

Article

Not peer-reviewed version

Enhancing Commercial Interactions through Comprehensive QA Fusion

Logan Wright , Aiden Carter * , [Wyne Nasir](#)

Posted Date: 8 April 2024

doi: 10.20944/preprints202404.0493.v1

Keywords: natural language processing; transformer; question answering



Preprints.org is a free multidiscipline platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Article

Enhancing Commercial Interactions through Comprehensive QA Fusion

Logan Wright, Aiden Carter and Wyne Nasir

Tufts University

* Correspondence: aidencarter@tufts.edu

Abstract: In the dynamic landscape of E-commerce, the proliferation of user-generated queries regarding products highlights the critical need for advanced automatic question answering (AQA) systems. These systems are instrumental in harnessing the vast array of information available online to deliver immediate and informative responses to potential buyers. Recognizing the complexity of such queries, which often require synthesis of reviews, product specifications, and responses to similar questions, we introduce a sophisticated model, the Comprehensive E-commerce Answer Generation System (CEAGS). This model adeptly navigates the challenges posed by irrelevant data and sentiment ambiguity in user-generated content. Our approach leverages a dual-phase process of relevance determination and sentiment clarity, setting the stage for a transformative response generation mechanism. Empirical analyses reveal the superiority of CEAGS, with our relevance assessment framework outstripping existing models by a notable margin in precision metrics, and our answer generation module achieving unprecedented gains in content preservation and coherence. Notably, CEAGS marks a pioneering contribution to the E-commerce domain by integrating disparate information sources to formulate responses that are not only accurate but also contextually rich and user-centric.

Keywords: natural language processing; transformer; question answering

1. Introduction

Question answering (QA) systems have emerged as a cornerstone technology in the field of artificial intelligence, striving to mimic human-like understanding and responses to queries. These systems span a wide array of applications [1,2], from digital assistants on smartphones to customer service bots on websites, fundamentally changing how humans interact with information technology. The evolution of QA systems has been propelled by advancements in natural language processing (NLP), machine learning, and deep learning, enabling these systems to process and interpret vast amounts of natural language data. At their core [3], QA systems aim to understand user queries in their natural language form, retrieve relevant information from a structured or unstructured data pool, and deliver concise, accurate answers. This process involves complex language understanding, context analysis, and semantic processing techniques, highlighting the interdisciplinary nature of the field that blends linguistics, computer science, and cognitive psychology.

In the early stages, question answering systems relied heavily on rule-based methods that required meticulous manual crafting of language rules and templates. However, the advent of machine learning and, more recently, deep learning models, especially Transformer-based architectures [9,10], has revolutionized the QA landscape. These technologies have introduced a level of understanding and flexibility previously unattainable, allowing for more nuanced interpretation of queries and the context surrounding them. Moreover, the introduction of large-scale datasets for training, such as SQuAD (Stanford Question Answering Dataset), and competitions like the NLP Progress leaderboard, have further spurred innovation and performance improvements in QA systems [11]. As a result, modern QA systems can now handle a broader range of question types, from simple factual queries to complex inferential questions requiring deep comprehension and synthesis of multiple information sources. This progress not only enhances the user experience by providing more accurate and contextually relevant answers but also opens new avenues for applications in education, healthcare, and beyond, where effective and intuitive access to information can have profound impacts.

The advent of automatic question answering (AQA) systems tailored for E-commerce settings marks a significant milestone in enhancing user engagement and satisfaction. These systems are designed to address the myriad of product-related queries posted by users, which, in the absence of responses from previous purchasers, remain unanswered [15–18]. The escalating demand for efficient and accurate AQA systems has spurred research into leveraging diverse data sources, including product reviews, specifications, and previously answered questions, to craft relevant responses. This endeavor has catalyzed the development of innovative methodologies in review-centric and multi-source answer generation, laying the groundwork for the Comprehensive E-commerce Answer Generation System (CEAGS).

Table 1. Enhanced examples from our answer generation dataset.

	Enhanced Example 1	Enhanced Example 2
Question	How responsive is the screen in <i>ABC</i> model?	Can the sound quality of the phone rival that of a home theater system?
Reference Answer	The display responsiveness of my device is subpar	Absolutely, the sound clarity and depth is exceptional
Duplicate Q&A (Enriched)	1) Between <i>XYZ</i> and <i>ABC</i> , which offers a better display experience? Definitely, <i>ABC</i> .	1) How does the audio fidelity compare to <i>MNO</i> ? <i>PQR</i> offers unparalleled sound quality.
Reviews	1) The <i>ABC</i> model performs well overall, but the display is not its strongest suit 2) Display performance of <i>ABC</i> leaves much to be desired	1) The audio output is disappointingly mediocre. 2) Lacks the expected audio quality.
Specifications	1) Display enhancements include...Slim Bezel: 2.05mm, Aspect Ratio: 2) Color Depth: 16.7 Million 3) Screen Size: 6.22 inches	1) Audio enhancements feature...Echo Cancellation: Dual-microphone system

Building an effective real-world AQA system for the E-commerce domain entails addressing the inherent challenges of dataset noise, including spelling mistakes, grammatical inconsistencies, and occasional code-switching, further complicated by the presence of irrelevant content and ambiguous sentiments in user reviews and queries. These issues underscore the critical need for a refined approach to information synthesis, as exemplified in our CEAGS model. Our system not only mitigates these challenges but also sets a new benchmark in the realm of AQA through its innovative use of Transformer-based models, renowned for their exceptional capabilities in handling natural language processing tasks. By harnessing the power of pre-trained Transformer [22] models, fine-tuned to our specific dataset, CEAGS achieves unparalleled performance in generating precise, context-aware responses to user queries, as detailed in our extensive experimental analysis [26–29].

The primary contributions of our work are multifaceted and can be delineated as follows:

- Introduction of a novel, comprehensive answer generation framework, CEAGS, that adeptly leverages three pivotal sources of information: product reviews, analogous queries, and detailed specifications, to inform its responses.
- A detailed exploration of the challenges inherent in synthesizing information from these diverse sources, namely the filtration of irrelevant data and the resolution of sentiment ambiguities, and the strategic methodologies employed by CEAGS to address these issues.
- The implementation of a cutting-edge, two-stage process within CEAGS that initially assesses relevance and clarifies sentiment, followed by a sophisticated mechanism for generating coherent and contextually rich answers.

- An exhaustive evaluation of CEAGS, demonstrating its superiority over existing models through comprehensive metrics such as BLEU and ROUGE, and corroborated by human assessment, showcasing an overall accuracy enhancement of 77.88%.

The remainder of this paper is structured to provide a thorough exposition of our research and findings. Section 2 delves into the related work, laying the groundwork for our innovations. Section 3 formally defines the problem statement, paving the way for the detailed description of CEAGS in Section 4. Section 5 presents a comprehensive analysis of our experimental results, highlighting the efficacy of CEAGS, followed by a conclusive summary in Section 6.

2. Related Work

In the rapidly evolving domain of natural language processing (NLP), substantial progress has been made in developing sophisticated models capable of understanding and generating human-like text. This section embarks on an exploration of the seminal advancements in large language models, with a particular emphasis on their application in relevancy prediction and the nuanced task of answer generation. Among the notable contributions, BERT (Bidirectional Encoder Representations from Transformers) emerges as a pioneering model, leveraging a bidirectional transformer encoder architecture to excel in tasks such as masked language modeling and Next Sentence Prediction (NSP) [26]. Following BERT's trajectory, RoBERTa refines this approach by excluding the NSP component and introducing dynamic masking, thereby enhancing the model's training effectiveness [32]. Additionally, T5 (Text-to-Text Transfer Transformer) adopts an encoder-decoder framework, distinguishing itself through its versatility across diverse NLP tasks, facilitated by a comprehensive transfer learning methodology [33]. These models, embodying the forefront of transformer-based technologies, form the cornerstone of our investigation into automatic question answering within the E-commerce domain, underpinning the development of the Comprehensive E-commerce Answer Generation System (CEAGS).

The quest for accurate relevancy prediction in question answering systems has spurred a multitude of studies, each aiming to harness and interpret the wealth of information encapsulated in user-generated content. Innovations in this area often revolve around the strategic ranking of potential answers, drawing upon insights from user reviews and similar queries. The model proposed by Yu et al. [34] exemplifies this trend, focusing on the retrieval of analogous questions to facilitate answer selection. Similarly, Cui et al. [35] introduced SuperAgent, a chatbot adept at synthesizing information from various sources to pinpoint the most fitting response. This body of work underscores the pivotal role of candidate ranking in enhancing the relevance of selected answers. In alignment with these advancements, recent explorations leveraging transformer-based models, as demonstrated by Mittal et al. [36], have further refined the process of relevancy determination. Building upon these insights, our approach integrates a novel ambiguity filtering mechanism, addressing the challenge of inconsistent answers prevalent in E-commerce interactions. This refinement serves as a precursor to our answer generation model, ensuring the contextual integrity of the responses generated by CEAGS.

The realm of answer generation has witnessed a dynamic interplay between recurrent neural network (RNN)-based and transformer-based methodologies, each contributing unique perspectives to the art of text generation. The inception of attention mechanisms and self-attention paradigms has significantly bolstered the performance of sequence-to-sequence models [22,41]. Despite these advances, the E-commerce sector often gravitates towards RNN-based frameworks for answer generation, utilizing product reviews as a primary information source [15,16]. Among the pioneering efforts, McAuley et al. [45] employed a Mixture of Experts (MOE) model, leveraging reviews to forecast binary responses. Further enhancements in this domain have been achieved through the application of attention mechanisms, as illustrated by Chen et al. [16], and the innovative use of BERT for binary question answering [50]. The integration of opinion mining with answer generation, as explored by Deng et al. [15], and the encoding of reviews and product specifications through dual encoders, as proposed by Gao et al. [18], represent key milestones in this evolution. Distinguishing itself from

prior endeavors, our work introduces CEAGS, a holistic pipeline designed to amalgamate insights from product specifications, similar queries, and reviews to craft coherent natural language responses. This initiative marks a significant departure from traditional methodologies, offering a comprehensive framework for leveraging diverse information sources in the E-commerce sector. By embracing the inherent noise within user-submitted answers, CEAGS ventures beyond the constraints of clean supervised data, pioneering a scalable and robust solution for E-commerce question answering.

3. Preliminary

In this section, we delineate the foundational challenge that our work, the Comprehensive E-commerce Answer Generation System (CEAGS), aims to tackle. The essence of this problem lies in the realm of automated natural language response generation, where the objective is to craft an articulate and informative answer, y , to a user-posed question, Q , regarding a product. This task is underscored by the need to judiciously select and utilize pertinent information from a composite set of information candidates $\{x_1, \dots, x_k\}$, which encompasses a myriad of data types relevant to the product in question.

Expanding on this framework, our dataset D comprises N instances, with each instance, denoted as d^i , encapsulating a multifaceted collection of data elements. These include the question Q^i posed by the user, an array of product reviews $\{r_1^i, \dots, r_k^i\}$, a compilation of similar or duplicate question-answer pairs $\{(q_1^i, a_1^i), \dots, (q_l^i, a_l^i)\}$, and a detailed list of product specifications $\{s_1^i, \dots, s_m^i\}$. Each of these elements contributes a unique perspective, enriching the context and depth of the information available for answer generation. The corresponding correct answer y^i serves as the ground truth for each query. Formally, the dataset D is represented as follows:

$$D = \left(Q^i, \{r_1^i, \dots, r_k^i\}, \{(q_1^i, a_1^i), \dots, (q_l^i, a_l^i)\}, \{s_1^i, \dots, s_m^i\}, y^i \right)_{i=1}^N \quad (1)$$

Our primary aim is to devise a mechanism for generating an answer y^i that not only draws upon the most relevant snippets of information but also presents this information in a manner that is both coherent and precise. This endeavor requires an intricate balance of information retrieval, relevance determination, and language generation capabilities, all of which are central to the objectives of CEAGS. The challenge extends beyond mere data aggregation; it demands an understanding and integration of diverse data types, ensuring that the generated response accurately reflects the user's query while adhering to the contextual and factual nuances of the product information.

4. Methodology

In this section, we outline the architecture and workflow of our innovative approach, the Comprehensive E-commerce Answer Generation System (CEAGS), designed to tackle the complex challenge of generating accurate and relevant answers to product-related questions by sifting through and synthesizing information from multiple sources. This method aims to distill and utilize the most pertinent information while eliminating any irrelevant or ambiguous data. CEAGS framework is comprised of three distinct yet interrelated components: (1) Relevancy Detection, (2) Ambiguity Resolution, and (3) Natural Language Answer Synthesis. Each of these components is meticulously engineered and fine-tuned to perform its specific function within the pipeline, thereby ensuring the generation of coherent, concise, and contextually appropriate responses.

4.1. Relevancy Detection

In the vast expanse of data typically associated with E-commerce queries, the sheer volume and diversity of information candidates present a significant challenge. The relevancy detection component of CEAGS employs a sophisticated Next Sentence Prediction (NSP) model to ascertain the relevance of each information candidate to the query. This model treats the query as the initial sentence and evaluates each candidate's potential to follow logically in a textual sequence, thus filtering out irrelevant or redundant information. The underlying hypothesis is that a candidate relevant to answering the query should logically succeed it in a coherent narrative. The process

involves concatenating the query Q^i with each candidate x_k^i , followed by tokenization and embedding. These embeddings, incorporating both lexical and positional information, are then processed by a transformer encoder, which predicts the relevance of the candidate as follows:

$$w_k^i = \text{tokenizer}([Q^i; x_k^i]) \quad \hat{y}_k^i = \text{transformer}(w_k^i) \quad (2)$$

The semicolon signifies the methodical concatenation of input sentences tailored to the pre-trained transformer's requirements. The objective of this relevancy detection phase is to optimize the cross-entropy loss, facilitating the accurate ranking of candidates based on their pertinence and contribution to generating an informed response.

4.2. Enhanced Ambiguity Resolution

The phenomenon of ambiguity within user-generated content on E-commerce platforms, notably from subjective reviews and Q&A pairs, introduces significant challenges in crafting clear and unambiguous responses. The CEAGS framework's ambiguity resolution module takes an innovative approach to mitigate this by meticulously sifting through the plethora of information to identify and exclude data that may conflict with the intended sentiment of the response. This mechanism is particularly invaluable for binary (yes/no) queries, where the presence of contrasting sentiments could significantly impair the model's capacity to learn and generate coherent responses. By implementing a sophisticated sentiment analysis algorithm, CEAGS is capable of discerning the underlying sentiment of each candidate information piece, ensuring that only those aligned with the query's context are considered. This not only streamlines the answer generation process but also substantially elevates the precision and relevance of the output, guaranteeing that the final response accurately mirrors the nuanced sentiment embedded within the input data.

4.3. Advanced Natural Language Answer Synthesis

Following the meticulous refinement of information candidates through relevancy detection and ambiguity resolution phases, CEAGS embarks on the task of synthesizing the final answer. Leveraging a state-of-the-art transformer-based encoder-decoder architecture, as initially conceptualized by Vaswani et al. [22], this module is engineered to navigate the intricacies of natural language processing. With an assembly of self-attention mechanisms and sophisticated feed-forward networks, the architecture adeptly parses and processes the input, thus ensuring the generation of answers that are not only contextually pertinent but also linguistically coherent. The inclusion of cutting-edge NLP techniques allows CEAGS to adeptly handle the nuances of natural language generation, empowering the system to deliver responses that are accurate, engaging, and seamlessly aligned with the user's original query. This capability marks a significant leap forward in the domain of automated response generation, positioning CEAGS as a frontrunner in delivering high-quality, linguistically refined answers in the E-commerce sphere.

4.4. Comprehensive Integration within CEAGS: Elevating E-Commerce Answer Generation

The CEAGS architecture exemplifies a paradigm of seamless integration, merging critical components of relevancy detection, ambiguity resolution, and advanced natural language synthesis into a unified answer generation pipeline. This integrated approach facilitates a holistic evaluation and utilization of all pertinent information sources, including product reviews, detailed specifications, and analogous Q&A pairs. By doing so, CEAGS ensures that every facet of available data is meticulously analyzed and employed to craft responses that epitomize clarity, relevance, and informational value. Furthermore, CEAGS' adoption of advanced NLP and machine learning methodologies significantly augments its capability to navigate the complexities of language and sentiment nuances, thereby enhancing the quality and accuracy of generated answers. This comprehensive framework not only sets a new benchmark in automated question answering but also heralds a new era of efficiency, accuracy, and user satisfaction in the E-commerce industry. Through its innovative design and sophisticated

data processing capabilities, CEAGS stands as a monumental advancement in the field, redefining the standards of customer interaction and support in digital commerce platforms.

5. Experiments

This section delves into the empirical examination designed to evaluate the Comprehensive E-commerce Answer Generation System (CEAGS), addressing key inquiries concerning its performance against baseline models, the precision and coherence of its generated answers, and the individual contributions of each pipeline variant to the answer generation process.

5.1. Dataset and Model Training

Our analysis begins with the relevancy prediction model, trained on a curated dataset, *D1*, of mobile-related inquiries. This dataset encompasses a broad spectrum of user-posted questions, each paired with candidates from reviews, similar Q&A pairs, and product specifications. To ensure a high degree of precision in relevancy prediction, each candidate was meticulously labeled to reflect its relevance to the corresponding question. The dataset *D1* comprises a total of 2000 questions, presenting a balanced mix of relevant and irrelevant candidates, as highlighted in Table 2.

Table 2. Dataset for Predicting Information Relevance.

	Training Set	Validation Set
Number of Queries	1638	362
Total Information Candidates	15122	3268
Relevant Candidates	8670	1736
Average Relevance of Specifications	0.308	0.253
Average Relevance of Q&A	0.668	0.634
Average Relevance of Reviews	0.626	0.573

Simultaneously, the answer generation model was fine-tuned using dataset *D2*, consisting of over 200K user-submitted questions accompanied by their respective answers. Prior to answer generation, candidates from all three sources were filtered through the relevancy and ambiguity prediction models to retain only the most pertinent information, typically resulting in seven or fewer candidates per question. The dataset's statistics, before and after filtering, are summarized in Table 3. This rigorous preparation ensures that the model is trained on highly relevant, concise datasets, enhancing its ability to generate precise and coherent answers.

Table 3. Dataset for Generating Answers.

	Training Set	Validation Set
Total Number of Queries	217086	11153
Number of WH (Who, What, Where, etc.) Questions	66075	3459
Average Candidates per Query	9.750	9.486
Average Number of Specifications per Query	2.381	2.381
Average Number of Reviews per Query	2.637	2.344
Average Number of Duplicate Questions per Query	4.732	4.762

5.2. Baseline Models and Evaluation Metrics

Our experimental setup pits CEAGS against a series of baseline models and evaluates its performance using both quantitative metrics and qualitative assessments. To comprehensively understand CEAGS's efficacy, we investigate its capabilities in relevancy prediction, ambiguity resolution, and natural language answer generation, as detailed in Table 4.

Table 4. Datasets Employed for Model Training.

Dataset	Model
Relevancy Dataset (D1)	RoBERTa-Answers
	BERT-Answers
	RoBERTa-QA Enhanced
	BERT-QA Enhanced
Answer Generation Dataset (D2)	Seq2Seq Enhanced
	HSSC-q Enhanced
	T5-QA Enhanced

- **Seq2Seq Enhanced** [41] - An advanced implementation of the sequence-to-sequence model, augmented with attention mechanisms for improved context understanding. This model processes concatenated queries and candidates to produce coherent responses.
- **HSSC-q Enhanced** - A variant of the HSSC model [52] tailored for question answering, which leverages sentiment analysis alongside answer generation for sentiment-coherent outputs.
- **T5-QA Enhanced** [33] - A fine-tuned version of the T5 model, specially adapted for answering E-commerce queries. This model exemplifies the fusion of extensive pre-training with task-specific tuning for optimal answer generation.

These models are evaluated against a suite of metrics, including ROUGE and BLEU scores for linguistic quality and human assessments for evaluating response relevance and coherence. The relevancy prediction model's accuracy is further gauged using precision, recall, and F1-score, providing a multifaceted view of CEAGS's performance.

5.3. Implementation Details

Leveraging the Transformers [53] package and PyTorch [54], we meticulously trained each component of the CEAGS framework.

5.3.1. Enhancing Relevancy Prediction

Both BERT and RoBERTa models were fine-tuned on the relevancy dataset *D1*, with variations introduced to assess the impact of including answers alone versus incorporating both questions and answers in the input. This distinction led to the development of the BERT-Answers and RoBERTa-Answers variants for focusing solely on the answers, and BERT-QA Enhanced and RoBERTa-QA Enhanced for a comprehensive approach. These models underwent extensive training over five epochs, with a batch size set at 32, aiming to refine the system's ability to discern relevant information from a sea of candidates.

5.3.2. Refining Ambiguity Prediction

The T5 model, pre-trained on a wide array of tasks, was specifically adapted for ambiguity prediction within our framework. This adaptation involved training the model to identify and filter out sentimentally conflicting candidates, ensuring that the generated answers remained consistent with the overall sentiment of the input. This step was crucial for maintaining the integrity and coherence of responses, particularly for binary (yes/no) questions.

5.3.3. Advancing Answer Generation

For answer generation, we implemented an enhanced Seq2Seq model with pre-trained GloVe embeddings and fine-tuned the T5 model to cater to the nuances of E-commerce queries. Additionally, the HSSC-q Enhanced model was trained following specifications detailed in the original HSSC paper

but was adapted to better suit the requirements of question answering. All models were trained with a focus on maximizing linguistic accuracy and relevance, supported by a batch size of 32 and a consistent learning rate, over a course of 25 epochs.

Through this comprehensive experimental setup, we aim to rigorously evaluate CEAGS's performance, highlighting its advancements in generating accurate, coherent, and contextually relevant answers in the E-commerce domain.

5.4. Experimental Findings

5.4.1. Evaluating Relevancy Prediction Performance

Our approach's relevancy prediction capabilities are demonstrated in Table 5, where CEAGS's various configurations are compared against baseline models. Notably, the BERT-QA Enhanced model excels in accuracy, precision, and F1-score, underscoring its effectiveness in discerning relevant information from a pool of candidates. This ability to filter relevant content significantly contributes to the subsequent accuracy and coherence of the answer generation phase of CEAGS.

Table 5. Performance Comparison on Relevancy Prediction.

Model	Accuracy	Precision	Recall	F1-Score
BERT-base	0.635	0.637	0.996	0.777
RoBERTa-Answers	0.708	0.767	0.7778	0.772
BERT-Answers	0.749	0.806	0.797	0.802
RoBERTa-QA Enhanced	0.764	0.832	0.789	0.810
BERT-QA Enhanced	0.838	0.873	0.872	0.873

5.4.2. Answer Generation Pipeline Effectiveness

The answer generation capabilities of CEAGS and its comparison against various baseline models are summarized in Table 6. Here, we observe that CEAGS, particularly the T5-QA Enhanced model, outperforms all baselines across metrics for both binary and WH questions. The comprehensive pipeline of CEAGS (Full Pipeline) demonstrates superior ability in generating contextually relevant and coherent answers, although it slightly trails the T5-QA Enhanced model in some metrics, illustrating the importance of integrating relevancy and ambiguity resolution in improving answer quality.

Table 6. Methodological Evaluation on Answer Generation.

	Binary Questions				WH Questions			
	R1	R2	RL	B1	R1	R2	RL	B1
T5-Base	9.74	1.89	9.18	0.22	8.14	2.20	7.80	0.77
Seq2Seq Enhanced	22.87	6.57	22.09	1.50	14.50	3.74	13.77	0.10
HSSC-q Enhanced	24.19	8.65	23.46	1.91	15.43	4.85	14.68	0.90
T5-QA Enhanced	31.27	12.85	29.65	5.48	22.69	8.77	20.71	3.21
CEAGS (Relevancy Only)	31.17	12.79	29.58	5.41	22.48	8.47	20.60	3.18
CEAGS (Full Pipeline)	30.47	12.16	28.88	5.18	22.64	8.73	20.85	3.28

Table 7 showcases diverse examples of generated answers by CEAGS and other models, reflecting the nuanced understanding and context-awareness that CEAGS brings to answer generation. The examples highlight CEAGS's ability to generate precise and accurate responses, closely aligned with the reference answers and the contextual details of the queries.

Table 7. Selected Examples of Generated Answers.

Query: Is this device equipped with gorilla glass protection? Reference Answer: No, additional protection is advised
T5-Base: Does it have gorilla glass? Seq2Seq Enhanced: No, it comes with gorilla glass HSSC-q Enhanced: Yes, it features gorilla glass 3
T5-QA Enhanced: No, it lacks gorilla glass. CEAGS (Relevancy): No gorilla glass is included. CEAGS (Full Pipeline): No, gorilla glass is not a feature.
Query: How does the ABC game perform in terms of graphics and gameplay? Reference Answer: HD mode runs smoothly, but HDR is too demanding for this model
T5-Base: Performance is lacking Seq2Seq Enhanced: ABC can be played with basic settings HSSC-q Enhanced: HD gameplay is possible on ABC
T5-QA Enhanced: Smooth gameplay and high-quality graphics are noted on ABC. CEAGS (Relevancy): ABC provides excellent performance on high settings. CEAGS (Full Pipeline): ABC achieves fluid gameplay in HD graphics.
Query: How is the sound quality and are there any heating issues? Reference Answer: Sound is impressive with no heating concerns
T5-Base: Device overheats Seq2Seq Enhanced: Good sound quality HSSC-q Enhanced: High sound quality
T5-QA Enhanced: Sound quality is satisfactory, though heating can be an issue CEAGS (Relevancy): Excellent sound, minimal heating detected CEAGS (Full Pipeline): Sound quality is exceptional with no heating problems

5.4.3. Insights from Human Evaluation

Human evaluation, as depicted in Table 8, provides a deeper insight into the qualitative aspects of the answers generated by CEAGS, affirming its superiority in creating contextually and factually accurate responses. Particularly, the full pipeline of CEAGS demonstrates a significant advantage over the T5-QA Enhanced model in terms of correctness against the context and reference answers, underlining the efficacy of CEAGS's integrated approach in enhancing the quality of generated answers across both binary and WH questions.

Table 8. Results from Human Evaluation.

	Binary Questions		WH Questions	
	Correct vs. Context	Correct vs. Reference	Correct vs. Context	Correct vs. Reference
T5-QA Enhanced	0.919	0.628	0.833	0.537
CEAGS (Full Pipeline)	0.943	0.845	0.869	0.656

In conclusion, the comprehensive experimental analysis confirms the effectiveness of the CEAGS framework in advancing the state of automated question answering within the E-commerce domain. By intricately balancing relevancy prediction, ambiguity resolution, and sophisticated natural language processing techniques, CEAGS sets a new benchmark for generating informative, coherent, and contextually aligned answers, thereby enriching the customer experience on E-commerce platforms.

6. Conclusion and Future Directions

In the dynamic and ever-evolving landscape of E-commerce, the ability to automatically generate accurate and informative answers to customer queries stands as a pivotal element in enhancing user

experience and facilitating informed purchasing decisions. This study pioneers in the E-commerce domain by introducing a comprehensive solution, the Comprehensive E-commerce Answer Generation System (CEAGS), designed to automate the generation of natural language responses based on a synthesis of diverse product-related information sources, including product specifications, similar questions, and user reviews.

The paper meticulously outlines the inherent challenges associated with leveraging noisy and inconsistent user-generated content and proposes CEAGS as a novel answer generation pipeline adept at navigating these complexities. Our approach significantly advances the state-of-the-art, with the CEAGS relevancy prediction model demonstrating a notable **12.36%** improvement in F1 score over existing baselines. Furthermore, we delve into the nuances of handling ambiguity within user data, employing a pretrained model to discern and adjust for sentiment discrepancies among answer candidates, thereby enhancing the coherence and relevance of generated answers.

A critical insight gleaned from this work is the potential to utilize user-submitted answers as a source of training data, thereby mitigating the reliance on labor-intensive supervised annotations traditionally required for training sophisticated Question Answering models. The efficacy of CEAGS is underscored by its performance across key content preservation metrics, such as BLEU and ROUGE, where it surpasses baseline models with significant margins. Moreover, human evaluations corroborate the superior accuracy of CEAGS, evidencing an **30.7%** overall improvement in accuracy compared to the generation model, thereby underscoring the effectiveness of our full pipeline approach in generating precise answers.

Looking ahead, our future endeavors aim to explore the integration of answer generation and ambiguity resolution within a unified training framework, seeking to further streamline the answer generation process. Additionally, we anticipate extending the applicability of our natural language generation techniques to a broader array of E-commerce contexts, encompassing queries related to offers, delivery specifics, and beyond, thus broadening the scope of automated customer support and engagement. This forward-looking vision not only promises to refine the quality of automated responses available to E-commerce users but also opens avenues for applying these advancements across various facets of digital commerce, potentially transforming the way online businesses interact with and serve their customer base.

References

1. Zijun Sun, Chun Fan, Qinghong Han, Xiaofei Sun, Yuxian Meng, et al. Self-explaining Structures Improve NLP Models. 2020. URL <https://arxiv.org/abs/2012.01786>.
2. Hao Fei, Meishan Zhang, and Donghong Ji. Cross-lingual semantic role labeling with high-quality translated training corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7014–7026, 2020.
3. A. Golubev and N. Loukachevitch. Transfer Learning for Improving results on Russian Sentiment Datasets. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog”*, pages 268–277, 2021.
4. Shengqiong Wu, Hao Fei, Fei Li, Meishan Zhang, Yijiang Liu, Chong Teng, and Donghong Ji. Mastering the explicit opinion-role interaction: Syntax-aided neural transition system for unified opinion role labeling. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence*, pages 11513–11521, 2022.
5. Wenxuan Shi, Fei Li, Jingye Li, Hao Fei, and Donghong Ji. Effective token graph modeling using a novel labeling strategy for structured sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4232–4241, 2022.
6. Hao Fei, Yue Zhang, Yafeng Ren, and Donghong Ji. Latent emotion memory for multi-label emotion classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7692–7699, 2020.
7. Fengqi Wang, Fei Li, Hao Fei, Jingye Li, Shengqiong Wu, Fangfang Su, Wenxuan Shi, Donghong Ji, and Bo Cai. Entity-centered cross-document relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9871–9881, 2022.

8. Ling Zhuang, Hao Fei, and Po Hu. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100:101919, 2023.
9. Jiangang Bai, Yujing Wang, Yiren Chen, Yaming Yang, Jing Bai, Jing Yu, and Yunhai Tong. Syntax-bert: Improving pre-trained transformers with syntax trees. *arXiv preprint arXiv:2103.04350*, 2021.
10. Hao Fei, Yafeng Ren, and Donghong Ji. Retrofitting structure-aware transformer language model for end tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2151–2161, 2020.
11. Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. Opinion word expansion and target extraction through double propagation. *Computational linguistics*, 37(1):9–27, 2011.
12. Hao Fei, Shengqiong Wu, Jingye Li, Bobo Li, Fei Li, Libo Qin, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Lasuie: Unifying information extraction with latent adaptive structure-aware generative language model. In *Proceedings of the Advances in Neural Information Processing Systems, NeurIPS 2022*, pages 15460–15475, 2022.
13. Hao Fei, Yafeng Ren, Yue Zhang, Donghong Ji, and Xiaohui Liang. Enriching contextualized language model from knowledge graph for biomedical information extraction. *Briefings in Bioinformatics*, 22(3), 2021.
14. Shengqiong Wu, Hao Fei, Wei Ji, and Tat-Seng Chua. Cross2StrA: Unpaired cross-lingual image captioning with cross-lingual cross-modal structure-pivoted alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2593–2608, 2023.
15. Yang Deng, Wenxuan Zhang, and Wai Lam. Opinion-aware answer generation for review-driven question answering in e-commerce, 2020.
16. Shiqian Chen, Chenliang Li, Feng Ji, Wei Zhou, and Haiqing Chen. Review-driven answer generation for product-related questions in e-commerce. WSDM '19, page 411–419, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359405. doi:10.1145/3289600.3290971. URL <https://doi.org/10.1145/3289600.3290971>.
17. Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023.
18. Shen Gao, Zhaochun Ren, Yihong Eric Zhao, Dongyan Zhao, Dawei Yin, and Rui Yan. Product-aware answer generation in e-commerce question-answering. *CoRR*, abs/1901.07696, 2019. URL <http://arxiv.org/abs/1901.07696>.
19. Shengqiong Wu, Hao Fei, Yafeng Ren, Donghong Ji, and Jingye Li. Learn from syntax: Improving pair-wise aspect and opinion terms extraction with rich syntactic knowledge. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 3957–3963, 2021.
20. Bobo Li, Hao Fei, Lizi Liao, Yu Zhao, Chong Teng, Tat-Seng Chua, Donghong Ji, and Fei Li. Revisiting disentanglement and fusion on modality and context in conversational multimodal emotion recognition. In *Proceedings of the 31st ACM International Conference on Multimedia, MM*, pages 5923–5934, 2023.
21. Hao Fei, Qian Liu, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Scene graph as pivoting: Inference-time image-free unsupervised multimodal machine translation with visual scene hallucination. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5980–5994, 2023.
22. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL <http://arxiv.org/abs/1706.03762>.
23. Jingye Li, Kang Xu, Fei Li, Hao Fei, Yafeng Ren, and Donghong Ji. MRN: A locally and globally mention-based reasoning network for document-level relation extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1359–1370, 2021.
24. Hao Fei, Shengqiong Wu, Yafeng Ren, and Meishan Zhang. Matching structure for dual learning. In *Proceedings of the International Conference on Machine Learning, ICML*, pages 6373–6391, 2022.
25. Hu Cao, Jingye Li, Fangfang Su, Fei Li, Hao Fei, Shengqiong Wu, Bobo Li, Liang Zhao, and Donghong Ji. OneEE: A one-stage framework for fast overlapping and nested event extraction. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1953–1964, 2022.
26. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL <http://arxiv.org/abs/1810.04805>.

27. Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237, 2019. URL <http://arxiv.org/abs/1906.08237>.
28. Shengqiong Wu, Hao Fei, Yixin Cao, Lidong Bing, and Tat-Seng Chua. Information screening whilst exploiting! multimodal relation extraction with feature denoising and multimodal topic modeling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14734–14751, 2023.
29. Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. *CoRR*, abs/1909.11942, 2019. URL <http://arxiv.org/abs/1909.11942>.
30. Hao Fei, Fei Li, Bobo Li, and Donghong Ji. Encoder-decoder based unified semantic role labeling with label-aware syntax. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12794–12802, 2021.
31. Bobo Li, Hao Fei, Fei Li, Yuhan Wu, Jinsong Zhang, Shengqiong Wu, Jingye Li, Yijiang Liu, Lizi Liao, Tat-Seng Chua, and Donghong Ji. DiaASQ: A benchmark of conversational aspect-based sentiment quadruple analysis. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13449–13467, 2023.
32. Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL <http://arxiv.org/abs/1907.11692>.
33. Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683, 2019. URL <http://arxiv.org/abs/1910.10683>.
34. Jianfei Yu, Minghui Qiu, Jing Jiang, Jun Huang, Shuangyong Song, Wei Chu, and Haiqing Chen. Modelling domain relationships for transfer learning on retrieval-based question answering systems in e-commerce. *CoRR*, abs/1711.08726, 2017. URL <http://arxiv.org/abs/1711.08726>.
35. Lei Cui, Shaohan Huang, Furu Wei, Chuanqi Tan, Chaoqun Duan, and Ming Zhou. SuperAgent: A customer service chatbot for E-commerce websites. In *Proceedings of ACL 2017, System Demonstrations*, pages 97–102, Vancouver, Canada, July 2017. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/P17-4017>.
36. Happy Mittal, Aniket Chakrabarti, Belhassen Bayar, Animesh Anant Sharma, and Nikhil Rasiwasia. Distantly supervised transformers for e-commerce product qa, 2021.
37. Hao Fei, Shengqiong Wu, Yafeng Ren, Fei Li, and Donghong Ji. Better combine them together! integrating syntactic constituency and dependency representations for semantic role labeling. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021*, pages 549–559, 2021.
38. Shengqiong Wu, Hao Fei, Hanwang Zhang, and Tat-Seng Chua. Imagine that! abstract-to-intricate text-to-image synthesis with scene graph hallucination diffusion. *Advances in Neural Information Processing Systems*, 36, 2024.
39. Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, and Tat-Seng Chua. Empowering dynamics-aware text-to-video diffusion with large language models. *arXiv preprint arXiv:2308.13812*, 2023.
40. Leigang Qu, Shengqiong Wu, Hao Fei, Liqiang Nie, and Tat-Seng Chua. Layoutllm-t2i: Eliciting layout guidance from llm for text-to-image generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 643–654, 2023.
41. Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate, 2016.
42. Hao Fei, Fei Li, Chenliang Li, Shengqiong Wu, Jingye Li, and Donghong Ji. Inheriting the wisdom of predecessors: A multiplex cascade framework for unified aspect-based sentiment analysis. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*, pages 4096–4103, 2022.
43. Hao Fei, Yafeng Ren, and Donghong Ji. Boundaries and edges rethinking: An end-to-end neural model for overlapping entity relation extraction. *Information Processing & Management*, 57(6):102311, 2020.
44. Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. Unified named entity recognition as word-word relation classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10965–10973, 2022.
45. Julian McAuley and Alex Yang. Addressing complex and subjective product-related queries with customer reviews. In *Proceedings of the 25th International Conference on World Wide Web, WWW '16*, page 625–635,

- Republic and Canton of Geneva, CHE, 2016. International World Wide Web Conferences Steering Committee. ISBN 9781450341431. doi:10.1145/2872427.2883044. URL <https://doi.org/10.1145/2872427.2883044>.
46. Hao Fei, Tat-Seng Chua, Chenliang Li, Donghong Ji, Meishan Zhang, and Yafeng Ren. On the robustness of aspect-based sentiment analysis: Rethinking model, data, and training. *ACM Transactions on Information Systems*, 41(2):50:1–50:32, 2023.
 47. Yu Zhao, Hao Fei, Yixin Cao, Bobo Li, Meishan Zhang, Jianguo Wei, Min Zhang, and Tat-Seng Chua. Constructing holistic spatio-temporal scene graph for video semantic role labeling. In *Proceedings of the 31st ACM International Conference on Multimedia, MM*, pages 5281–5291, 2023.
 48. Hao Fei, Yafeng Ren, Yue Zhang, and Donghong Ji. Nonautoregressive encoder-decoder neural framework for end-to-end aspect-based sentiment triplet extraction. *IEEE Transactions on Neural Networks and Learning Systems*, 34(9):5544–5556, 2023.
 49. Yu Zhao, Hao Fei, Wei Ji, Jianguo Wei, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Generating visual spatial description via holistic 3D scene understanding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7960–7977, 2023.
 50. Daria Dzendzik, Carl Vogel, and Jennifer Foster. Is it dish washer safe? automatically answering “yes/no” questions using customer reviews. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 1–6, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi:10.18653/v1/N19-3001. URL <https://www.aclweb.org/anthology/N19-3001>.
 51. Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. Reasoning implicit sentiment with chain-of-thought prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1171–1182, 2023.
 52. Shuming Ma, Xu Sun, Junyang Lin, and Xuancheng Ren. A hierarchical end-to-end model for jointly improving text summarization and sentiment classification. *CoRR*, abs/1805.01089, 2018. URL <http://arxiv.org/abs/1805.01089>.
 53. Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface’s transformers: State-of-the-art natural language processing, 2020.
 54. Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.