Marketing Science Institute Working Paper Series 2026

Report No. 26-103

From Reviews to Responses: Bridging Pre- and Post-Purchase Consumers through AI-Enhanced QA with RAG

XingJian You, Ishita Chakraborty and Neeraj Arora

From Reviews to Responses:
Bridging Pre- and Post-Purchase Consumers through AI-Enhanced QA with RAG

XingJian You, PhD Candidate, Wisconsin School of Business, United States, xingjian.you@wisc.edu
Ishita Chakraborty, Assistant Professor of Marketing, Wisconsin School of Business, United States, ishita.chakraborty@wisc.edu
Neeraj Arora, Professor of Marketing, Wisconsin School of Business, United States, neeraj.arora@wisc.edu

Corresponding author: XingJian You (xingjian.you@wisc.edu)
Author notes:
Acknowledgments: We thank the participants at the *Wisconsin School of Business Symposium on AI and Marketing 2025, Madison, WI*, the *Marketing Science Conference 2025, Washington, DC*, and the *Biz AI Conference 2025, University of Texas at Dallas, TX* for their valuable feedback. We also thank Shrabastee Banerjee for her thoughtful comments.
Regarding AI usage, since the central theme of this paper is AI, we employed large language models (LLMs) to generate answers, conduct analyses (e.g., topic modeling), and assist with data annotation. The procedures for these applications are detailed in the paper. In addition, LLMs were used during the writing process for grammar checking.

From Reviews to Responses: Bridging Pre- and Post-Purchase Consumers through AI-Enhanced QA with RAG

Abstract

Question Answering (QA) on customer-facing platforms (e-commerce, travel, education, and brand websites) often suffers from delayed, low-quality responses and limited user participation. While customer reviews are abundant, their unstructured nature limits direct use to answer specific questions, and Large Language Models (LLMs) alone lack product-specific details and subjective insights. To address these challenges, we propose and evaluate a novel QA framework that integrates LLMs with reviews through Retrieval-Augmented Generation (RAG). Our framework incorporates three components: (1) RAG for dynamically retrieving relevant reviews at inference time; (2) a question–review type matching module that enhances topical alignment; (3) an answerability classifier that determines whether a reliable answer can be generated. Using a data set of 500 Amazon questions, 2000+ human responses, and 14,000 review sentences, we systematically evaluate different model variants for both lexical similarity metrics (ROUGE-L) and human judgments. Our full model improves lexical similarity scores by 50% from baseline LLM answers, matches or exceeds 72% of human responses, and approaches the best human answers in clarity, relevance, and informativeness. In particular, human evaluations show that our full model performs particularly well on subjective questions and lexical similarity metrics fail to capture this performance gain. Overall, our findings show how LLMs and reviews can be combined to build scalable QA systems, while also revealing the limits of lexical similarity metrics and highlighting the importance of human-centered evaluation.

Keywords:
Question Answering, Generative AI, Natural Language Processing, RAG, E-Commerce

Question Answering (QA) has become a significant form of user-generated content (UGC) on customer-facing platforms, spanning e-commerce sites (e.g., JD.com, Target), travel forums (e.g., Tripadvisor), and brand support communities (e.g., Apple Support). We define customer-facing QA as contexts where customers pose product- or service-related questions and receive responses from peers, brand representatives or community members, thereby reducing pre-purchase uncertainty. They play a critical role in both customer experience and platform outcomes. According to recent industry reports, about 74% of online shoppers regularly read QA sections on a product page (CustomerThink 2024) before making a purchase, and up to 53% of shoppers abandon their purchase if common questions are not answered (ConvertCart 2025). In addition, effective QA can enhance perceived product fit and quality by reducing uncertainty about the product, leading to an average increase of .5 to 1 star in product ratings (Banerjee, Dellarocas, and Zervas 2021).

Another prevalent form of UGC that prospective buyers rely on before making a purchase is reviews or feedback from past customers, which describe their experiences with different aspects of the product. Although both QAs and reviews inform customer decision-making, they are structurally different: QAs typically occur pre-purchase and address specific questions, whereas reviews are written post-purchase and tend to be broader in scope. Figure 1 is an illustration of these two types of content in the context of e-commerce websites. Platforms explicitly separate these two content types to facilitate distinct stages of the purchase cycle.

From the standpoint of a potential buyer, both forms of UGC have some limitations. Post-purchase customer answers are often delayed by around two days, risking lost sales. Furthermore, engagement is low; only one in five customers contributes to the QA section, resulting in a 4:1 ratio of reviews to answers (Gupta et al. 2019). The quality of answers also varies widely: while a clear and relevant response can effectively reduce customer uncertainty, a vague or misleading answer may increase confusion or frustration.

In contrast to QA, reviews are abundant, but are not tailored to address specific cus-

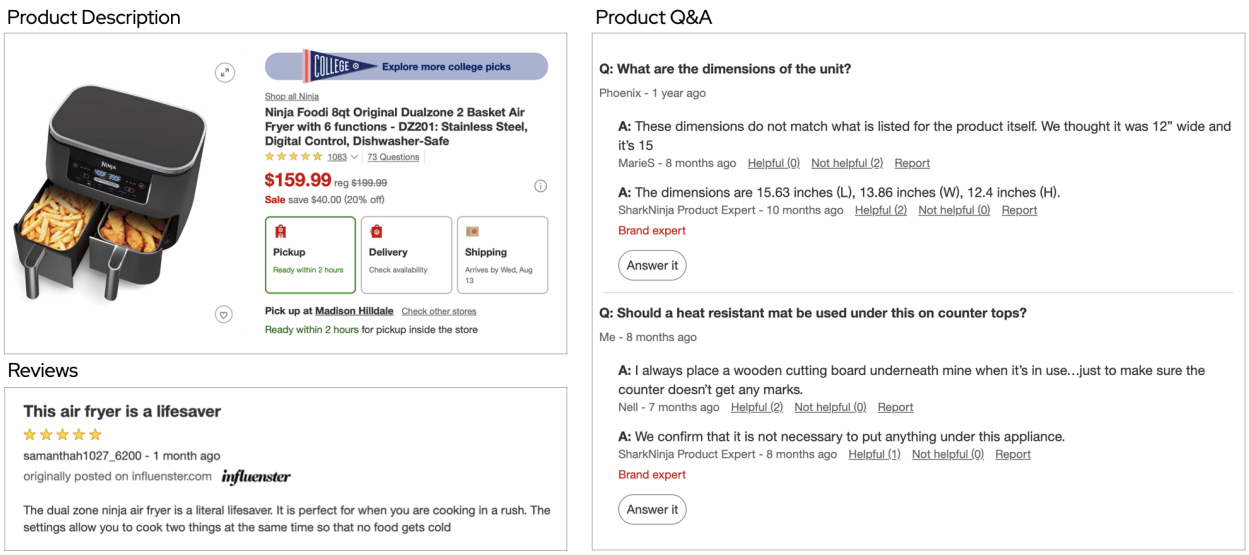

Figure 1: A Product on an E-Commerce Site with Description, Reviews and QA

tomer inquiries. They tend to be driven by self-expression, often containing sentiment-laden and multi-themed content that can be lengthy and difficult to parse. While this richness provides valuable context, their unstructured, free-flowing nature limits their effectiveness in supporting efficient, query-specific information retrieval; an essential requirement for a robust QA system. Table 1 illustrates this difference with the question of whether an air fryer is too large for two people. Answer 1 directly addresses the question, whereas answer 2 is less informative. In contrast, the review covers a broad spectrum of product features, but relevant information ("perfect for 1 to 2 people") is buried in a long discussion.

These limitations in existing user-generated content have sparked a growing interest in the use of advanced AI technologies to improve QA. Recent developments in Large Language Models (LLMs), such as GPT-4 and Gemini, present promising opportunities to automate question-answering systems on different platforms, significantly reducing response times. These models are capable of producing human-like responses and have demonstrated strong performance in general-purpose text generation and reasoning tasks (Bommasani et al. 2021; Achiam et al. 2023).

More recently, marketing and social science scholars have begun applying LLMs to diverse research contexts, such as qualitative and quantitative market analysis (Blanchard et al. 2025;

Table 1: Example of Product Description, QA, and Review

Component Type	Example Content
Product Name	Ninja Foodi 8qt Original Dualzone 2 Basket Air Fryer with 6 functions
Product Description	The first air fryer with two independent baskets that let you cook two foods at once, not back-to-back like a traditional single-basket air fryer.
Question	Is it too big for two people?
Answer 1	There is only two in my house and it works perfect in making both aspects of supper/lunch/breakfast at the same time. For example chicken breasts in one and potatoes or veggies in the other and a salad on the side. You will have to increase cooking time.
Answer 2	Not perfect :)
Relevant Review	I use my air fryer multiple times a week, some weeks daily, and it has completely transformed how I cook. It's fast, cooks incredibly well (especially using one basket at a time), and delivers consistently crispy, delicious results. The controls are simple and clear, though I do have to be a bit careful with bumping the temp knob by accident. That's a tiny issue compared to all the wins. It's absolutely perfect for 1 to 2 people, and I often don't even need both baskets. I especially love making fish in it, and it has been a game changer for my diet. Meat turns out so juicy and easy, and even veggies and leftovers reheat like a dream. It's one of my favorite appliances and my second most-used piece of cookware. So fun and simple to use that I actually feel like a real chef. Worth every penny. I'd buy it again in a heartbeat.

Arora, Chakraborty, and Nishimura 2024), perceptual mapping (Li et al. 2024), discovering consumer needs (Timoshenko, Mao, and Hauser 2025), ideation (De Freitas, Nave, and Puntoni 2025), and the examination of textual behavioral data (Feuerriegel et al. 2025), with encouraging early results. However, LLMs also have some important shortcomings; they struggle with product-specific fine-grained queries due to limitations in their training data (Kayali et al. 2024; OpenAI 2025; Blanchard et al. 2025) and can hallucinate when context is missing (Borji 2023). In addition, language models often underperform in subjective questions shaped by personal preferences or individual experiences, such as "Is this lotion greasy?" (Bjerva et al. 2020).

Evaluating the quality of the LLM answers presents another challenge. Traditional QA metrics emphasize lexical overlap between a *benchmark human answer* and a model-generated response. Yet, since a question may be answered well in multiple ways (e.g., relevance, depth, etc.), lexical similarity alone often fails to reflect how users perceive or interpret an answer. In fact, previous work finds that many lexical overlap metrics show a limited correlation with actual comprehension (Guo et al. 2025), making the automation of

QA with LLMs even more challenging.

Against this backdrop of QA, reviews, and LLMs, in this paper we seek to answer three research questions.

1. How can one form of UGC be used to enrich another with the help of LLMs? Specifically, how might the generative strengths of LLMs, when combined with grounded, experience-rich post-purchase reviews, enhance the QA system on platforms?

2. Can this hybrid LLM-enabled approach deliver high-quality answers across question types, particularly for subjective vs. objective questions?

3. How should we evaluate LLM-generated answers to customer queries? Are lexical overlap metrics sufficient, or is consumer-centered human evaluation essential? How do human and automatic assessments differ and what insights arise from their divergence?

To answer our research questions, we propose a framework that complements an LLM with consumer reviews to significantly improve the quality of answers to questions. In this framework, the LLM bridges pre-purchase and post-purchase customers, drawing on both internal knowledge of the LLM and user-generated content to create informed responses. The central piece of our framework is Retrieval-Augmented Generation (RAG) (Lewis et al. 2020) which is a hybrid approach that augments the output of a language model with information retrieved from an external knowledge source. In this paper, this external knowledge source is consumer reviews. By injecting important contextual knowledge, RAG enables LLMs to dynamically incorporate the most relevant, up-to-date review content at inference time. This makes it particularly suitable for enhancing customer-facing QAs with transient, large volumes of review data.

Although RAG improves responses by incorporating contextual information, a persistent challenge lies in retrieval precision. While several platforms typically have a large pool of reviews, much of this content does not directly address the specific customer question (Deng et al. 2023). Standard RAG implementations often rely solely on the lexical similarity

between questions and reviews, which can fail to capture answer relevance. To address this limitation, we propose a question–review matching module in our framework that re-ranks the reviews based on how similar the content is to the question, and whether the review talks about the same topic as the question (e.g. fit, usability, customer service). In other words, if the question is asking about how well a product fits, the module gives higher priority to reviews that also discuss fit, rather than unrelated aspects.

Finally, while the RAG and matching modules improve contextual grounding and relevance, not all questions are answerable, even when extensive information is available. Attempting to answer such unanswerable questions can lead to hallucinated, irrelevant, or low-quality responses. This behavior can undermine user trust. To mitigate this natural behavior of LLMs, we introduce an answerability classifier in our framework that evaluates whether a given question can be reliably answered using the available knowledge base. Its sole purpose is to determine the presence or absence of sufficient supporting information before any response is generated.

To recap, our answer generation framework consists of three elements: RAG that uses consumer reviews as its knowledge source, a matching module to map appropriate review dimensions with a specific question, and an answerability classifier that evaluates the ability of the module to generate a reliable answer and only generates answers that are meaningful.

We test our proposed QA framework empirically by using data collected from the product page of Amazon (Gupta et al. 2019), which includes 500 consumer questions, 2,049 human answers, and 14,083 review sentences. Each question is typically answered by multiple individuals, thus resulting in more answers than questions. For each question, we generate an answer using a variety of models (LLM-only, LLM+RAG, LLM+RAG+Matching, etc.) and benchmark the quality of these answers against human answers. To evaluate the quality of the generated answers, we adopt a dual approach: automatic lexical similarity metrics and human evaluation[1].

[1] *Human answer* refers to a response produced by a person, such as a customer. *Human evaluation* denotes the assessment of human and LLM answers by human raters along dimensions such as relevance and clarity.

We begin our analysis with automatic evaluation using a widely used metric (ROGUE-L, Lin 2004) that measures the longest common subsequence between generated and reference answers to capture lexical overlap. Although this metric captures the similarity at the word level, it overlooks deeper aspects of the answer quality. To complement this analysis, we perform a theory-driven human evaluation that focuses on five key dimensions: clarity, relevance, informativeness, trustworthiness, and helpfulness. We recruited 582 human annotators from Prolific to evaluate both model-generated and human-written answers, allowing us to benchmark model-based (LLM-only, LLM+RAG, LLM+RAG+Matching, etc.) answers against human answers.

The results of automatic evaluation demonstrate the value of our framework. Incorporating reviews through RAG increases the lexical similarity by 30%—from .105 (LLM-only) to .143 ($p < .001$). Adding the answerability classifier further improves performance by an additional 10%, although it limits the system to answering only 62.6% of the questions. The matching of the type of question to the review slightly increases the answer rate to 64.4%, with greater gains in review-sparse categories such as customer service.

The human evaluation results reveal several important insights. Our full model achieves an average answer quality score of 4.14 on a 5-point scale, which is significantly higher than the average score of human answers (3.50, $p < .001$), and begins to approach the best-human score of 4.34, though the difference remains significant ($p = .007$). Notably, the model meets or exceeds 72% of human answers. Across dimensions such as relevance, clarity, and informativeness, the model's responses are closer to the best human answers. For example, on the relevance dimension, the best-human score is 4.51, while the model achieves 4.32 ($p = .028$), with similar gains observed for clarity and informativeness. Surprisingly, LLM answers are also perceived as more trustworthy than average human answers when the source of the response is unknown to the rater.

Interestingly, our full model performs better on subjective questions versus objective questions; this is contrary to our finding based on lexical similarity and the previous liter-

ature that found that language models struggle in such domains (Bjerva et al. 2020). We demonstrate that this occurs because model-based answers to subjective questions combine the machine-like precision of LLM with "word of mouth" of reviews, making them particularly suitable for subjective questions. In fact, model-based answers often surpass human responses, but lexical similarity metrics penalize any divergence from the benchmark human answers, regardless of whether the deviation reflects higher or lower quality from the benchmarks. This highlights a limitation of lexical metrics, which penalize higher-quality responses when they diverge from weaker human benchmarks. Overall, our findings underscore the importance of human-centered evaluation, as similarity to benchmark answers on QA platforms should not be treated as the gold standard.

In summary, this paper makes three key contributions in the QA space, which is an underresearched area in the literature in marketing. First, we propose and empirically test a novel LLM-enabled QA framework which bridges pre-purchase and post-purchase customers to generate answers to questions that individuals ask. We demonstrate how firms can strategically combine proprietary consumer-generated content with LLM capabilities to enhance pre-purchase experience for potential buyers. Second, we show that lexical similarity-based metrics, although widely used, are flawed. We recommend human-centered evaluation when assessing answer quality. We show that building effective and trusted QA systems requires not only strong AI models but also platform-level strategies that foster human-AI collaboration. Finally, we show that answers generated by our framework can surpass the quality of a large majority of human answers, especially for subjective questions.

The paper is organized as follows. We begin with a background section, which reviews existing research on automating QA and introduces our approach of using an LLM, enhanced with reviews, as an information source for answer generation. We then outline our proposed QA framework and an illustration of how it works. This is followed by a section on the description of data and a section on the methodological foundations for the paper. The results section describes our findings in detail, following which we conclude with a summary

of the main findings, practical implications of our work, limitations, and future research.

BACKGROUND

Customer-facing QA sits at the intersection of marketing and artificial intelligence. The setting allows us to contribute to three key streams of literature: foundational language models, user-generated content, and consumer-centered evaluation. First, in the context of QA, we move beyond traditional machine learning approaches by leveraging LLMs to automate the task, offering a more flexible and scalable solution. Second, we improve LLMs using a customized RAG framework for QA, introducing a new way to leverage consumer reviews as an active source of information. Third, we shift the evaluation lens from a purely language-processing task to a consumer-centered perspective by combining similarity-based metrics with human judgments.

In this section, we briefly outline our contributions to the existing literature along these dimensions: (1) how LLMs are used to automate QA, (2) how can RAG be leveraged with reviews to improve LLMs, and (3) how should LLM-generated answers to customer questions be evaluated.

How LLMs Are Used to Automate QA

Automating question answering is a core task in natural language processing (NLP), designed to generate relevant answers based on supporting content. Early research, most of which appeared in the literature outside marketing, focused on structured tasks—settings where both the questions and supporting contexts are well-formatted, and drawn from clean sources like Wikipedia. Benchmark data sets such as SQuAD (Rajpurkar et al. 2016), Natural Questions (Kwiatkowski et al. 2019), and TriviaQA (Joshi et al. 2017) reflect this paradigm[2].

In this paper, we focus on customer-facing QA—a specialized form of QA that directly

[2]For example, given a paragraph like "The Apollo program was the third United States human spaceflight program carried out by NASA," and a factual question—"Who carried out the Apollo program?"—the correct answer "NASA" is an exact phrase in the paragraph.

supports buying decisions. Unlike traditional QA, which typically involves objective and factual queries grounded in structured sources, customer-facing QA addresses open-ended and experiential questions that require nuanced interpretation of consumer perceptions. The supporting evidence often comes from UGC, such as reviews. As a result, it demands new approaches in retrieval, information sources, and evaluation (Deng et al. 2023).

In the past, researchers have proposed several methods based on machine/deep learning to automate answer generation. Moqa (McAuley and Yang 2016) approached the task as a classification problem, treating each retrieved review as an 'expert' that votes on binary (yes/no) questions, while for open-ended queries it surfaces relevant reviews to aid users. Kim et al. (2022) framed the problem as a retrieval task, with the aim of extracting the most relevant review snippets or previously posted answers to approximate a response. Gupta et al. (2019) adopted a generative approach in addition to retrieval models, training sequence-to-sequence to generate free-form answers based on input questions. Although these approaches differ in design, they share key limitations: they are static and struggle with logical and context-dependent reasoning in product-specific scenarios.

LLMs present a promising alternative to traditional machine learning approaches. Unlike static models, LLMs are trained on large-scale internet corpora, enabling them to capture consumer needs. They also possess internal knowledge, reasoning capabilities, and the ability to summarize information, which makes them well suited to generate answers. By shifting from conventional machine learning models to LLM-based approaches, our work fills a gap in the literature and offers a more adaptive and scalable solution for automating customer-facing QA. However, as noted in the Introduction, LLMs have limitations. They may lack product-specific knowledge because the data on which they are trained lack a firm's private data (Kayali et al. 2024; OpenAI 2025), struggle with subjective or context-rich questions (Bjerva et al. 2020; Deng et al. 2023), and are perceived as less trustworthy than human responses (Seymour et al. 2024). In the following section, we discuss how we enhance LLMs to overcome these challenges.

How to Enhance LLMs with RAG Using Reviews

In the context of customer-facing QA, the missing product information and consumer perceptions that LLMs lack could be embedded in reviews. Although reviews were not written to answer targeted questions, they often contain precisely the details that shoppers are looking for. The difficulty is that these details are buried in lengthy, unstructured, sentiment-laden text, mixed with irrelevant topics. However, with appropriate structuring—such as attribute extraction and semantic retrieval—reviews can be transformed into a reliable information source (Chakraborty, Kim, and Sudhir 2022).

The early papers to investigate reviews focused mainly on their valence. For example, linking online review ratings and length to economic outcomes such as sales (Chevalier and Mayzlin 2006). This was followed by work on the distributional properties of reviews, exploring how the shape of rating distributions influences consumer behavior, how topic trends in reviews correlate with consumer perceptions and market performance (Tirunillai and Tellis 2014) and how the extremity is shaped by review writing biases (Brandes, Godes, and Mayzlin 2022). The development of language models enabled the extraction of a richer set of attributes from reviews, such as staff courtesy, wait time, and cleanliness, to gain a more nuanced understanding of customer satisfaction (Puranam, Kadiyali, and Narayan 2021; Peter, Chakraborty, and Banerjee 2022).

Our study builds on this evolution by leveraging reviews more dynamically, not to extract fixed attributes, but as an information source to generate question-specific answers that help mitigate consumer uncertainty. This demonstrates a novel use of UGC. To fully realize this potential, LLMs must be paired with mechanisms that extract question-relevant insights from unstructured reviews. This calls for a hybrid approach that generates answers to consumer questions using context-specific, review-based evidence in conjunction with an LLM.

To operationalize this idea, we adopt a RAG-based approach over alternatives like prompt engineering, which is limited to the knowledge base on which the LLM is trained, and fine-tuning, which is static and less adaptable. In our framework, RAG serves as the backbone

for bridging pre-purchase questions with post-purchase consumer experiences. By retrieving relevant review snippets and using an LLM to synthesize them into fluent, question-specific responses (Finardi et al. 2024), we transform reviews from passive background content into active, context-aware support for answer generation.

Although RAG has been applied in domains such as clinical decision making and book review generation (Arslan et al. 2024), its application to customer-facing QA is a novel aspect of this study and presents unique challenges. In this setting, reviews are written informally, are loosely structured, and the questions are not always answerable. To overcome these challenges, we extend the standard RAG framework with two key enhancements: a question–review matching module to strengthen context alignment, and an answerability classifier to filter out questions that cannot be addressed. Together, these adaptations make RAG better suited for customer-facing QA and other applications that involve unstructured UGC.

How to Evaluate QA Answers

Evaluating the quality of the answers is another critical yet complex component in the construction of QA systems. Traditionally, this task has been evaluated with similarity metrics like ROUGE (Recall-Oriented Understudy for Gisting Evaluation; Lin 2004) and BLEU (Bilingual Evaluation Understudy; Papineni et al. 2002). ROUGE measures the longest common subsequence between generated and reference texts, while BLEU assesses how many word sequences are shared, both emphasizing lexical overlap. For example, for a reference sentence, "The headphones support Bluetooth 5.0 and have a 30-hour battery life" and a generated sentence "These wireless headphones last around 30 hours and work well with Bluetooth", although the core content aligns, the wording and the order differences yield a medium ROUGE-L score. However, these lexical similarity metrics are often biased (Yang et al. 2018), failing to capture dimensions that truly matter to consumers, such as clarity, informativeness, and practical relevance and even tend to penalize concise and readable re-

sponses (Sulem, Abend, and Rappoport 2018). Based on such concerns, Guo et al. (2025) argue that evaluation should be guided by real-world effectiveness, not just technical benchmarks.

To address these challenges in our specific context, we adopt a dual evaluation strategy that combines similarity-based automatic metrics such as ROUGE with human judgments, allowing us to better capture user-centered aspects of answer quality. Human judgments allow us to measure quality on dimensions that include answer clarity, informativeness, relevance, trust, and helpfulness. Previous research has shown that consumer satisfaction can significantly affect company performance and growth (Gustafsson, Johnson, and Roos 2005). Building on this insight, our study is among the first to reframe evaluation as a consumer-centric task, using the precision and scalability of similarity-based metrics while establishing assessments in dimensions that matter to users.

PROPOSED QA FRAMEWORK

In this section, we begin by describing our proposed framework, how it works, and an illustration of answers generated by this framework versus human answers. Figure 2 provides an overview of our QA framework. Two consumer groups are central to this framework: post-purchase consumers, who have used the product and shared their experiences through written reviews (populating the review database), and pre-purchase consumers, who ask the questions. The LLM serves as a bridge, connecting pre-purchase consumers, who typically pose specific yet uncertain queries, with the rich, experience-based content generated by post-purchase reviewers.

Our framework consists of three key modules: a RAG component to incorporate reviews; a question-review matching module, which enhances RAG by improving retrieval quality; and an answerability classifier to filter out unanswerable questions. First, to obtain product-specific information, we incorporate RAG into the QA setting to use reviews as the context. RAG relies on a similarity score to retrieve the most relevant reviews for a given question.

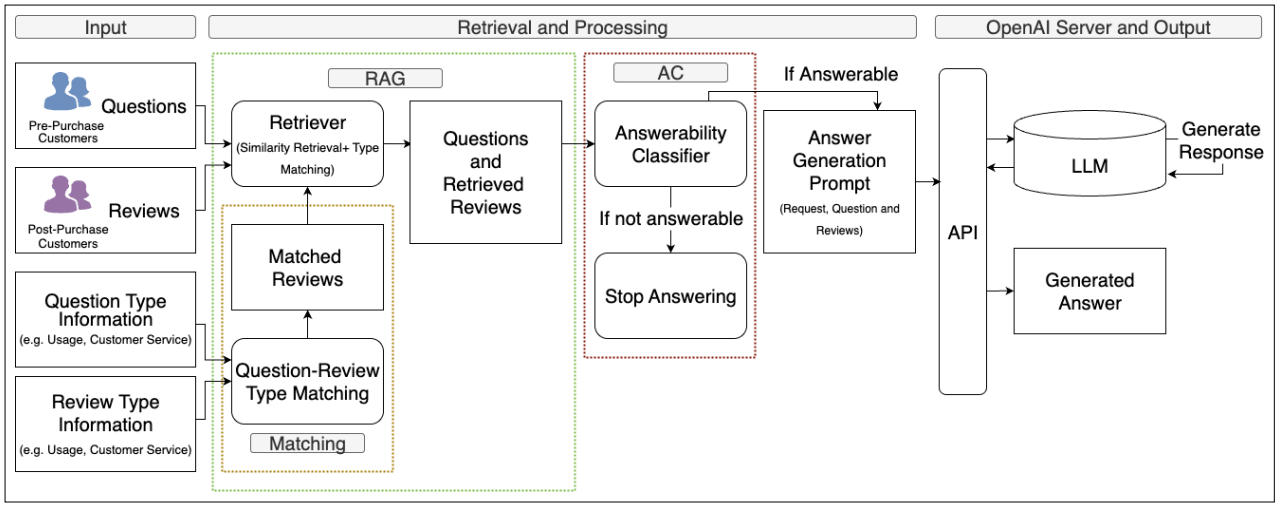

Figure 2: Our Proposed Answer Generation Process

Both the question and associated post-purchase reviews are embedded into dense semantic representations. These reviews are then concatenated with the original question and passed on to the language model, which generates an answer based on user experiences.

Second, in addition to similarity-based retrieval that we leverage using RAG, we enhance the alignment between questions and relevant consumer reviews by incorporating shared category structures between questions and reviews. This "matching" module pairs a question related to product specifications with reviews that also pertain to product specifications.

Third, the proposed framework precludes the possibility that the model generates an unreliable answer (e.g., when the QA does not have the required information). The answerability classifier ensures that the QA system "admits" that it is unable to answer a question. Details pertaining to the technical implementation of each module are provided in the *Methodological Foundations* section.

Figure 3 demonstrates how each component incrementally improves answer quality and how our framework harnesses post-purchase reviews to provide more accurate responses. In this illustration (an actual example from our data), a consumer is asking whether comfort pads for jackets come as a pair or are sold as singletons ("Is this for 1 or 2?"). Because the question concerns order delivery and order fulfillment, it is classified under Customer Service.

In the figure, *Answer Source* indicates the model used to produce the answer and retrieve reviews. *Generated Answer* contains the model's response, with "No Answer" shown when the question is deemed unanswerable. *Subset of Retrieved Reviews* lists illustrative examples from the 25 reviews extracted by the model with their types.

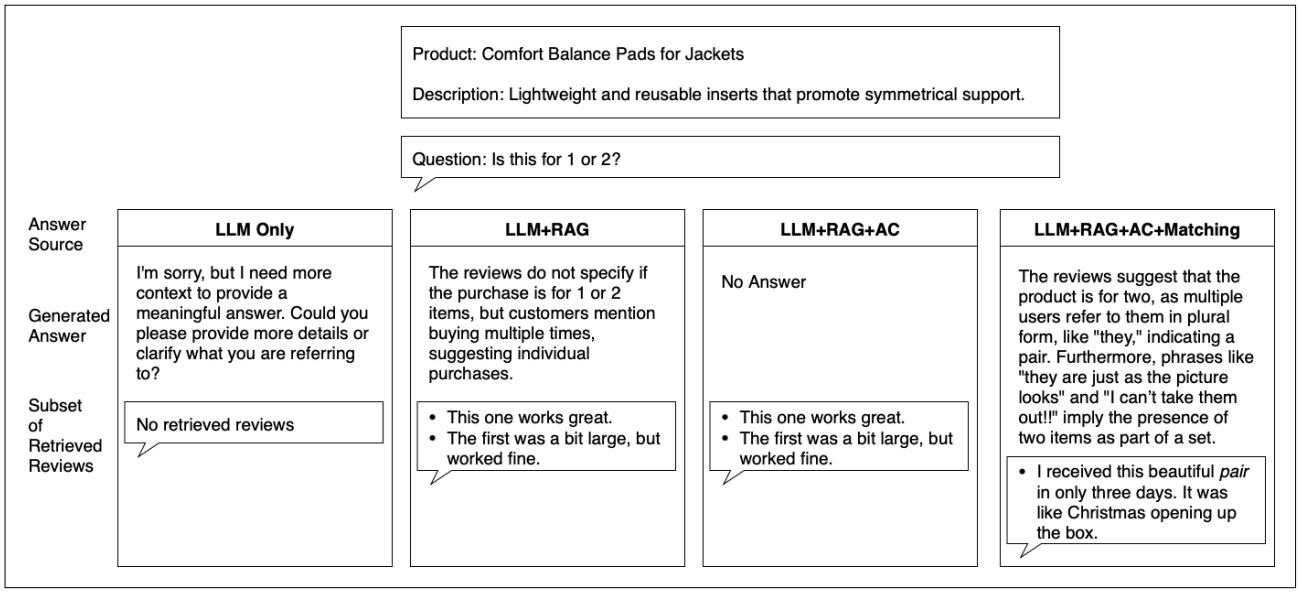

Figure 3: Example LLM-Generated Answers and Retrieved Reviews for a Product Quantity Question

The LLM-only model, without any extracted reviews, fails to provide a response, requesting additional context, possibly because it did not quite understand the question, which is terse and difficult to comprehend. Incorporating RAG improves the LLM's ability to grasp the question by relying on consumer reviews. However, the answer with reviews retrieved by RAG remains inconclusive, as RAG retrieves reviews that are semantically related to the notion of quantity, such as "this *one* works great" and "The *first* was a bit large", which is lexically similar to the question: "Is this for *1* or *2*?" but these reviews do not clarify whether the product includes one or two pads.

The answerability classifier helps here by filtering out questions lacking sufficient evidence, reducing the risk of hallucinated or speculative responses. However, because it does not improve retrieval, the supporting reviews remain unchanged; thus, LLM+RAG+AC refuses to answer when the available evidence is insufficient.

Finally, with matching, the system retrieves more directly relevant reviews under the same customer service type on order fulfillment, such as "I received this beautiful *pair*," enabling a clear and accurate response. Together, these three modules, RAG, answerability classifier, and type matching, enhance the accuracy and reliability of the generated answers in QA contexts.

Answers from Our QA Framework: An Illustration

We begin by describing the characteristics of a high-quality answer in the QA context. To illustrate these characteristics, we present two representative examples, one objective and one subjective, that include both human and LLM responses. For LLM-generated responses, we include answers from two models: the LLM-only baseline and our full model. For human responses, all answers were rated on a 1–5 scale by individual raters, from which we selected the highest-rated (best), lowest-rated (worst), and the response with a rating closest to the average score (average) to capture the range of human answer quality.

Table 2 illustrates an objective question a potential buyer asked for an Internet radio: 'Does this model have an Ethernet port?' A total of 10 humans answered this question. The best human answer is clear, direct, and supported by contextual details, including configuration options and personal use. In contrast, average human answers are factually correct but brief, while low-quality ones can be off-topic. For AI-generated responses, the LLM-only response is vague and drifts into unrelated device categories. In contrast, our full model delivers a concise and accurate response based on retrieved product information. It directly addresses the user's concern and provides a clear, confident statement that mirrors the effectiveness of the best human answer. This comparison suggests that a high-quality answer to objective questions should be clear, informative, and relevant to the question. In addition, contextual details can enhance credibility.

Table 3 presents a subjective question a potential buyer asked for a nail polish kit: "Do your own nails grow with this product? Mine are very soft." Unlike factual queries, this

Table 2: Comparison of Answers for an Objective Question Example

Objective Question Example	
Product and Question	**Product**: SiriusXM TTR1 Internet Radio **Introduction**: SiriusXM TTR1 Tabletop Internet Radio streams SiriusXM via Wi-Fi/Ethernet, with dual alarms, a large display, and remote control. **Question:** Does this model have an Ethernet port? I do not have WiFi.
Human Answers	**Best Human Answer:** Yes, you can set it up for either wired or not wired from the menu. It has an Ethernet connection in the back. I have two, one wired and one Wi-Fi. Works great either way. Enjoy! **Average Human Answer:** Yes. It does have an Ethernet port on the back. **Worst Human Answer:** Not sure. I had mine hooked up wirelessly. I had problems with both of them. Finally threw them away. I love Sirius. I now have a radio that is probably meant for a car. It is connected to a boom box with an antenna ran outside. Works perfect.
LLM-only Answer:	To determine whether a specific model has an Ethernet port, you would need to provide the exact model name or number of the device you're asking about. Generally, many desktop computers and some laptops come with built-in Ethernet ports, while many modern ultrabooks, tablets, and smartphones do not.
Full Model (LLM + Reviews):	Yes, this model has an Ethernet port, allowing it to connect to a wired network if you do not have WiFi.

Notes. The quality of human answers in these two examples was assessed by a group of raters. We describe this in greater detail in the evaluation section.

type of question invites personal opinions and experiential insights. The best human answer is direct, detailed, and empathetic: it explains how the product supports nail growth. In contrast, the average human answer is relevant but lacks context or depth, while the worst answer is vague and disconnected. The LLM-only response offers generic health advice; however, the lack of product-specific context makes it unhelpful for consumers looking for reassurance about this particular item. In comparison, our full model produces an answer that is direct, informative, and empathetic, synthesizing the feedback of multiple users into a balanced, supportive response. This comparison highlights that effective answers to subjective questions should be relevant, grounded in experience, and conveyed in a conversational yet balanced tone. Although personal anecdotes can enhance credibility, synthesized insights from multiple users often provide greater representativeness.

Across the two examples, a high-quality answer in the QA setting—whether responding to objective or subjective questions, should be direct, relevant, clear, informative, and contextually grounded. LLMs are capable of producing fluent and coherent responses, while reviews offer rich product-specific insights based on real-world experiences. The three main

Table 3: Comparison of Answers for a Subjective Question Example

Subjective Question Example	
Product and Question	**Product:** SensatioNail Starter Kit **Introduction:** SensatioNail Gel Nail Polish Kit offers long-lasting gel manicures with an LED lamp, suitable for various gel polish brands. Compact, efficient, salon-quality results. **Question:** Do your own nails grow with this product? Mine are very soft. Thanks.
Human Answers	**Best Human Answer:** Your nails will be hard due to the GEL so they will not rip/tear as long as the GEL is on. When you remove the GEL you will need to be careful because your natural nail will actually be a bit weaker/thinner than usual... (truncated for brevity) **Average Human Answer:** Unfortunately that happened to my nails too. I think it is the acetone. :(**Worst Human Answer:** Mine were very soft and brittle after removing the product.
LLM-only:	To help your nails grow stronger, you might want to look for nail products that contain ingredients known to promote nail health, such as biotin, keratin, or calcium... (truncated)
Full Model (LLM + Reviews):	Yes, many users have reported that their nails have grown longer and stronger with the use of this gel product. Reviewers noted that their nails feel stronger with the gel protection, and the product helped prevent issues like peeling and splitting.

Notes. The quality of human answers in these two examples was assessed by a group of raters. We describe this in greater detail in the evaluation section.

modules of our framework, RAG, answerability classifier, and type matching, enhance the accuracy and reliability of the generated answers in QA contexts. We demonstrate this in the empirical section. Next, we describe the data we use in this paper to test our framework.

DATA DESCRIPTION

We employ the QA dataset introduced by Gupta et al. (2019), which contains question–answer pairs collected from the QA sections of Amazon product pages. The dataset covers a broad range of products, from utilitarian products to hedonic, and includes objective and subjective questions, making it representative of real-world consumer inquiries.

From this dataset, we randomly selected 500 questions in four product categories: beauty, electronics, clothing, and cell phones and accessories while ensuring that the sample is balanced across categories. These questions are accompanied by a total of 2,049 human-provided answers. Each question was answered by multiple users, averaging around 4 responses per question. For each question, we identified the corresponding product and collected all avail-

able consumer reviews for that product, we then segmented the reviews into individual sentences, as a single review can address multiple topics. This resulted in a total of 14,083 single-sentence reviews.

To support question–review type matching and deeper analysis of the data, we categorize questions into a set of *types* (e.g., usage, customer service) that can also be applied to reviews. Since questions typically define the thematic focus of an interaction, identifying their type enables consistent alignment with reviews under the same classification scheme. A key challenge in this process is that questions often span two distinct dimensions: the product category (e.g., Electronics, Beauty) and the user's intent (e.g., Usage, Fit).[3] Previous studies have addressed this issue by leveraging pre-established categories provided by platforms or users, such as "product," "delivery," or "returns" (Kim et al. 2022; Cao et al. 2012). In contrast, our dataset does not contain predefined question types.

To capture both dimensions in a data-driven manner, we employ Top2Vec (Angelov 2020), an unsupervised topic modeling approach that clusters text into latent semantic topics. Running it on 12,000 questions from the same AmazonQA dataset, we first extract 83 fine-grained clusters. We then consolidate these into 10 broader topics using the hierarchical structure provided by Top2Vec (details of the topic analysis are provided in Web Appendix A) and further group them into five core question types. Based on these definitions, we use GPT-4o to annotate our sampled 500 questions and the corresponding reviews, following the prompt structure in Web Appendix A. Notably, a single question can be assigned to multiple types, whereas each review sentence—being a single unit of text—is assigned to only one type. Prior work has shown that LLMs can serve as accurate and efficient annotators with higher accuracy and inter-coder agreement than crowd workers at a fraction of the cost in general text-annotation tasks including topic identification tasks (Gilardi, Alizadeh, and Kubli 2023) and some domain-specific labeling tasks (Aguda et al. 2024).

Table 4 summarizes the definitions and distribution of the type of questions. Approx-

[3]For example, the question "How can I use e-sim cards on iPhone 16?" belongs to the product category Cell Phones and Accessories and the question type could be Usage.

imately 70% of the questions fall into a single type, with 30% having multiple types. To further characterize the questions, we annotate each as subjective or objective, depending on whether it seeks internal impressions or externally verifiable facts (Bjerva et al. 2020). The final column of Table 4 reports the proportion of subjective questions for each type. For subsequent analysis, we use Specs and Features as the exemplar objective category of questions (Percentage of subjective questions=8.67%) and User Experience and Feel as the exemplar subjective category of questions (Percentage of subjective questions=85.58%).

Table 4: Summary of Question Types

Question Type	**Definition**	**Question Count**	**Number of Review Sentences**	**% of Subjective Questions**
Specs and Features	The technical attributes and functional characteristics that describe what the product is and what it can do.	150	2,078	8.67%
Usage	Guidance on how to use the product, including setup, troubleshooting, and advice on buying and usage.	145	2,722	33.10%
Fit and Compatibility	Information about the product's compatibility with other items and systems, and clothing fit.	197	1,870	21.32%
Customer Service	Contents about customer service and support, including delivery, authenticity, pricing, warranty, and returns.	50	1,232	10.00%
User Experience and Feel	Subjective feedback on the product's performance and sensory appeal, including looks, feel, and other aesthetic impressions.	104	6,181	85.58%
Total	Total Counts across all the questions	500	14,083	27.60%

Notes. "Question Count" reflects the number of questions classified under each type; because questions may belong to multiple types while each review sentence is assigned to only one type, the totals exceed the 500 unique questions in the dataset. "Review Sentences" refers to single-sentence reviews.

METHODOLOGICAL FOUNDATIONS

Building on the three key components of our framework, we implement five model variants, progressing from a baseline LLM-only setup to a fully enhanced system. Figure 2 illustrates the architectures of the basic and full models. We start with the LLM-only model, where the question is sent directly to GPT-4o without any external context, relying solely on the model's internal knowledge. All model outputs are generated with a fixed temperature of .5.

We then gradually add components until reaching our full model.

Retrieval-Augmented Generation (RAG)

We begin by enhancing the LLM-only model by adopting a retrieval-augmented generation, which enhances language model output by retrieving relevant external information. In our customer-facing QA setting, we use post-purchase consumer reviews as this external knowledge source. This adaptation results in the LLM+RAG model, where responses are generated based on both the question and the reviews retrieved (Gao et al. 2023).

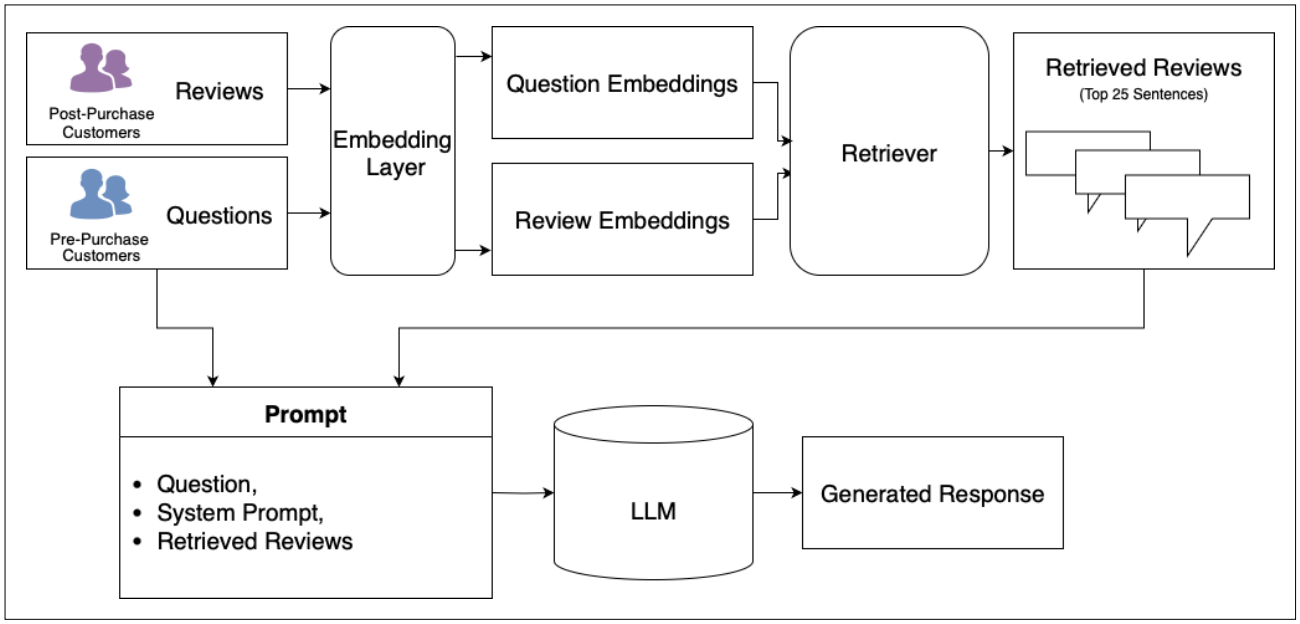

Figure 4: Detailed Process of RAG

Figure 4 shows the detailed process of RAG, we begin by passing both the input questions and reviews to the retriever. Specifically, we embed each question and its associated reviews into embeddings using `text-embedding-ada-002`, a state-of-the-art model optimized for semantic similarity. Let $Q = \{q_i\}$ denote the set of all questions and $R = \{r_j\}$ denote the set of all available reviews. For each question $q_i \in Q$ and review $r_j \in R$, we compute the inner product similarity between their embeddings as follows:

$$\text{Similarity}_{\text{IP}}(q_i, r_j) = \frac{q_i \cdot r_j}{|q_i||r_j|}, \tag{1}$$

Reviews with the highest similarity scores indicate stronger semantic alignment. Using Equation 1, we select the top 25 reviews, concatenate them with the original question, and pass the combined input to OpenAI API to generate the answer. Table 5 shows the structure of our prompt.

Table 5: The prompt structure of our RAG system

Type	**Prompt**
Question	Does this phone take decent pictures?
System Prompt	Based on the following customer reviews, can you provide an answer to the question below? Summarize the reviews and ensure the response is clear and brief while addressing the key points.
Select reviews (from RAG)	- the camera is good quality, also the front camera. You can take panoramic photos. - At first, my pics were coming out too dark. I went to the tools and found I could lighten them with the brightness which was of course good. - This phone is one of the best! is fast, takes great pictures and is thin, not heavy. - ...

Notes. The RAG system retrieves 25 review sentences for each question; only a subset is shown here for illustration.

Answerability Classifier (AC)

To further enhance answer quality, we extend the LLM+RAG model with an answerability classifier, resulting in the LLM+RAG+AC model variant. In this setting, after obtaining relevant reviews through RAG, the model first determines whether the combined information, comprising the question, the retrieved reviews, and the internal knowledge of the LLM, is sufficient to generate a reliable answer. If not, it refrains from responding.

Formally, given a question q_i, and our set of reviews $R = \{r_j\}$, we define $RS_i \subset R$ as the subset of reviews retrieved for q_i. Together with internal knowledge of the LLM D, the model estimates the probability of generating a valid answer a_i as:

$$P(a_i \mid q_i, RS_i, D) > \theta \tag{2}$$

where θ is an internal threshold judged by LLM itself. If this threshold is not met, the model returns 'No Answer' to avoid generating unsupported or potentially misleading content. In practical deployment, such unanswered queries can be escalated to human agents or surfaced

to sellers and post-purchase consumers for follow-up.

Our approach leverages the generative flexibility of the LLM to perform both answerability prediction and answer generation within a unified framework. Unlike previous work, such as Gupta et al. (2019), which trains a separate machine learning classifier using a fixed set of reviews, AC offers several advantages. First, it eliminates the need to pre-train a dedicated classifier. Second, it ensures consistency by using the same retrieved reviews for both decision and generation. Third, it remains highly adaptable: as LLMs continue to evolve, their internal knowledge and generation quality improve, making dynamic answerability assessment more effective than fixed-rule systems.

Question Review Type Matching

Although the answerability classifier ensures that only high-confidence responses are generated, the overall effectiveness of a RAG system is still fundamentally constrained by the quality of retrieved reviews. If the retrieved context is irrelevant or uninformative, even a capable generation model may fail to produce a meaningful answer. This highlights a key limitation of relying solely on lexical similarity for review retrieval: reviews may share surface-level wording with a question but fail to address its underlying intent or information need.

To address this challenge, Cao et al. (2012) show that incorporating category information can enhance the retrieval of similar questions in Community Question Answering systems like Yahoo! Answers. Building on this insight, we extend their approach to the customer-facing QA setting by improving the alignment between questions and consumer reviews. Rather than using fixed, pre-defined categories, we introduce a scalable annotation and matching framework that identifies and aligns the underlying types of both questions and reviews. This allows us to enhance retrieval quality by prioritizing semantically relevant reviews, improving RAG performance in diverse and marketing-relevant contexts.

Building on the LLM+RAG+AC framework and the annotated question/review types

introduced in the Data Description section, we implement two type-aware matching strategies that transform the similarity score in Eq. (1) into a type-adjusted relevance score. We then select the top 25 review sentences based on this relevance score.

Our primary approach is weighted matching, which adjusts the similarity score based on whether the review and question share the same type. Let $Q = \{q_i\}$ and $R = \{r_j\}$ denote the sets of all questions and reviews, respectively. Let $T = \{t_1, t_2, t_3, t_4, t_5\}$ represent the predefined set of five types of question or review (e.g. specs and features, usage, etc.). Each question $q_i \in Q$ and the review $r_j \in R$ is associated with a type label from T, denoted $\text{type}(q_i) \in T$ and $\text{type}(r_j) \in T$, respectively.

$$\text{Relevance}_{\text{weighted}}(q_i, r_j) = \begin{cases} w + \frac{q_i \cdot r_j}{|q_i||r_j|}, & \text{if type}(q_i) = \text{type}(r_j) \\ \frac{q_i \cdot r_j}{|q_i||r_j|}, & \text{otherwise} \end{cases} \tag{3}$$

Intuitively, weighted matching increases the score of type-consistent reviews, ensuring that more relevant reviews are ranked higher. We set $w = .1$ based on a pilot grid search, where it achieved the best empirical performance.

For comparison, we implement *exact matching*, a stricter model that retains only reviews of the same type as the question:

$$\text{Relevance}_{\text{exact}}(q_i, r_j) = \begin{cases} \frac{q_i \cdot r_j}{|q_i||r_j|}, & \text{if type}(q_i) = \text{type}(r_j) \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

Although exact matching enforces strict alignment between question and review types, it risks discarding useful evidence, particularly when the question type boundaries are ambiguous or the annotations are noisy. In contrast, weighted matching strikes a balance between precision and recall by softly prioritizing type-aligned content without excluding semantically relevant material that may fall outside rigid type definitions. By boosting the similarity scores of type-consistent reviews, weighted matching improves retrieval precision, especially

in cases where relevant content may be overshadowed by lexically similar but topically irrelevant text. This type-aware reweighting improves alignment between the retrieved evidence and the intent of the question.

For example, consider the question *"Does this jacket have inner pockets?"* (type: *Specs and Features*). Without question type matching, retrieval based solely on semantic similarity may rank higher an unrelated review such as "I love wearing this jacket when hiking in the mountains" (type: *Usage*) because it shares more surface words (e.g., "jacket") with the question. In contrast, a truly relevant review "It has two deep inside storage space" (type: *Specs and Features*) uses different wording for "inner pockets," lowering its similarity score. The advantage of weighted matching is that it boosts the score of type-consistent reviews, ensuring that relevant evidence is ranked higher.

These final enhancements yield our final two models: LLM+RAG+AC+Weighted Match and LLM+RAG+AC+Exact Match, which integrate generation, answerability assessment, and type-aware retrieval into a unified and adaptive QA system.

Evaluation

Having described our different models, we next outline two different ways in which we evaluate their performance.

Automatic Evaluation (ROUGE-based)

We begin our evaluation using ROUGE-L to assess how closely LLM-generated responses align with high-quality benchmark human responses. ROUGE-L is a widely used automatic evaluation metric that captures sentence-level overlap based on the longest common subsequence (LCS) (Lin 2004). A higher ROUGE-L score indicates a greater similarity between the generated and benchmark responses.

To compute ROUGE-L, each generated answer a_g is compared with its corresponding reference answer a_f by identifying the longest common subsequence (LCS) between them.

The LCS captures the longest sequence of words that appear in both texts in the same order, though not necessarily contiguously (details of the LCS computation are provided in Web Appendix B). Based on LCS, ROUGE-L is computed using:

$$\text{ROUGE-L} = \frac{(1+\beta^2) \cdot \text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}}, \tag{5}$$

where precision and recall are calculated as:

$$\text{Precision} = \frac{\text{LCS}(a_g, a_f)}{|a_g|} \tag{6}$$

$$\text{Recall} = \frac{\text{LCS}(a_g, a_f)}{|a_f|} \tag{7}$$

Here, $|a_g|$ and $|a_f|$ represent the lengths of the generated answer and the reference answer, respectively. The parameter β adjusts the relative importance of recall versus precision, and we set $\beta = 1$ in our study, which balances precision and recall equally. In most QA settings, each question is answered by many individuals. To establish a good benchmark human response against which an LLM answer could be compared, we initially considered using the most upvoted answer per question as the reference. However, this approach proved unreliable due to the highly skewed nature of our dataset: 72% of answers received zero upvotes, and an additional 21% received only one. To make matters worse, many top-voted responses were off-topic or uninformative.[4] Given these limitations, we adopted a model-assisted strategy to identify a high-quality human answer benchmark. Using GPT-4o, we determined the benchmark human answer for each question from all available human responses. The LLM answer was then evaluated against this benchmark human answer selected using the ROUGE-L measure. This approach automated the entire evaluation process.

[4]For instance, in response to the question "Is it free shipping the TV to Ireland?", the most upvoted answer reads: "Unfortunately, they have discovered that it is actually cheaper to ship Ireland to the television." While humorous, it provides no informative value.

Human Evaluation

Although automatic metrics like ROUGE-L provide a standardized way to compare LLM-generated and human answers, they may be biased and fall short in capturing critical dimensions such as nuance, tone, and user-centric value (Yang et al. 2018; Sulem, Abend, and Rappoport 2018). To address these limitations and better reflect user perceptions, we perform a human evaluation in addition to automatic evaluation.

The evaluation of LLM-generated answers to product-related questions is closely related to the work in community QA and chatbot evaluation. To establish our evaluation criteria, we draw on insights from these literature streams as well as our own concept development. In community QA, Kim and Oh (2009) identify content quality and emotional resonance as the drivers of the selection of the "best answer" The SERVQUAL framework (Ladhari 2009) emphasizes trust and reliability. Chatbot research highlights the importance of informativeness, communicability, and social appropriateness (Chaves and Gerosa 2021; Han, Yin, and Zhang 2023). In addition, Deng, Zhang, and Lam (2020) and Banerjee, Dellarocas, and Zervas (2021) underscore the role of helpfulness in shaping user evaluations of AI-generated content.

Based on this literature we use Clarity, Relevance, and Informativeness to capture overall content quality. We include Trust to reflect emotional resonance and perceived reliability, and Helpfulness to measure the answer's ability to reduce uncertainty. The definitions and references for these dimensions and how we measure them are provided in Table 6. In a study we describe later, human annotators evaluated both consumer-written and LLM-generated answers on these five dimensions using a 1–5 Likert scale (strongly disagree–strongly agree).

Table 6: Evaluation Metrics for Assessing Answer Quality

Metric	Definition	Item measurement (1-5 scale)
Relevance (Liu et al. 2020)	The degree to which an answer is related to the question and addresses its core aspects.	How well does the answer address the question and its key aspects?
Clarity (Kim and Oh 2009)	Concise and to the point with a clear explanation.	Is the answer concise, to the point, and well-explained?
Informativeness (Deng, Zhang, and Lam 2020)	Informativeness refers to how detailed and useful the information provided is.	How detailed and useful is the information provided?
Trust (Seymour et al. 2024; Ladhari 2009)	A belief or assessment that an entity (e.g., AI or human) is worthy of trust.	Does the answer inspire confidence in its accuracy and credibility?
Helpfulness (Deng, Zhang, and Lam 2020; Banerjee, Dellarocas, and Zervas 2021)	How helpful is the generated answer to the user?	How well does the answer clear confusion or provide useful guidance?

RESULTS

Automatic Evaluation Results

The answers we evaluated were drawn from six sources: (1) human responses, and (2 to 6) five model variants: LLM only, LLM + RAG, LLM + RAG + AC, LLM + RAG + AC + Weighted Match, and LLM + RAG + AC + exact match. These variants start with an LLM-only base model and progressively incorporate RAG, answerability classification, and matching components to form the full model. Our automatic evaluation covers 500 questions and 2,500 QA pairs for five LLM-based models, using the best human answers as references. For systems equipped with an Answerability Classifier (AC), ROUGE-L scores are calculated only for questions predicted to be answerable. We also report the proportion of questions each model answers, providing insight into its selectivity.

The results of the automatic evaluation in Table 7 highlight the contribution of each component in our framework. Compared to the LLM-only baseline, incorporating RAG substantially improves the ROUGE-L scores (from .105 to .143, with $p < .001$), demonstrating the value of grounding answers in post-purchase reviews. Adding an answerability classifier further significantly improves performance by filtering out unsupported questions, though it reduces the answer rate to 62.6%.

Table 7: ROUGE-L Scores and Answered Question Percentage for LLM Answers

Answer Type	ROUGE-L	% Answered	p-value
LLM+ Only	.105 (.002)	100%	/
LLM+RAG	.143 (.003)	100%	.000
LLM+RAG+AC	.159 (.005)	62.60%	.004
LLM+RAG+AC+ Weighted Match	.158 (.005)	64.40%	.008
LLM+RAG+AC+ Exact Match	.163 (.005)	56.20%	.001

Notes. Values are reported as mean (standard error). For p-value of LLM answers, LM+RAG is compared with LLM only, the remaining variants with RAG are compared with LLM+RAG. For answers with AC, only the part it answered is compared.

Adding the weighted module increases coverage to 64.4% while maintaining comparable ROUGE-L scores with other model variants, indicating greater reach without sacrificing quality. Importantly, much of this added coverage comes from questions that other models leave unanswered, typically harder questions, suggesting the gain is due to better retrieval. Although the exact match yields the highest ROUGE-L score, it answers only 56.2% of the questions. Given this trade-off, we focus on the weighted match variant in subsequent analyses, as it offers the best balance between answer quality and coverage. In general, our full model—with RAG, answerability classifier, and weighted matching—delivers the best performance across metrics.

Results by Question Types

We also analyze the results reported in Table 7 by question type to learn how our model performs across various kinds of questions, from objective to subjective. Figure 5 shows ROUGE-L scores across question types, and Web Appendix C provides detailed results. A consistent pattern across question types is that our proposed approach leveraging reviews improves performance on the similarity measure (ROUGE-L). Here LLM-only model serves as the baseline against which we compare our models. However, the gains are uneven; the question type category of Specs and features (mostly objective questions) shows greater

gains, while lesser gains accrue for the question type category of User experience and feel (mostly subjective questions).

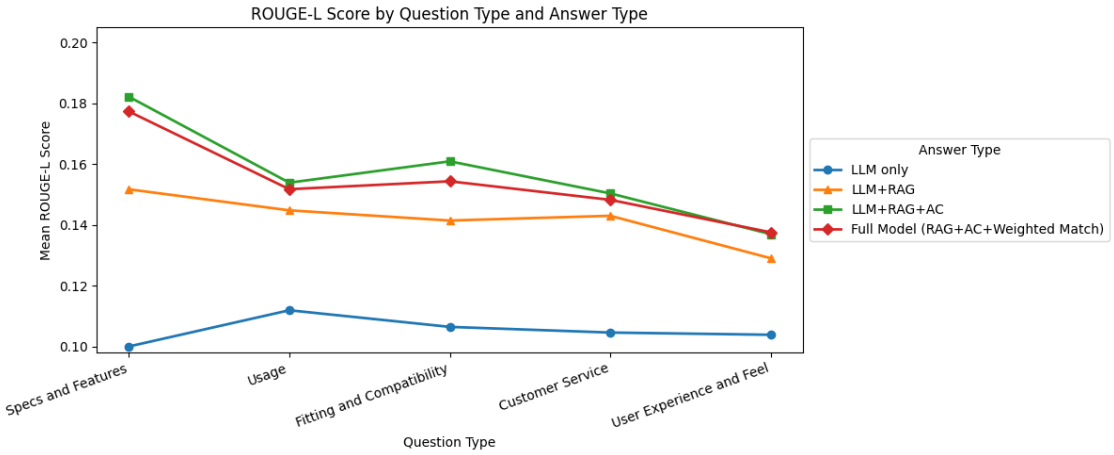

Figure 5: ROUGE-L Score across Question Types

For Specifications and features, for which most questions are objective, the ROUGE-L scores improve significantly from .100 (LLM only) to .177 (full model), representing a 77% increase ($p < .001$). This indicates substantial gains from our RAG and weighted matching. In contrast, for User Experience and Feel, for which most questions are subjective, the ROUGE-L scores show a smaller, but still significant, improvement from .104 to .138—a 33% increase ($p < .001$).

These results based on a lexical similarity metric suggest that our full model is more effective for objective questions, likely because relevant information is more consistently available and easier to retrieve. However, for subjective questions, reviews are likely fragmented or emotionally expressive, making answerability harder and lexical overlap more limited. This aligns with previous findings that LLMs tend to perform better on objective tasks than on subjective ones (Bjerva et al. 2020; Longoni and Cian 2022).

Although ROUGE-L results reinforce the view that LLMs perform better on objective questions, they may underestimate answer quality for subjective questions. In open-ended, opinion-driven questions—such as those in User Experience and Feel—low lexical overlap may reflect stylistic diversity rather than lower performance. These limitations underscore

the importance of complementing automatic metrics with human evaluation, which can better capture qualities such as nuance, tone, and contextual relevance. We report on these in the next section.

Human Evaluation Results

At this point, we have answers from six sources: (1) human responses (with multiple answers per question), and (2–6) five model variants—LLM-only, LLM+RAG, LLM+RAG+AC, LLM+RAG+AC+Weighted Match, and LLM+RAG+AC+Exact Match. In the study we describe next, we focus on data from five of these sources. We dropped LLM+RAG+AC+ Exact Match from further consideration because LLM+RAG+AC+Weighted Match performed better in the analyses reported in the previous section. We treat these as five experimental conditions and ask humans to evaluate the answers from these conditions. To accomplish this, we conducted a human evaluation study between subjects on Prolific with these five conditions. To manage evaluation costs, we randomly selected 200 questions, 40 from each of the five types of questions, resulting in a total of 1,588 responses from human and model sources.

We recruited a total of 582 participants and each respondent was asked to evaluate five randomly selected QA pairs on the five evaluation dimensions listed in Table 6. Because several participants sometimes evaluated the same QA pair, we combined their ratings by taking the average score for that QA pair. To minimize bias, the source of each answer (human vs. LLM) was not disclosed to the participants. Each QA pair was presented alongside the product name and a brief product description to provide sufficient context, an example question from the survey is shown in Appendix A.

For each condition, we report the average score on the five evaluation dimensions. In addition, we define the Over-Human Ratio as the percentage of times a model-generated answer is rated higher than or equal to human answers to the same question. For each of the five LLM variants, we compare its answer against all available human responses per

question and compute the proportion it outperforms. For instance, if an LLM's answer is rated better than four out of five human answers, its Over-Human Ratio for that question is 80%. Averaging this percentage across all questions yields the Over-Human Ratio for that model. This metric offers a fine-grained assessment of how frequently each model produces responses that exceed the quality of typical human answers.

Because most questions are associated with multiple human answers, we further contextualize model performance by reporting three reference points: the overall average score across all human answers, the average score of the highest-rated human answer per question (best human), and the average score of the lowest-rated human answer per question (worst human).

Table 8 presents the results of the human evaluation of a wide variety of QA answers. Focusing first on the "Average score" column, the quality of the human answer varies widely, with scores ranging from 2.42 for the worst human responses to 4.34 for the best. The average human score is 3.50, which is nearly identical to the LLM-only score of 3.50 ($p = .998$).

Table 8: Human Evaluation Results by Answer Source

Answer Source	Average Score	Clarity	Informativity	Relevance	Trust	Helpfulness	% Over Human
Best Human Answer	4.34 (.05)	4.41 (.06)	4.29 (.06)	4.51 (.05)	4.11 (.06)	4.38 (.06)	-
Human Answer	3.50 (.04)	3.50 (.04)	3.43 (.04)	3.71 (.04)	3.37 (.04)	3.46 (.04)	-
Worst Human Answer	2.42 (.07)	2.38 (.08)	2.35 (.08)	2.67 (.08)	2.42 (.07)	2.27 (.08)	-
LLM Only	3.50 (.08)	3.61 (.08)	3.58 (.09)	3.46 (.09)	3.54 (.07)	3.29 (.10)	53%
LLM+RAG	4.09 (.06)	4.10 (.07)	4.12 (.07)	4.23 (.06)	3.86 (.06)	4.13 (.07)	72%
LLM+RAG+AC	4.11 (.06)	4.16 (.07)	4.17 (.06)	4.36 (.06)	3.80 (.07)	4.06 (.08)	69%
LLM+RAG+AC+Weighted Match (Full Model)	4.14 (.06)	4.17 (.07)	4.17 (.06)	4.32 (.07)	3.84 (.07)	4.19 (.07)	72%

Notes. Values are reported as mean (standard error). For methods incorporating AC, the comparison is based only on the portion of the question they answered.

In contrast, models that incorporate RAG consistently outperform the human average. Our full model, which combines RAG, an answerability classifier, and Weighted Match,

achieves an average score of 4.14, which is significantly higher than the average human baseline of 3.50 ($p < .001$) and begins to approach the best-human performance of 4.34. For the 64.4% questions that it answers (see Table 7), considering coverage, automatic and human evaluation, our full model does an excellent job.

Looking at the five dimensions of evaluations in Table 8, the results show that our full model is capable of generating responses that are relevant, clear, and informative - comparable to the best human answers. For relevance in particular, low-rated responses are often off-topic. This is reflected in the scores: the best human answers scored 4.51 on relevance, while the worst scored only 2.67. The LLM-only model, lacking access to reviews, scored 3.46, which is below the average human score of 3.71. In contrast, our full model reached 4.32, surpassing the human average ($p < .001$) while remaining just below the best human performance. Similar improvements in clarity and informativity are observed.

Our full model scored 3.84 on the trust dimension, outperforming the human average of 3.37 ($p < .001$). This suggests that, when identity is anonymized (that is, we don't reveal who answered the question, a human or an LLM) high-quality LLM answers can be perceived as more trustworthy than human responses, highlighting their potential for consumer acceptance.

In the last column of Table 8, we report the Over-Human Ratio, a metric that quantifies how often a model-generated answer outperforms human answers to the same question. The results show that the LLM-only model answers meet or exceed human responses in 53% of cases, and our full model significantly increases this ratio to 72% with $p < .001$, thus exceeding the majority of human responses. This, we believe, is an impressive result.

Analysis by Question Types

To better understand the performance of the model across different types of consumer queries, we analyze the results by question type, focusing on subjective versus objective questions. Table 9 provides detailed comparisons between these two representative question types (re-

sults for all question types are provided in Appendix B). For objective questions (specifications and features), the best human answers score 4.47, while our full model averages 3.95, with an over-human ratio of 67%. For subjective questions (User Experience and Feel), the best human score is 4.35, and our model achieves 4.26, statistically indistinguishable from the top human level ($p = .492$), and performing as well as or better than 81% of human responses. These results indicate that our model excels in subjective questions, diverging from both prior literature and automatic evaluation trends. This likely occurs because reviews often offer richer contextual information on subjective questions such as user experience than on objective questions such as product specifications.

Table 9: Human Evaluation Results by Answer Source for Subjective and Objective Question Types

Answer Source	Average Score	Clarity	Informativity	Relevance	Trust	Helpfulness	% Over Human
Objective questions: Specs and Features							
Best Human Answer	4.47 (.11)	4.50 (.15)	4.50 (.11)	4.58 (.12)	4.22 (.15)	4.58 (.11)	-
Human Answer	3.65 (.08)	3.58 (.10)	3.61 (.09)	3.88 (.09)	3.51 (.09)	3.68 (.10)	-
Worst Human Answer	2.49 (.17)	2.41 (.18)	2.52 (.21)	2.78 (.21)	2.42 (.18)	2.31 (.19)	-
LLM Only	3.11 (.18)	3.31 (.20)	3.20 (.22)	3.05 (.20)	3.20 (.19)	2.82 (.23)	39%
LLM+RAG+AC+Weighted Match (Full Model)	3.95 (.18)	4.06 (.19)	3.87 (.20)	4.19 (.22)	3.68 (.21)	3.96 (.21)	67%
Subjective questions: User Experience and Feel							
Best Human Answer	4.35 (.09)	4.49 (.10)	4.23 (.12)	4.50 (.11)	4.18 (.12)	4.36 (.11)	-
Human Answer	3.50 (.07)	3.54 (.09)	3.42 (.08)	3.71 (.09)	3.40 (.08)	3.43 (.09)	-
Worst Human Answer	2.63 (.15)	2.50 (.16)	2.57 (.16)	2.93 (.19)	2.67 (.16)	2.46 (.18)	-
LLM Only	3.61 (.15)	3.80 (.16)	3.74 (.20)	3.69 (.16)	3.52 (.15)	3.29 (.22)	55%
LLM+RAG+AC+Weighted Match (Full Model)	4.26 (.10)	4.32 (.13)	4.32 (.10)	4.39 (.13)	3.96 (.13)	4.30 (.12)	81%

Notes. Values are reported as mean (standard error). For methods incorporating AC, the comparison is based only on the portion of the question they answered.

To further understand why our findings diverge from the prior literature, suggesting that LLMs perform well on objective questions but poorly on subjective ones (Bjerva et al.

2020; Longoni and Cian 2022), we revisit the examples presented in Table 2 and 3. In the example involving an objective question (does the internet radio have an ethernet port), the best human response combines factual accuracy with light context, whereas the worst answer fails to address the question. Our full model offers a concise and correct answer, aligning with previous claims about the strength of the LLM in factual tasks. In the example involving a subjective question (do your nails grow with this product) in Table 3, the best human answer is derived from rich personal experience, while the worst answer is uninformative. The LLM-only model provides generic advice. In contrast, our full model synthesizes diverse post-purchase reviews into a balanced, representative response.

These comparisons point to several reasons for the strong performance of our model in subjective questions. First, subjective questions differ from objective ones in that they rarely have a single correct answer. Instead, users seek diverse, experience-based perspectives. Our model addresses this by aggregating multiple post-purchase reviews to generate representative responses, rather than relying on a single personal opinion. Second, the pattern of results also reflects a deeper distinction in how users form trust: For objective, fact-based questions, users often look for authoritative, machine-like precision; for subjective or experiential questions, they value authenticity and 'word of mouth'. By grounding its responses in real consumer experiences, our full model incorporates both aspects well.

Previous work has cautioned against overhumanizing or underhumanizing AI responses (Crolic et al. 2022; Schanke, Burtch, and Ray 2021). Our hybrid framework navigates this tension by combining the fluency of natural language generation with the factual grounding of the retrieved content, resulting in answers that are both expressive and evidence-based. By anchoring responses in consumer reviews, the model assumes the role of a neutral summarizer rather than mimicking human subjectivity, thereby avoiding the pitfalls of humanization. Grounding responses in consumer reviews also enables the model to be especially informative: on informativity within subjective questions, our full model gets a similar rating to the best human answers (4.32 vs. 4.23, $p = .588$), showing capability of generating

meaningful answers. Together, these design choices make our model particularly well-suited for subjective tasks, such as describing how a product feels, smells, or performs in everyday use.

A puzzling aspect of our results so far is that our full model performs very well for objective questions (Specs and features) on the similarity measure (ROUGE-L) in Figure 5 and the opposite is true in the results based on human evaluation in Table 9, where it performs better for subjective questions (User Experience and Feel).

To understand this discrepancy, we examine the over human ratio of the two question types. Table 9 shows that our full model achieves an over-human ratio of 81% of human answers in subjective questions, compared to 67% in objective ones. This suggests that our full model produces particularly strong responses to subjective queries. However, these high-quality answers sometimes diverge from the benchmark because they provide more complete or more useful information than the human responses, reducing lexical similarity (ROUGE-L: .138 vs. .177 for objective questions). Appendix C further illustrates this difference visually: plotting all model outputs and human responses reveals that higher answerability and greater variability in subjective queries make it more difficult for the model to align with human responses.

Importantly, divergence from the reference human answer does not always indicate a lower answer quality. This shows a limitation for the ROUGE-L measure: When an LLM answer is *better* than the human baseline, which occurs for our full model, it results in lower ROUGE-L scores. Intuitively, this occurs because the answer is different from the human answer. Our full model matches or exceeds 72% of the human responses (81% for subjective questions), but this high-quality performance can be *penalized* by similarity-based metrics. We therefore conclude that similarity to human benchmark answer may not always be the aspirational goal of an LLM, and aligning with poor human responses is certainly not desirable. This explains why human evaluation and similarity-based evaluation diverge and illustrates the need to go beyond lexical overlap and rely on human evaluation when

assessing AI-generated responses.

CONCLUSION

Summary

We begin this paper by noting that, from a consumer's point of view, though useful in reducing uncertainties, the answers on consumer-facing QA sections have limitations: Responses are often delayed, only a small fraction of customers contribute and the quality of the responses varies. In contrast, post-purchase reviews are abundant but not tailored to specific customer questions. Their sentiment-driven, unstructured nature makes them rich in context but hard to parse for efficient, query-specific information retrieval.

To address the challenges of customer-facing QA, we introduce and empirically validate a novel LLM-enabled framework that connects pre-purchase inquiries with post-purchase experiences to generate high-quality answers. This framework enriches foundational LLMs by incorporating consumer reviews through Retrieval-Augmented Generation (RAG), enabling responses that are both accurate and contextually grounded. RAG allows the LLM to dynamically retrieve relevant review content at the time of inference. To improve retrieval precision, we add a question–review type-matching module that prioritizes reviews addressing the same topic as the question (e.g., fit, usability). Recognizing that not all questions are answerable, our framework also includes an answerability classifier that assesses whether sufficient supporting information exists before generating a response. Together, these three components—RAG, type-matching, and answerability filtering—form a robust system for generating reliable, review-informed product answers.

We empirically evaluate our proposed customer-facing QA framework using a dataset from Amazon that includes 500 questions, over 2,000 human answers, and around 14,000 single-sentence reviews. We benchmark multiple model variants—ranging from LLM-only to LLM combined with Retrieval-Augmented Generation (RAG), question–review type-matching, and an answerability classifier—using both lexical similarity scores (ROUGE-L)

and human evaluations. While RAG boosts lexical similarity scores by 30%, further gains come from the answerability module and review-type matching, increasing both answer quality and coverage.

Human evaluation provides deeper insights. Our full model achieves an average score of 4.14 on a 5-point Likert scale—surpassing the average human baseline of 3.50 and performing as well as or better than 72% of human responses. It begins to approach the best human answers on metrics such as clarity, relevance, and informativeness, and performs better on subjective (versus objective) questions. One of our key findings is that lexical similarity metrics have shortcomings—they penalize high-quality answers from our framework that deviate from human references. These human references, we show, vary in quality. Some are excellent and many are not. This reinforces the importance of human-centered evaluation in assessing LLM-driven QA systems.

Practical Implications

Traditional customer-facing QA sections, such as those used by Target, JD.com and Tripadvisor, rely on communities of past consumers to respond to questions from potential consumers. These systems suffer from key limitations: responses are often delayed, and the quality of answers can vary widely, from highly informative to misleading or irrelevant. To address these shortcomings, we propose a framework that LLMs trained on extensive public data, enhanced with proprietary review data available to retailers. Although consumers could separately consult LLMs and reviews, causing significant search costs, our framework unifies these two information sources to deliver answers that are more relevant, informative, and user-friendly. Empirically, our system outperforms both the standalone LLM responses and 72% of human-provided answers.

More broadly, this paper demonstrates how firms can strategically combine proprietary consumer-generated content with LLM capabilities to enhance the pre-purchase experience. The effectiveness of our framework is closely tied to the volume and quality of available post-

purchase reviews, underscoring the importance of cultivating a robust review ecosystem. By better matching consumers with products that meet their needs, our approach also has the potential to reduce return rates, a major operational challenge for retailers. Our system is flexible in terms of human involvement. At one end of the spectrum, it can operate autonomously, only generating answers when sufficient evidence is available. On the other hand, it can produce draft responses for human moderators to review and approve. For questions deemed unanswerable by the model, platforms can solicit crowd input, for example, by featuring unanswered questions in review prompts or embedding them in the review-writing interface. This hybrid design preserves space for human judgment in areas where automation is insufficient.

To sustain long-term value, platforms should incentivize users to contribute both reviews and questions. These user-generated texts not only support future buyers, but also enhance model performance by enriching the underlying training and grounding data. Our proposed framework, comprising RAG, an answerability classifier, and a weighted question–review matching module, is particularly effective when review content is diverse and informative.

Finally, platforms should move beyond traditional similarity-based metrics like ROUGE when evaluating customer-facing QA performance. As our findings show, such metrics fail to capture critical dimensions of answer quality, including clarity, trustworthiness, and helpfulness. A hybrid evaluation strategy that combines automated metrics with human-centered assessments is essential. Additionally, prompting users to rate existing answers can improve transparency and provide valuable training signals for future improvements. Ultimately, building effective and trusted QA systems requires not only strong AI models, but also platform-level strategies that foster human-AI collaboration, encourage user engagement, and embrace multidimensional evaluation.

Limitations and Future Research

A positive aspect of the proposed framework is that it does not speculate on an answer. However, this could also be a source of frustration for a potential buyer. A QA framework should strive for an answerability rate greater than 64.4% that our data permit. Here, there are two paths to consider. First, future research could examine how both the quantity and the quality of available reviews influence answerability. As shown in Figure 8 (Appendix D), there is suggestive evidence that the percentage of questions answered within a type increases with the number of review sentences available for that type. However, this evidence is based on five question types, and further work is needed to better understand how both the volume of reviews (overall and within each question type), their quality, and their relevance shape answerability outcomes. Second, the threshold used to classify a question as "answerable" warrants further investigation. A more nuanced or context-dependent threshold may help balance the trade-off between providing precise, evidence-based answers and meeting buyers' expectations for responsiveness.

Another factor that deserves attention is the timing of reviews. Over time, customers' perceptions of a product may shift, and in some cases the product or service itself may change—for instance, a restaurant might hire a new chef, altering the answer to the same question. To address this issue, more recent reviews could be assigned greater weight in retrieval. Furthermore, when a substantive change in the product or service occurs, the platform could set a temporal boundary so that reviews written after the change are prioritized in RAG-based retrieval. Implementing such an approach would require the availability of precise timestamp data in the dataset, which is absent in our case.

In addition to reviews, our framework can be readily extended to incorporate previous QA exchanges as contextual data for answer generation. This extension is promising, as prior answers may provide complementary evidence beyond what reviews alone can offer. However, two challenges arise. First, semantically similar questions (e.g., "how heavy is iPhone 15" and "what is the weight of iPhone 15") should ideally be merged to avoid redundancy, ensuring

that new questions remain meaningfully distinct from prior ones. Second, the temporal validity of earlier QA must be considered, since answers may become outdated once the product or service has been updated.

Taken together, these limitations underscore that future research should not only refine the technical dimensions of retrieval and classification, but also account for the evolving, contextual, and complementary nature of product information.

REFERENCES

Achiam, Josh, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat et al. (2023), "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*.

Aguda, Toyin, Suchetha Siddagangappa, Elena Kochkina, Simerjot Kaur, Dongsheng Wang, Charese Smiley, and Sameena Shah (2024), "Large language models as financial data annotators: A study on effectiveness and efficiency," *arXiv preprint arXiv:2403.18152*.

Angelov, Dimo (2020), "Top2vec: Distributed representations of topics," *arXiv preprint arXiv:2008.09470*.

Arora, Neeraj, Ishita Chakraborty, and Yohei Nishimura (2024), "Express: Ai-human hybrids for marketing research: Leveraging llms as collaborators," *Journal of Marketing*, page 00222429241276529.

Arslan, Muhammad, Hussam Ghanem, Saba Munawar, and Christophe Cruz (2024), "A Survey on RAG with LLMs," *Procedia computer science*, 246, 3781–3790.

Banerjee, Shrabastee, Chrysanthos Dellarocas, and Georgios Zervas (2021), "Interacting user-generated content technologies: How questions and answers affect consumer reviews," *Journal of Marketing Research*, 58 (4), 742–761.

Bjerva, Johannes, Nikita Bhutani, Behzad Golshan, Wang-Chiew Tan, and Isabelle Augenstein (2020), "Subjqa: A dataset for subjectivity and review comprehension," *arXiv preprint arXiv:2004.14283*.

Blanchard, Simon J, Nofar Duani, Aaron M Garvey, Oded Netzer, and Travis Tae Oh (2025), "EXPRESS: New Tools, New Rules: A Practical Guide to Effective and Responsible GenAI Use for Surveys and Experiments Research," *Journal of Marketing*, page 00222429251349882.

Bommasani, Rishi, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill et al. (2021), "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*.

Borji, Ali (2023), "A Categorical Archive of ChatGPT Failures," *arXiv preprint arXiv:2302.03494*.

Brandes, Leif, David Godes, and Dina Mayzlin (2022), "Extremity bias in online reviews: The role of attrition," *Journal of Marketing Research*, 59 (4), 675–695.

Cao, Xin, Gao Cong, Bin Cui, Christian S Jensen, and Quan Yuan (2012), "Approaches to exploring category information for question retrieval in community question-answer archives," *ACM Transactions on Information Systems (TOIS)*, 30 (2), 1–38.

Chakraborty, Ishita, Minkyung Kim, and K Sudhir (2022), "Attribute sentiment scoring with online text reviews: Accounting for language structure and missing attributes," *Journal of Marketing Research*, 59 (3), 600–622.

Chaves, Ana Paula and Marco Aurelio Gerosa (2021), "How should my chatbot interact? A survey on human-chatbot interaction design," *International Journal of Human-Computer Interaction*, 37 (8), 729–758.

Chevalier, Judith A and Dina Mayzlin (2006), "The effect of word of mouth on sales: Online book reviews," *Journal of marketing research*, 43 (3), 345–354.

ConvertCart "Product Page Statistics Every eCommerce Pro Should Know (2025 Update)," (2025) https://www.convertcart.com/blog/ecommerce-product-page-statistics, accessed: 2025-04-03.

Crolic, Cammy, Felipe Thomaz, Rhonda Hadi, and Andrew T Stephen (2022), "Blame the bot: Anthropomorphism and anger in customer–chatbot interactions," *Journal of Marketing*, 86 (1), 132–148.

CustomerThink "From questions to conversions: The surprising impact of QA sections on e-commerce success," (2024) https://customerthink.com/from-questions-to-conversions-the-surprising-impact-of-qa-sections-on-e-commerce-success/, accessed: 2025-08-19.

De Freitas, Julian, Gideon Nave, and Stefano Puntoni (2025), "Ideation with generative AI—in consumer research and beyond," *Journal of Consumer Research*, 52 (1), 18–31.

Deng, Yang, Wenxuan Zhang, and Wai Lam "Opinion-aware answer generation for review-driven question answering in e-commerce," "Proceedings of the 29th ACM International Conference on Information & Knowledge Management," pages 255–264 (2020).

Deng, Yang, Wenxuan Zhang, Qian Yu, and Wai Lam (2023), "Product question answering in e-commerce: A survey," *arXiv preprint arXiv:2302.08092*.

Feuerriegel, Stefan, Abdurahman Maarouf, Dominik Bär, Dominique Geissler, Jonas Schweisthal, Nicolas Pröllochs, Claire E Robertson, Steve Rathje, Jochen Hartmann, Saif M Mohammad et al. (2025), "Using natural language processing to analyse text data in behavioural science," *Nature Reviews Psychology*, 4 (2), 96–111.

Finardi, P., L. Avila, R. Castaldoni, P. Gengo, C. Larcher, M. Piau, P. Costa, and V. Caridá (2024), "The chronicles of RAG: The retriever, the chunk and the generator," *arXiv preprint arXiv:2401.07883* https://arxiv.org/abs/2401.07883.

Gao, Yunfan, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Haofen Wang, and Haofen Wang (2023), "Retrieval-augmented generation for large language models: A survey," *arXiv preprint arXiv:2312.10997*, 2.

Gilardi, Fabrizio, Meysam Alizadeh, and Maël Kubli (2023), "ChatGPT outperforms crowd workers for text-annotation tasks," *Proceedings of the National Academy of Sciences*, 120 (30), e2305016120.

Guo, Yue, Jae Ho Sohn, Gondy Leroy, and Trevor Cohen (2025), "Are LLM-generated plain language summaries truly understandable? A large-scale crowdsourced evaluation," *arXiv preprint arXiv:2505.10409*.

Gupta, Mansi, Nitish Kulkarni, Raghuveer Chanda, Anirudha Rayasam, and Zachary C Lipton (2019), "Amazonqa: A review-based question answering task," *arXiv preprint arXiv:1908.04364*.

Gustafsson, Anders, Michael D Johnson, and Inger Roos (2005), "The effects of customer satisfaction, relationship commitment dimensions, and triggers on customer retention," *Journal of marketing*, 69 (4), 210–218.

Han, Elizabeth, Dezhi Yin, and Han Zhang (2023), "Bots with feelings: Should AI agents express positive emotion in customer service?," *Information Systems Research*, 34 (3), 1296–1311.

Joshi, Mandar, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer (2017), "Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension," *arXiv preprint arXiv:1705.03551*.

Kayali, Moe, Fabian Wenz, Nesime Tatbul, and Çağatay Demiralp (2024), "Mind the data gap: Bridging llms to enterprise data integration," *arXiv preprint arXiv:2412.20331*.

Kim, Eun, Hyejung Yoon, Jungeun Lee, and Misuk Kim (2022), "Accurate and prompt answer-

ing framework based on customer reviews and question-answer pairs," *Expert Systems with Applications*, 203, 117405.

Kim, Soojung and Sanghee Oh (2009), "Users' relevance criteria for evaluating answers in a social Q&A site," *Journal of the American society for information science and technology*, 60 (4), 716–727.

Kwiatkowski, Tom, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee et al. (2019), "Natural questions: a benchmark for question answering research," *Transactions of the Association for Computational Linguistics*, 7, 453–466.

Ladhari, Riadh (2009), "A review of twenty years of SERVQUAL research," *International journal of quality and service sciences*, 1 (2), 172–198.

Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel et al. (2020), "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in neural information processing systems*, 33, 9459–9474.

Li, Peiyao, Noah Castelo, Zsolt Katona, and Miklos Sarvary (2024), "Frontiers: Determining the validity of large language models for automated perceptual analysis," *Marketing Science*, 43 (2), 254–266.

Lin, Chin-Yew "Rouge: A package for automatic evaluation of summaries," "Text summarization branches out," pages 74–81 (2004).

Liu, Xiaomo, G Alan Wang, Weiguo Fan, and Zhongju Zhang (2020), "Finding useful solutions in online knowledge communities: A theory-driven design and multilevel analysis," *Information Systems Research*, 31 (3), 731–752.

Longoni, Chiara and Luca Cian (2022), "Artificial intelligence in utilitarian vs. hedonic contexts: The "word-of-machine" effect," *Journal of Marketing*, 86 (1), 91–108.

McAuley, Julian and Alex Yang "Addressing complex and subjective product-related queries with customer reviews," "Proceedings of the 25th International Conference on World Wide Web," pages 625–635 (2016).

McInnes, Leland, John Healy, and James Melville (2018), "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*.

OpenAI "Enterprise Privacy at OpenAI," (2025) https://openai.com/enterprise-privacy, accessed: 2025-08-14, Updated: 2025-07-04.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu "Bleu: a Method for Automatic Evaluation of Machine Translation," Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, "Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics," pages 311–318, Philadelphia, Pennsylvania, USA: Association for Computational Linguistics (2002) https://aclanthology.org/P02-1040/.

Peter, Lee, Ishita Chakraborty, and Shrabastee Banerjee (2022), "AI applications to customer feedback research: A review," *Review of Marketing Research: Artificial Intelligence and Marketing*.

Puranam, Dinesh, Vrinda Kadiyali, and Vishal Narayan (2021), "The impact of increase in minimum wages on consumer perceptions of service: A transformer model of online restaurant reviews," *Marketing Science*, 40 (5), 985–1004.

Rajpurkar, Pranav, Jian Zhang, Konstantin Lopyrev, and Percy Liang (2016), "Squad: 100,000+ questions for machine comprehension of text," *arXiv preprint arXiv:1606.05250*.

Schanke, Scott, Gordon Burtch, and Gautam Ray (2021), "Estimating the impact of "humanizing" customer service chatbots," *Information Systems Research*, 32 (3), 736–751.

Seymour, Mike, Lingyao Yuan, Kai Riemer, and Alan R Dennis (2024), "Less artificial, more intelligent: understanding affinity, trustworthiness, and preference for digital humans," *Information Systems Research*.

Sulem, Elior, Omri Abend, and Ari Rappoport (2018), "BLEU is not suitable for the evaluation of text simplification," *arXiv preprint arXiv:1810.05995*.

Timoshenko, Artem, Chengfeng Mao, and John R Hauser (2025), "Can Large Language Models Extract Customer Needs as well as Professional Analysts?," *arXiv preprint arXiv:2503.01870*.

Tirunillai, Seshadri and Gerard J Tellis (2014), "Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent dirichlet allocation," *Journal of marketing research*, 51 (4), 463–479.

Yang, An, Kai Liu, Jing Liu, Yajuan Lyu, and Sujian Li (2018), "Adaptations of ROUGE and BLEU to better evaluate machine reading comprehension task," *arXiv preprint arXiv:1806.03578*.

APPENDIX

Appendix A. Human Evaluation Form

In Figure 6, we present an example of the survey interface. The product, its description, and the corresponding question are displayed, along with an answer drawn from either a human or one of the LLM model variants. Participants are then asked to rate the answer across the evaluation metrics provided.

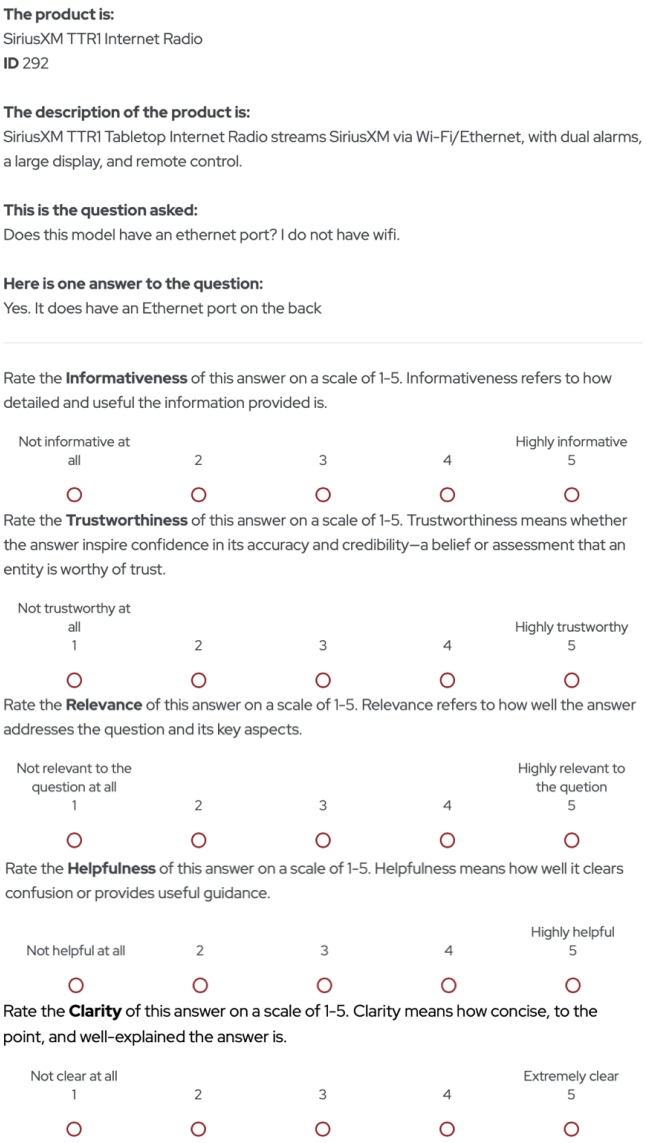

Figure 6: Human Evaluation Example (Answer Rating Task)

Appendix B. Human Evaluation Results by Question Type

Tables 10 and 11 report the detailed human evaluation results for all five question types. Values are reported as mean (standard error).

Table 10: Human Evaluation Results by Question Type (Specs and Features, User Experience and Feel)

Answer Source	Average Score	Clarity	Informativity	Relevance	Trust	Helpfulness	% Over Human
Specs and Features							
Best Human Answer	4.47 (.11)	4.50 (.15)	4.50 (.11)	4.58 (.12)	4.22 (.15)	4.58 (.11)	-
Human Answer	3.65 (.08)	3.58 (.10)	3.61 (.09)	3.88 (.09)	3.51 (.09)	3.68 (.10)	-
Worst Human Answer	2.49 (.17)	2.41 (.18)	2.52 (.21)	2.78 (.21)	2.42 (.18)	2.31 (.19)	-
LLM only	3.11 (.18)	3.31 (.20)	3.20 (.22)	3.05 (.20)	3.20 (.19)	2.82 (.23)	39%
LLM+RAG	4.03 (.13)	3.98 (.15)	4.02 (.18)	4.15 (.12)	3.93 (.12)	4.06 (.18)	65%
LLM+RAG+AC	4.00 (.19)	4.05 (.27)	3.93 (.21)	4.60 (.11)	3.49 (.20)	3.95 (.26)	54%
LLM+RAG+AC+ Weighted Match	3.95 (.18)	4.06 (.19)	3.87 (.20)	4.19 (.22)	3.68 (.21)	3.96 (.21)	67%
User Experience and Feel							
Best Human Answer	4.35 (.09)	4.49 (.10)	4.23 (.12)	4.50 (.11)	4.18 (.12)	4.36 (.11)	-
Human Answer	3.50 (.07)	3.54 (.09)	3.42 (.08)	3.71 (.09)	3.40 (.08)	3.43 (.09)	-
Worst Human Answer	2.63 (.15)	2.50 (.16)	2.57 (.16)	2.93 (.19)	2.67 (.16)	2.46 (.18)	-
LLM only	3.61 (.15)	3.80 (.16)	3.74 (.20)	3.69 (.16)	3.52 (.15)	3.29 (.22)	55%
LLM+RAG	4.32 (.09)	4.26 (.13)	4.40 (.10)	4.50 (.12)	3.97 (.12)	4.46 (.11)	80%
LLM+RAG+AC	4.16 (.10)	4.19 (.14)	4.23 (.10)	4.44 (.11)	3.86 (.13)	4.09 (.13)	76%
LLM+RAG+AC+ Weighted Match	4.26 (.10)	4.32 (.13)	4.32 (.10)	4.39 (.13)	3.96 (.13)	4.30 (.12)	81%

Notes. Values are reported as mean (standard error). For methods incorporating AC, the comparison is based only on the portion of the question they answered.

Table 11: Human Evaluation Results by Question Type (Customer Service, Fitting & Compatibility, and Usage)

Answer Source	Average Score	Clarity	Informativity	Relevance	Trust	Helpfulness	% Over Human
Customer Service							
Best Human Answer	4.15 (.10)	4.26 (.12)	4.21 (.14)	4.34 (.12)	3.79 (.12)	4.15 (.14)	-
Human Answer	3.42 (.08)	3.41 (.09)	3.38 (.09)	3.63 (.09)	3.26 (.08)	3.41 (.09)	-
Worst Human Answer	2.37 (.13)	2.43 (.18)	2.26 (.15)	2.48 (.18)	2.39 (.16)	2.31 (.16)	-
LLM only	3.42 (.17)	3.58 (.17)	3.34 (.22)	3.33 (.19)	3.63 (.15)	3.22 (.22)	50%
LLM+RAG	3.89 (.14)	3.98 (.16)	3.93 (.15)	4.09 (.15)	3.63 (.14)	3.85 (.20)	71%
LLM+RAG+AC	4.00 (.15)	4.12 (.18)	4.10 (.17)	4.10 (.17)	3.78 (.15)	3.91 (.22)	73%
LLM+RAG+AC+ Weighted Match	3.89 (.17)	3.97 (.21)	3.91 (.20)	4.06 (.21)	3.68 (.18)	3.83 (.21)	69%
Fitting and Compatibility							
Best Human Answer	4.29 (.12)	4.26 (.16)	4.19 (.14)	4.51 (.13)	4.17 (.13)	4.30 (.15)	-
Human Answer	3.40 (.09)	3.39 (.10)	3.33 (.10)	3.65 (.10)	3.29 (.09)	3.31 (.10)	-
Worst Human Answer	2.24 (.14)	2.21 (.18)	2.08 (.14)	2.67 (.18)	2.20 (.17)	2.04 (.15)	-
LLM only	3.55 (.16)	3.49 (.18)	3.68 (.19)	3.58 (.19)	3.65 (.15)	3.35 (.22)	57%
LLM+RAG	4.27 (.11)	4.20 (.14)	4.41 (.11)	4.33 (.12)	4.01 (.13)	4.39 (.13)	78%
LLM+RAG+AC	4.24 (.11)	4.25 (.13)	4.26 (.12)	4.39 (.13)	4.01 (.13)	4.29 (.14)	73%
LLM+RAG+AC+ Weighted Match	4.22 (.09)	4.21 (.13)	4.27 (.10)	4.43 (.11)	3.89 (.13)	4.28 (.12)	71%
Usage							
Best Human Answer	4.43 (.11)	4.52 (.14)	4.30 (.17)	4.62 (.11)	4.19 (.11)	4.52 (.13)	-
Human Answer	3.50 (.09)	3.58 (.10)	3.40 (.10)	3.68 (.10)	3.38 (.09)	3.48 (.11)	-
Worst Human Answer	2.36 (.13)	2.34 (.16)	2.32 (.17)	2.51 (.16)	2.40 (.16)	2.22 (.17)	-
LLM only	3.79 (.17)	3.88 (.15)	3.94 (.18)	3.64 (.23)	3.69 (.17)	3.78 (.22)	64%
LLM+RAG	3.93 (.15)	4.09 (.16)	3.84 (.17)	4.07 (.15)	3.76 (.15)	3.88 (.18)	64%
LLM+RAG+AC	4.06 (.12)	4.12 (.14)	4.22 (.16)	4.26 (.14)	3.73 (.16)	3.97 (.15)	61%
LLM+RAG+AC+ Weighted Match	4.24 (.10)	4.17 (.15)	4.29 (.10)	4.44 (.10)	3.89 (.16)	4.40 (.11)	70%

Notes. Values are reported as mean (standard error). For methods incorporating AC, the comparison is based only on the portion of the question they answered.

Appendix C. Visualizing Model vs. Human Answers

To illustrate the difference in automatic and human evaluation, we use UMAP (McInnes, Healy, and Melville 2018) to visualize the semantic distribution of all model outputs from our full model and all answers from the benchmark human answers across objective (specs and features) and subjective (user experience) questions in Figure 7.

Specifically, we first encode each answer into a high-dimensional embedding that captures its semantic meaning. We then apply UMAP—a dimensionality reduction technique that projects these embeddings into two dimensions while preserving their relative distances. To focus on within-type variation, we run UMAP separately for each question type, since projecting all types together would emphasize broad differences across types and obscure finer type-specific semantic distinctions. Because each UMAP is computed independently, the axes differ between plots and are chosen to best represent the structure of that question type.

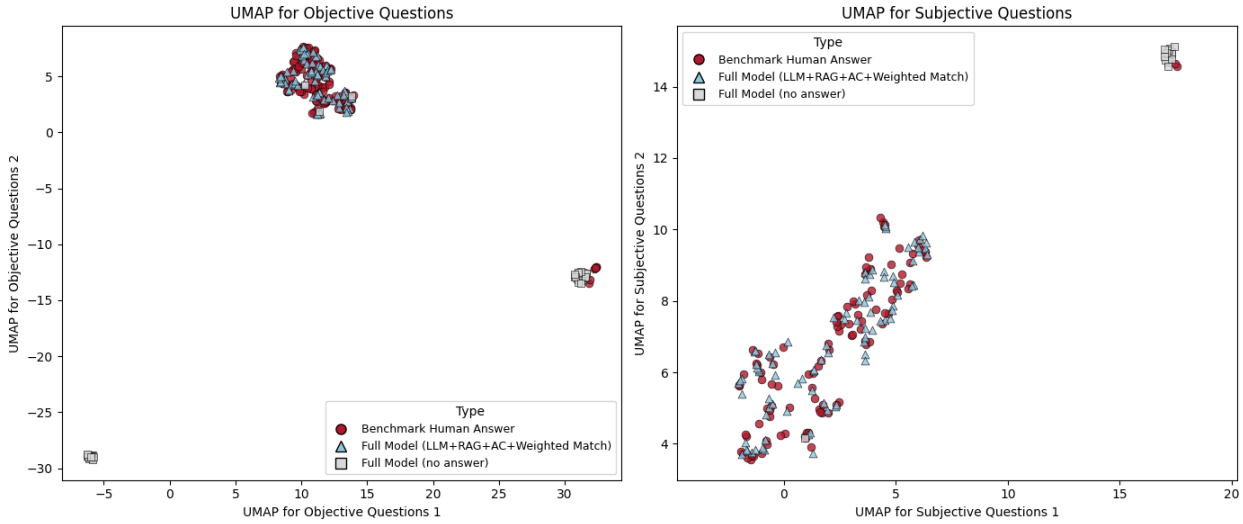

Figure 7: UMAP for Objective and Subjective Question Types

As shown in Figure 7, objective questions (Specs and Features) display a larger outlier island corresponding to "no answer" cases, consistent with their lower answerability (51.3% vs. 83.7% for subjective questions). Besides "no answer" cases, answers to the objective

questions (Specs and Features) are relatively tightly clustered, indicating a strong lexical and semantic alignment between LLM and the human output with a ROUGE-L of .177. In contrast, for subjective questions (user experience and feel) with a .138 ROUGE-L, the answers are more dispersed, reflecting the subjective and varied nature of these responses. The higher dispersion combined with high answerability could increase variability, making it more challenging for the model to generate answers that closely align with human responses.

Appendix D. Answerability vs. Review Sentence Count

Figure 8 plots review count against the percentage of questions answered by the full model. While based on a limited number of question types, the figure suggests a positive association between the availability of reviews and answerability. This pattern indicates that platforms may potentially improve question coverage by encouraging more user reviews.

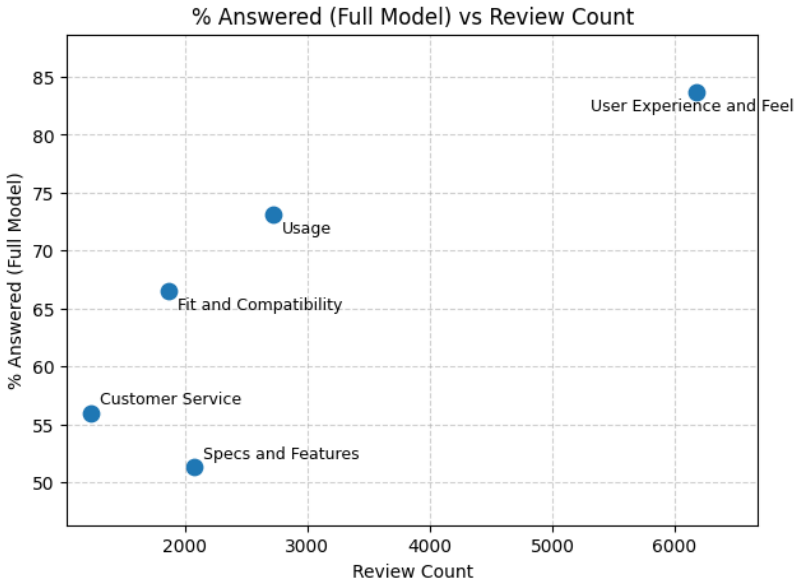

Figure 8: Answerability vs. Review Sentence Count

AMA Journals Anonymized Web Appendix

Table of Contents

WEB APPENDIX A. IDENTIFYING AND ANNOTATING QUESTION TYPES

Before conducting question–review type matching, we first identify and annotate types for both questions and reviews, since our dataset does not contain type information.

Consumer questions often reflect both product categories (e.g., clothing, electronics) and question intents (e.g., specifications, usage). To separate these dimensions and extract the underlying question types, we apply Top2Vec to a sample of 12,000 AmazonQA questions. This unsupervised topic modeling approach initially produces 83 fine-grained topics, which we consolidate into 10 high-level ones using Top2Vec's hierarchical structure. As shown in Table W1, these 10 topics represent a mixture of product categories and question intents.

Since our goal is to extract question types (i.e., intents rather than product-specific themes), we manually examine each of the 10 high-level topics. Based on their representative examples and informed by prior literature, we distill them into five overarching question types: Specs and Features, Fitting and Compatibility, Usage, Customer Service, and User Experience and Feel.

Table W1 summarizes this process. Each row corresponds to one of the 10 Top2Vec-derived topics, showing representative questions, their frequency, and the final question types extracted from each topic. This mapping serves as the foundation for our downstream task of aligning questions with relevant reviews.

After establishing the question types, we use the OpenAI API (GPT-4o) to annotate questions and review sentences. Questions may be assigned to multiple types, whereas each review sentence belongs to a single type. Table W2 presents the prompt used for classification, which specifies the annotator role, the definitions of dimensions, and illustrative examples. The same structure is applied to both reviews and questions, except that questions can receive multiple labels while reviews only belong to one.

Table W1: Categorization of Product-Related Questions

Topic	Example	Extracted Question Types
QT1	Will this one fit a Macbook Pro 13in laptop as well? Does it have a disk drive?	Specs and Features, Fitting and Compatibility
QT2	How do I use it? Do you ship to Canada? Anyone know if this mount can be used upside down?	Usage, Customer Service
QT3	Where is it made? What are these glass or plastic?	Specs and Features
QT4	Does this watch fit on any of the NATO watch straps? Can you wear these socks with regular shoes?	Fitting and Compatibility
QT5	Is this blow dryer good for fine hair? Can this be used to dry regular nail polish as well? What are the ingredients? I recently put a permanent color in that turned out lighter than expected.	Usage, Specs and Features, User Experience and Feel
QT6	Does this work with the iPad mini? Can this be used on an iPhone 4?	Fitting and Compatibility
QT7	Does it come with a charger? Does the battery have to be charged? If so, how do you charge it?	Specs and Features, Usage
QT8	Is this phone unlocked? Does this phone use a SIM card?	Specs and Features
QT9	Will it fit a 4-year-old boy? What size should I order for a 38-inch waist?	Fitting and Compatibility
QT10	How long is the cord? What are the dimensions?	Specs and Features, Usage

Table W2: Prompt structure used for review classification and evaluation

Type	Prompt
System	You are an experienced data evaluating engineer with extensive experience in classifying reviews. Your task is to categorize the reviews based on the following types: `{Dimensions}` For the following reviews in the category, evaluate the answer in the dimensions we specified.
Dimensions	The dimensions and their corresponding definitions.
Examples	*User Experience and Feel: "I really like the Xtreme's rugged case and feel like it can hold up well in everyday use and on my camping trips."* *Other Dimensions: ...*

WEB APPENDIX B. LCS CALCULATION

In automatic evaluation, ROUGE-L relies on the Longest Common Subsequence (LCS). Based on our PQA setting, the longest common subsequence (LCS) between an LLM generated answer A_{LLM} and a reference human answer A_h is computed recursively as follows:

$$LCS(A_h, A_{\text{LLM}}) = \begin{cases} 0, & \text{if } m = 0 \text{ or } n = 0 \\ 1 + LCS(A_{h,m-1}, A_{\text{LLM},n-1}), & \text{if } a_{h,m} = a_{\text{LLM},n} \\ \max(LCS(A_{h,m-1}, A_{\text{LLM}}), LCS(A_h, A_{\text{LLM},n-1})), & \text{otherwise} \end{cases} \tag{1}$$

Here, $A_{h,m-1}$ and $A_{\text{LLM},n-1}$ denote the sequences excluding their last elements, and $a_{h,m}$, $a_{\text{LLM},n}$ are the final tokens in each sequence. Intuitively, the recursion adds 1 whenever the last tokens match, otherwise it propagates the maximum subsequence length from the shorter prefixes. This formulation allows ROUGE-L to capture sentence-level similarity while preserving word order without requiring exact matches.

WEB APPENDIX C. AUTOMATIC EVALUATION RESULTS BY QUESTION TYPE

Table W3 reports the automatic evaluation results by question type. The performance patterns are consistent with the overall results. The table presents the ROUGE-L score (for models with AC, only answered questions are included), the percentage of questions answered, and the corresponding p-values. Note that LLM+RAG is compared with LLM only, while LLM+RAG+AC and LLM+RAG+AC+Weighted Match are compared with LLM+RAG.

Table W3: ROUGE-L Scores and Answered Question Percentage by Question Type

Answer Type	ROUGE-L	% Answered	p-value
Specs and Features			
LLM Only	.100 (.004)	100%	/
LLM+RAG	.152 (.008)	100%	.000
LLM+RAG+AC	.182 (.012)	48.67%	.033
LLM+RAG+AC+Weighted Match	.177 (.012)	51.33%	.073
User Experience and Feel			
LLM Only	.104 (.004)	100%	/
LLM+RAG	.129 (.004)	100%	.000
LLM+RAG+AC	.137 (.006)	82.69%	.265
LLM+RAG+AC+Weighted Match	.138 (.005)	83.65%	.215
Fitting and Compatibility			
LLM Only	.106 (.003)	100.00%	/
LLM+RAG	.141 (.005)	100.00%	.000
LLM+RAG+AC	.161 (.006)	65.48%	.013
LLM+RAG+AC+Weighted Match	.154 (.007)	66.50%	.110
Usage			
LLM Only	.112 (.004)	100.00%	/
LLM+RAG	.145 (.005)	100.00%	.000
LLM+RAG+AC	.154 (.006)	71.03%	.258
LLM+RAG+AC+Weighted Match	.152 (.006)	73.10%	.387
Customer Service			
LLM Only	.105 (.006)	100.00%	/
LLM+RAG	.143 (.011)	100.00%	.003
LLM+RAG+AC	.150 (.023)	52.00%	.774
LLM+RAG+AC+Weighted Match	.148 (.021)	56.00%	.827

Notes. Values are reported as mean (standard error). For p-value of LLM answers, LM+RAG is compared with LLM only, the remaining variants with RAG are compared with LLM+RAG. For answers with AC, only the part it answered is compared.