contagion: A brief overview and future directions. *Front. Psychol.*, 12:712606.

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*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Ming-Hui Huang and Roland T. Rust. 2024. Feeling ai for customer care. *Journal of Marketing*, 88(1):23–44.

HuggingFaceH4. 2023. Zephyr-7b-beta. https://huggingface.co/HuggingFaceH4/zephyr-7b-beta.

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

Sang-Hyun Kim, Kyung Hoon Park, and Jiyoung Park. 2019. Emotion as a signal of product quality: Its effect on purchase decision. *Internet Research*, 29(5):1103–1124.

Youn-Kyung Kim and Pauline Sullivan. 2019. Emotional branding speaks to consumers' heart: the case of fashion brands. *Fashion Text.*, 6(1).

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*.

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.

Sophia Yat Mei Lee, Ying Chen, and Chu-Ren Huang. 2010. A text-driven rule-based system for emotion cause detection. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 45–53, Los Angeles, CA. Association for Computational Linguistics.

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

Jiwei Li, Myle Ott, Claire Cardie, and Eduard Hovy. 2014. Towards a general rule for identifying deceptive opinion spam. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1566–1576, Baltimore, Maryland. Association for Computational Linguistics.

Xiangju Li, Shi Feng, Daling Wang, and Yifei Zhang. 2019. Context-aware emotion cause analysis with multi-attention-based neural network. *Knowledge-Based Systems*, 174:205–218.

Nadia Malik and Muhammad Bilal. 2024. Natural language processing for analyzing online customer reviews: a survey, taxonomy, and open research challenges. *PeerJ Computer Science*, 10:e2203.

Dina Mayzlin, Yaniv Dover, and Judith A Chevalier. 2012. Promotional reviews: An empirical investigation of online review manipulation. *SSRN Electron. J.*

Meta AI. 2024a. Llama-3.1-8b-instruct: Instruction-tuned multilingual llm. https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct.

Meta AI. 2024b. Llama-3.2-1b-instruct: Instruction-tuned multilingual llm. https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct.

Meta AI. 2024c. Llama-3.3-70b-instruct: Instruction-tuned multilingual llm. https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct.

Microsoft. 2024a. Phi-3.5-mini-instruct. https://huggingface.co/microsoft/Phi-3.5-mini-instruct.

Microsoft. 2024b. Phi-4-mini-4k-instruct. https://huggingface.co/microsoft/Phi-4-mini-4k-instruct.

Mistral AI. 2023. Mistral-7b-instruct-v0.3: Instruction-fine-tuned llm. https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3.

Mistral AI. 2024. Mistral-small-24b-instruct-2501. https://huggingface.co/mistralai/Mistral-Small-24B-Instruct-2501.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. 2009. Compositionality principle in recognition of fine-grained emotions from text. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, pages 278–285. Association for the Advancement of Artificial Intelligence (AAAI).

OpenAI. 2024. Gpt-4o model card. https://huggingface.co/openai/gpt-4o.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.

Michel Tuan Pham. 2007. Emotion and rationality: A critical review and interpretation of empirical evidence. *Review of general psychology*, 11(2):155–178.

Robert Plutchik. 1980. *Emotion: A Psychoevolutionary Synthesis*. Harper & Row, New York.

Robert Plutchik. 2001. The nature of emotions. *American Scientist*, 89(4):344–350.

Qwen Team. 2024. Qwen2.5-72b-instruct: 72b-parameter instruction-fine-tuned llm. https://qwenlm.github.io/blog/qwen2.5/.

Qwen Team. 2025. Qwq-32b: 32b-parameter specialized llm. https://huggingface.co/Qwen/QwQ-32B.

Ana Babić Rosario, Francesca Sotgiu, Kristine De Valck, and Tammo H.A. Bijmolt. 2016. The effect of electronic word of mouth on sales: A meta-analytic review of platform, product, and metric factors. *Journal of Marketing Research*, 53(3):297–318.

James A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.

Klaus R. Scherer and Harald G. Wallbott. 1994. The ISEAR dataset: International survey on emotion antecedents and reactions. *Technical Report*.

Aaditya Singh, Aditya Joshi, Sachin Pawar, Soumen Chakrabarti, and Pushpak Bhattacharyya. 2021. An end-to-end network for emotion-cause pair extraction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1456–1461. Association for Computational Linguistics.

Smriti Singh, Cornelia Caragea, and Junyi Jessy Li. 2024. Language models (mostly) do not consider emotion triggers when predicting emotion. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 603–614, Mexico City, Mexico. Association for Computational Linguistics.

Tiberiu Sosea and Cornelia Caragea. 2020. CancerEmo: A dataset for fine-grained emotion detection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8892–8904, Online. Association for Computational Linguistics.

Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74, Prague, Czech Republic. Association for Computational Linguistics.

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

Shanshan Wang, Lingling Zhan, and Rui Liu. 2022. Emotional responses in online reviews: The role of emotional clarity and intensity. *SSRN Electronic Journal*.

Zhongyu Wei, Jiangxia Li, Yaojie Zhang, Zhendong Mao, Yulan He, and Bingquan Wang. 2020. A unified sequence labeling model for emotion cause pair extraction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1449–1460, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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 M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Rui Xia and Zixiang Ding. 2019. Emotion-cause pair extraction: A new task to emotion analysis in texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1003–1012, Florence, Italy. Association for Computational Linguistics.

Chunchang Xie, Junxi Jin, and Xiaoling Guo. 2022. Impact of the critical factors of customer experience on well-being: Joy and customer satisfaction as mediators. *Front. Psychol.*, 13:955130.

Yelp. 2025. Yelp open dataset. Accessed: 2025-02-16.

Dezhi Yin, Samuel D. Bond, and Han Zhang. 2014. Anxious or angry? effects of discrete emotions on the perceived helpfulness of online reviews. *MIS Q.*, 38(2):539–560.

Yujie Yuan, Min Zhang, and Yang Liu. 2020. Emotion-cause pair extraction with graph convolutional networks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1010–1020, Online. Association for Computational Linguistics.

# A LLMs Utilized

In our experiments, we adopt a range of recent widely-used Large Language Models (LLMs). For our comprehensive experimental evaluation, we leveraged `NVIDIA H100 GPUs` to execute all open-source models. Proprietary models were evaluated via their respective vendor APIs.

As stated in the main manuscript, we evaluated 23 language models, which are stratified into two principal categories: general-purpose instruction-following models and models explicitly architected or fine-tuned for sophisticated, multi-step reasoning.

**Model Inventory** The complete inventory of the 23 models benchmarked in Tables 2 and 3 is as follows. Models marked with a † are reasoning-enhanced.

- `LLaMA-3.2-1B-Instruct` (Meta AI, 2024b)
- `LLaMA-3.2-3B-Instruct` (AI@Meta, 2024)
- `LLaMA-3.1-8B-Instruct` (Meta AI, 2024a)
- `LLaMA-3.3-70B-Instruct` (Meta AI, 2024c)
- `Mistral-7B-Instruct-v0.2` (Jiang et al., 2023)
- `Mistral-7B-Instruct-v0.3` (Mistral AI, 2023)
- `Mistral-Small-24B-Instruct-2501` (Mistral AI, 2024)
- `Zephyr-7b-beta` (HuggingFaceH4, 2023)
- `Qwen2.5-0.5B-Instruct` (Qwen Team, 2024)
- `Qwen2.5-3B-Instruct` (Qwen Team, 2024)
- `Qwen2.5-7B-Instruct` (Qwen Team, 2024)
- `Qwen2.5-32B-Instruct` (Qwen Team, 2024)
- `Qwen2.5-72B-Instruct` (Qwen Team, 2024)
- `QwQ-32B†` (Qwen Team, 2025)
- `Phi-3.5-mini-instruct` (Microsoft, 2024a)
- `Phi-4` (Microsoft, 2024b)
- `Gemma-2-2b-it` (Team et al., 2024)
- `Gemma-2-9b-it` (Team et al., 2024)
- `Gemma-2-27b-it` (Team et al., 2024)
- `DeepSeek-R1†` (DeepSeek AI, 2025)
- `GPT-4o` (OpenAI, 2024)
- `o1-mini†` (OpenAI, 2024)
- `Claude 3.5 Sonnet` (Anthropic, 2024)

Access to open-source models was facilitated via the HuggingFace library (Wolf et al., 2020).

| Metric | Beauty | Clothing | Home | Electronics | TripAdvisor | Yelp |
|---|---|---|---|---|---|---|
| *Overall Statistics* | | | | | | |
| Total Emotions | 957 | 897 | 916 | 913 | 1183 | 1084 |
| Avg Emotions per Review | 2.39 | 2.24 | 2.29 | 2.28 | 2.96 | 2.71 |
| *Emotion: Anger* | | | | | | |
| Count (Percentage) | 58 (6.1%) | 41 (4.6%) | 73 (8.0%) | 82 (9.0%) | 86 (7.3%) | 66 (6.1%) |
| Total Triggers | 148 | 106 | 221 | 261 | 371 | 270 |
| Avg Triggers per Emotion | 2.55 | 2.59 | 3.03 | 3.18 | 4.31 | 4.09 |
| Avg Trigger Length (words) | 6.80 | 8.70 | 8.20 | 8.52 | 9.73 | 9.12 |
| *Emotion: Anticipation* | | | | | | |
| Count (Percentage) | 105 (11.0%) | 100 (11.1%) | 82 (9.0%) | 89 (9.7%) | 180 (15.2%) | 196 (18.1%) |
| Total Triggers | 148 | 135 | 122 | 133 | 279 | 300 |
| Avg Triggers per Emotion | 1.41 | 1.35 | 1.49 | 1.49 | 1.55 | 1.53 |
| Avg Trigger Length (words) | 10.30 | 10.62 | 10.24 | 11.23 | 10.62 | 10.40 |
| *Emotion: Disgust* | | | | | | |
| Count (Percentage) | 139 (14.5%) | 111 (12.4%) | 117 (12.8%) | 120 (13.1%) | 172 (14.5%) | 101 (9.3%) |
| Total Triggers | 292 | 215 | 256 | 277 | 506 | 268 |
| Avg Triggers per Emotion | 2.10 | 1.94 | 2.19 | 2.31 | 2.94 | 2.65 |
| Avg Trigger Length (words) | 7.81 | 8.65 | 8.38 | 9.02 | 8.92 | 9.10 |
| *Emotion: Fear* | | | | | | |
| Count (Percentage) | 14 (1.5%) | 4 (0.4%) | 16 (1.7%) | 16 (1.8%) | 36 (3.0%) | 12 (1.1%) |
| Total Triggers | 18 | 5 | 20 | 28 | 64 | 20 |
| Avg Triggers per Emotion | 1.29 | 1.25 | 1.25 | 1.75 | 1.78 | 1.67 |
| Avg Trigger Length (words) | 13.00 | 12.80 | 10.30 | 9.71 | 11.12 | 10.80 |
| *Emotion: Joy* | | | | | | |
| Count (Percentage) | 294 (30.7%) | 318 (35.5%) | 305 (33.3%) | 282 (30.9%) | 312 (26.4%) | 328 (30.3%) |
| Total Triggers | 829 | 846 | 809 | 716 | 1498 | 1353 |
| Avg Triggers per Emotion | 2.82 | 2.66 | 2.65 | 2.54 | 4.80 | 4.12 |
| Avg Trigger Length (words) | 7.04 | 6.43 | 6.97 | 7.14 | 8.14 | 7.86 |
| *Emotion: Sadness* | | | | | | |
| Count (Percentage) | 112 (11.7%) | 100 (11.1%) | 103 (11.2%) | 103 (11.3%) | 95 (8.0%) | 76 (7.0%) |
| Total Triggers | 221 | 204 | 234 | 201 | 262 | 167 |
| Avg Triggers per Emotion | 1.97 | 2.04 | 2.27 | 1.95 | 2.76 | 2.20 |
| Avg Trigger Length (words) | 7.91 | 8.80 | 7.84 | 8.79 | 9.57 | 10.16 |

**Table 5:** Gold Standard Aggregated Dataset: Emotion Statistics across Domains.

**Filtering Low-Capacity Models** Pilot tests showed that some earlier Pre-trained Language Models (PLMs), such as BART (Lewis et al., 2019), could not follow the multi-step instructions in our prompt. They would either echo the template verbatim or loop incoherently, which is behavior typical of low-parameter models on long, structured inputs. We therefore excluded such models from our final benchmark to ensure meaningful comparisons.

## B Inference Configuration Details

For our experiments, we implemented a standardized inference configuration across both closed-source and open-source LLMs to ensure consistent and reliable outputs. While different LLM providers recommend varying default settings (e.g., GPT-4o suggests `temperature=1.0`, DeepSeek recommends (`temperature=0.6`), we conducted extensive parameter tuning to optimize for our specific task requirements.

### B.1 Parameter Settings

We empirically determined the following optimal configuration: `inference_config = {'temperature': 0.2, 'top_p': 0.95, 'top_k': 25, 'max_tokens': 2500, 'n': 1}`

### B.2 Rationale for Parameter Choices

`Temperature` $= 0.2$ : We deliberately chose a low temperature setting for several reasons:

- Our task focuses on *detection* and *extraction* rather than creative generation.
- Lower temperature produces more deterministic outputs, which is crucial for consistent emotion detection.
- Reduces variation in *trigger span* identification, leading to more reliable extractions.
- Aligns with our need for *precise*, focused responses rather than diverse creative generations.

**TOP-P (0.95) and TOP-K (25)**: These settings provide a balanced approach to sampling:

- `top-p = 0.95` ensures consideration of the most probable tokens while maintaining some flexibility.
- `top-k = 25` limits the selection pool to the most relevant tokens.
- The combination helps maintain coherence while capturing *emotional nuances* in the reviews.

**MAX TOKENS (2500)**: This limit was set to accommodate:

- Comprehensive *emotion analysis*.
- Multiple *trigger span* extractions.
- *Reasoning steps* in the chain-of-thought prompting.
- *Self-reflection* mechanisms in the EOT-DETECT framework.

These parameters were validated through extensive testing across our evaluation suite, demonstrating consistent performance in both *emotion detection* and *trigger extraction* tasks. The configuration prioritizes *reliability* and *precision* over creativity, aligning with our task's objective of accurate analysis rather than generative capability.

# C Implementation Details of Fine-tuned Models

We provide comprehensive details about our experimental setup, including data preprocessing, model architecture, and training configuration. All experiments were conducted using the Unsloth framework.

## C.1 Data Preprocessing

We randomly split EOT-X into training (80%), validation (10%), and test (10%) sets, ensuring that reviews from the same product (for Amazon) or experience/service (for TripAdvisor and Yelp) were kept within the same split to prevent data leakage.

| **Hyperparameter** | **Value** |
|---|---|
| *Dataset Statistics* | |
| `Training Set` | 1,919 examples |
| `Validation Set` | 239 examples |
| `Test Set` | 241 examples |
| *Model Architecture* | |
| `Maximum Sequence Length` | 2,048 tokens |
| `Data Type` | bfloat16 |
| *LoRA Configuration* | |
| `Rank (r)` | 128 |
| `Alpha` | 128 |
| `Dropout` | 0.05 |
| `Use RSLoRA` | True |
| *Training Configuration* | |
| `Number of Epochs` | 5 |
| `Batch Size (per device)` | 2 |
| `Gradient Accumulation Steps` | 8 |
| `Effective Batch Size` | 16 |
| `Total Training Steps` | 600 |
| `Optimizer` | AdamW |
| `Learning Rate` | 2e-5 |
| `Weight Decay` | 0.01 |
| `Learning Rate Scheduler` | Cosine |
| `Warmup Ratio` | 0.05 |
| `Early Stopping Patience` | 2 |
| `Early Stopping Threshold` | 0.005 |
| `Save/Evaluation Frequency` | Every 200 steps |
| *Inference Configuration* | |
| `Maximum New Tokens` | 2,048 |
| `Temperature` | 0.2 |
| `Top-p` | 0.95 |
| `Top-k` | 25 |
| `Repetition Penalty` | 1.1 |

**Table 6:** Hyperparameters and configuration details for our experiments.

# D Annotation Guidelines

### Introduction for Annotators

*Dear Annotators,*

Thank you for participating in this crucial annotation task for EMOTION AND OPINION TRIGGER DETECTION in e-commerce reviews. Your expertise is vital for constructing the high-quality

EOT-X dataset, which will significantly advance our understanding of customer emotional responses and their underlying triggers in product reviews.

In this task, you will identify emotions expressed in reviews and extract the exact text spans (OPINION TRIGGERS) that explain why each emotion is present. Please read these guidelines thoroughly before beginning the annotation process.

### D.1 Task Overview

Your annotation task consists of two main components:

**Emotion Detection:** Identify all emotions expressed in a review using PLUTCHIK'S 8 primary emotions.
**Opinion Trigger Extraction:** Extract the precise text spans from the review that explain the presence of each detected emotion.

### D.2 Emotion Framework: PLUTCHIK'S 8 Primary Emotions

Focus exclusively on these 8 primary emotions:

- Joy
- Trust
- Fear
- Surprise
- Sadness
- Disgust
- Anger
- Anticipation

**Note:** If no clear emotion is detected in the review, label it as *Neutral*.

### D.3 Opinion Triggers

An OPINION TRIGGER is defined as a direct extract from the review that explains why the customer experienced a particular emotion. Each trigger must be:

- **Extractive:** Exactly as it appears in the review.
- **Clearly Linked:** Directly associated with the identified emotion.
- **Self-Contained:** Understandable without needing additional context.

### D.4 Annotation Process

**General Instructions**

- **Read Thoroughly:** Carefully read the entire review before starting your annotations.
- **Identify Emotions:** Detect all emotions expressed in the review.
- **Extract Triggers:** For each detected emotion, extract all relevant opinion triggers as exact text spans.
- **Documentation:** Record your annotations in the provided format.
- **Confidence Level:** Only mark emotions and triggers you are confident about.

### D.5 Specific Guidelines for Emotion Detection

- **Multiple Emotions:** Reviews may express more than one emotion. Annotate every clearly expressed emotion.
- **Emotion Intensity:** Focus solely on the presence or absence of an emotion; do not attempt to annotate the intensity.
- **Implicit vs. Explicit:** Annotate only those emotions that are clearly expressed. Do not infer or speculate.
- **Contextual Consideration:** Use the entire review context to guide your decision.

**Edge Cases and Special Considerations** Trigger Boundary Cases:

**Nested Triggers:** When a shorter trigger is nested within a longer one, choose the span that is most informative and coherent.
**Discontinuous Triggers:** Do not annotate discontinuous spans. Choose the most representative continuous span that best explains the emotion.
**Comparative Expressions:** For comparative statements (e.g., *Better than all other brands I've tried*), include the full comparison if it is directly relevant to explaining the emotion.

### D.6 Special Review Types

**Very Short Reviews:** Even in very short reviews, if clear emotional signals are present, annotate them. In such cases, the entire review might serve as the trigger.
**Technical Reviews:** Focus on the emotional content, not on technical details, unless the technical detail directly causes an emotional reaction.
**Sarcastic Reviews:** Identify and annotate the intended emotion rather than the literal meaning. If sarcasm is detected, note it in the comments field (if needed).

### D.7 Annotation Format

For each review, provide the following information:

- **Review ID**
- **List of Detected Emotions:** For each emotion, include:
  - **Emotion Label:** One of PLUTCHIK'S 8 primary emotions or *Neutral*.
  - **List of Triggers:** Exact text spans from the review.

### D.8 Format Example

```
{
  "review_id": "123",
  "emotions": [
    {
      "emotion": "Joy",
      "triggers": [
        "works perfectly every time",
        "excellent battery life"
      ]
    }
  ]
}
```

**Quality Control Measures** Self-Check Before Submission:

- **Emotion Consistency:** All annotated emotions belong to PLUTCHIK'S set or are labeled as Neutral.
- **Trigger Accuracy:** Every OPINION TRIGGER is an exact text span from the review.
- **Clear Linkage:** Each emotion-trigger pair is clearly and unambiguously connected.
- **Completeness:** No clear emotional expressions are left unannotated.

**Final Notes** (*1*) **Attention to Detail:** Your meticulous attention to detail is crucial for the success of this dataset. (*2*) **Feedback:** If you encounter any ambiguities or issues, please reach out to the research team immediately. (*3*) **Consistency:** Maintain high consistency across all annotations to ensure the dataset is robust and reliable. (*4*) **Expert Judgment:** Rely on your expertise in both emotion analysis and consumer psychology to make nuanced decisions.

*Thank you for your dedication and careful work in this annotation task. Your contributions are instrumental in advancing research on emotion-opinion trigger detection in e-commerce reviews.*

## E Model Responses

In this section, we present the original customer review, followed by example emotions and opinion triggers detected by various LLMs in Table 7 and 8. These examples provide a clear understanding of how different models interpret the same input review to detect emotions and triggers.

By presenting the review first, we facilitate a direct comparison between the input text and the output generated by different models.

`Review:` `I've been buying these for my mom for 2 years now. She works in food service, on her feet 10+ hours at a time and these are the only shoes she will wear to work! A touch wider than normal shoes (like most skechers) so she can wear thicker socks in winter and in summer she has room in case her feet swell a little. Non slip, great arch support; my mom replaces about every 6mo as she works 6 days a week.`

| MODELS | EMOTION | OPINION TRIGGERS |
|---|---|---|
| `Deepseek-R1` | Trust | • The only shoes she will wear to work<br>• Non slip, great arch support |
| | Anticipation | • She can wear thicker socks in winter<br>• In summer she has room in case her feet swell a little |
| `GPT-4o` | Joy | • I've been buying these for my mom for 2 years now<br>• These are the only shoes she will wear to work<br>• Non slip, great arch support |
| | Trust | • These are the only shoes she will wear to work<br>• Non slip, great arch support<br>• A touch wider than normal shoes (like most Skechers) so she can wear thicker socks in winter and in summer she has room in case her feet swell a little |
| | Anticipation | • My mom replaces about every 6 months as she works 6 days a week |
| `LLaMA-3.3-70B-Instruct` | Trust | • The only shoes she will wear to work<br>• Great arch support |
| | Joy | • She can wear thicker socks in winter and in summer she has room in case her feet swell a little |
| | Anticipation | • My mom replaces about every 6 months as she works 6 days a week |

**Table 7:** Example model responses generated using our EOT-DETECT framework. The table showcases the models' ability to jointly detect emotions and their corresponding opinion triggers from a sample review. Each entry illustrates how specific text spans are identified as triggers, including instances where multiple triggers are associated with a single emotion.

| MODELS | EMOTION | OPINION TRIGGERS |
|---|---|---|
| Claude 3.5 Sonnet | Trust | • I've been buying these for my mom for 2 years now<br>• These are the only shoes she will wear to work<br>• My mom replaces about every 6 months |
| | Joy | • Non slip, great arch support<br>• She has room in case her feet swell a little |
| o1-mini | Trust | • These are the only shoes she will wear to work<br>• Non slip<br>• Great arch support<br>• She replaces about every 6 months as she works 6 days a week<br>• A touch wider than normal shoes<br>• She can wear thicker socks in winter and in summer she has room in case her feet swell a little |
| | Joy | • I've been buying these for my mom for 2 years now<br>• These are the only shoes she will wear to work |
| | Anticipation | • She replaces about every 6 months as she works 6 days a week |
| Mistral-Small-24B-Instruct | Joy | • She works in food service, on her feet 10+ hours at a time and these are the only shoes she will wear to work!<br>• Non slip, great arch support; my mom replaces about every 6 months as she works 6 days a week. |
| | Trust | • I've been buying these for my mom for 2 years now. |
| | Anticipation | • A touch wider than normal shoes (like most Skechers) so she can wear thicker socks in winter and in summer she has room in case her feet swell a little. |

**Table 8:** Example model responses generated using our EOT-DETECT framework. The table showcases the models' ability to jointly detect emotions and their corresponding opinion triggers from a sample review. Each entry illustrates how specific text spans are identified as triggers, including instances where multiple triggers are associated with a single emotion.