Mind, Matter, and Markets: A Survey of Human-Centered LLM Applications in E-commerce

Arnav Attri, Anuj Attri, Pushpak Bhattacharyya
Computer Science and Engineering, IIT Bombay, India
{arnavcs, ianuj, pb}@cse.iitb.ac.in

Abstract

The rapid growth of e-commerce platforms has created overwhelming product volumes and diverse consumer needs, leading to significant decision-making challenges for users. Traditional recommendation systems struggle to process the complexity and nuance in reviews, which contain rich emotional and contextual information crucial for purchase decisions. Large Language Models (LLMs) have emerged as powerful tools for addressing this complexity through sophisticated opinion summarization and product analysis capabilities. This survey synthesizes recent LLM-based innovations in e-commerce, focusing on both emotional and factual aspects of customer feedback processing. We identify five core research directions: **(1) Multi-Source Opinion Summarization**, which integrates diverse product metadata and reviews; **(2) Emotion-Aware Opinion Summarization**, which prioritizes affective information in customer feedback; **(3) Query-Focused Comparative Summarization**, enabling tailored product comparisons; **(4) Opinion Trigger Detection**, identifying text spans that evoke specific emotional responses; and **(5) Query-Focused Explainable Recommendation**, providing transparent rationales for product suggestions. We also examine the emerging use of LLMs as evaluators for reducing human annotation requirements while maintaining alignment with user preferences.

1 Introduction

E-commerce platforms have fundamentally transformed global retail landscapes, with worldwide online sales exceeding \$5.7 trillion in 2022 and projected to surpass \$8 trillion by 2026 (eMarketer, 2023). This exponential growth has generated unprecedented volumes of heterogeneous data, including product `metadata`, user-generated reviews, and diverse purchasing trajectories. While this information abundance theoretically empowers consumers, it paradoxically induces significant cognitive burden—a phenomenon extensively studied as ***choice overload*** (Scheibehenne et al., 2010). Contemporary consumers encounter substantial friction (Chen et al., 2019b; Wang and Benbasat, 2022) when navigating complex product catalogs, interpreting subjective feedback, performing multi-attribute comparisons, and comprehending algorithmic recommendation rationales.

Traditional computational approaches to e-commerce information processing—encompassing rule-based sentiment classification, aspect-based opinion mining, and matrix factorization-based collaborative filtering—demonstrate limited *efficacy* in addressing these multifaceted challenges (Wang et al., 2016; Chu and Wang, 2019). These methodologies typically operate in silos, focusing exclusively on either sentiment polarity detection, feature extraction, or recommendation generation, without modeling the intricate cognitive, affective, and contextual dimensions underlying consumer decision-making processes (Brazinskas, 2020; Tay et al., 2019). Consequently, users continue experiencing substantial friction in their purchasing journeys (Bagozzi and Dholakia, 2003; Duan et al., 2008b), manifesting as cart abandonment, decision fatigue, and diminished post-purchase satisfaction.

The emergence of Large Language Models (LLMs) presents significant computational advances for mitigating these architectural and interaction design challenges (Brown et al., 2020a; Ouyang et al., 2022c). LLMs exhibit substantial proficiency in contextual representation learning, few-shot adaptation, and complex reasoning across multi-modal data streams, facilitating the processing of heterogeneous, high-dimensional information while generating outputs that correspond to `human-interpretable cognitive frameworks` (Wei et al., 2022a; Liu et al., 2023a). These computational properties establish LLMs as foundational architectural primitives for enhancing user-platform interaction paradigms within e-commerce computational ecosystems (Fu et al., 2023b; Chiang et al., 2023).

Recent research has leveraged LLMs to develop more intuitive, human-centric approaches to `e-commerce` information processing (Li et al., 2020a). These methodologies represent a paradigmatic shift from traditional product-centric architectures toward user-centric systems that prioritize emotional intelligence, query relevance, and decision transparency (Im et al., 2021; Zhang and Chen, 2020a). By incorporating these design principles, contemporary approaches address critical pain points in the consumer journey—information overload, affective uncertainty, personalization deficits, and trust erosion (Greifeneder et al., 2007; Pham, 2007).

Mind, Matter, and Markets: A Tripartite Taxonomic Framework: This survey presents a structured analytical framework—`Mind, Matter, and Markets`—to systematically examine the impact of LLM-based innovations on e-commerce information processing. This *tripartite* taxonomy categorizes recent methodological advances based on their primary computational focus, providing insights into how LLMs influence cognitive processing (`Mind`), computational infrastructure and data synthesis (`Matter`), and economic interaction mechanisms (`Markets`).

MIND encompasses methodologies that prioritize the psychological and affective dimensions of consumer experience. These approaches recognize that purchasing decisions are not purely rational optimization problems but are significantly influenced by emotional responses to products, reviews, and contextual factors (Damasio, 2004; Lerner et al., 2015). By capturing and interpreting the affective content embedded in customer feedback, these methodologies provide emotionally resonant information that aligns with empirical models of human decision-making (Kim et al., 2019; Wang et al., 2023c).

Example: Consider a user researching wireless headphones who encounters reviews expressing "*frustration*" with battery life versus "*delight*" with sound quality. Mind-focused approaches would explicitly model these emotional dimensions, generating summaries that convey not just factual information ("battery lasts 8 hours") but affective context ("users express frustration with the 8-hour battery life, particularly for long commutes").

MATTER refers to techniques that synthesize and contextualize factual product information according to specific user information needs. These approaches acknowledge that different consumers require different information subsets about identical products, contingent upon their particular queries and use-case requirements (Angiolillo et al., 2022; Ankit et al., 2022). By integrating heterogeneous product metadata (`technical specifications, marketing descriptions, feature lists`) with user-generated content (`reviews, ratings, Q&A`), these methods generate comprehensive yet targeted information representations (Li et al., 2020a; Im et al., 2021).

Example: For a query "*best laptop for video editing*", a Matter-focused system would synthesize technical specifications (GPU memory, CPU cores), marketing descriptions ("professional-grade performance"), and relevant review excerpts ("rendered 4K video in 20 minutes") into a coherent, query-specific summary. This contrasts with generic product descriptions that may emphasize irrelevant attributes like portability for gaming use cases. Multi-source opinion summarization creates significantly more informative and contextually relevant product overviews than review-only approaches.

MARKETS focuses on operationalizing these computational insights within commercial platforms to enhance real-world consumer decision-making workflows. These approaches address practical deployment challenges in e-commerce ecosystems, including multi-product comparison interfaces, recommendation justification mechanisms, and purchase confidence optimization (Wang et al., 2018b; Chen et al., 2018c). By presenting information in formats that facilitate direct comparison and transparent algorithmic reasoning, these methods reduce decision friction and enhance user trust (Le et al., 2021a; Echterhoff et al., 2023a).

Example: A `Markets`-focused system might present comparative tables showing how three recommended smartphones perform across user-specified criteria (camera quality, battery life, price), accompanied by natural language explanations: "*Phone A excels in low-light photography based on 200+ user reviews, while Phone B offers superior battery performance for heavy usage* *patterns.*" Research on query-focused comparative summaries demonstrates that such comparative explanations significantly improve decision confidence and purchase satisfaction.

Five Pioneering Research Directions: Within this taxonomic framework, we identify five pioneering research directions that collectively transform e-commerce information processing architectures:

QUERY-FOCUSED COMPARATIVE EXPLAINABLE SUMMARIZATION (QF-CES) enables systematic comparison of multiple recommended products within the context of specific user queries. By presenting information in structured tabular formats alongside natural language "final verdict" explanations, QF-CES facilitates efficient cross-product comparison while maintaining query relevance. This approach bridges the `Matter` and Markets dimensions by contextualizing product information according to user needs and facilitating practical decision-making workflows. Empirical evaluations demonstrate that QF-CES reduces inference latency by approximately 40% compared to direct prompt-based approaches while maintaining output quality.

EMOTION-AWARE OPINION SUMMARIZATION (EAOS) captures both cognitive (explicit opinions) and affective (associated emotions) dimensions of customer reviews. Grounded in `Plutchik`'s (Plutchik, 1988) circumplex model of eight primary emotions—joy, trust, fear, surprise, sadness, disgust, anger, and anticipation—EAOS generates summaries that reflect not only ***what*** customers think but ***how*** they emotionally respond. This approach directly addresses the MIND dimension by recognizing the crucial role of affective states in purchasing decisions. ***Controlled user studies demonstrate that 82% of participants prefer emotion-aware summaries over traditional opinion summaries***, with significant improvements in decision confidence metrics.

EMOTION AND OPINION TRIGGER DETECTION (EOT) identifies not only what emotions are expressed in reviews but specifically which textual spans trigger those emotional responses. By explicitly modeling causal relationships between opinion triggers (textual evidence) and affective dimensions (`emotion categories`), EOT provides deeper insights into product-experience relationships. This approach primarily addresses the Mind dimension by elucidating causal mechanisms between product attributes and emotional responses. Comprehensive benchmarking across 23 contemporary LLMs demonstrates the effectiveness of structured reasoning approaches for this causal modeling task.

MULTI-SOURCE OPINION SUMMARIZATION (M-OS) extends traditional review-based opinion summarization by integrating product metadata (titles, descriptions, features, specifications) with customer reviews. This approach recognizes that comprehensive product understanding requires synthesizing both objective manufacturer-provided attributes and subjective user experiences. M-OS addresses the Matter dimension by providing holistic product representations that combine disparate information sources into coherent narratives. Experimental results demonstrate that M-OS significantly enhances user engagement, with 87% of study participants preferring multi-source summaries over traditional review-only approaches.

QUERY-FOCUSED EXPLAINABLE RECOMMENDATION (QF-ER) generates natural language explanations that justify product recommendations based on specific user queries rather than historical user profiles. Unlike traditional collaborative filtering systems that rely on user-item interaction matrices, QF-ER focuses exclusively on current query context, enhancing privacy while maintaining personalization effectiveness. This approach primarily addresses the `Markets` dimension by building user trust through transparent justification of algorithmic recommendations. The methodology employs reference-free evaluation metrics to assess explanation quality across multiple dimensions including `clarity, fluency, coherence, and query relevance`.

These five research directions, while methodologically distinct, exhibit significant interconnections and complementarities within the proposed framework. M-OS serves as a foundational component for comprehensive product representation, which can be enhanced with emotional dimensions (EAOS), adapted for comparative scenarios (QF-CES), enriched with causal insights (EOT), or leveraged for recommendation justification (QF-ER). Collectively, they represent a paradigmatic shift from isolated technical solutions toward integrated approaches that address multiple dimensions of the e-commerce user experience simultaneously. The `formal definitions, including input and output specifications` for each of these research directions, are provided in **(Section 4)**.

2 Language Models

Understanding the architectural evolution of language models is essential for contextualizing the e-commerce applications explored in this survey. We examine the progression from foundational pre-trained models to large language models, along with the optimization techniques that make them practical for deployment in commercial systems.

PLMs: Before the widespread adoption of modern, large-scale LLMs, several foundational sequence-to-sequence models established the viability of transformers for complex generative tasks. Among the most influential are BART, T5, and PEGASUS, which have served as critical baselines in summarization research.

BART: (*Bidirectional and Auto-Regressive Transformers*) (Lewis et al., 2020a) is a sequence-to-sequence model specifically pre-trained as a denoising autoencoder. Its architecture consists of a bidirectional encoder to read and corrupt input text and a left-to-right auto-regressive decoder to reconstruct the original text. During pre-training, an uncorrupted text sequence X is transformed by a noise function g into a corrupted version $\tilde{X}$. The model is then trained to reconstruct X by maximizing the likelihood $P(X|\tilde{X})$. This pre-training objective, which corrupts text by masking tokens or permuting sentences, compels the model to learn robust bidirectional representations, making it highly effective for abstractive summarization.

T5: (*Text-to-Text Transfer Transformer*) (Raffel et al., 2020) introduced a unified framework that casts every NLP task as a text-to-text problem. Instead of having task-specific architectures, T5 uses a standard encoder-decoder transformer that is trained to generate a target text string given an input text string. To specify the task, a short prefix is added to the original input. For summarization, the input is formatted as follows:

```
summarize: <document text>
```

The model is then fine-tuned to generate the corresponding summary. This versatile approach allows a single model to perform a wide array of tasks—from translation to question answering to summarization—simply by changing the input prefix, demonstrating state-of-the-art performance and greatly simplifying the transfer learning pipeline.

PEGASUS: (*Pre-training with Extracted Gap-sentences for Abstractive Summarization*) (Zhang et al., 2020a) is a transformer-based encoder-decoder model specifically architected for abstractive summarization. Its key innovation lies in its pre-training objective, known as **Gap-Sentence Generation** (GSG). Instead of masking random tokens, PEGASUS masks entire sentences from a document and trains the model to generate them from the remaining context. Specifically, given a document D with sentences $\{s_1, s_2, \ldots, s_n\}$, a subset of "important" sentences $S_{\text{gsg}} \subseteq \{s_1, s_2, \ldots, s_n\}$ is selected to be masked. The model is then trained to maximize the conditional likelihood $P(S_{\text{gsg}} \mid D \setminus S_{\text{gsg}})$, where $D \setminus S_{\text{gsg}}$ represents the document with the gap-sentences removed. Because these important sentences often function as a pseudo-summary, this pre-training task closely mirrors the downstream task of summarization, enabling the model to achieve strong performance with minimal fine-tuning.

Large Language Models (LLMs): The emergence of Large Language Models (LLMs) represents a significant paradigm shift from the foundational models discussed previously. This shift is characterized by an unprecedented increase in scale—both in model parameters and training data—leading to the development of remarkable emergent capabilities, such as the ability to perform complex tasks in a zero-shot or few-shot manner without task-specific training (Brown et al., 2020b).

The evolution of LLMs began with a focus on autoregressive pre-training, where a model is trained to predict the next token in a sequence. While powerful, these base models were not inherently aligned with human intent. The breakthrough came with the introduction of **instruction-tuning** and **Reinforcement Learning from Human Feedback (RLHF)** (Ouyang et al., 2022b). In this multi-stage process, a pre-trained model is first fine-tuned on a dataset of curated instructions and responses (*supervised fine-tuning*, SFT). Subsequently, a reward model r_θ is trained to predict human preferences, and the SFT model is further fine-tuned to optimize this reward. The objective for the policy π_{RL} is to maximize the expected reward while not deviating too far from the base model, typically constrained by a KL-divergence penalty:

$$\text{maximize } \mathbb{E}_{y\sim\pi_{\text{RL}}(\cdot|x)}[r_\theta(x,y) - \beta D_{\text{KL}}(\pi_{\text{RL}}(\cdot|x)||\pi_{\text{SFT}}(\cdot|x))] \quad (1)$$

where x is the prompt, y is the completion, and β is the KL coefficient. This alignment process has been fundamental to the success of modern conversational agents and has given rise to a dynamic ecosystem of both proprietary and open-source models.

Key LLM Families: The contemporary landscape of LLMs is characterized by several prominent model families that have fundamentally shaped the trajectory of natural language processing research and applications. The selection of model families examined herein reflects their substantial impact on both academic research and practical applications, as evidenced by their widespread adoption, extensive fine-tuning variants, and influence on subsequent architectural innovations. Furthermore, these families demonstrate varying approaches to critical challenges in large-scale language modeling, including computational efficiency, multilingual capabilities, reasoning enhancement, and the balance between model capacity and inference costs. The comparative analysis of these architectures provides essential context for understanding current state-of-the-art capabilities and identifying promising directions for future research endeavors.

`Llama` (Meta AI): The `Llama` series of models (Grattafiori et al., 2024) has been pivotal in democratizing access to high-performance LLMs. Their release catalyzed a wave of innovation in the open-source community, leading to the development of numerous variants and fine-tunes. Subsequent releases, like `Llama 3`, have continued to close the performance gap with proprietary counterparts.

`Mistral` (Mistral AI): This family of models is notable for its architectural efficiency. Models like `Mixtral-8x7B` (Jiang et al., 2023) popularized the use of a sparse **Mixture-of-Experts (MoE)** architecture in open-source models. In an MoE layer, the output y is a weighted sum of the outputs from a set of "expert" networks $\{E_1, \ldots, E_n\}$, where the weights are determined by a gating network $g(x)$:

$$y = \sum_{i=1}^{n} g(x)_i \cdot E_i(x) \tag{2}$$

This allows the model to have a very large number of parameters while only activating a fraction of them for any given input, significantly reducing computational cost during inference.

`Gemma` (Google): Developed by Google and derived from the same research and technology used to create the `Gemini` models, the `Gemma` family provides another high-quality, open-source option for researchers and developers (Team et al., 2024).

`Qwen` (Alibaba): The `Qwen` series of models has demonstrated particularly strong performance on a wide range of benchmarks, with a notable strength in multilingual capabilities and instruction-following across diverse languages and domains (Qwen et al., 2025).

Frontier Models and Advanced Reasoning: At the cutting edge are `proprietary models` explicitly architected for complex, multi-step reasoning, moving beyond standard instruction-following.

OpenAI Models: `OpenAI`'s models, such as the **`GPT`** series (OpenAI, 2023), have consistently pushed the boundaries of LLM capabilities. Recent advancements have focused on enhancing reasoning. As suggested by the papers in this survey, advanced reasoning-enhanced variants, conceptually referred to as models like **`o3`**, are designed to handle complex, structured prompts and perform systematic analysis, setting the benchmark for tasks requiring deep inference.

Anthropic Models: `Claude` family, particularly models like **`Claude 4 Opus`**, has been developed with a strong emphasis on reliability and sophisticated reasoning. As seen in the surveyed research, these models can execute a "`thinking`" process, which is an explicit implementation of **Chain-of-Thought (CoT)** reasoning (Wei et al., 2022d). In this process, the model is prompted to generate a sequence of intermediate, logical steps before arriving at a final answer, significantly improving its performance on tasks that require complex deliberation.

Parameter-Efficient Fine-Tuning (PEFT): While instruction-tuning and RLHF create powerful general-purpose models, adapting them to specialized tasks or domains often requires further fine-tuning. However, full fine-tuning of a multi-billion parameter LLM is computationally prohibitive, requiring immense memory and yielding a separate, full-sized model for each task. To overcome this, the field has widely adopted (**PEFT**).

The core principle of PEFT is to freeze the vast majority of the pre-trained model's weights and train only a small number of new or adapted parameters. A leading PEFT method is **Low-Rank Adaptation (LoRA)** (Hu et al., 2021). LoRA posits that the change in a weight matrix during adaptation, ΔW, has a low "intrinsic rank." It therefore approximates this change by decomposing it into two much smaller, low-rank matrices, B and A:

$$W_0 + \Delta W \approx W_0 + BA \tag{3}$$

where $W_0 \in \mathbb{R}^{d\times k}$ are the frozen pre-trained weights, while $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$ are trainable low-rank matrices, with the rank $r \ll \min(d, k)$. By training only A and B, the number of trainable parameters is drastically reduced, making it feasible to fine-tune massive models on consumer-grade hardware. Frameworks like **`Unsloth`** (Daniel Han and team, 2023) have further optimized these techniques, enabling even faster and more memory-efficient fine-tuning. This has been instrumental in creating the specialized, high-performing open-source models evaluated throughout this survey.

Model Quantization for Efficient Inference: Alongside PEFT, which addresses the memory demands of *training*, model quantization tackles the computational and memory costs of *inference*. The core challenge is that LLMs are typically trained using 32-bit floating-point precision (FP32), resulting in massive memory footprints (e.g., a 7B parameter model requires 28GB of VRAM). Quantization reduces this burden by converting the model's weights from high-precision data types to low-precision ones, such as 8-bit or 4-bit integers (INT8/INT4).

The fundamental principle is an affine transformation that maps a high-precision floating-point weight tensor W to a lower-precision integer tensor W_q. This is achieved using a scaling factor S and a zero-point Z:

$$W_q = \text{round}\left(\frac{W}{S} + Z\right) \tag{4}$$

During inference, the weights are de-quantized back to an approximation of the original floating-point values: $\tilde{W} \approx S \cdot (W_q - Z)$. This process significantly reduces the model's size and can accelerate computation on hardware with native support for low-precision arithmetic, albeit with a potential trade-off in model performance. For example, a weight value of 0.5 in FP32 might be mapped to the integer 192 in an INT8 representation that spans the range $[-1.0, 1.0]$.

Early post-training quantization (PTQ) methods focused on simple rounding, but modern techniques are far more sophisticated to preserve model fidelity. Methods like GPTQ (Frantar et al., 2022) and AWQ (Lin et al., 2023) use calibration data to identify and preserve salient weights, minimizing the performance degradation. The introduction of **QLoRA** (Dettmers et al., 2023) was a major breakthrough, enabling 4-bit fine-tuning by introducing a new data type, 4-bit NormalFloat (NF4), which is information-theoretically optimal for normally distributed weights. QLoRA also employs Double Quantization, which *quantizes* the quantization constants themselves for further memory savings.

This progress has been operationalized through community-driven formats, most notably **GGUF** (GPT-Generated Unified Format). Evolving from the earlier GGML format, GGUF is a binary file format designed specifically for storing and rapidly loading quantized models for inference, particularly on CPUs via frameworks like `llama.cpp`. By packaging the model's architecture, metadata, and quantized weights into a single portable file, GGUF has been instrumental in making state-of-the-art LLMs accessible on consumer-grade hardware. Together, quantization and PEFT form a powerful toolkit for the development and deployment of large-scale language models.

Inference Optimization Frameworks: While quantization reduces the static memory footprint of an LLM, another critical challenge is maximizing inference throughput and efficiently managing memory during runtime, especially for dynamic batching of requests with variable lengths. To address this, specialized serving frameworks have become essential. **vLLM** (Kwon et al., 2023) is a high-throughput serving engine that introduced **PagedAttention**, a novel algorithm inspired by virtual memory and paging in operating systems. Instead of pre-allocating a contiguous memory block for the Key-Value (KV) cache of a sequence, PagedAttention partitions the KV cache into blocks that can be stored non-contiguously, mitigating internal fragmentation and enabling near-optimal memory usage. This allows for significantly higher batch sizes and boosts GPU utilization, leading to dramatic improvements in serving throughput. Many such high-performance systems are built on distributed computing frameworks like **RAY** (Moritz et al., 2018), which provides a simple, universal API for building and scaling distributed applications, handling complex tasks like parallel processing and distributed memory management.

Beyond optimizing the inference engine itself, a higher-level ecosystem of orchestration frameworks has emerged to simplify the development of complex, multi-step applications. **LANGCHAIN** (Chase, 2022) provides a comprehensive toolkit for "chaining" LLM calls with other components, such as external APIs, databases, and memory modules. It abstracts common patterns for building agents, retrieval-augmented generation (RAG) pipelines, and other composite AI systems. More recently, **LANGGRAPH** (LangChain, 2023) has extended this paradigm by representing application logic as a cyclic graph instead of a simple Directed Acyclic Graph (DAG). This allows developers to build more sophisticated and robust agents that can loop, self-correct, and manage complex state over multiple steps, more closely mimicking human-like deliberation and planning. These frameworks act as crucial middleware, bridging the gap between a raw LLM and a deployable, production-grade application.

Model Ecosystem: For *brevity*, this survey has focused on a limited number of prominent open-source model families. However, the field is characterized by a vibrant and rapidly expanding ecosystem. Platforms like **Hugging Face**[1] (Wolf et al., 2020) serve as a central hub, hosting tens of thousands of pre-trained models, datasets, and tools, fostering collaborative development and reproducibility. To simplify programmatic access across this landscape, services like **OpenRouter**[2] have emerged. These platforms act as inference aggregators, providing a unified API endpoint that allows developers to interact with dozens of different models—from proprietary ones like `GPT-4o` and `Claude 4 Opus` to a wide array of open-source variants—through a single, standardized interface. This abstraction layer facilitates rapid experimentation and helps manage costs by routing requests to the most suitable model.

To navigate the performance of this vast collection of models, community-driven leaderboards have become essential. The **LMSys Chatbot Arena**[3] (Zheng et al., 2023), for instance, provides a continuously updated ranking of models based on crowdsourced, anonymous, side-by-side human preference comparisons, using an `Elo` rating system to quantify performance. This dynamic ecosystem ensures that the state-of-the-art is constantly being challenged and that increasingly powerful models are becoming accessible to the entire research community.

[1] https://huggingface.co/models

[2] https://openrouter.ai

3 Prompting Techniques

Prompting serves as the primary interface for interacting with LARGE LANGUAGE MODELS (LLMS), allowing users to elicit desired behaviors through textual instructions. This section provides an overview of prompting techniques, from basic to advanced approaches, highlighting their evolution and impact on model performance.

Zero-shot Prompting: `Zero-shot prompting` involves directly querying an LLM with an instruction without providing any examples in the prompt. This approach relies solely on the model's vast pre-trained knowledge to understand and execute the task (Brown et al., 2020b). For instance, in e-commerce applications, a simple instruction like "`Generate a comprehensive summary of the following product reviews highlighting key features and customer sentiments`" exemplifies zero-shot prompting for opinion summarization. Similarly, "`Identify the emotions expressed in this product review using Plutchik's emotion categories`" demonstrates zero-shot emotion detection. While straightforward, this approach can yield suboptimal performance on complex or specialized tasks that fall outside the model's training distribution. Nevertheless, modern LLMs demonstrate remarkable zero-shot capabilities across diverse domains, a phenomenon often described as an "emergent ability" that appears as model scale increases (Wei et al., 2022b).

Few-shot Prompting: `Few-shot prompting` enhances model performance by including demonstration examples within the prompt, a technique known as in-context learning (Brown et al., 2020b). By providing several input-output pairs before the target query, the model can better understand the task's pattern and expected output format. For example, in e-commerce emotion detection:

[3] https://lmarena.ai

```
Identify emotions and their triggers in product reviews:
Input: “This wireless headphone has amazing sound quality but the battery dies quickly.” Output: Joy (amazing sound quality), Disappointment (battery dies quickly)
Input: “The delivery was delayed and the packaging was damaged.” Output: Anger (delivery was delayed), Disgust (packaging was damaged)
Input: “I’m so excited to try this new skincare routine!” Output: ?
```

Few-shot prompting offers several advantages: it requires no model parameter updates, provides explicit task guidance, and can significantly boost performance with only a handful of examples. However, its effectiveness is highly dependent on factors like example selection, formatting, and ordering, as the model's performance is sensitive to the distribution and structure of the demonstrations (Min et al., 2022).

Chain-of-Thought (CoT) Prompting: CoT encourages the model to generate intermediate reasoning steps before producing a final answer (Wei et al., 2022d; Kojima et al., 2023). By decomposing complex problems into manageable sub-steps, CoT prompting dramatically improves performance on tasks requiring *multi-step reasoning*, such as product comparison, query-focused summarization, and recommendation justification. For example, in product comparison:

```
Compare these smartphones for a photography enthusiast. Think step by step:
Step 1: Analyze camera specifications (megapixels, aperture, lens quality)
Step 2: Review customer feedback on photo quality
Step 3: Consider additional photography features (night mode, portrait mode)
Step 4: Evaluate price-to-performance ratio
Final recommendation: Based on superior camera hardware and positive photography reviews...
```

The prompt typically includes the instruction to “think step by step” or provides examples with explicit reasoning chains.

Research has shown that CoT prompting is particularly effective for larger language models (typically >100B parameters), demonstrating that reasoning capabilities emerge at scale (Wei et al., 2022b). Variations such as **Zero-shot CoT** (Kojima et al., 2023) use simple prompts like “Let's think step by step” to elicit reasoning without examples, while **Few-shot CoT** (Wei et al., 2022d) provides demonstration examples with reasoning steps.

Self-Consistency: Self-consistency (Wang et al., 2023e) extends CoT prompting by generating multiple reasoning paths and selecting the most consistent answer through majority voting. This approach mitigates reasoning errors by aggregating results across different solution attempts, leading to more reliable outputs. For example, when generating product recommendations, an LLM might generate several reasoning paths:

```
Path 1: User wants durability → Check build quality reviews → Recommend Product A
Path 2: Budget constraints → Compare price points → Recommend Product A
Path 3: Feature requirements → Match specifications → Recommend Product A
Final decision: Product A (consistent across all reasoning paths)
```

This approach is particularly valuable in e-commerce applications where recommendation confidence and explanation consistency are crucial for user trust.

Tree of Thoughts (ToT): Tree of Thoughts (Yao et al., 2023a) expands on CoT by exploring multiple reasoning branches simultaneously, enabling more systematic problem-solving. ToT implements a search algorithm (breadth-first or depth-first) over intermediate reasoning steps, evaluating promising paths while pruning unpromising ones. This approach allows for backtracking and exploration of alternative solution strategies, particularly valuable for tasks like game playing, planning, and complex problem-solving where considering multiple possibilities is *beneficial*.

Least-to-Most Prompting: Least-to-Most prompting (Zhou et al., 2023a) breaks down complex problems into simpler subproblems that build upon each other. The approach first solves easier components and progressively leverages these solutions to tackle more challenging aspects. This technique has shown particular effectiveness for compositional reasoning and programming tasks where incremental progress facilitates solving the overall problem.

ReAct Prompting: ReAct (Reasoning and Acting) (Yao et al., 2023b) interleaves reasoning steps with actions in environments where interaction is necessary. The framework combines natural language reasoning with the ability to take actions (such as searching for information, using tools, or executing operations) and then observing outcomes to inform subsequent reasoning. In e-commerce applications, ReAct enables dynamic product research:

```
Thought: User needs a laptop for
gaming. I should check current
gaming laptop reviews.
Action: Search["best gaming
laptops 2024 reviews"]
Observation: Found reviews
mentioning RTX 4080, high
refresh rate displays...
Thought: Now I should compare
specific models mentioned in
reviews.
Action: Compare["ASUS ROG vs MSI
Gaming laptop specs"]
Observation: ASUS has better
cooling, MSI has superior
display...
```

This approach has proven effective for tasks requiring dynamic interaction with product databases, review aggregation, and real-time price comparison.

Self-Verification: Self-verification techniques (Weng et al., 2023) prompt LLMs to critically evaluate their own outputs for correctness, consistency, and comprehensiveness. By explicitly asking models to check their reasoning, identify potential errors, and verify factual claims, self-verification improves output reliability. Common implementations include multi-stage prompting where an initial solution is followed by a verification phase that checks for errors before producing a final, refined answer.

Self-Refinement: Self-refinement (Madaan et al., 2023) extends self-verification by enabling models to iteratively improve their outputs based on self-identified issues. The model generates an initial response, critically evaluates it, and then produces an improved version. This process can iterate multiple times, with each cycle addressing previously identified shortcomings. Self-refinement has shown particular promise for tasks requiring high quality and precision, such as code generation, essay writing, and complex reasoning.

Meta-Prompting: Meta-prompting (Suzgun and Kalai, 2024) involves guiding LLMs to generate their own prompts or refine existing ones. By leveraging the model's capabilities to design effective instructions, meta-prompting can optimize task performance without human intervention. This approach often includes generating multiple candidate prompts, evaluating their quality, and selecting the most effective version for the target task.

Automatic Prompt Engineering: Automatic Prompt Engineer (APE) (Zhou et al., 2023b) systematically optimizes prompts using search algorithms or learning-based approaches. These methods explore the prompt space to identify instructions that maximize performance on specific tasks, often surpassing human-designed prompts. APE techniques include gradient-based optimization, evolutionary algorithms, and reinforcement learning from model outputs.

Constitutional AI: Constitutional AI (Bai et al., 2022) enhances LLM behavior through a two-stage process: *supervised learning* from human feedback and *reinforcement learning* from AI feedback. The approach uses a set of principles (a "constitution") to guide model responses, enabling models to critique and revise their own outputs according to specified ethical and behavioral guidelines. This technique has proven particularly effective for reducing harmful outputs while maintaining helpfulness, achieving up to 25% improvement in safety metrics compared to standard **RLHF** approaches (Ouyang et al., 2022b).

Tree of Thoughts Variations: Building upon the foundational ToT framework (Yao et al., 2023a), several advanced variations have emerged. **Graph of Thoughts** (GoT) (Besta et al., 2024) extends tree-based reasoning to arbitrary graph structures, enabling more complex reasoning patterns and information aggregation. **Algorithm of Thoughts** (AoT) (Sel et al., 2024) incorporates algorithmic examples to guide the search process, achieving 10%-15% performance improvements on complex reasoning tasks. **Skeleton-of-Thought** (SoT) (Ning et al., 2024) first generates a reasoning skeleton before filling in details, reducing inference time by up to 40% while maintaining quality. These variations demonstrate the continued evolution of structured reasoning approaches, with each addressing specific limitations of the original ToT framework through enhanced search strategies, computational efficiency, or reasoning flexibility.

4 Problem Formulation

This section presents formal problem definitions for five key research directions in e-commerce NLP. Each task addresses distinct challenges in product recommendation and review analysis, requiring specialized natural language understanding and generation capabilities.

4.1 Query-Focused Comparative Explainable Summarization (QF-CES)

Users often struggle with decision paralysis when comparing multiple recommended products without consolidated, query-specific insights. The QF-CES task addresses this challenge by generating comparative summaries that directly respond to user information needs.

Task Definition: Given a user query $q_i \in \mathcal{Q}$ and the top-k recommended products $\mathcal{P}_i = \{p_{ij}\}_{j=1}^{k}$ where $k = 3$, the goal is to generate a comparative summary through the mapping:

$$\mathcal{H} : \mathcal{Q} \times \mathcal{P}^k \rightarrow \mathcal{C} \times \mathcal{V}$$

where $\mathcal{H}(q_i, \mathcal{P}_i) = (c_i, v_i)$ produces a structured comparison table $c_i \in \mathcal{C}$ and a natural language verdict $v_i \in \mathcal{V}$.

Input: A natural language query q (e.g., *"best wireless headphones for running under ₹12,000"*) and three recommended products $\mathcal{P}_i = \{p_1, p_2, p_3\}$, where each product p_j contains metadata including title, description, specifications, customer reviews, ratings, and pricing information.

Output: The system generates: (1) a structured comparison table c_i highlighting key attributes relevant to the query, and (2) a natural language explanation v_i providing a final recommendation verdict.

Example: For the query *"best wireless headphones for running under ₹12,000"*, the comparison table focuses on key attributes such as *battery life*, *comfort and fit*, *durability*, *connectivity*, and *overall value*. The verdict offers a nuanced recommendation based on performance, price, and user needs.

The comparison table c_i is structured as follows:

Attribute	**Prod A**	**Prod B**	**Prod C**
Base Price	₹10,499	₹7,999	₹11,499
Final Price	₹4,299	₹5,499	₹8,999
Battery Life	40h	22h	30h
Comfort	Secure, sweat-proof	Lightweight, loose fit	Snug, over-ear
Durability	Rugged, IPX5	Moderate, IPX4	High, IPX5
Bluetooth Version	5.3	5.0	5.2
Rating	4.6/5	4.1/5	4.7/5
Pros	Best battery, durable, secure fit	Budget option, decent sound	Balanced performance, high-quality audio
Cons	Slightly bulky	Shorter battery, lower rating	Higher cost

Table 1: QF-CES comparison table: Wireless headphones for running under ₹12,000

Final Verdict Summary: **Product A** emerges as the top choice for runners seeking endurance and reliability — it offers an unmatched 40-hour battery life, rugged water resistance (IPX5), and secure fit at a discounted price of ₹4,299. If **comfort and affordability** are bigger priorities, **Product B** provides a lightweight option under ₹5,500, though with trade-offs in battery and rating. Meanwhile, **Product C** is ideal for **audio enthusiasts** who want premium sound and build quality, but it comes at a higher price.

Evaluation Framework: We evaluate QF-CES outputs using a five-dimensional assessment framework. Given the quadruple $(c_i, v_i, q_i, \mathcal{P}_i)$, the evaluation function:

$$\mathcal{E}_{\text{QF-CES}} : \mathcal{C} \times \mathcal{V} \times \mathcal{Q} \times \mathcal{P}^k \to \mathcal{L}_5^5$$

returns quality scores across five dimensions: *clarity* (readability and organization), *faithfulness* (accuracy to source data), *informativeness* (coverage of relevant details), *format adherence* (structural compliance), and *query relevance* (alignment with user intent).

4.2 Emotion-Aware Opinion Summarization (EAOS)

Traditional opinion summarization often reduces complex customer emotions to simple positive/negative polarities, losing nuanced affective information. EAOS addresses this limitation by generating summaries that capture the full spectrum of customer emotions while maintaining factual accuracy.

Task Definition: Given a product $p \in \mathcal{P}$ and a collection of customer reviews $R = \{r_i\}_{i=1}^m$ where $m = 10$, the objective is to generate an emotion-aware summary through:

$$\mathcal{G} : \mathcal{P} \times \mathcal{R} \to \mathcal{S} \times \mathcal{E}^8 \tag{5}$$

This mapping produces both a textual summary $s \in \mathcal{S}$ and emotion annotations $\boldsymbol{e} \in \mathcal{E}^8$ based on Plutchik's emotion wheel: {*joy, trust, fear, surprise, sadness, disgust, anger, anticipation*}.

Input: A product title p (e.g., *"Samsung Galaxy Bluetooth Speaker"*) and 10 customer reviews $R = \{r_1, r_2, ..., r_{10}\}$, where each review r_i contains a title and review text with 10-100 tokens.

Output: An emotion-aware summary s (125 words) that integrates factual product aspects with emotional customer responses, along with emotion intensity mappings across eight primary emotions.

Example: For a Bluetooth speaker priced at ₹4,999 with mixed reviews, the EAOS output might be: *"Customers express strong joy and trust regarding the speaker's exceptional bass quality and reliable wireless connectivity up to 10 meters. However, several users report anger and frustration about the battery lasting only 4-5 hours instead of the advertised 12 hours at ₹4,999 price point. The compact design generates anticipation for outdoor activities, though some express fear about the speaker's durability after reports of volume button malfunctions within 6 months of purchase."*

The system employs a `four-step` reasoning process: (1) aspect-emotion mapping to identify which product features trigger specific emotions, (2) emotional balance assessment to ensure fair representation, (3) narrative integration to create coherent text, and (4) refinement and validation for quality assurance.

Evaluation Framework: The evaluation function $\mathcal{E} : \mathcal{S} \times \mathcal{P} \times \mathcal{R} \to \mathcal{L}_5^7$ assesses summaries across seven dimensions: *fluency*, *coherence*, *faithfulness*, *emotional accuracy*, *emotional spectrum coverage*, *emotional bias mitigation*, and *contextual emotional relevance*.

4.3 Emotion-Opinion Trigger Detection (EOT)

Understanding not just *what* emotions customers express, but *why* they feel that way, is crucial for actionable business insights. EOT addresses this challenge by jointly detecting emotions and identifying the specific textual spans that trigger those emotions.

Task Definition: Given a customer review $R = \{R_i\}_{i=1}^N$ as a sequence of N tokens, the objective is to identify emotion-trigger pairs through:

$$\mathcal{M} : \mathcal{R} \to 2^{\mathcal{E}_P \times \mathcal{T}_e} \tag{6}$$

where $\mathcal{E}_P = \mathcal{P} \cup \{\text{Neutral}\}$ represents the emotion space based on Plutchik's eight primary emotions:

$$\mathcal{P} = \{\text{Joy, Sadness, Anger, Fear,} \\ \text{Trust, Disgust, Surprise, Anticipation}\}$$

and $\mathcal{T}_e$ contains extractive opinion triggers explaining each detected emotion.

Input: A single customer review R with $10 - 100$ tokens.

Output: Emotion-trigger mappings $O(R) = \{(e, T_e)\}$ where each emotion e is paired with its triggering text spans T_e.

Example: Consider the review: *"I love the sleek aluminum design and 6GB RAM performance at ₹18,999, but I'm disappointed by the poor customer service response time when I reported screen flickering issues."*

The EOT system would output:

- (*joy*, *"love the sleek aluminum design and 6GB RAM performance at ₹18,999"*)
- (*sadness*, *"disappointed by the poor customer service response time"*)
- (*anger*, *"screen flickering issues"*)

This joint modeling approach enables *interpretable emotion analysis* by establishing explicit causal relationships between customer feelings and their textual manifestations, moving beyond simple emotion classification to explanatory emotion understanding.

4.4 Multi-Source Opinion Summarization (M-OS)

E-commerce product information is typically fragmented across multiple sources (`descriptions`, `specifications`, `reviews`, `ratings`), creating cognitive overload for users. M-OS addresses this challenge by synthesizing heterogeneous information sources into comprehensive, unified summaries.

Task Definition: Given a product $p \in \mathcal{P}$ with complete metadata tuple $p = (\tau, \mathfrak{d}, \mathcal{K}, \mathcal{S}, \rho_a, \mathcal{R})$, where τ is the title, $\mathfrak{d}$ is the description, $\mathcal{K}$ represents key features, $\mathcal{S}$ contains specifications, ρ_a is the average rating, and $\mathcal{R}$ is the review collection, the objective is to generate a comprehensive summary through:

$$\mathcal{F}_{\text{M-OS}} : \mathcal{P} \to \mathcal{S}$$

Input: Complete product metadata including: product title, manufacturer description, key feature list, technical specifications, average rating score, and customer review collection.

Output: A unified summary s that coherently integrates objective product information with subjective user experiences, providing holistic product understanding.

Example: For a smartphone priced at ₹32,999, M-OS might integrate:

- **Technical specs:** *"6.7-inch AMOLED display, 128GB storage, 50MP triple camera"*
- **Marketing description:** *"Premium flagship with AI-powered photography"*
- **User reviews:** *"Excellent camera performance but heating issues during gaming"*
- **Ratings:** *"4.3/5 stars from 2,847 reviews"*

Into a comprehensive summary: *"This ₹32,999 smartphone features a* 6.7*-inch* AMOLED *display and* 128*GB storage with premium flagship positioning. Users consistently praise the* 50*MP triple camera's AI-powered photography capabilities, validating marketing claims about advanced imaging. However, multiple reviewers report heating issues during intensive gaming sessions, suggesting thermal management limitations despite the overall positive 4.3/5 star rating from nearly* $3,000$ *customers. The device offers solid premium features but potential buyers should consider usage patterns."*

Evaluation Framework: The evaluation function $\mathcal{E}_{\text{M-OS}} : \mathcal{S} \times \mathcal{P} \to \mathcal{L}_5^7$ assesses summaries across seven quality dimensions: *fluency*, *coherence*, *relevance*, *faithfulness*, *aspect coverage*, *sentiment consistency*, and *specificity*.

4.5 Query-Focused Explainable Recommendation (QF-ER)

Recommendation systems often suffer from algorithmic opacity, providing products without explaining why they match user needs. QF-ER addresses this transparency gap by generating natural language explanations that justify recommendation relevance to specific user queries.

Task Definition: Given a user query $q \in \mathcal{Q}$ and a recommended product $p \in \mathcal{P}$, the objective is to generate an explanation through:

$$\mathcal{G} : \mathcal{Q} \times \mathcal{P} \to \mathcal{T} \times \Gamma \times \Delta$$

This produces an explanation text $e \in \mathcal{T}$, confidence level $\gamma \in \Gamma = \{\text{Low, Medium, High}\}$, and relevance assessment $\delta \in \Delta = \{\text{Yes, No}\}$.

Input: A user query q (e.g., *"laptop for video editing under ₹80,000"*) and a recommended product p with comprehensive metadata including title, description, features, specifications, ratings, reviews, and pricing information.

Output: A natural language explanation e (75-100 words) with confidence score γ and binary relevance assessment δ, justifying why the product matches the query requirements.

Example: For the query *"laptop for video editing under ₹80,000"* and a recommended laptop priced at ₹74,999:

Explanation: *"This laptop excellently matches your video editing requirements with its Intel i7-12700H processor and 16GB DDR4 RAM, providing sufficient computational power for 4K video processing. The dedicated NVIDIA RTX 3060 graphics card with 6GB VRAM accelerates video rendering in Adobe Premiere Pro and DaVinci Resolve. The 1TB NVMe SSD ensures fast project file access. At ₹74,999, it fits comfortably within your ₹80,000 budget. Customer reviews consistently highlight 90% satisfaction for creative applications, with average 4.4/5 stars for video editing performance."*

Recommendation Confidence: `High`
Correct Recommendation: `Yes`

This framework provides transparent recommendations by justifying the product-query fit. The **Recommendation Confidence: High** score indicates a strong alignment between the laptop's technical specifications (Intel i7, RTX 3060, 16GB RAM) and the explicit needs of video editing. The **Correct Recommendation: Yes** verdict offers a clear, binary confirmation that the product is a suitable choice, building user trust through explainable alignment.

5 Related Work

Our work intersects several key research areas each addressing critical gaps in current e-commerce decision-support technologies.

5.1 Query-Focused Comparative Explainable Summarization (QF-CES)

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

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

LLM-based EVALUATORS as traditional metrics like ROUGE (Lin, 2004a) and BLEU (Papineni et al., 2002a) often misalign with human judgments for opinion summaries. Recent NLP advancements, particularly in LLMs, offer promising alternatives. Studies have explored LLM-based evaluation methods (Fu et al., 2023a; Chiang and Lee, 2023a; Wang et al., 2023a; Kocmi and Federmann, 2023), including CHAIN-OF-THOUGHT approaches (Liu et al., 2023b; Wei et al., 2022c) and reference-free evaluation (Chiang and Lee, 2023c). proposed two prompt strategies for opinion summary evaluation on 7 metrics.

QF-CES differs from existing work through: (**1**) Consolidated Comparison of three products simultaneously; (**2**) Query-Based Personalization, preserving privacy; (**3**) Dynamic Attribute Generation tailored to user queries; (**4**) Category-Agnostic approach applicable across product domains; (**5**) Recommendation-Engine Agnostic, functioning with any ranking system; and (**6**) Multi-Source Integration, generating comprehensive summaries beyond user reviews. These features collectively offer a more versatile, privacy-conscious, and informative comparative summarization solution.

5.2 Emotion-Aware Opinion Summarization (EAOS)

The development of EMOTION-AWARE OPINION SUMMARIZATION addresses a long-standing and critical limitation in traditional opinion analysis: the *affective blind spot*. For decades, research in opinion summarization has evolved significantly, yet has largely failed to capture the rich emotional dimensions that fundamentally shape consumer perception and purchasing decisions (Chen et al., 2022; Felbermayr and Nanopoulos, 2016). This has resulted in the perpetuation of shallow summaries that, while factually grounded, lack the emotional depth required for genuine user understanding.

The trajectory of `opinion summarization` began with extractive methods, which focused on identifying and concatenating salient sentences from source reviews (Erkan and Radev, 2004; Kim et al., 2011). The field later transitioned to more sophisticated neural and abstractive approaches (Bražinskas et al., 2020; Amplayo and Lapata, 2020), which enabled the generation of novel, more fluent summaries. Research further specialized into aspect-specific (Amplayo et al., 2021) and multi-source summarization (Li and Lam, 2020). Despite these advances, the core focus remained on distilling rudimentary sentiment polarity (i.e., positive, negative, neutral). Even recent, large-scale summarization efforts have primarily centered on sentiment, thereby overlooking the crucial nuances of discrete emotions like *joy*, *trust*, or *disappointment* that are potent determinants of consumer behavior (Bhaskar et al., 2023; Hosking et al., 2023; Pappas and Androutsopoulos, 2014).

Concurrently, but largely in isolation, the field of Emotion Analysis in NLP has matured significantly. Grounded in foundational psychological frameworks such as Plutchik's wheel of emotions and Ekman's basic emotions (Plutchik, 1988, 2000; Ekman, 1992), researchers have developed robust models for both emotion classification (Mohammad and Bravo-Marquez, 2017; Felbo et al., 2017) and emotion extraction (Ding et al., 2020; Ying et al., 2019; Li et al., 2023b). These efforts have successfully equipped machines to identify and categorize a wide spectrum of human emotions expressed in text. However, this line of research has predominantly focused on *analysis* and *extraction*, rather than the generative task of synthesizing these emotional insights into a coherent, human-readable summary.

This separation of disciplines created what can be termed the *unaddressed frontier*: the generative task of synthesizing affectively nuanced summaries remained almost entirely unexplored. The advent of modern Large Language Models (LLMs) has been the primary catalyst enabling this new research direction. With their emergent capabilities in affective reasoning (Tse-Hsun et al., 2024) and `abstractive` compression (Deroy et al., 2023), LLMs provide the first technically viable tools to bridge the gap between cognitive opinion and affective experience.

The ***novelty*** of EAOS also introduces new challenges, particularly in evaluation. It is well-documented that traditional metrics like ROUGE (Lin, 2004b), BLEU (Papineni et al., 2002a), and BERTSCORE (Zhang et al., 2020b) often correlate poorly with human judgments for nuanced summarization tasks (Shen and Wan, 2023). This inadequacy is magnified when assessing the fidelity of emotional representation. Consequently, a parallel line of research has emerged on leveraging LLMs themselves as scalable and effective evaluators (Fu et al., 2023b; Chiang and Lee, 2023a,b; Wang et al., 2023a). Methodologies such as Chain-of-Thought (CoT) prompting (Liu et al., 2023b; Wei et al., 2022c) and reference-free evaluation (Chiang and Lee, 2023c) are being developed to create more reliable and human-aligned assessment protocols. The EAOS framework is the *first* to systematically address this long-ignored gap, introducing a comprehensive, theoretically-grounded methodology for both the generation and multi-dimensional evaluation of summaries that truly reflect the customer's emotional journey.

5.3 Emotion and Opinion Trigger Detection (EOT)

While identifying the emotion expressed in a review is valuable, understanding *why* that emotion was elicited is crucial for generating truly interpretable and actionable insights. This has given rise to the task of EMOTION AND OPINION TRIGGER DETECTION (EOT), which involves the joint identification of an emotion and the specific text span (*opinion trigger*) that caused it. This task directly addresses the fundamental question of causality in user feedback, a dimension that has remained largely under-explored in e-commerce contexts.

The broader field of Emotion Analysis has long been central to NLP, demonstrating how affective signals shape online discourse and influence consumer decisions (Mohammad and Turney, 2013). Initial research focused on classifying text into coarse sentiment categories (positive, neutral, negative). Recognizing the limitations of this approach, researchers soon adopted more nuanced emotion taxonomies, such as those proposed by (Russell, 1980), (Ekman, 1992), and (Plutchik, 2001), to capture the complexity of human emotional expression.

The more specific subfield of Emotion-Trigger Analysis, or Emotion-Cause Extraction (ECE), has evolved through several methodological phases. Early approaches utilized rule-based systems (Neviarouskaya et al., 2009; Lee et al., 2010) and statistical methods (Gui et al., 2016; Xia and Ding, 2019) to identify the causes of emotions in text. More recent studies have employed sophisticated deep learning techniques, including graph-based models and attention mechanisms, to perform joint emotion-cause extraction (Wei et al., 2020; Fan et al., 2021; Singh et al., 2021), as well as context-aware models for more accurate trigger identification (Li et al., 2019).

However, this body of work has two critical limitations concerning e-commerce. First, *prior research has almost exclusively focused on genres like news articles and social media*, leaving the domain of product reviews virtually unexplored. The unique linguistic style and structure of reviews present distinct challenges not found in other text types. A recent study by (Singh et al., 2024) on the social media dataset EMOTRIGGER highlighted the limitations of modern LLMs in trigger identification, reinforcing the fact that this remains an unsolved problem, especially in new domains. To date, **emotion-trigger analysis remains an unexplored research area in e-commerce**.

Second, progress in this field has been heavily reliant on dataset development. Key resources like SemEval (Strapparava and Mihalcea, 2007), GoEmotions (Demszky et al., 2020), and domain-specific benchmarks like CancerEmo (Sosea and Caragea, 2020) have propelled emotion analysis forward. Yet, as of now, **no existing dataset provides annotations for both fine-grained emotions and their corresponding opinion triggers specifically for e-commerce platforms**. This lack of a foundational benchmark has been a major barrier to research.

The recent advancements in Large Language Models (LLMs), with their powerful capabilities in contextual understanding and generating emotionally nuanced text (Brown et al., 2020a; Ouyang et al., 2022a), offer a promising avenue to address this gap. While their potential for general emotion analysis is being actively investigated (Acheampong et al., 2023; Huang and Rust, 2024), the joint task of EOT in e-commerce represents a novel application that this survey identifies as a key research direction.

5.4 Multi-Source Opinion Summarization (M-OS)

MULTI-SOURCE OPINION SUMMARIZATION (M-OS) represents a critical evolution beyond traditional summarization techniques, which have historically focused on a single source of information: customer reviews (Wang and Ling, 2016; Chu and Liu, 2019). While valuable, review-only summaries provide a purely subjective and often incomplete perspective. M-OS addresses this by creating holistic summaries that integrate objective product attributes with subjective user opinions, thereby facilitating more informed and confident consumer decision-making.

The *progression of opinion summarization* has seen a steady increase in the diversity of information sources. Early methods relied on extractive (Erkan and Radev, 2004) or abstractive (Brazinskas, 2020) techniques applied solely to review texts. The first step towards a multi-source paradigm involved incorporating easily accessible textual data. For instance, (Zhao et al., 2020) enhanced summaries by utilizing product descriptions in addition to reviews. This was followed by research into multimodality, where supervised methods were developed to combine textual information with visual data like product images (Li et al., 2020b).

The emergence of Large Language Models (LLMs) enabled more sophisticated multi-source integration. Recent work by (Siledar et al., 2024) introduced a structured approach (MEDOS) that fused information from three distinct sources: product reviews, descriptions, and question-and-answer (Q&A) pairs. These advancements significantly improved the factual grounding and comprehensiveness of generated summaries.

Despite this progress, a significant gap has persisted in the literature: **the comprehensive integration of all available product metadata, especially detailed technical specifications, remains largely unexplored**. While prior work has incorporated high-level descriptions or key features, the dense, structured information contained within product specification tables is often overlooked. This oversight is primarily due to the technical challenge of processing and coherently synthesizing such diverse and lengthy data types.

Modern LLMs, with their vastly expanded context windows and superior reasoning capabilities, are uniquely positioned to close this gap. They enable the development of M-OS systems that can process the *entire* product context—including titles, descriptions, features, ratings, reviews, and detailed specifications—to dynamically generate a single, unified summary. As demonstrated by recent work, this approach reduces the cognitive load on users by eliminating the need to manually parse and cross-reference multiple information sources.

This increased complexity of the M-OS task also exposes the limitations of traditional evaluation metrics like ROUGE (Lin, 2004b) and BERTSCORE (Zhang et al., 2020b), which are known to correlate poorly with human judgments for such multifaceted outputs (Shen and Wan, 2023). Consequently, advancing M-OS is intrinsically linked to developing robust, reference-free evaluation paradigms that leverage LLMs as scalable and nuanced critics (Fu et al., 2023b; Chiang and Lee, 2023c).

5.5 Query-Focussed Explainable Recommendation (QF-ER)

Recommender systems traditionally focus on predicting what users will like, but explainable recommendation addresses the ***why*** behind these predictions to enhance user trust and satisfaction. The initial explorations were driven by understanding that mere accuracy was insufficient for a positive user experience. (Herlocker et al., 2000) conducted the first user study examining how explanations affect user acceptance, identifying the benefits of explanations for building user trust and established a framework for evaluating explanation effectiveness, while (Papadimitriou et al., 2012) proposed a taxonomy of explanation styles (user-based, item-based, feature-based), providing vocabulary influencing subsequent frameworks like (Zhang and Chen, 2020b) "5W" categorization.

A significant shift occurred with the increasing availability of *user-generated reviews*, which provided rich textual content for generating explanations. Feature-based explanations emerged with (Zhang et al., 2014) Explicit Factor Model (EFM), leveraging phrase-level sentiment analysis on user reviews to generate feature-based explanations. Building on this, (He et al., 2015) developed TriRank, constructing a heterogeneous tripartite graph of User-Item-Aspect relationships weighted by review sentiment, while (Chen et al., 2018b) advanced this with NARRE, combining rating prediction with explanation extraction through attention mechanism that identified important review text.

Knowledge graphs offered richer context for more comprehensive explanations. (Ai et al., 2018) leveraged heterogeneous knowledge base embeddings for explainable recommendations, while (Catherine et al., 2017) demonstrated KG-based explanation generation even without review text. (Wang et al., 2018c) proposed KPRN, which generated traceable reasoning paths to explain recommendations. (Xian et al., 2019) extended this with PGPR, employing reinforcement learning to explore large graphs efficiently, later refined by (Xian et al., 2020) with CAFE, which used user profiles to guide path search.

Advanced *neural approaches* further enhanced explanation quality through attention mechanisms and sophisticated text generation. (Chen et al., 2019c) developed CAML, employing co-attention between user and item review representations to simultaneously perform rating prediction and explanation generation, while (Gao et al., 2019) introduced DEAML, combining hierarchical concept graphs with attention to mitigate the accuracy-explainability trade-off. Natural language generation techniques enabled more sophisticated explanations. (Li et al., 2017) proposed NRT, which generated concise explanations while predicting ratings through multi-task learning. Later approaches like (Li et al., 2021b) introduced PETER, which leveraged transformer architectures to learn joint representations of users, items, and context, generating explanations conditioned on these representations. Building on this work, (Li et al., 2023a) developed PEPLER, which used prompt-enhanced personalized generation to improve the fluency and contextual alignment of explanations. (Ni et al., 2019b) tackled the challenge of generating explanations without supervised human-written examples by distantly labeling review sentences as aspect mentions. (Cheng et al., 2023) further advanced this area with ERRA, a model combining personalized review retrieval and aspect learning to generate more accurate and informative explanations.

LLMs transformed explainable recommendation research with unprecedented capabilities for generating nuanced explanations. (Ma et al., 2024) introduced XRec, a model-agnostic framework using LLMs to generate comprehensive explanations guided by collaborative filtering signals. (Yang et al., 2024) proposed LLM2ER-EQR, addressing challenges of personalization through a novel reinforcement learning framework that fine-tuned LLMs with explainable quality rewards. (Luo et al., 2023a) explored LLMXRec using instruction tuning for LLM-generated explanations, while (Wang et al., 2024) proposed LLM-PKG, building product knowledge graphs with LLMs for e-commerce explanations. In the e-commerce domain, explainability is particularly crucial for purchasing decisions and user trust. The EFM demonstrated improved user engagement on the JingDong platform. (Wang et al., 2022) showed Fast Fine-grained Sentiment for Explainable Recommendation (FSER) combined sentiment analysis of user reviews to provide explanations for recommendations, highlighting positive attributes that resonated with user preferences—particularly important in e-commerce where opinions and emotional responses influence purchases. These approaches generally require user profiles or historical interaction data to generate explanations, potentially compromising privacy while limiting flexibility across recommendation systems.

Temporal and Dynamic Approaches: Recognizing that user preferences evolve over time, Chen et al. (2019a) introduced a time-aware neural model combining recurrent neural networks with attention mechanisms to generate dynamic explanations that adapt to recent user behavior. This approach improved the sequential modeling of explainable user preferences, capturing the temporal dynamics of user-item interactions more effectively than static models.

Despite significant advances, existing approaches have predominantly focused on *user-centric* explanations that are contingent upon historical interactions and user profiles. This paradigm presents two critical limitations: (1) privacy risks associated with the extensive collection and use of personal data, and (2) an inability to satisfy the immediate, context-specific information needs articulated in user queries.

Our proposed **Query-Focused Explainable Recommendation** (QF-ER) framework addresses these shortcomings by generating explanations that respond directly to a user's query, eliminating any reliance on historical data. This design simultaneously preserves user privacy and delivers personalization grounded in the immediate query context. Our approach marks a paradigm shift in explanation generation, moving from a model based on *who you are* to one driven by *what you need*. Because the method requires *neither* user history nor proprietary ranking signals, it can be integrated with *any* recommendation engine and is capable of identifying and flagging commercially-driven placements.

5.6 The Unaddressed Frontier

While the preceding analysis highlights substantial advancements within each respective area, a critical analysis reveals a persistent fragmentation. Research has traditionally progressed in isolated silos: EXPLAINABLE RECOMMENDATION systems focused on justification without deep emotional context; OPINION SUMMARIZATION distilled sentiment polarity but often overlooked objective product metadata and causal triggers; and EMOTION ANALYSIS identified affective states without synthesizing them into actionable, coherent narratives for decision-making. This *siloed* methodology fundamentally fails to capture the holistic, multifaceted nature of the consumer's decision-making journey in the e-commerce ecosystem.

This survey addresses this *unaddressed frontier* by proposing a fundamental paradigm shift away from disjointed, product-centric tasks towards an integrated, human-centric synthesis. Our `Mind, Matter, and Markets` framework provides the conceptual backbone for this shift, and the five pioneering research directions we formalize—QF-CES, EAOS, EOT, M-OS, and QF-ER—are not merely incremental improvements. They represent a new class of e-commerce NLP problems that explicitly model the interplay between a user's cognitive and affective states (`Mind`), objective factual information (`Matter`), and the practical contexts of commercial platforms (`Markets`). Collectively, our work pioneers a unified vision that ***no prior work*** has articulated for the e-commerce space, one that prioritizes contextual relevance, emotional nuance, and causal reasoning to chart a clear course for the next generation of truly human-centered systems.

6 Datasets

This section presents an overview of `key datasets` utilized in LLM-based e-commerce information processing research, including novel contributions from proprietary industrial datasets that address critical resource gaps in the field.

6.1 Flipkart Q2P Dataset:

The research directions discussed in this survey introduce unique data requirements: large-scale collections of real-world user queries directly mapped to recommended products, accompanied by comprehensive multi-modal metadata for each product. Prior to this work, no such comprehensive resource existed, creating significant barriers to research progress in query-focused e-commerce applications. To address this fundamental gap, we introduce a foundational dataset that enables systematic investigation of user query understanding and product recommendation in realistic settings.

Q2P Dataset Overview: This dataset represents the first large-scale collection containing $7,500$ unique, real-world user queries sourced from a major e-commerce platform. Each query is systematically mapped to the top-3 products recommended by the platform's production recommendation system, yielding a total of $22,500$ query-product pairs. The dataset structure can be formally represented as:

$$\mathcal{D} = \{(q_i, P_i)\}_{i=1}^{7500}$$

where q_i denotes the i-th user query and $P_i = \{p_1, p_2, p_3\}$ represents the set of top-3 recommended products for that query.

Each product p_j in the dataset contains exceptionally rich metadata spanning multiple information modalities:

$$p_j = \{\texttt{title}, \texttt{description}, \texttt{key\_features}, \texttt{price}, \texttt{rating\_count}, \texttt{average\_rating}, \texttt{reviews}, \texttt{specifications}\} \quad (6)$$

The dataset encompasses 10 diverse e-commerce categories, including *Mobile Phones*, *Clothing*, *Electronics*, *Home & Kitchen*, and *Books*, ensuring broad domain coverage and cross-category generalizability. The distribution maintains balanced representation across categories:

$$|\mathcal{D}| = \sum_{k=1}^{10} |\mathcal{D}_k| = 7{,}500$$

where $\mathcal{D}_k$ represents the query subset for category k.

Dataset Statistic	Value
Unique user queries	7,500
Total products	22,500
Average reviews per product	10.0
Avg. specification length (tokens)	242.6
Avg. review length (tokens)	17.99
Avg. description length (tokens)	105.79
Avg. key features length (tokens)	24.64

Table 2: Statistical Overview of the Q2P Dataset

Data Quality and Annotation Standards: The dataset undergoes rigorous quality control procedures to ensure annotation consistency and reliability. All queries represent genuine user search intentions collected from production systems, while product metadata is extracted directly from vendor-provided information and user-generated content. This approach ensures ecological validity and real-world applicability of research findings derived from this resource.

6.2 Multi-Domain Datasets and Sampling Methodologies

For comprehensive cross-domain evaluation and generalizability assessment, we leverage established multi-domain datasets spanning diverse industries and user interaction patterns. These datasets enable systematic analysis of how LLM-based approaches perform across different domains, user populations, and linguistic variations.

The **Amazon Product Reviews dataset** (Hou et al., 2024) provides extensive coverage across multiple product categories including Beauty, Home & Garden, Electronics, Clothing, and Automotive. This dataset offers rich user-generated content with temporal spans covering multiple years, enabling longitudinal analysis of opinion evolution and seasonal trends.

For product sampling within each domain, we employ *Simple Random Sampling Without Replacement (SRSWOR)* to ensure unbiased selection (Cochran, 1977):

$$P_{\text{sample}} = \text{SRSWOR}(P_{\text{domain}}, n)$$

where P_{domain} represents all products in a specific domain and n denotes the desired sample size. This sampling strategy ensures that each product has equal probability of selection, eliminating potential selection biases.

The **TripAdvisor Reviews dataset** (Li et al., 2014) complements our analysis by providing hospitality and travel domain perspectives, while the **Yelp Business Reviews dataset** (Yelp Inc., 2025) contributes local business and restaurant review data. These `datasets` collectively enable cross-domain validation of proposed methodologies across fundamentally different service categories.

Review Filtering and Quality Control: When processing review texts, we apply systematic length-based filtering to control for content quality and informativeness (Kim and Lee, 2019; Herrando et al., 2021; Xie and Lee, 2022):

$$R_{\text{filtered}} = \{r \in R_{\text{original}} \mid L_{\text{min}} \leq |r| \leq L_{\text{max}}\}$$

where $|r|$ denotes the token count of review r, and L_{min} and L_{max} represent minimum and maximum length thresholds, respectively. Typical values are $L_{\text{min}} = 10$ and $L_{\text{max}} = 500$ tokens to ensure meaningful content while excluding extremely verbose reviews.

For temporal representation in longitudinal studies, we employ stratified sampling to maintain proportional representation across time periods:

$$R_{\text{stratified}} = \bigcup_{t \in T} \text{SRSWOR}(R_t, n_t)$$

where T represents the set of time periods, R_t denotes reviews from period t, and $n_t = n \cdot \frac{|R_t|}{|R_{\text{total}}|}$ ensures proportional allocation.

This sampling approach provides several methodological advantages: (1) uniform coverage probability $\pi = \frac{n}{|P|}$ across products, (2) temporal representativeness through stratification, and (3) statistical independence across domains for valid cross-domain comparisons.

Data Quality Considerations and Limitations: Researchers working with e-commerce datasets should exercise caution when incorporating numerical ratings (Mayzlin et al., 2014; de Langhe et al., 2016; Guo et al., 2020) and helpfulness votes (Yin et al., 2014; Lappas and Terzi, 2016; Deng et al., 2020). These signals are subject to well-documented biases including:

- **Fake Review Injection**: Systematic manipulation through incentivized positive reviews and competitor-targeted negative reviews
- **Rating Inflation**: Temporal drift toward higher ratings due to platform recommendation algorithms favoring highly-rated products
- **Selection Bias**: Non-random patterns in which users choose to leave reviews, creating skewed representations of product quality
- **Helpfulness Gaming**: Strategic voting on review helpfulness that may not reflect genuine utility assessments

To `mitigate` these issues, we recommend focusing primarily on textual content analysis while treating numerical signals as auxiliary features requiring careful validation. Additionally, temporal analysis of rating distributions can help identify potential manipulation patterns and inform appropriate filtering strategies.

Ethical Considerations and Privacy: All datasets used in this survey comply with platform terms of service and applicable privacy regulations. User-identifying information has been removed or `anonymized`, and all analysis focuses on aggregate patterns rather than individual user behaviors. Researchers utilizing these datasets should ensure compliance with institutional review board requirements and data protection regulations in their respective jurisdictions.

7 Evaluation Metrics

Evaluating the quality of generated opinion summaries is a critical and multifaceted task. The methodologies for this assessment are broadly categorized into two paradigms: *reference-based* metrics, which compare system-generated summaries against human-written ground truths, and *reference-free* metrics, which evaluate summary quality without requiring a gold-standard reference. This section details the key approaches within each paradigm.

7.1 Reference-Based Evaluation:

Reference-based evaluation has long been the standard for assessing summarization quality. This approach encompasses three primary methods: automated metrics that quantify textual similarity, direct human evaluation that captures subjective quality, and faithfulness metrics that measure factual consistency.

7.2 Automatic Evaluation

These metrics provide scalable and reproducible scores by algorithmically comparing a candidate summary to one or more reference summaries.

ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) (Lin, 2004c) is a set of metrics based on n-gram recall. It measures how many n-grams from the human-written reference summaries are found in the system-generated summary. The most common variants are:

- **ROUGE-N:** Measures the overlap of n-grams. For unigrams (N=1), the recall formula is:

$$\text{ROUGE-1} = \frac{\sum_{g \in R} \min(\text{Count}(g, C), \text{Count}(g, R))}{\sum_{g \in R} \text{Count}(g, R)} \tag{7}$$

  where R is the reference, C is the candidate, and g represents each unigram. This formulation computes recall-based overlap, which is the standard ROUGE-1 metric. ROUGE-2 uses the same principle for bigrams.

- **ROUGE-L:** Measures the longest common subsequence (LCS) to evaluate structural similarity, rewarding longer contiguous matches. The score is calculated as a ratio of the LCS length to the reference length.

BLEU (*Bilingual Evaluation Understudy*) (Papineni et al., 2002b) evaluates summaries based on n-gram *precision*, measuring how many n-grams in the candidate summary appear in the reference. While unigram precision assesses *adequacy* (content capture), higher n-grams assess *fluency*. BLEU's key feature is its **Brevity Penalty** (BP), which penalizes candidate summaries that are shorter than the reference, calculated as:

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases} \tag{8}$$

where c is the candidate length and r is the reference length.

BERTScore (Zhang et al., 2020c) moves beyond lexical overlap by measuring the semantic similarity between candidate and reference summaries using contextual embeddings from BERT. It computes precision, recall, and an F1 score by matching tokens based on their cosine similarity. The recall is calculated as:

$$\mathcal{R}_{\text{BERT}} = \frac{1}{|R|} \sum_{x \in R} \max_{y \in C} \mathbf{x}^T \mathbf{y} \tag{9}$$

where $\mathbf{x}$ and $\mathbf{y}$ are the normalized embeddings for tokens in the reference R and candidate C. The final score is the harmonic mean of this recall and a similarly computed precision.

METEOR (*Metric for Evaluation of Translation with Explicit ORdering*) (Banerjee and Lavie, 2005) enhances simple precision and recall by incorporating stemming and synonym matching. Its score is based on a harmonic mean of precision and recall (weighted towards recall) and a fragmentation penalty that penalizes non-contiguous matches to better assess fluency. The final score is computed as:

$$M = F_{\text{mean}} \cdot (1 - \text{Penalty}) \tag{10}$$

where F_{mean} represents the harmonic mean of precision and recall. This simplified formulation captures the essential components of METEOR appropriate for survey-level discussion.

7.3 Human Evaluation

Direct assessment by human annotators remains the *gold* standard for judging subjective qualities like coherence and usefulness.

Best-Worst Scaling (BWS) (Flynn and Marley, 2014) is a robust comparative judgment method. Annotators are shown a set of summaries (e.g., from 4 different models) and asked to identify the single best and single worst summary. Scores are aggregated across many judgments, providing a more reliable preference ranking than traditional rating scales (Kiritchenko and Mohammad, 2017).

Likert Scales (Likert, 1932) are widely used to rate summaries on specific dimensions (e.g., *fluency*, *coherence*, *faithfulness*) using an ordinal scale, typically from 1 to 5 (e.g., Very Poor to Excellent). This allows for granular, multi-dimensional feedback on a summary's performance.

7.4 Faithfulness Evaluation

Faithfulness, or factual consistency with the source document, is a critical dimension of summary quality. Specialized metrics have been developed to assess it:

- **SummaC** (Laban et al., 2022): An NLI-based model designed to detect inconsistencies at various levels of granularity between a summary and its source.
- **CTC** (Deng et al., 2021): A framework that evaluates information alignment to gauge both consistency and relevance.
- **FactCC** (Kryscinski et al., 2020): A BERT-based classification model trained to verify the factual consistency of a generated summary against its source article.
- **FactGraph** (Ribeiro et al., 2022): Enhances factuality evaluation by encoding both the source and summary into structured graphs and comparing their representations.

7.5 Metrics for Emotion and Opinion Trigger Detection

Evaluating the joint task of Emotion and Opinion Trigger Detection (EOT) requires assessing performance on two distinct sub-tasks: the classification of emotions and the extraction of their corresponding trigger spans. Therefore, a combination of classification and text-overlap metrics is employed:

Precision (P): For the emotion detection sub-task, this measures the accuracy of the predicted emotions. It is the fraction of correctly identified emotions (True Positives, TP) out of all emotions predicted by the model (TP + False Positives, FP).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{11}$$

Recall (R): This measures the model's ability to find all relevant emotions. It is the fraction of correctly identified emotions (TP) out of all actual emotions present in the ground truth (TP + False Negatives, FN).

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{12}$$

F1-Score (F1): As the harmonic mean of Precision and Recall, the F1-score provides a single, balanced measure of performance for the emotion detection sub-task, which is particularly useful when dealing with imbalanced emotion distributions (van Rijsbergen, 1979).

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{13}$$

For the opinion trigger extraction sub-task, this is the most stringent metric. It requires that the predicted text span is an identical character-for-character match with the ground-truth trigger span. A score of 1 is given for a perfect match, and 0 otherwise.

This is a more lenient metric for trigger extraction that considers token-level overlap. A match is counted if the set of tokens in the predicted span has a non-empty intersection with the set of tokens in the ground-truth span, acknowledging cases where the model correctly identifies the core trigger but with slightly different boundaries.

Based on the metric from (Lin, 2004b), this evaluates trigger quality by measuring the recall of unigrams (individual words) between the predicted trigger (T_{pred}) and the ground-truth trigger (T_{gt}).

$$\text{R1} = \frac{\sum_{u \in T_{gt}} \min(\text{Count}(u, T_{pred}), \text{Count}(u, T_{gt}))}{\sum_{u \in T_{gt}} \text{Count}(u, T_{gt}))} \tag{14}$$

Also from (Lin, 2004b), this metric evaluates trigger quality by identifying the longest common subsequence (LCS) of words between the predicted and ground-truth spans, rewarding structural similarity and the preservation of word order.

$$\text{RL} = \frac{\text{LCS}(T_{pred}, T_{gt})}{\text{length}(T_{gt})} \tag{15}$$

7.6 Reference-Free Evaluation

A significant limitation of reference-based metrics is their dependency on a human-written "gold standard." These metrics often penalize summaries that are semantically correct but use different wording or phrasing, a common characteristic of advanced LLMs. In fact, studies have shown that summaries generated by models like GPT can be preferred by humans over the original human-written references (Luo et al., 2023b).

This has spurred the development of **reference-free** metrics, which evaluate summary quality based on intrinsic characteristics or by using a powerful LLM as a proxy for a human evaluator (Liu et al., 2023b; Fu et al., 2023a). This approach allows for assessment across various dimensions—such as fluency, coherence, faithfulness (to the source), and specificity—without the need for a reference summary, offering a more scalable and potentially more aligned evaluation paradigm for modern generative models (Chiang and Lee, 2023b).

7.7 Metrics for QF-CES

For a complex, query-focused task like QF-CES, which generates multi-faceted outputs, a specialized set of `reference-free metrics` is required for comprehensive evaluation. The following dimensions are crucial for assessing the quality of such comparative summaries:

1. **`clarity` (CL)**- Clarity measures the degree to which the information in the Comparative Summary is clearly presented, avoiding ambiguity and ensuring that comparisons are easy to understand. The summary should be clear, concise, and easy to comprehend, using simple language and avoiding technical jargon whenever possible. It should be well-structured and well-organized, presenting comparison of the three products in a straightforward manner. The metric evaluates the readability of the entire summary, ensuring it is free from grammatical errors and has a logical flow between different sections and points. Additionally, the clarity of the tabular data is assessed to ensure it clearly conveys the comparisons between three products.

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

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

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

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

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

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

7.8 Metrics for EAOS

Evaluating the affective dimensions of a summary requires a specialized set of metrics that go beyond standard linguistic quality. For a nuanced task like EAOS, the following seven reference-free dimensions provide a comprehensive framework for assessment:

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.

7.9 Metrics for M-OS

Evaluating summaries that fuse information from multiple diverse sources—including objective metadata and subjective reviews—requires a robust set of reference-free metrics. The following seven dimensions are used to assess the quality and utility of M-OS outputs:

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

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

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

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

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

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

7. `specificity` (SP)- Specificity measures the level of detail and precision in the information and opinions presented in the summary. A specific summary provides concrete facts, measurements, or detailed descriptions about the product's features, performance, and user experiences. It avoids vague or general statements and instead offers precise information that gives readers a clear and thorough understanding of the product's characteristics and performance. Summaries should be penalized

for missing out details and should be awarded if they are specific.

7.10 Metrics for QF-ER

The evaluation of query-focused explanations requires a set of metrics that assess not only the linguistic quality and factual accuracy of the text but also its direct utility in answering a user's specific question. The following dimensions are used to provide a holistic assessment of QF-ER systems:

1. `clarity` **(CL)**- Clarity measures how well the explanation conveys information without ambiguity or confusion. A clear explanation presents product information relevant to the query in a straightforward, easily understandable manner, avoiding vague language, unexplained technical terms, or confusing descriptions. It ensures that users can immediately grasp how specific product features relate to their query requirements without having to decipher complex or unclear statements.

2. `fluency` **(FL)**- Fluency measures the quality of the explanation in terms of grammar, spelling, punctuation, capitalization, word choice, and sentence structure. The explanation should be easy to read, follow, and comprehend without any errors that hinder understanding, while maintaining a natural flow between query-specific information and product details.

   Note: When evaluating fluency, focus specifically on the linguistic quality and readability of the explanation, not whether the information is factually accurate or relevant to the query (which are covered by other metrics).

3. `coherence` **(CO)**- Coherence measures how well-structured and logically connected the explanation is. A coherent explanation should build from sentence to sentence, forming a unified and organized narrative that clearly relates the product to the user's query. It should avoid contradictions, irrelevant details, or abrupt jumps in reasoning, and instead present information in a smooth, logically progressive manner that helps users follow the explanation effortlessly.

4. `faithfulness` **(FA)**- Faithfulness measures the extent to which every piece of information mentioned in the explanation is verifiable, supported, present, or can be reasonably inferred from the product metadata. The explanation should be grounded in the product's metadata (including title, description, key features, specifications, reviews, and average rating) and should not introduce hallucinated or incorrect information. When discussing how the product relates to the user's query, all claims should be directly supported by the available product information.

5. `informativeness` **(INF)**- Informativeness measures the depth, breadth, and utility of product information provided in the explanation. It evaluates how well the explanation covers important product attributes and presents decision-critical details that would help a user make an informed choice, regardless of query specifics. High informativeness means the explanation provides rich, useful product insights.

6. `query relevance` **(QR)**- Query relevance evaluates how directly the explanation addresses the specific user query intent. It measures whether the explanation focuses on the explicit and implicit requirements expressed in the query, without introducing irrelevant information. High query relevance means the explanation precisely targets what the user was asking for. A relevant explanation not only addresses what was asked but provides information that would genuinely help users make better purchasing decisions based on their specific needs.

7. `conciseness` **(CON)**- Conciseness assesses whether the explanation is succinct and avoids unnecessary information, without being overly verbose. A concise explanation provides all query-relevant information efficiently, without redundancy, digressions, or excessive detail that doesn't contribute to addressing the user's query. It balances brevity with completeness, ensuring all necessary information is included without superfluous content.

8. `specificity` **(SP)**- Specificity measures the level of detail and precision in the information presented in the explanation. A specific explanation provides concrete facts, measurements,

or detailed descriptions about the product's features, performance, and user experiences that are relevant to the query. It avoids vague or general statements and instead offers precise information that gives readers a clear and thorough understanding of how the product's characteristics relate to their specific query.

9. **`sentiment consistency` (SC)**- Sentiment consistency measures how well the explanation's sentiment aligns with the sentiment expressed in the product reviews and ratings while remaining appropriate for the query context. The explanation should accurately reflect the balance of positive, negative, and neutral opinions from actual users' experiences with the product, particularly for aspects relevant to the query. An explanation with high sentiment consistency will neither be overly positive when reviews express concerns nor overly negative when reviews are predominantly positive.

8 Challenges and Open Problems

The integration of Large Language Models (LLMs) into e-commerce applications has demonstrated substantial potential for opinion mining and product summarization. However, numerous critical challenges and limitations persist across data acquisition, model reliability, evaluation methodologies, and deployment considerations. Resolving these fundamental issues remains essential for advancing robust and responsible human-centered LLM applications. This section systematically examines the most pressing challenges confronting the field.

Data Quality and Privacy Constraints: The efficacy of contemporary systems fundamentally depends on high-quality, large-scale data resources, presenting multifaceted challenges across the development pipeline.

Scarcity of Annotated Data: Gold-standard training and evaluation require extensive human-annotated datasets. However, developing such resources demands substantial financial investment and labor-intensive annotation processes, particularly for fine-grained tasks such as Emotion and Opinion Trigger Detection (EOT) or Emotion-Aware Opinion Summarization (EAOS). Recent benchmark developments including EOT-X and M-OS-EVAL demonstrate significant contributions to the field, yet their creation highlights the considerable effort required, potentially constraining broader research community participation.

Synthetic Data Dependency: To mitigate human annotation costs, researchers increasingly employ LLMs for synthetic training data generation, as exemplified by the `EAOS-SUMM` dataset. While this methodology provides enhanced scalability, it introduces risks of model-inherent biases and reduced linguistic diversity. Excessive reliance on synthetic data may yield models that excel on self-generated content while failing to generalize to the inherently unpredictable and nuanced characteristics of authentic human language use (Shumailov et al., 2023; Siledar et al., 2023).

Privacy-Personalization Tension: Traditional personalization approaches have relied extensively on comprehensive user profiling and historical behavioral data, raising substantial privacy considerations. While innovative methodologies such as Query-Focused Customer Experience Summarization (QF-CES) and Query-Focused Emotion Recognition (QF-ER) demonstrate the viability of query-based personalization strategies, broader challenges persist. Systems requiring comprehensive user understanding must carefully navigate the fundamental tension between delivering personalized experiences and preserving user privacy—a consideration increasingly critical within contemporary privacy-conscious digital environments.

Factual Accuracy and Hallucination Mitigation: Ensuring factual correctness represents a fundamental challenge across all generative applications. For Multi-perspective Opinion Summarization (M-OS), this necessitates accurate representation of technical specifications without fabricating or omitting critical details. For EAOS and EOT applications, it requires grounding all emotional interpretations within source textual evidence. LLMs demonstrate susceptibility to "hallucination"—generating plausible yet factually incorrect information. While structured prompting frameworks such as `EOT-DETECT` incorporating built-in verification mechanisms can mitigate these issues, maintaining complete faithfulness, particularly with complex and contradictory source materials, remains an unresolved challenge.

Retrieval-Augmented Generation Complexity: Modern e-commerce applications increasingly adopt `Retrieval-Augmented Generation` (RAG) architectures to access real-time product information, dynamic pricing data, and evolving inventory status (Lewis et al., 2020b; Guu et al., 2020). However, RAG systems introduce substantial complexity in maintaining retrieval quality and relevance. The dynamic nature of e-commerce data—where product specifications, availability, and user reviews change continuously—poses significant challenges for retrieval systems that must balance recency, relevance, and computational efficiency. Furthermore, the integration of retrieved information with generated summaries requires sophisticated fusion mechanisms to ensure coherence and prevent contradictory information propagation (Shuster et al., 2021; Yu et al., 2022).

Algorithmic Bias and Representational Fairness: LLMs are trained on extensive internet text corpora containing inherent societal biases (Bender et al., 2021). These biases can manifest in generated summaries through mechanisms such as over-representing majority perspectives while minimizing or excluding minority viewpoints (Sheng et al., 2021). The EAOS framework's incorporation of an *Emotional Bias Mitigation* metric directly acknowledges this risk. Ensuring these systems deliver fair, equitable, and representative summaries constitutes a critical ethical challenge requiring sustained research attention and methodological vigilance.

Continual Pre-training and Model Adaptation Challenges: E-commerce domains exhibit rapidly evolving characteristics, including emerging product categories, shifting consumer preferences, and evolving linguistic patterns in user-generated content. Traditional static pre-training approaches prove insufficient for capturing these temporal dynamics (Qin et al., 2022; Ke et al., 2022). `Continual Pre-training (CPT)` methodologies offer promising solutions but introduce significant computational overhead and catastrophic forgetting risks. The challenge lies in developing efficient incremental learning strategies that can incorporate new e-commerce knowledge—such as novel product attributes, emerging brand terminology, and evolving review patterns—without degrading performance on previously learned tasks (Jin et al., 2023; Wang et al., 2023b). Additionally, determining optimal update frequencies and data selection criteria for continual pre-training in dynamic e-commerce environments remains an open research question.

Evaluation Complexity: As demonstrated throughout this survey, traditional metrics including ROUGE (Lin, 2004b) and BERTScore (Zhang et al., 2020b) prove inadequate for evaluating nuanced system outputs. The field increasingly adopts multi-dimensional human evaluations and LLM-based assessment approaches. However, these methodologies introduce *novel* complexities. Human evaluation exhibits inherent subjectivity and limited scalability, while LLM-based evaluation, despite demonstrating strong correlation with human judgments, can ***manifest distinct biases***, including preferential treatment of summaries generated by models within the same architectural family. Developing robust, scalable, and unbiased evaluation protocols represents a substantial research challenge requiring dedicated investigation.

Computational and Accessibility Barriers: Training and deploying state-of-the-art LLMs demands significant computational resources, typically requiring access to high-performance hardware including `NVIDIA A100` or `H100` GPUs. Moreover, the most capable models, including `OpenAI's GPT-4o` and `Anthropic's Claude 3.5 Sonnet`, remain proprietary and accessible exclusively through cost-prohibitive API services. These requirements create substantial barriers for academic institutions and smaller organizations, potentially limiting innovation across the research community. While developing efficient, fine-tuned models such as `EOT-LLAMA` provides promising alternatives, performance gaps with frontier models frequently persist.

Cross-Domain Generalization Limitations: The surveyed research methodologies are predominantly optimized for e-commerce applications. The domain-specific linguistic patterns, specialized terminology, and information source characteristics remain particular to product review contexts. The generalizability of these frameworks to alternative domains—including medical patient feedback summarization, legal document analysis, or financial report processing—remains an open empirical question. Each novel domain would likely necessitate substantial adaptation and domain-specific fine-tuning procedures.

Limited Interactive Capabilities: The majority of described systems operate through "single-shot" generation paradigms, accepting input and producing static summaries. Truly human-centered systems would incorporate interactive and conversational capabilities, enabling users to pose follow-up queries ("Provide additional details regarding battery performance"), refine scope parameters ("Exclude price-focused reviews"), or resolve ambiguities. Integrating the sophisticated summarization capabilities demonstrated by these frameworks into dynamic, conversational interfaces represents a significant developmental step that remains largely unexplored.

Information Volume and Processing Complexity: Contemporary e-commerce platforms encompass millions of products, each associated with numerous reviews and comprehensive specifications. Processing this information volume and complexity presents substantial computational and methodological challenges (Bagozzi et al., 1999; Duan et al., 2008a).

User Preference Subjectivity and Diversity: Individual users demonstrate varying preferences, priorities, and information requirements, even when evaluating identical products. Developing personalized summaries and explanations addressing this diversity without extensive user profiling presents ongoing challenges (Kim et al., 2019; Wang and Benbasat, 2022).

Information versus Conciseness Trade-offs: Delivering comprehensive information while maintaining readability and relevance requires careful optimization. Excessive detail can overwhelm users, while insufficient information may impede informed decision-making processes (Greifeneder et al., 2007).

Privacy and Personalization Balance: Traditional personalization methodologies depend on extensive user profiling, raising privacy concerns. Developing approaches delivering personalized information based exclusively on current query contexts rather than historical behavioral data presents both methodological challenges and research opportunities (Damasio, 2004; Lerner et al., 2015).

Evaluation Framework Subjectivity: Evaluating summaries, comparisons, and explanations involves inherently subjective and multidimensional considerations. Developing robust evaluation frameworks achieving alignment with human judgment represents a significant methodological challenge (Chiang et al., 2023; Fu et al., 2023b).

Domain and Linguistic Generalization: E-commerce encompasses diverse product categories featuring domain-specific terminology and contextual considerations. Developing approaches that generalize across domains and languages while capturing domain-specific nuances presents substantial methodological challenges (Li et al., 2020a).

Given these *fundamental challenges*, continued advancement of LLM-driven systems for e-commerce applications will require both technical innovation and principled design considerations. We *conclude* this survey with a synthesis of key insights and concluding observations.

9 Summary and Conclusion

The exponential growth of e-commerce has introduced significant challenges in information processing, confronting consumers with substantial volumes of product information and user-generated content. This survey has examined recent advancements in applying Large Language Models (LLMs) to address information overload through systematic transformation into actionable, user-centered insights. The field has progressed from traditional, isolated approaches toward integrated systems that comprehensively address consumer decision-making processes.

A primary contribution of this survey is the formalization of the **Mind, Matter, and Markets** framework, a systematic conceptual structure for categorizing and analyzing these developments. This framework delineates innovations addressing the *Mind* (cognitive and emotional dimensions of user feedback), the *Matter* (factual and objective product characteristics), and the *Markets* (practical deployment of insights in commercial systems). Through this analytical lens, we have conducted comprehensive examination of five research directions that demonstrate significant impact on e-commerce applications.

These five methodological approaches—Multi-Source Opinion Summarization (M-OS), Emotion-

Aware Opinion Summarization (EAOS), Query-Focused Comparative Explainable Summarization (QF-CES), Emotion and Opinion Trigger Detection (EOT), and Query-Focused Explainable Recommendation (QF-ER)—collectively represent a fundamental methodological shift from product-centric data processing toward human-centric information synthesis. Rather than exclusively extracting features or sentiment classifications, these approaches generate summaries that demonstrate factual completeness, emotional awareness, contextual relevance, and transparent justification. Empirical validation through user studies demonstrates consistent preference for these enhanced summarization approaches compared to baseline methods.

From a methodological perspective, this survey has documented concurrent evolution in system development and evaluation techniques. The field has transitioned from basic zero-shot prompting strategies to sophisticated, structured reasoning frameworks incorporating self-reflection mechanisms and multi-step analytical processes. Additionally, evaluation methodologies are undergoing substantial transformation, moving from lexical-overlap metrics such as ROUGE toward more robust, reference-free assessment approaches that utilize LLMs as evaluators, demonstrating improved correlation with human judgment.

Future research directions emerging from this analysis include several promising areas for investigation. The **integration** of these five distinct approaches into unified systems capable of generating emotionally-aware, multi-source, comparative summaries responsive to specific queries represents a natural progression. Additional research opportunities include **multimodal** extensions incorporating visual and audio data from video reviews, development of **interactive and conversational** summary interfaces, and investigation of **cross-domain generalization** of these frameworks to information-intensive domains including healthcare and financial services.

In conclusion, the research developments surveyed in this paper establish foundations for advanced e-commerce platforms that extend beyond data presentation to provide intelligent, contextually-aware consumer assistance. Through continued development of systems that align with human cognitive and emotional processing patterns, these approaches demonstrate potential to address information complexity while supporting informed decision-making processes. The systematic application of LLMs to consumer-facing summarization tasks represents a significant step toward more effective human-computer interaction in commercial environments.

References

Francisca Adoma Acheampong, Henry Nunoo-Mensah, and Wancheng Chen. 2023. A survey on emotion analysis using large language models. *arXiv preprint arXiv:2310.19839*.

Qingyao Ai, Vahid Azizi, Xu Chen, and Yongfeng Zhang. 2018. Learning heterogeneous knowledge base embeddings for explainable recommendation. *Algorithms*, 11(9).

Reinald Kim Amplayo and Mirella Lapata. 2020. Unsupervised aspect-based opinion summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4776–4788.

Reinald Kim Amplayo, Stefanos Thomson, and Mirella Lapata. 2021. Aspect-controllable opinion summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2621–2633.

Marco Angiolillo, John O'Donovan, and Naren Sastry. 2022. What are you looking for? a new approach to query-based product recommendation. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '22, pages 253—-258.

Kumar Ankit, Balu A, and Sreejith S, G. 2022. A hybrid recommender system for e-commerce based on user reviews and product specifications. In *2022 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC)*, pages 1–6. IEEE.

Richard P Bagozzi and Utpal M Dholakia. 2003. Working to make a difference: a goal-theoretic perspective on volunteering. *Psychology & Marketing*, 20(5):377–404.

Richard P. Bagozzi, Mahesh Gopinath, and Prashanth U. Nyer. 1999. The role of emotions in marketing. *Journal of the Academy of Marketing Science*, 27(2):184–206.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert

Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional ai: Harmlessness from ai feedback.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21', pages 610–623. Association for Computing Machinery.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2024. Graph of thoughts: Solving elaborate problems with large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17682–17690.

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

Arthur Bražinskas, Reinald Kim Amplayo, and Mirella Lapata. 2020. Unsupervised opinion summarization with noising and attending. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5557–5568.

Jyrki Brazinskas. 2020. The consumer contextual decision-making model. *Frontiers in Psychology*, 11:570430.

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. 2020a. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33*, pages 1877–1901.

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. 2020b. Language models are few-shot learners.

Rose Catherine, Kathryn Mazaitis, Maxine Eskenazi, and William Cohen. 2017. Explainable entity-based recommendations with knowledge graphs.

Harrison Chase. 2022. Langchain. https://github.com/langchain-ai/langchain. Accessed on June 10, 2024.

Chong Chen, Min Zhang, Yiqun Liu, and Shaoping Ma. 2018a. Neural attentional rating regression with review-level explanations. In *Proceedings of the 2018 World Wide Web Conference*, pages 1583–1592. International World Wide Web Conferences Steering Committee.

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

Min Chen, Xiangnan Liu, Yifan Dou, Hong Yin, and Jing Li. 2018c. A reinforcement learning framework for e-commerce recommendation. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 917–922. IEEE.

Xu Chen, Yongfeng Zhang, and Zheng Qin. 2019a. Dynamic explainable recommendation based on neural attentive models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 53–60.

Yubo Chen, Yifei Wang, and Sivan Nevo. 2019b. The impact of user-generated content on consumer purchase decisions: A meta-analysis. *Journal of Marketing*, 83(6):94–113.

Yuxiang Chen, Preslav Nakov, and Min Zhang. 2022. A survey on opinion summarization. *arXiv preprint arXiv:2212.09411*.

Zhongxia Chen, Xiting Wang, Xing Xie, Tong Wu, Guoqing Bu, Yining Wang, and Enhong Chen. 2019c. Co-attentive multi-task learning for explainable recommendation. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 2137–2143. International Joint Conferences on Artificial Intelligence Organization.

Hao Cheng, Shuo Wang, Wensheng Lu, Wei Zhang, Mingyang Zhou, Kezhong Lu, and Hao Liao. 2023. Explainable recommendation with personalized review retrieval and aspect learning. In *Proceedings* of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL '23, pages 51–64, Toronto, Canada. Association for Computational Linguistics.

Cheng-Han Chiang and Hung-yi Lee. 2023a. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*.

Cheng-Han Chiang and Hung-yi Lee. 2023b. A survey on language model-based evaluation. *arXiv preprint arXiv:2311.08282*.

Wei-Lin Chiang and Hung-yi Lee. 2023c. A closer look at reference-free evaluation for dialogue. *arXiv preprint arXiv:2301.07693*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. https://lmsys.org/blog/2023-03-30-vicuna/.

En-Chuan Chu and Bing Liu. 2019. Aspect-based opinion summarization with gated convolutional networks. In *The World Wide Web Conference*, WWW '19, pages 273–283. Association for Computing Machinery.

Yang Chu and Jian Wang. 2019. A survey on e-commerce information processing technology. *Journal of Physics: Conference Series*, 1187(4):042048.

William G. Cochran. 1977. *Sampling techniques*, 3rd edition. John Wiley & Sons, New York.

Etienne Colas, Béatrice Fromy, and Thomas Hiebel. 2023. Knowrec: A knowledge-grounded model for explainable recommendation. In *Proceedings of the Seventeenth ACM Conference on Recommender Systems*, pages 202–212.

Antonio R. Damasio. 2004. *Descartes' error: Emotion, reason, and the human brain*. Penguin books, New York, NY.

Michael Han Daniel Han and Unsloth team. 2023. Unsloth.

Bart de Langhe, Philip M. Fernbach, and Donald R. Lichtenstein. 2016. Navigating by the stars: Investigating the actual and perceived validity of online user ratings. *Journal of Consumer Research*, 42(6):817–833.

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.

Mingkai Deng, Bowen Tan, Zhengzhong Liu, Eric Xing, and Zhiting Hu. 2021. Compression, transduction, and creation: A unified framework for evaluating natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7580–7605, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhaoxuan Deng, Jiaming Li, and Julian McAuley. 2020. Everything is not equal: An adversarial learning approach for fair helpfulness prediction of online product reviews. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, WSDM '20, pages 153–161. Association for Computing Machinery.

Oshin Deroy, Toan Ton-That, Pankaj Kumar, An Le, and Hady W. Lauw. 2023. Abstractive summarization of product reviews with various encoders. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2200–2210, Dubrovnik, Croatia. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs.

Zixiang Ding, Hui Zhang, and Jun Liu. 2020. ECPE-2D: Emotion-cause pair extraction based on joint two-dimensional representation, learning and decoding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3733–3742, Online. Association for Computational Linguistics.

Wenjing Duan, Bin Gu, and Andrew B. Whinston. 2008a. Do online reviews matter?–an empirical investigation of panel data. *International Journal of Electronic Commerce*, 13(2):29–57.

Wenjing Duan, Bin Gu, and Andrew B Whinston. 2008b. The dynamics of online word-of-mouth and product sales—an empirical investigation of the movie industry. *Journal of Retailing*, 84(2):233–242.

Tim Echterhoff, Malin Eiband, Andreas Klenk, and Heinrich Hussmann. 2023a. Explainable ai and trust: A systematic review. *ACM Computing Surveys*, 56(3):1–37.

Tim Echterhoff, Andreas Klenk, Malin Eiband, and Heinrich Hussmann. 2023b. Comparing products using shared aspects and their values from online reviews. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 155–169.

Paul Ekman. 1992. An argument for basic emotions. *Cognition Emotion*, 6(3-4):169–200.

eMarketer. 2023. Global ecommerce forecast 2023. https://www.insiderintelligence.com/content/global-ecommerce-forecast-2023. Accessed on June 10, 2024.

Güneş Erkan and Dragomir R. Radev. 2004. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479.

Chuang Fan, Hong-Yu Wang, Jiachen Du, and Maosong Sun. 2021. A multi-task learning framework for emotion-cause pair extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8345–8356, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Armin Felbermayr and Alexandros Nanopoulos. 2016. A survey of opinion summarization methods. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1686–1696, Austin, Texas. Association for Computational Linguistics.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji to learn representations for emotion analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1615–1625, Copenhagen, Denmark. Association for Computational Linguistics.

T.N. Flynn and A.A.J. Marley. 2014. Best-worst scaling: theory and methods. In Stephane Hess and Andrew Daly, editors, *Handbook of Choice Modelling*, Chapters, chapter 8, pages 178–201. Edward Elgar Publishing.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2022. GPTQ: Accurate post-training quantization for generative pre-trained transformers.

Jinlan Fu, See-Kiong Chen, and Anthony Tiong. 2023a. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.01397*.

Yinlin Fu, Kaize Shi, Guandong Xu, and Qing Li. 2023b. A survey on large language models for e-commerce. *arXiv preprint arXiv:2308.06946*.

Jingyue Gao, Xiting Wang, Yasha Wang, and Xing Xie. 2019. Explainable recommendation through attentive multi-view learning. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, AAAI '19, pages 3622–3629.

Weihua Gao, Jia Zhang, Yanchi Liu, Haoyu Wang, Bowen Li, Shirui Liu, and Hua Wang. 2024. Dre-rec: A de-biased, robust, and explainable recommender system based on large language models. *arXiv preprint arXiv:2402.19010*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models.

Rainer Greifeneder, Herbert Bless, and Klaus Fiedler. 2007. On the role of affect in the endowment effect. *Emotion*, 7(4):847.

Lin Gui, Jiachen Hu, Yulan He, Qin Ru, and Ruifeng Lu. 2016. A question answering approach to emotion cause extraction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1593–1602, Austin, Texas. Association for Computational Linguistics.

Rui Guo, Jingjing Zhang, and Mi Zhou. 2020. The effect of rating system design on consumer behavior: A natural experiment on yelp. *Marketing Science*, 39(2):386–405.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: Retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20, pages 3929–3938.

Xiangnan He, Tao Chen, Min-Yen Kan, and Xiao Chen. 2015. Trirank: Review-aware explainable recommendation by modeling aspects. In *Proceedings of the*

24th ACM International on Conference on Information and Knowledge Management, pages 1661–1670. ACM.

Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl. 2000. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work*, CSCW '00, page 241–250, New York, NY, USA. Association for Computing Machinery.

Carolina Herrando, Julio Jimenez-Martinez, and Maria J. Martin-De Hoyos. 2021. Understanding the new wave of digital consumers: A study of the determinants of review helpfulness. *Journal of Business Research*, 125:34–42.

Thomas Hosking, Hao Jiang, and Steffen Eger. 2023. Hierarchical summarization of user reviews. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10183–10195.

Rui-jie Hou, Jin-ji Li, and Wayne Xin Zhao. 2024. Large-scale multi-domain amazon product reviews for recommender systems. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Torino, Italia. ELRA and ICCL.

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.

Zhen Huang and John Rust. 2024. Emotionally-aware chatbots: a survey. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Papers)*, pages 101–112.

Giel Im, Sang-Gun Lee, and Young-Chan Lee. 2021. The role of emotional intelligence in the intention to use artificial intelligence-based financial services. *Sustainability*, 13(16):9177.

Hayato Iso, Takuma SATO, Kenta Oku, Fumito Nishino, Satoshi Akasaki, and Atsushi Suzuki. 2022. Comparative study of abstractive and extractive summarization of user reviews. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 4119–4123.

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.

Xisen Jin, Xiang Dong, Sean Welleck, Pengfei Li, and Chao An. 2023. Catastrophic forgetting in continual language model pre-training. *arXiv preprint arXiv:2310.05739*.

Zixuan Ke, Haowei He, Mikel Artetxe, Graham Neubig, and Kyunghyun Cho. 2022. Continual pre-training of language models. In *The Tenth International Conference on Learning Representations*.

Jihye Kim, Jae-Gil Lee, and Sang-Wook Park. 2019. Modeling user emotion in online reviews for recommendation. In *Proceedings of the 2019 IEEE/WIC/ACM International Conference on Web Intelligence*, WI '19, pages 90–97.

Min-Young Kim and Eun-Jung Lee. 2019. The effect of review quality on purchase intention: The role of the length of online reviews. *Journal of Theoretical and Applied Electronic Commerce Research*, 14(2):80–92.

Soo-Min Kim, Ani Nenkova, Michael Gamon, and Miri Raskino. 2011. Extractive summarization of customer reviews using a fancy random walk model. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1246–1256.

Svetlana Kiritchenko and Saif Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 465–470, Vancouver, Canada. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. *arXiv preprint arXiv:2302.14520*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th ACM Symposium on Operating Systems Principles*, SOSP '23, pages 611–626.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

LangChain. 2023. Langgraph: Multi-agent workflows. https://blog.langchain.dev/langgraph/. Accessed on June 10, 2024.

Theodoros Lappas and Evimaria Terzi. 2016. The impact of helpfulness votes on the content of online reviews. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 129–138. International World Wide Web Conferences Steering Committee.

An Le, Tuan Nguyen, and Hady W. Lauw. 2021a. Explainable recommendation with comparative summaries. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management*, CIKM '21, pages 959—-968.

An Le, Tuan Nguyen, and Hady W Lauw. 2021b. Explainable recommendation with comparative summaries. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 959–968.

Sung-Hee Lee, Gum-Won Lee, Sang-Kyu Lee, and Geun-Bae Lee. 2010. Finding causal relations of emotions in text. In *2010 2nd International Conference on Information Technology for Application*, pages 1–4, Gyeongju, Korea (South). IEEE.

Jennifer S. Lerner, Ye Li, Piercarlo Valdesolo, and Karim S. Kassam. 2015. Emotion and decision making. *Annual Review of Psychology*, 66:799–823.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandrina Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen-tau Yih, Tim Rocktaschel, Sebastian Riedel, and Douwe Kiela. 2020b. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Ang Li and Chandan K Reddy. 2020. Generate, filter, and rank: a new framework for review-based recommendations. In *Proceedings of the 13th international conference on web search and data mining*, pages 315–323.

Lei Li, Yongfeng Zhang, and Li Chen. 2021a. Personalized transformer for explainable recommendation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 494–503. Association for Computational Linguistics.

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

Lei Li, Yongfeng Zhang, and Li Chen. 2023a. Personalized prompt learning for explainable recommendation. *ACM Trans. Inf. Syst.*, 41(4).

Lin Li, Di Wu, Jing Wang, and Wei Li. 2020a. Bert-based sentiment analysis for stock price prediction. In *2020 15th international conference on computer science education (ICCSE)*, pages 569–574. IEEE.

Linzi Li and Wai Lam. 2020. Leveraging graph-based content representations for multi-source document summarization. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5580–5591.

Peng Li, Zhen-Zhen Ouyang, Hong-Song Wang, De-Zheng Xu, Xiao-Li Zhang, and Xiao-Ling Yang. 2014. Personalized point-of-interest recommendation by mining users' preference on non-functional aspects from reviews. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 1967–1970. Association for Computing Machinery.

Piji Li, Zihao Wang, Zhaochun Ren, Lidong Bing, and Wai Lam. 2017. Neural rating regression with abstractive tips generation for recommendation. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, page 345–354, New York, NY, USA. Association for Computing Machinery.

Xiangju Li, Wei Song, Shoushan Li, Xiao Han, Guozheng Zhang, Le Sun, and Jonathan Zhu. 2019. A co-attention neural network model for emotion cause analysis with emotional context-awareness. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3697–3706, Hong Kong, China. Association for Computational Linguistics.

Yizhuo Li, Lu Chen, Kai-Wei Chang, and Kai He. 2020b. Multimodal summarization with guidance of multimodal reference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4603, Online. Association for Computational Linguistics.

Yuxiang Li, Zechang Xiao, Qing Ran, Zhaohua Jiang, and Jiaao Liu. 2023b. Emotion-cause pair extraction from text: A survey. *Computational Intelligence and Neuroscience*, 2023:e3782015.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of Psychology*, 22(140):1–55.

Chin-Yew Lin. 2004a. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, pages 74–81.

Chin-Yew Lin. 2004b. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, pages 74–81.

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

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Xingyu Dang, and Song Han. 2023. AWQ: Activation-aware weight quantization for LLM compression and acceleration.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, and James Glass. 2023b. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Yucong Luo, Mingyue Cheng, Hao Zhang, Junyu Lu, Qi Liu, and Enhong Chen. 2023a. Unlocking the potential of large language models for explainable recommendations. *arXiv preprint arXiv:2312.15661*.

Zhe Fitch Luo, Yumo Zhang, Si Chen, Yida Li, Wen-Guan Wang, Wei Li, and H. Howie Yu. 2023b. ChatGPT as a factual inconsistency evaluator for text summarization. *arXiv preprint arXiv:2303.11525*.

Qiyao Ma, Xubin Ren, and Chao Huang. 2024. XRec: Large language models for explainable recommendation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 391–402, Miami, Florida, USA. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback.

Dina Mayzlin, Yaniv Dover, and Judith Chevalier. 2014. Promotional reviews: An empirical investigation of online review manipulation. *The American Economic Review*, 104(8):2421–2455.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Saif Mohammad and Felipe Bravo-Marquez. 2017. WASSA-2017 shared task on emotion intensity. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 34–49, Copenhagen, Denmark. Association for Computational Linguistics.

Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. In *Computational Intelligence*, volume 29, pages 436–465. Wiley Online Library.

Philipp Moritz, Robert Nishihara, Stephanie Wang, Alexey Tumanov, Richard Liaw, Eric Liang, Melih Elibol, Zongheng Yang, William Paul, Michael I. Jordan, and Ion Stoica. 2018. Ray: A distributed framework for emerging AI applications. In *13th USENIX Symposium on Operating Systems Design and Implementation (OSDI 18)*, pages 561–577.

Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. 2009. E-mail emotion-cause analysis. In *2009 IEEE International Conference on Systems, Man and Cybernetics*, pages 3522–3529, San Antonio, TX, USA. IEEE.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019a. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 188–197, Hong Kong, China. Association for Computational Linguistics.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019b. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 188–197, Hong Kong, China. Association for Computational Linguistics.

Xuefei Ning, Zinan Lin, Zixuan Zhou, Zifu Wang, Huazhong Yang, and Yu Wang. 2024. Skeleton-of-thought: Prompting llms for efficient parallel generation.

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

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

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. 2022b. Training language models to follow instructions with human feedback.

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 E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022c. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35*, pages 27730–27744.

Athanasios Papadimitriou, Panagiotis Symeonidis, and Yannis Manolopoulos. 2012. A generalized taxonomy of explanations styles for traditional and social recommender systems. *Data Mining and Knowledge Discovery*, 24(3):555–583.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002a. 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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002b. 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.

Nikolaos Pappas and Ion Androutsopoulos. 2014. Gisting with people's opinions: A bayesian approach to opinion summarization. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 548–557.

Wenjie Peng, Yixin Wang, Junchi Li, Guangde Sun, and Jiansong Sun. 2024. Uncertainty-aware llm for explainable recommendation. *arXiv preprint arXiv:2402.13843*.

Michel Tuan Pham. 2007. The logic of feeling. *Journal of consumer psychology*, 17(3):149–153.

Robert Plutchik. 1988. *The emotions: Facts, theories, and a new model*. University Press of America, Lanham, MD.

Robert Plutchik. 2000. *Emotions in the practice of psychotherapy: Clinical implications of a psychoevolutionary theory*. American Psychological Association, Washington, DC.

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

Yujia Qin, Yankai Lin, Yidi Bai, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2022. ELLE: Efficient lifelong pre-training for emerging data. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10915–10927. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report.

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. *Journal of Machine Learning Research*, 21(140):1–67.

Leonardo F. R. Ribeiro, Mengwen Liu, Iryna Gurevych, Markus Dreyer, and Mohit Bansal. 2022. FactGraph: Evaluating factuality in summarization with semantic graph representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3238–3253, Seattle, United States. Association for Computational Linguistics.

James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

Benjamin Scheibehenne, Rainer Greifeneder, and Peter M. Todd. 2010. Can there ever be too many options? a meta-analytic review of choice overload. *Journal of Consumer Research*, 37(3):409–425.

Bilgehan Sel, Ahmad Al-Tawaha, Vanshaj Khattar, Ruoxi Jia, and Ming Jin. 2024. Algorithm of thoughts: enhancing exploration of ideas in large language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Yixin Shen and Xiaojun Wan. 2023. A survey on evaluation of story generation. *arXiv preprint arXiv:2305.10619*.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges. *ACM Computing Surveys*, 54(8).

Ilia Shumailov, Zakhar Shumaylov, Yura Zverev, Yiren Zhao, Nicolas Bär, Yarin Gal, Nicolas Papernot, and Ross Anderson. 2023. The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv:2305.17493*.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 630–645. Association for Computational Linguistics.

Tejpalsingh Siledar, Suman Banerjee, Amey Patil, Sudhanshu Singh, Muthusamy Chelliah, Nikesh Garera, and Pushpak Bhattacharyya. 2023. Synthesize, if you do not have: Effective synthetic dataset creation strategies for self-supervised opinion summarization in E-commerce. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13480–13491, Singapore. Association for Computational Linguistics.

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

Amrit Singh, Aakarsh Anand, Abhinav Kumar, and Ashutosh Modi. 2021. Graph-based modeling for emotion-cause pair extraction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1190–1200, Online. 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 12th language resources and evaluation conference*, pages 4967–4974.

Carlo Strapparava and Rada Mihalcea. 2007. Semeval-2007 task 14: Affective text. In *Proceedings of the fourth international workshop on semantic evaluations*, pages 70–74.

Mirac Suzgun and Adam Tauman Kalai. 2024. Meta-prompting: Enhancing language models with task-agnostic scaffolding.

Juntao Tan, Shuyuan Xu, Yingqiang Ge, Yunqi Li, Yongfeng Zhang, and Xu Chen. 2021. Counterfactual explainable recommendation. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 1818–1827.

Yi Tay, Anh Tuan Luu, and Siu Cheung Hui. 2019. A survey on deep learning for aspect-based sentiment analysis. In *Proceedings of the 2019 international conference on asian language processing (IALP)*, pages 25–30. IEEE.

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.

Huang Tse-Hsun, Jernite Yacine, Indra Genta, Kim Eun-Gung, Bowman Samuel, R., and Cho Kyunghyun. 2024. Affective reasoning in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Mexico City, Mexico. Association for Computational Linguistics.

C. J. van Rijsbergen. 1979. *Information Retrieval*, 2nd edition. Butterworths, London, UK.

Dandan Wang and Wang Ling. 2016. Neural network-based abstractive summarization for customer reviews. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 89–99, San Diego, California. Association for Computational Linguistics.

Jiazheng Wang, Zhepeng Wang, Lian-He Zheng, Yilei Shen, Yujie Zhang, Jing Gao, and Dongyan Zhao. 2023a. Is chatgpt a good evaluator? a preliminary study. *arXiv preprint arXiv:2303.04060*.

Liyuan Wang, Xingbo Wang, Kuo Zhang, Chenghao Liu, and Long Zhang. 2023b. A comprehensive survey on continual learning: From pre-training to utilizing large language models. *arXiv preprint arXiv:2310.16527*.

Menghan Wang, Yuchen Guo, Duanfeng Zhang, Jianian Jin, Minnie Li, Dan Schonfeld, and Shawn Zhou.

2024. Enabling explainable recommendation in e-commerce with LLM-powered product knowledge graph. *arXiv preprint arXiv:2412.01837*.

Nan Wang, Hongning Wang, Yilin Jia, and Hong Yin. 2018a. Explainable recommendation via multi-task learning in opinionated text data. In *Proceedings of the 41st international ACM SIGIR conference on research and development in information retrieval*, pages 165–174.

Xiang Wang, Dingxian Wang, Can Xu, Xiangnan He, Yixin Cao, and Tat-Seng Chua. 2018b. Explainable reasoning over knowledge graphs for recommendation. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, pages 2334–2341. AAAI Press.

Xiang Wang, Dingxian Wang, Canran Xu, Xiangnan He, Yixin Cao, and Tat-Seng Chua. 2018c. Explainable reasoning over knowledge graphs for recommendation.

Xiao Wang and Izak Benbasat. 2022. The effects of explanations on user trust in and the use of recommendation agents: A meta-analysis. *Journal of Management Information Systems*, 39(2):489–531.

Xin Wang, Yang Liu, Yichuan Liu, Zhaohui Wang, and Yong Yu. 2023c. A survey on large language models for recommendation. *arXiv preprint arXiv:2305.19860*.

Xinyu Wang, Jian Wang, Jiawei Chen, Yujie Li, and Chunyan Miao. 2023d. Gcre: A generative model for comparative relation extraction-based explainable recommendation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15103–15116.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023e. Self-consistency improves chain of thought reasoning in language models.

Yequan Wang, Minlie Huang, Li Zhao, and Xiaoyan Zhu. 2016. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 606–615.

Ying Wang, Xin He, Hongji Wang, Yudong Sun, and Xin Wang. 2022. Fast explainable recommendation model by combining fine-grained sentiment in review data. *Computational Intelligence and Neuroscience*, 2022:4940401.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022b. Emergent abilities of large language models.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022c. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 675–689. Curran Associates, Inc.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022d. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Zefeng Wei, Sujian Li, Jian Su, Yan Wen, and Yang Xu. 2020. Effective inter-clause modeling for end-to-end emotion-cause pair extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2291–2301, Online. Association for Computational Linguistics.

Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao. 2023. Large language models are better reasoners with self-verification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2550–2575, Singapore. Association for 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 1667–1676, Florence, Italy. Association for Computational Linguistics.

Yikun Xian, Zuohui Fu, S. Muthukrishnan, Gerard de Melo, and Yongfeng Zhang. 2019. Reinforcement knowledge graph reasoning for explainable recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 285–294, New York, NY, USA. Association for Computing Machinery.

Yikun Xian, Zuohui Fu, Handong Zhao, Yingqiang Ge, Xu Chen, Qiaoying Huang, Shijie Geng, Zhou Qin, Gerard de Melo, S. Muthukrishnan, and Yongfeng Zhang. 2020. Cafe: Coarse-to-fine neural symbolic reasoning for explainable recommendation. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 1645–1654, New York, NY, USA. Association for Computing Machinery.

Keli Xie and Young-Joon Lee. 2022. What makes a good online review? the role of review text length and emotional intensity in review helpfulness. *Journal of Vacation Marketing*, 28(2):199–213.

Aobo Yang, Bahare Rastegarpanah, and Kristina Lerman. 2022. Comparative explanations of recommendations. *ACM Transactions on Recommender Systems*, 1(1):1–22.

Liying Yang, Yitong Qiu, Qing Li, Lu Chen, Ruitong Zhang, and Yuming Liu. 2021. Explanation-based personalized review generation. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 2209–2213.

Mengyuan Yang, Mengying Zhu, Yan Wang, Linxun Chen, Yilei Zhao, Xiuyuan Wang, Bing Han, Xiaolin Zheng, and Jianwei Yin. 2024. Fine-tuning large language model based explainable recommendation with explainable quality reward. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, volume 38 of *AAAI '24*, pages 9250–9259. AAAI Press.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023a. Tree of thoughts: deliberate problem solving with large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023b. React: Synergizing reasoning and acting in language models.

Yelp Inc. 2025. Yelp open dataset. https://www.yelp.com/dataset. Accessed on June 10, 2025.

Hong Yin, Bin Cui, Li Chen, Zhiting Hu, and Chen Zhang. 2014. Modeling the propagation of app adoptions in a billion-scale app-user network. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 1266–1275. Association for Computing Machinery.

Wen-Bin Ying, Xin-You Chen, Zhang-Tao Wang, Chen-Ming Li, Jia-Jun Zhang, Zhong-Qiu Yang, and Hong-Han Chen. 2019. A survey on emotion-cause extraction. *SCIENCE CHINA Technological Sciences*, 62(10):1649–1664.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2022. Generate rather than retrieve: Large language models are strong context generators. *arXiv preprint arXiv:2209.10063*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020a. Pegasus: pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

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

Yongfeng Zhang and Xu Chen. 2020a. Explainable recommendation: A survey and new perspectives. In *Proceedings of the 29th international joint conference on artificial intelligence*, pages 4931–4939.

Yongfeng Zhang and Xu Chen. 2020b. Explainable recommendation: A survey and new perspectives. *Foundations and Trends® in Information Retrieval*, 14(1):1–101.

Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. 2014. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, page 83–92, New York, NY, USA. Association for Computing Machinery.

Wenxiao Zhao, Ziqiang Cao, Yang Song, Kun-Ta Chuang, and Heng-Tze Cheng. 2020. Improving abstractive opinion summarization with product descriptions. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4068–4073, Online. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng Liu, Siyuan Lee, Frank F Li, Hao Liu, Banghua Zhao, Yu Sheng, Ali Halawi, Danyang Li, et al. 2023. Judging llms is not as easy as you think: A case study on chatbot arena. *arXiv preprint arXiv:2310.05447*.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2023a. Least-to-most prompting enables complex reasoning in large language models.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023b. Large language models are human-level prompt engineers.