

# User Review Writing via Interview with Dialogue Systems

Yoshiki Tanaka and Michimasa Inaba  
The University of Electro-Communications  
{y-tanaka, m-inaba}@uec.ac.jp

## Abstract

User reviews on e-commerce and review sites are crucial for making purchase decisions, although creating detailed reviews is time-consuming and labor-intensive. In this study, we propose a novel use of dialogue systems to facilitate user review creation by generating reviews from information gathered during interview dialogues with users. To validate our approach, we implemented our system using GPT-4 and conducted comparative experiments from the perspectives of system users and review readers. The results indicate that participants who used our system rated their interactions positively. Additionally, reviews generated by our system required less editing to achieve user satisfaction compared to those by the baseline. We also evaluated the reviews from the readers' perspective and found that our system-generated reviews are more helpful than those written by humans. Despite challenges with the fluency of the generated reviews, our method offers a promising new approach to review writing.

## 1 Introduction

User reviews posted on e-commerce and review sites benefit both buyers and sellers. For users considering a purchase, these reviews provide product details and insights, thereby supporting the decision-making process of purchase. Additionally, they act as feedback for sellers, enabling product quality improvement. Reviews that contain specific product details or are based on personal user experiences are especially valuable. However, writing these insightful reviews can be time-consuming and labor-intensive, as the task requires careful language and proofreading for potential buyers.

To reduce users' burden of writing reviews, researchers (e.g., (Dong et al., 2017; Li and Tuzhilin, 2019)) have focused on automated review genera-

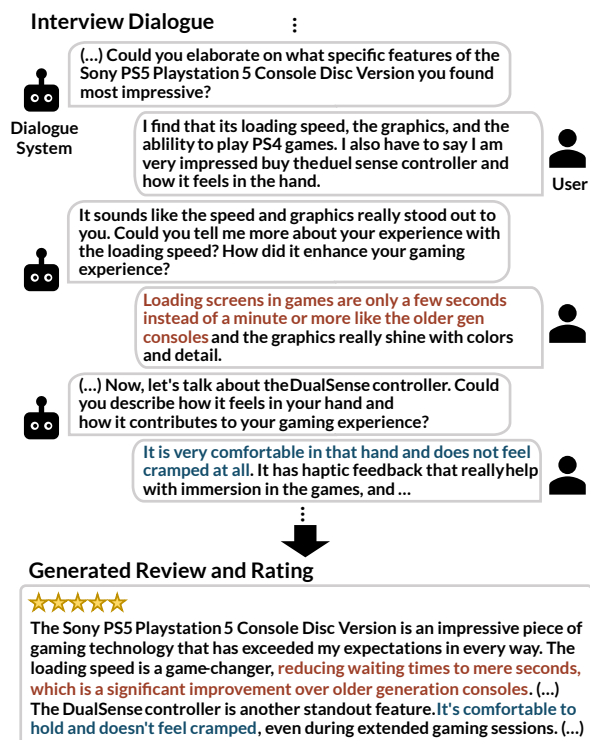


Figure 1: Example of review creation supported by the proposed method. In the dialogue and review, the contents of the same-colored text correspond to each other.

tion, thereby making the review creation process more efficient.

Despite these efforts, the ability of existing review generation systems to personalize for users and products remains limited. Previous studies have explored utilizing a wide range of information about products and users for generating reviews. However, available subjective information is limited. Specifically, without direct input from the user, incorporating the user's actual experiences with the product into the generated review is challenging. This constraint significantly limits the system's ability to personalize for the user. To overcome this problem, we focused on supporting the creation of reviews by directly eliciting infor-

mation about products from users.

In this study, we propose the novel utilization of dialogue systems for creating user reviews. Figure 1 shows an example of the review creation process supported by the dialogue system according to our proposed method. First, the dialogue system acts as an interviewer, eliciting user opinions on products through interview dialogues. Second, the review text generator generates review text based on the dialogue history. Finally, the rating predictor predicts a rating consistent with the generated review text. Our method allows users to easily create reviews by simply interacting with the system, thus reducing the effort involved in review creation.

To evaluate our method, we implemented a system incorporating our approach using GPT-4. Subsequently, we conducted experiments using our system, collecting data on dialogues between the system and users, the generated reviews, predicted ratings, and participants' feedback on our system. We discuss the effectiveness of our method after analyzing the collected data. In summary, our main contributions are as follows:

1. As a novel application of dialogue systems, we propose a method for supporting user review creation. Furthermore, we developed a system incorporating our approach using GPT-4.
2. We conducted a comprehensive survey from the perspectives of system users and review readers, showing that our method can provide high-quality and helpful reviews for both parties.

## 2 Related Work

### 2.1 Interview Dialogue System and Dataset

The interview dialogues are aimed at eliciting information from the interviewees. Prior research suggests that surveys conducted on chatbot platforms yield higher-quality responses than web survey platforms (Kim et al., 2019). This finding indicates that employing dialogue systems to collect user opinions and impressions is a promising approach.

Researchers have collected interview dialogue data on various topics, including radio (Majumder et al., 2020), news (Zhu et al., 2021), sports (Sun et al., 2022), and cooking (Okahisa et al., 2022).

The objectives of these collections vary from analyzing dialogue patterns (Majumder et al., 2020; Okahisa et al., 2022) to dialogue summarization (Zhu et al., 2021). Here, we utilize the interview dialogue system to support the creation of helpful reviews.

### 2.2 Review Generation

User reviews reflect user's opinions and requests regarding a product. These insights benefit buyers and sellers. Additionally, user reviews have a wide range of applications. Previous research has applied reviews to natural language processing tasks such as recommendations (Qiu et al., 2021), opinion summarization (Bražinskas et al., 2020), and task-oriented dialogue (Zhao et al., 2023).

User reviews that include detailed information about the product and user experiences are useful. However, writing these reviews is a labor-intensive task for humans. To increase the efficiency of this process, researchers have proposed automated review generation models, enhancing their review generation capabilities by utilizing information such as ratings (Dong et al., 2017; Sharma et al., 2018; Li et al., 2019; Kim et al., 2020), images (Truong and Lauw, 2019; Vu et al., 2020), past reviews written by the user (Li and Tuzhilin, 2019), and aspect-oriented features (Li and Tuzhilin, 2019). Unlike these studies, we focus on the collaborative writing of user reviews with the support of the dialogue system.

Some researchers have focused on supporting users in creating reviews, similar to our approach (Ni and McAuley, 2018; Bhat et al., 2023). For example, Ni and McAuley proposed utilizing short phrases related to products that are provided by customers, such as review summaries and product titles, as auxiliary data for generating reviews (Ni and McAuley, 2018). In their system, the user provides information in a unidirectional manner. In contrast, we utilize an interview-specific dialogue system to collect information from the user through interactive interaction. The dialogue system can ask follow-up questions to obtain additional details regarding a product although this information may be ambiguous. This capability supports the creation of detailed reviews.

### 2.3 Dialogue Summarization

In our method, we proposed to convert conversational data (i.e., interview dialogue history) into non-conversational data (i.e., review texts).

Therefore, our work is closely related to dialogue summarization research. To build an effective model for dialogue summarization, researchers have proposed diverse approaches to learning methods (Zou et al., 2021; Li et al., 2023; Zhong et al., 2022; Zhang et al., 2021). Additionally, researchers have built dialogue summarization datasets that can be used for training models; these datasets cover daily life conversations (Gliwa et al., 2019; Chen et al., 2021), meetings (Carletta et al., 2006; Zhong et al., 2021), TV series (Chen et al., 2022), and media dialogue (Zhu et al., 2021). While these studies aim to condense dialogue histories into brief texts, our approach takes a different direction. We focus on extracting useful product information for readers from interview dialogues and organizing it into a non-conversational data format, rather than compressing it into shorter text.

### 3 Methodology

To create useful reviews, reviewers must provide detailed product information. Interview dialogue systems are employed to effectively elicit this information. To enhance readability, we propose organizing the dialogue history into a non-dialogue format. Our method comprises three processes: interview dialogue, review text generation, and rating prediction. In this paper, the systems that perform these processes are referred to as the “interview dialogue system,” the “review text generator,” and the “rating predictor,” respectively. Our system utilizes these components in sequence to generate reviews as the output. An overview of our system is shown in Figure 2. We use the gpt-4-0613 model to implement our system.

#### 3.1 Interview Dialogue System

To assist potential buyers in making purchase decisions, guiding users to create helpful reviews is crucial. In our approach, therefore, our system should be designed to effectively collect information from the user. To achieve this, we propose utilizing an interview dialogue system. For the interview dialogue system, it is desirable to elicit both the pros and cons of a product in a balanced and detailed manner. Specifically, the system should be capable of asking follow-up questions about the content mentioned by the user or changing the topic to inquire about different aspects of the product.

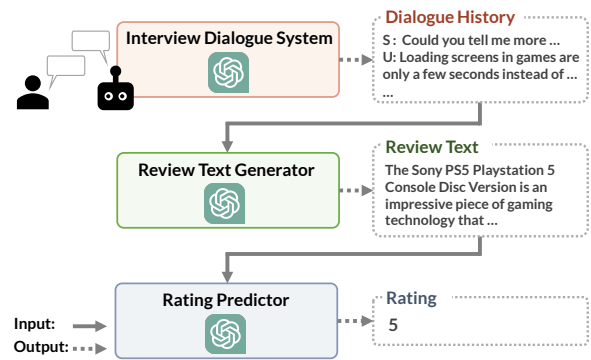


Figure 2: Overview of our system. First, the interview dialogue system interviews the user to elicit their impressions and requests about the product they used. Next, the review text generator uses the dialogue history as input to generate a review text. Finally, the rating predictor predicts a rating consistent with the sentiment of the generated review text.

We designed a prompt that incorporated instructions for the system to perform these behaviors. Moreover, aiming to both collect sufficient information for creating reviews and ensure users don’t become bored, we added constraints regarding the number of turns to the prompt. In our experiments, we adopted instructions to ask at least 8 questions and conclude the interview within 15 turns. Additionally, to ensure the interview does not continue indefinitely, we externally implemented a setting in the interview dialogue system to end the dialogue after 15 turns. The prompt template for the interview dialogue is shown in Appendix A.1.

#### 3.2 Review Text Generator

Although the dialogue history between interview dialogue systems and users offers useful and detailed product information, it often contains redundancies. Consequently, it is not appropriate to post it directly as a user review. Therefore, we propose transforming the dialogue history into a format suitable for reviews. Our review text generator aims to capture the essence of the interview dialogue history while generating review texts from the perspective of the user. To generate reviews that align with the user’s feedback, the system must faithfully reflect the content of the dialogue history in the review text. Our prompts include instructions to concisely summarize important information mentioned during the interview and generate the main body of the user review. The prompt template for the review generation is shown in Appendix A.2.

### 3.3 Rating Predictor

In e-commerce and review platforms, customer ratings are aggregated into a single score, providing other users with an initial impression of the product. For an aggregated score to be reliable, reviewers must assign ratings that accurately reflect the content of their review text. While the ratings impact the reputation widely, considering the potential for human error in assigning ratings, automating the task might be an effective solution. Our rating predictor automatically outputs a rating consistent with the sentiment of the input review text, ranging from 1 to 5 as an integer. Ratings consistent with the content of the review texts could reduce exaggerated scoring caused by user subjectivity. As a result, this can improve the reliability of the ratings.

We utilized GPT-4 to implement a rating predictor. To enhance predictive performance, we designed prompts that apply chain-of-thought prompting (Wei et al., 2022; Wang et al., 2023; Kojima et al., 2022), that feeds large language models not only examples of question-and-answer pairs but also examples of the thought processes leading to those answers. In this study, we collected five sets of product titles, review texts, and ratings from Amazon.com to create output exemplars, each corresponding to ratings from 1 to 5. Subsequently, for each set, we wrote descriptions of the reasoning paths leading to the prediction of the rating from the product title and review text. We used these as few-shot exemplars within the context. Similarly, for target reviews, GPT-4 is encouraged to generate a reasoning path and an answer.

## 4 Experiments

We aim to facilitate the review-writing process for reviewers and provide helpful reviews to readers. To investigate the practicality of our method, we conducted evaluations from the perspectives of system users (Section 4.1) and review readers (Section 4.2).

### 4.1 Participant Evaluation

To evaluate our method, we collected feedback through interviews, generated reviews, and questionnaires. Data collection was conducted through Amazon Mechanical Turk (MTurk)<sup>1</sup>.

<sup>1</sup><https://www.mturk.com>

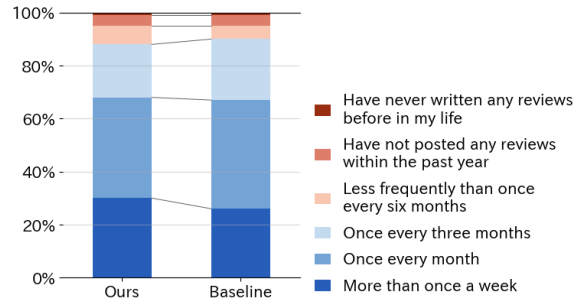


Figure 3: Participant responses to “In the past year, how often have you posted reviews?”

### 4.1.1 Experimental Setup

We tuned the temperature for each system. For the interview dialogue system, the temperature was set to 0.2. The review text generator and rating predictor generate outputs that are faithful to the input. Therefore, we set the temperature to 0 for these systems to suppress the diversity of the generated text.

### 4.1.2 Baseline System

To demonstrate the effectiveness of using interview dialogue systems that adapt questions based on the context, we constructed a baseline system. The baseline system replaces the interview dialogue system with one that asks manually created questions in a fixed order. To construct the baseline system, we manually created nine questions on topics such as the reason for purchasing the product and the evaluation of the product in comparison with other products. All questions asked by the baseline system are listed in Appendix B. We collected data using this system in the same manner as with our proposed system.

### 4.1.3 Evaluation Procedure

Initially, participants conducted an interview dialogue with our interview dialogue system. After the interview, they were presented with the generated reviews and ratings. Participants then completed a post-interview survey comprising multiple-choice and open-ended questions. For each setting, we recruited 100 participants located in AU, CA, NZ, GB, or the US and had a 95% approval rate with at least 500 previously approved HITs.

### 4.1.4 Post-Interview Survey

After the interview, participants responded to a post-interview survey. Several questions in this

Table 1: Likert Items in Post-interview Survey

Dimension	Labels in Figure 4	Statements
Interview	Enjoyable	<i>How fun was your interaction with the chatbot interviewer?</i>
	Skillful	<i>The interviewer skillfully elicited your impressions or opinions.</i>
	In-depth	<i>The chatbot interviewer attempted to elicit your impressions or opinions in depth.</i>
Review	Faithful	<i>The system-generated review faithfully reflects what you said during your interviews.</i>
	Concise	<i>The system-generated review offers a concise summary of the points you mentioned during the interview.</i>
System	Quality	<i>Please rate the overall quality of the system.</i>
	Burdened(I)	<i>I felt burdened to have an interview chat about the product.</i>
	Burdened(R)	<i>Writing a review with the support of the system is more burdensome than writing a review yourself.</i>

Participants' Responses: Ours (Upper) vs. Baseline System (Lower)

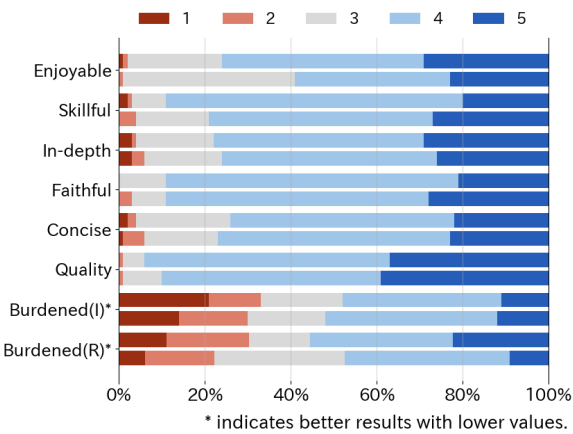


Figure 4: Participant responses to questions on a Likert scale from 1 (Strongly disagree) to 5 (Strongly agree) in a post-interview survey. For each question, the upper bar shows the results from our system and the lower bar shows the baseline results.

survey were answered using a 5-point Likert scale. These questions are related to the interview dialogue, the generated reviews, and the overall system (See Table 1).

We also asked participants how frequently they post reviews to compare with their usual review-writing experiences. As shown in Figure 3, 95% of the participants posted at least one user review in the past year. Additionally, participants were asked to rate the product they selected by responding to the question: "If you were to rate this product again, what rating would you give it?" and provided a rating from 1 to 5.

#### 4.1.5 Participant Feedback Analysis

Figure 4 illustrates the distribution of responses to eight questions<sup>2</sup>. Regarding the dimensions of

<sup>2</sup>For Burdened(R), we excluded responses from participants who selected the "Have never written any reviews before in my life" option to the question in Figure 3.

the interview and review, most participants evaluated two components positively: our interview dialogue system and our review text generator. Participants showed a similar positive trend across two settings for the four items: In-depth, Faithful, Concise, and Quality. Notably, for Quality, 90% or more of the participants rated the overall quality of the systems positively.

Our system provided users with more enjoyable interviews and higher satisfaction regarding the generated reviews compared with the baseline system. As shown in Figure 4, when using our system based on GPT-4, more participants agreed that interacting with the system was fun. Moreover, the difference in the methods used to elicit information—our interview dialogue system and the baseline—impacts users' enjoyment, with statistically significant differences (Mann–Whitney U test,  $p < 0.05$ ). Participants also responded to the multiple-choice question, "If you had to edit and post a system-generated review to your satisfaction, how much of it would you need to rewrite?". Figure 5 shows that different types of systems resulted in varied response distributions. In particular, 38% and 27% of participants using the baseline system and our system, respectively, responded that they needed to rewrite more than 50% of the review. These results indicate that our system can provide reviews with higher satisfaction than the baseline.

Our system imposed a greater burden on participants. Figure 4 shows that a higher percentage of participants agreed that *writing a review with the support of our system is more burdensome than writing alone*, compared to the baseline. We argue that the response time of the system is one of the reasons for this difference. Our GPT-4-based system, which generates responses based on users' utterances, takes a longer time to gener-

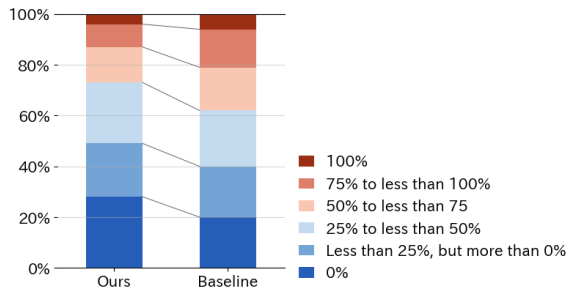


Figure 5: Participant responses to “If you had to edit and post a system-generated review to your satisfaction, how much of it would you need to rewrite?”

ate responses than a baseline that asks predefined questions. Notably, several participants suggested that the response speed of our system should be improved. In response to the free-form question “What is one enhancement that can be made to improve this system?”, we received answers such as “more fast replies” and “need quick reply.” In our experiments, unlike the ChatGPT interface<sup>3</sup>, we did not employ real-time response generation using streaming functionality. Adding this feature would be an effective modification to enhance our system’s response speed, which is expected to significantly improve user experience.

#### 4.1.6 Case Study

Our interview dialogue system can generate follow-up questions that explore the content of users’ ambiguous responses in depth. Table 2 shows an example of the data collected, comprising the dialogue history regarding an electric shaver and the corresponding review text generated. During the interview, our system initially asked about the participant’s overall satisfaction with the product, to which the participant replied, “... well satisfied but with few minor issues.” Based on this response, our system posed follow-up questions to clarify the aspects that the participant was satisfied with and the issues they mentioned. As demonstrated in this example, our system can elicit deeper information about products from users compared with the baseline system.

Additionally, our interview dialogue system can change topics during an interview to collect information on various aspects of a product. For the interview dialogues shown in Table 2, the first three turns focus on the participant’s satisfaction with the product. Subsequently, the system changes the

topic by saying “Now, let’s go back to the issue” and thus shifting focus to the issues that the user mentioned in the first turn. In the next turn, our system shifts the topic again to highlight the product’s impressive features. By switching topics in this manner, our system can acquire information on various aspects of the product.

Our system can also generate review texts and ratings that faithfully reflect the content mentioned in the interview. As shown in the lower part of Table 2, the generated review text closely aligns with user comments such as “i would say well satisfied ...” and “some times small hair from the beard gets stuck.” Moreover, our system accounts for issues mentioned by the user and predicts a rating of 4, one less than the maximum.

## 4.2 Third-Party Evaluation

Helpful user reviews offer effective decision-making support for potential buyers and sellers. To assess the quality of the reviews generated by our system from the reader’s perspective, we conducted multidimensional evaluations.

### 4.2.1 Experimental Procedure

This evaluation was conducted in a pairwise setting. In this experiment, annotators were presented with two reviews: one generated by our system and another for comparison. We compared our system-generated reviews with Amazon reviews to evaluate the quality difference from human-written reviews. Additionally, to assess the differences from the perspective of the readers, we compared our system-generated reviews with those generated by a baseline system, which were collected in Section 4.1.

Annotators chose either the superior option or a tie for each of the eight dimensions, as shown in Table 3. We recruited the annotators through MTurk.

### 4.2.2 Selection Criteria for Comparative Reviews

We collected human-written reviews from the Amazon Review Dataset (Ni et al., 2019). To facilitate comparison for the annotators, we initially removed any human-written reviews where the category or rating did not match those generated by our system. After this filtering, to avoid selecting poor-quality reviews, we retained reviews in the top 5% of human-written reviews based on helpfulness votes, using these as compar-

<sup>3</sup><https://chatgpt.com/>

Table 2: Example of a dialogue between our system and the user (top, S: system, U: user) and generated review text (bottom). In this example, our rating predictor predicted a rating of 4.

---

S (...) To start off, could you tell me about your overall satisfaction with this product?

U **i would say well satisfied** but with few minor issues.

S I see, you're generally satisfied but there are a few minor issues. Could you elaborate on what you particularly like about the Braun Series 9 9370cc Electric Shaver?

U **some times small hair from the beard gets stuck**s in particular spot on the blade area in the machine.

S I understand that small hairs getting stuck in the blade area is one of the issues you've encountered. But before we delve into that, could you first share what you find satisfying or impressive about the product? What are its positive points in your opinion?

U the postivie sides is smooth shaving experience and battery capacity which is good

S It's great to hear that you appreciate the smooth shaving experience and the battery capacity. Now, let's go back to the issue you mentioned earlier about small hairs getting stuck. Could you tell me more about this problem? How often does it happen and how does it affect your shaving experience?

U i wouls say not often, ifaced this issue only once.

S I see, (...) That's good to know. Now, aside from the smooth shaving experience and the battery capacity, are there any other features of the Braun Series 9 9370cc Electric Shaver that you find beneficial or impressive?

...

---

**I am generally satisfied** with the Braun Series 9 9370cc Rechargeable Wet & Dry Men's Electric Shaver with Clean & Charge Station. The product offers a smooth shaving experience, which I find impressive. The battery capacity is also commendable, lasting for a good amount of time. I appreciate the versatility of the product, as it can be used for both wet and dry shaving. The Clean & Charge Station that comes with the shaver is quite efficient and useful. It effectively cleans and charges the shaver, making it more convenient compared to other products I've used before. **However, I did encounter a minor issue where small hairs from my beard got stuck** in a particular spot on the blade area. This happened only once and did not significantly affect my overall shaving experience. Considering all the features and my experience with the product, I believe it offers good value for its price. I would recommend the Braun Series 9 9370cc Electric Shaver to others, mainly for its battery performance and its variety of features, including its wet and dry usage.

---

ison candidates. Finally, among these candidates, we selected the review of the product whose title achieved the highest ROUGE-L score (Lin, 2004).

We applied the same process when matching reviews generated by the baseline system; however, for some reviews collected using our system, there were no corresponding reviews in the baseline data that matched both category and rating. Therefore, we excluded these unmatched data. Consequently, while we conducted comparisons for all 100 reviews written by humans, only 96 baseline-generated reviews met the criteria.

### 4.2.3 Results and Discussions

The overall results are presented in Table 4. The annotators prefer the reviews generated by our system to those written by humans or generated by the baseline system. Notably, the reviews generated by our system are helpful, provide a balanced view of pros and cons, and offer comprehensive information. These findings indicate that our interview dialogue system is capable of eliciting a wide range of information about products from users through topic transitions.

The reviews generated by our system lack the fluency of human-written reviews. For instance, our review text generator tends to use the formal product title when referring to the product. Addi-

tionally, human-written reviews contain more individual experiences compared with those generated by our system. Despite these limitations, our system has high scalability, offering the potential for improvement. Specifically, our system's output could be enhanced by refining the prompts to generate texts that are more human-like and elicit detailed usage experiences from users.

By replacing the baseline system, which uses fixed questions, with our interview dialogue system, we observe improvements across all metrics. Notably, our system can generate reviews that are rich in experience-based information, contain more detailed information, and cover a broader range of topics. This demonstrates that our system can elicit more detailed and extensive information from users through follow-up questions and topic transitions.

### 4.3 Discussion on Predicted Ratings

To further explore the characteristics of the reviews and ratings generated by our system, we analyze them along two axes: the difference based on the source of the ratings (comparing ratings assigned by humans to those predicted by our system) and the difference based on the annotators (comparing the ratings given by system users to those assigned by third parties). To obtain ratings

Table 3: Questions in comparative evaluation

Labels in Table 4	Questions
Helpfulness	<i>Which review would be more helpful for making a purchase decision?</i>
Fluency	<i>Which review exhibits a more fluent and human-like writing style?</i>
Conciseness	<i>Which review is more concise and to the point?</i>
Experience	<i>Which review provides more information based on the actual usage experience of the product?</i>
Balance	<i>Which review presents a more balanced view of the product’s pros and cons?</i>
Depth	<i>Which review provides more in-depth information about any specific aspect of the product?</i>
Coverage	<i>Which review mentions a more comprehensive range of product aspects?</i>
Overall	<i>Which review is overall more preferable?</i>

Table 4: Results of third-party evaluation. The values represent the percentage of votes each received.

Reviews	Helpfulness	Fluency	Conciseness	Experience	Balance	Depth	Coverage	Overall
Human	38.0	<b>47.0</b>	37.0	<b>57.0</b>	37.0	43.0	40.0	41.0
Tie	6.0	15.0	6.0	9.0	15.0	10.0	5.0	7.0
Ours	<b>56.0</b>	38.0	<b>57.0</b>	34.0	<b>48.0</b>	<b>47.0</b>	<b>55.0</b>	<b>52.0</b>
Baseline	38.5	28.1	45.8	21.9	34.4	35.4	35.4	37.5
Tie	12.5	33.3	6.2	16.7	15.6	12.5	10.4	17.7
Ours	<b>49.0</b>	<b>38.5</b>	<b>47.9</b>	<b>61.5</b>	<b>50.0</b>	<b>52.1</b>	<b>54.2</b>	<b>44.8</b>

Table 5: Average absolute difference in ratings between Amazon customers and Turkers (top-left), between system-predicted ratings and Turkers’ ratings for system-generated reviews (top-right), and between system-predicted ratings and participants’ ratings (bottom-right, see Section 4.1.4).

Annotator/Source	Human-written	System-generated
Turkers	0.59	0.12
Participants in Section 4.1	-	0.57

assigned by third parties, we newly recruited annotators from MTurk and asked them to assign ratings to both the human-written reviews (left column)<sup>4</sup> and those generated by our system (right column). We also collected ratings assigned by participants from the experiments in Section 4.1. Note that these participants, unlike the Turkers, had seen the ratings predicted by our system.

The results in the top row of Table 5 demonstrate that the difference between the ratings predicted by our system and those assigned by third parties is remarkably smaller than the difference found in human-written reviews. This finding indicates that the sentiment of the reviews generated by our system is easily comprehensible to readers.

The ratings predicted by our system, as shown in the right column of Table 5, align more closely with those assigned by third-party annotators than with those of system users. This finding indicates that our system emphasizes objectivity over sub-

jectivity in its ratings.

The aforementioned observations indicate that our system generates review texts that are easy for humans to understand and provide more objective ratings. This finding suggests that our interview dialogue system and review text generator can generate reviews that accurately capture reviewers’ sentiments, thereby supporting informed purchasing decisions, while the rating predictor also provides highly objective and reliable ratings.

## 5 Conclusion

In this study, we present a novel method for utilizing dialogue systems to facilitate user review creation. Our approach involves three processes: interview dialogue, review text generation, and rating prediction. Although ensuring the fluency of the system-generated reviews remains a challenge, our method provides high-quality and helpful reviews for both reviewers and their readers.

Our method possesses high scalability. For instance, feeding product descriptions into our interview dialogue system could lead to deeper interview dialogues about more detailed information. However, our experiments have shown that even without such extensions, our system is capable of providing reviews that are more helpful than human-written ones. Furthermore, adapting our dialogue system’s strategies to user preferences during review writing could improve user experience. Further research can accomplish this objective by conducting a more detailed analysis of user preferences.

<sup>4</sup>For the annotations, we used 100 human-written reviews selected in Section 4.1.



## References

- Advait Bhat, Saaket Agashe, Parth Oberoi, Niharika Mohile, Ravi Jangir, and Anirudha Joshi. 2023. [Interacting with next-phrase suggestions: How suggestion systems aid and influence the cognitive processes of writing](#). In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23, page 436–452, New York, NY, USA. Association for Computing Machinery.
- Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020. [Unsupervised opinion summarization as copycat-review generation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5151–5169, Online. Association for Computational Linguistics.
- Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2006. The ami meeting corpus: A pre-announcement. In *Machine Learning for Multimodal Interaction*, pages 28–39, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Mingda Chen, Zewei Chu, Sam Wiseman, and Kevin Gimpel. 2022. [SummScreen: A dataset for abstractive screenplay summarization](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8602–8615, Dublin, Ireland. Association for Computational Linguistics.
- Yulong Chen, Yang Liu, and Yue Zhang. 2021. [DialogSum challenge: Summarizing real-life scenario dialogues](#). In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 308–313, Aberdeen, Scotland, UK. Association for Computational Linguistics.
- Li Dong, Shaohan Huang, Furu Wei, Mirella Lapata, Ming Zhou, and Ke Xu. 2017. [Learning to generate product reviews from attributes](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 623–632, Valencia, Spain. Association for Computational Linguistics.
- Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. [SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization](#). In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.
- Jihyeok Kim, Seungtaek Choi, Reinald Kim Amplayo, and Seung-won Hwang. 2020. [Retrieval-augmented controllable review generation](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2284–2295, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Soomin Kim, Joonhwan Lee, and Gahgene Gweon. 2019. [Comparing data from chatbot and web surveys: Effects of platform and conversational style on survey response quality](#). In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–12, New York, NY, USA. Association for Computing Machinery.
- Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. [Large language models are zero-shot reasoners](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.
- Junyi Li, Wayne Xin Zhao, Ji-Rong Wen, and Yang Song. 2019. [Generating long and informative reviews with aspect-aware coarse-to-fine decoding](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1969–1979, Florence, Italy. Association for Computational Linguistics.
- Pan Li and Alexander Tuzhilin. 2019. [Towards controllable and personalized review generation](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3237–3245, Hong Kong, China. Association for Computational Linguistics.
- Yu Li, Baolin Peng, Pengcheng He, Michel Galley, Zhou Yu, and Jianfeng Gao. 2023. [DIONYSUS: A pre-trained model for low-resource dialogue summarization](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1368–1386, Toronto, Canada. Association for Computational Linguistics.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Bodhisattwa Prasad Majumder, Shuyang Li, Jianmo Ni, and Julian McAuley. 2020. [Interview: Large-scale modeling of media dialog with discourse patterns and knowledge grounding](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8129–8141, Online. Association for Computational Linguistics.
- Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. [Justifying recommendations using distantly-labeled reviews and fine-grained aspects](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 188–197, Hong Kong, China. Association for Computational Linguistics.

- Jianmo Ni and Julian McAuley. 2018. [Personalized review generation by expanding phrases and attending on aspect-aware representations](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 706–711, Melbourne, Australia. Association for Computational Linguistics.
- Taro Okahisa, Ribeka Tanaka, Takashi Kodama, Yin Jou Huang, and Sadao Kurohashi. 2022. [Constructing a culinary interview dialogue corpus with video conferencing tool](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3131–3139, Marseille, France. European Language Resources Association.
- Zhaopeng Qiu, Xian Wu, Jingyue Gao, and Wei Fan. 2021. [U-bert: Pre-training user representations for improved recommendation](#). volume 35, pages 4320–4327.
- Vasu Sharma, Harsh Sharma, Ankita Bishnu, and Labhesh Patel. 2018. [Cyclegen: Cyclic consistency based product review generator from attributes](#). In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 426–430, Tilburg University, The Netherlands. Association for Computational Linguistics.
- Hanfei Sun, Ziyuan Cao, and Diyi Yang. 2022. [SPORTSINTERVIEW: A large-scale sports interview benchmark for entity-centric dialogues](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5821–5828, Marseille, France. European Language Resources Association.
- Quoc-Tuan Truong and Hady Lauw. 2019. [Multimodal review generation for recommender systems](#). In *The World Wide Web Conference, WWW ’19*, page 1864–1874, New York, NY, USA. Association for Computing Machinery.
- Xuan-Son Vu, Thanh-Son Nguyen, Duc-Trong Le, and Lili Jiang. 2020. [Multimodal review generation with privacy and fairness awareness](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 414–425, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. [Self-consistency improves chain of thought reasoning in language models](#). In *The Eleventh International Conference on Learning Representations*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.
- Xinyuan Zhang, Ruiyi Zhang, Manzil Zaheer, and Amr Ahmed. 2021. [Unsupervised abstractive dialogue summarization for tete-a-tetes](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14489–14497.
- Chao Zhao, Spandana Gella, Seokhwan Kim, Di Jin, Devamanyu Hazarika, Alexandros Papangelis, Behnam Hedayatnia, Mahdi Namazifar, Yang Liu, and Dilek Hakkani-Tur. 2023. [“what do others think?”: Task-oriented conversational modeling with subjective knowledge](#). In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 309–323, Prague, Czechia. Association for Computational Linguistics.
- Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2022. [Dialoglm: Pre-trained model for long dialogue understanding and summarization](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11765–11773.
- Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. [QMSum: A new benchmark for query-based multi-domain meeting summarization](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921, Online. Association for Computational Linguistics.
- Chenguang Zhu, Yang Liu, Jie Mei, and Michael Zeng. 2021. [MediaSum: A large-scale media interview dataset for dialogue summarization](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5927–5934, Online. Association for Computational Linguistics.
- Yicheng Zou, Bolin Zhu, Xingwu Hu, Tao Gui, and Qi Zhang. 2021. [Low-resource dialogue summarization with domain-agnostic multi-source pretraining](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 80–91, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

## A Prompt Template

### A.1 Prompt for Interview Dialogue

Table 6 shows a prompt template for interview dialogues. [PRODUCT\_NAME] is a placeholder for the product title, which will be replaced with the product title selected by the participant. [MAX\_QUESTION] and [MIN\_QUESTION] are placeholders for the maximum and minimum number of dialogue turns. In our experiments, we used 15 and 8, respectively.

## **A.2 Prompt for Review Generation**

Table 7 shows a prompt template for review generation. Similar to that for interviewing, [PRODUCT\_NAME] is a placeholder for the product title. [DIALOGUE] is a placeholder for the dialogue history, into which the interview dialogue history between our system and the participants is inserted.

## **B Baseline Details**

Table 8 shows a prompt template for review generation. Similar to that for interviewing, [PRODUCT\_NAME] is a placeholder for the product title. [DIALOGUE] is a placeholder for the dialogue history, into which the interview dialogue history between our system and the participants is inserted.

Table 6: Prompt template for interviewing.

---

Your role is “interviewer” and my role is “interviewee”.  
About the product I am going to present, please elicit my impressions and opinions from me when I have touched it.

Note the following statements.

1. The interviewer elicits the interviewee’s satisfaction and dissatisfaction (the positive and negative points) with the product in a well-balanced and detailed.
2. In response to the interviewee’s response, the interviewer asks more in-depth questions about the aspect or elicits feedback about other aspects of the product.
3. Be sure to attach the name of your role at the beginning of your utterance. Since your role is “interviewer”, your generation should begin with “Interviewer:”.
4. Don’t generate interviewee’s utterances.
5. Add “[Wait\_for\_Response]” at the end of your utterance and wait for my response.
6. You must ask at least [MIN\_QUESTION] questions. In other words, the dialogue must continue for [MIN\_QUESTION] or more turns.
7. Having fulfilled the 6th statement, you can terminate the interview at your discretion. However, the interview must be completed within [MAX\_QUESTION] turns.
8. When you terminate the interview, add “[End\_of\_Interview]” at the end of your utterance.

Now, please elicit my impressions and opinions about the following product from me.  
[PRODUCT\_NAME]

---

Table 7: Prompt template for review generation.

---

[DIALOGUE]

The above is a dialogue about “[PRODUCT\_NAME]” between the interviewer and the interviewee who has touched on this product.

Write a customer review about the product as if written by the interviewee, by briefly summarizing the important information mentioned in the above interview, such as the good and bad points of the product and the interviewee’s experience with it.  
Do not output the review’s title.  
The following is a body of the product review of the product written by the interviewee:

---

Table 8: Questions asked by the baseline system

---

Q-1	First, could you tell me about the features and functions of this product? What kind of product is this?
Q-2	What made you decide to purchase this product?
Q-3	If you have any points that you like or are satisfied with this product, please tell me in detail.
Q-4	What are the advantages of this product compared to other products?
Q-5	If you have any dissatisfaction with this product or areas for improvement for this product, please tell me in detail.
Q-6	What are the disadvantages of this product compared to other products?
Q-7	Who would this product be suitable for?
Q-8	Is this product worth the price? Also, why do you think so?
Q-9	Finally, do you have any requests or impressions about the product?

---