

Generating Comparative Table by LLM-based Product Review Summarization

Kanako Nakai¹[0009-0003-2219-4475], Takehiro Yamamoto¹[0000-0003-0601-3139],
and Hiroaki Ohshima¹[0000-0002-9492-2246]

University of Hyogo, Kobe, Japan
ad24c041@guh.u-hyogo.ac.jp
t.yamamoto@sis.u-hyogo.ac.jp
ohshima@ai.u-hyogo.ac.jp

Abstract. In this study, we propose a method for generating a comparative table by summarizing product reviews using a Large Language Model (LLM). A comparative table summarizes, for each aspect of two products, the evaluations present in reviews and the number of reviews for each evaluation. By using an LLM to create the table, it becomes possible to generate the table on demand and to change the aspects to be compared depending on the products. The proposed method employs an LLM to extract aspects and their associated evaluations from reviews. These evaluations are summarized for each aspect to create a table mapping evaluations to their respective aspects. We used a review dataset from Rakuten Ichiba and automatically summarized reviews using an LLM to generate comparative tables. We conducted a user study to verify whether the tables are useful for comparing products. Based on the results of the user study, the proposed method was found to be more helpful for comparison than the baseline.

Keywords: comparative table · LLM · contrastive summarization

1 Introduction

When comparing products, reviews are an important source of information. However, some products have a large number of reviews, and reading all of them takes a lot of time. Also, reviews on websites are often presented in no particular order relevant to specific feature comparisons, with various evaluations mixed together, and a single review might contain evaluations of multiple aspects. Because of this, it is not easy to find features that are common to both products, features that appear only in one product, or how evaluations differ for the same aspect. Even though reviews contain important information, the way they are shown on product pages is not suitable for comparing products.

We generate a **comparative table** that shows how each product is evaluated for the same aspect. When creating the comparative table, we also show how many evaluations there are for each evaluation. This makes it easy to see which evaluations are most common at a glance. The comparative table to be created

		Product A	Product B	
Aspect	Noise cancelling	○Noise cancelling is excellent(10) • I was surprised by how good the noise cancelling performance was. • The noise cancelling was also quite impressive. ⋮ × Noise isolation is not great(3) • Compared to noise-cancelling earphones from popular brands, a lot of outside noise gets in. Not so great... ⋮	○Satisfied with performance(2) • I'm satisfied with the noise cancelling performance. ⋮ × Performance is not high(8) • I don't think the noise cancelling function is even worth comparing to the AirPods. • Compared to models that allow noise cancelling mode switching, this one is less effective. ⋮	Summary sentence
	Performance of noise cancelling			Evaluation
	Battery life	○Good battery life(4) • The battery lasts a long time, and I don't need to charge it for over a week, so I can use it for long hours without worry. ⋮ × Doesn't last a month(1) • The battery on these earphones doesn't last even a month.	○Good battery life(2) • The battery lasts well, and I had no issues even after using it for about six month. ⋮ × Poor battery life(2) • It would have been perfect if the battery lasted a bit longer. ⋮	
battery	Battery level display		○Battery level display is easy to check(7) • The remaining vattery is shown on the case, making it easy to check. ⋮	

Fig. 1. An example image of a comparative table generated by review summarization.

is shown in Fig. 1, using earphones as an example. For example, the top row in Fig. 1 summarizes reviews about **“Noise cancelling.”** The table shows that Product A has 10 positive reviews saying *“Noise cancelling is excellent.”* and 3 negative reviews saying *“Sound isolation is not good.”* Similarly, for Product B, there are 2 positive reviews saying *“Satisfied with the performance.”* and 8 negative reviews saying *“performance is not high.”* From this, we can see that Product A has more positive opinions about “Noise cancelling” than Product B. By generating a comparative table like this, we believe it becomes possible to compare products in a short time based on actual user reviews.

To generate a comparative table for products, we automatically summarize reviews using LLMs. There are two main advantages to using LLMs. The first advantage is that the table can be generated on demand. This means we can change the aspects used for comparison depending on the products. For example, even within the same earphone category, when comparing earphones A and B, we might use aspects like “Sound quality,” “Design,” and “Size.” When comparing earphones C and D, we might use aspects like “Sound quality,” “Noise cancelling,” and “Bluetooth features.” This allows us to change the comparison aspects based on what is mentioned in the reviews. The second advantage is that it is not necessary to train the model separately for each domain. This makes it easier to apply the method to other product domains. Our proposed method does not require retraining the model for each product and can generate comparative tables for products from various categories.

To investigate whether the proposed method is useful for comparing products, we conducted a user study comparing it with the baseline method. In the user study, we asked three questions to measure information satisfaction, variety of

aspects, and discovery of new aspects. The results showed that the proposed method scored significantly higher than the baseline in all three questions.

2 Related Work

2.1 Multi Document Summarization

In this study, we generate a comparative table based on multiple reviews. This task is similar to multi-document summarization, which means summarizing several documents together.

Mehta et al. proposed a method to improve summary quality by using several summarization systems and focusing on the content that overlaps or comes from high-performing systems [10]. Atri et al. proposed a model called FABRIC, which is an encoder-decoder model designed to improve topic consistency and relationships between documents when summarizing multi-document [2]. Yan et al. proposed a model called SC^2 , which generates intermediate comparative tables to help LLMs improve the accuracy of comparative reasoning [13]. Zhang et al. proposed a method called TOMDS that focuses on topics to create summaries that match user needs and preferences [14]. Another study by Zhang et al. introduced fairness evaluation metrics and methods to create fair summaries across social attributes such as gender, political party, and emotion [15].

Although this study also summarizes multiple reviews, it is different from previous studies because it focuses on comparison.

2.2 Review Summarization and Sentiment Analysis

In this study, we extract aspects and evaluations from multiple reviews, then divide the reviews into positive and negative ones for each aspect and summarize them. This task is similar to review summarization and to the task of classifying reviews as positive or negative.

Dave et al. proposed a method to automatically find product features in reviews and classify them as positive or negative [4]. Tripathy et al. proposed a method that combines N-gram models and machine learning to classify reviews as either positive or negative [12]. Amplayo et al. proposed a method that creates synthetic training data using review-summary pairs. Their model predicts aspects at different levels of detail and generates summaries that focus on specific aspect queries [1]. A similar service to this research is Amazon’s “Customer Review Highlights,” which uses generative AI to create short summaries for each aspect in reviews ¹.

Our study also summarizes reviews by aspect, but it is different because it creates summaries that are aligned for two products, allowing for direct comparison.

¹ <https://www.aboutamazon.com/news/amazon-ai/amazon-improves-customer-reviews-with-generative-ai>

2.3 Contrastive Summarization

Recently, there has been research on contrastive summarization. Contrastive summarization is creating a summary that focuses on the differences between two documents based on the same topics. This study can be seen as one type of contrastive summarization because it summarizes reviews of two products by focusing on both similarities and differences based on the same aspects. Some related works have explored contrastive summarization using product reviews. Lerman et al. proposed contrastive summarization on customer reviews and showed that it is more useful than summarizing just one product [9]. Iso et al. proposed CoCoSUM, a framework that generates both contrastive summaries and common summaries from two sets of reviews for comparison [8]. Gunel et al. proposed a method called STRUM, which uses a pre-trained T5 model to extract aspects and their values from web pages and create a structured summary that helps compare multiple entities [6]. They also proposed STRUM-LLM, which uses LLMs to summarize information from several web pages by attribute and rank these attributes in comparative tables [7]. STRUM-LLM is similar to our study because it can also take two entities as input, collect information from the web, and generate comparative tables. However, STRUM-LLM does not provide details about how the system works, so it is hard to reproduce. Also, while our study uses a two-level structure to represent aspects, STRUM-LLM uses a single-level structure.

3 Proposed Method

In this section, we first explain the comparative table created in this study. Then, we describe our method. Finally, we explain the prototype we developed.

3.1 Generating Comparative Tables by Summarizing Reviews

In this study, we create a comparative table consisting of aspects, sub-aspects, evaluations, and summary sentences, as shown in Fig. 1. An **aspect** is a word that describes a feature of the product, and an **evaluation** is the part of the review that mentions the aspect. For earphones, for instance, aspects may include “Sound quality,” “Bluetooth,” or “Noise cancellation.” An evaluation could be a sentence like “*the sound quality is good.*” for the aspect “Sound quality,” or “*Bluetooth connection is easy.*” for the aspect “Bluetooth.” We call the more detailed categories under each aspect **sub-aspect**, and the sentences that summarize positive or negative evaluations for each sub-aspect are called **summary sentence**. For example, the aspect “Noise cancelling” includes a sub-aspect called “Noise cancelling performance.” In Fig. 1, the positive evaluations of Product A for this sub-aspect are summarized in the sentence “Noise cancelling is excellent.” In the comparative table, the number of evaluations included in each summary is shown next to the summary, and the evaluations are listed below it.

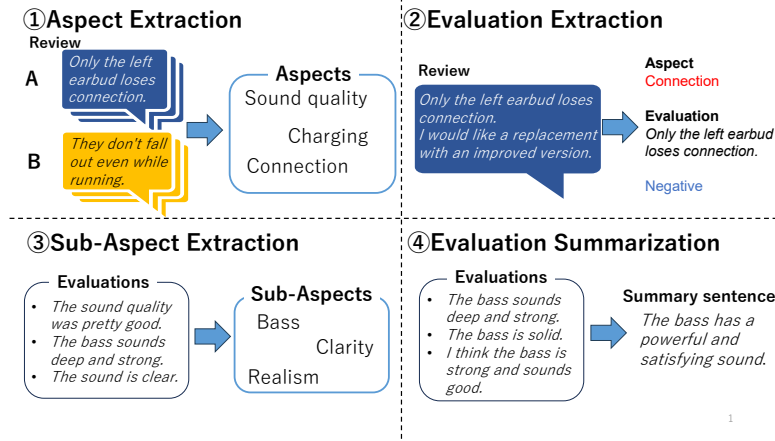


Fig. 2. Overview of the proposed method.

3.2 Overview of the Proposed Method

Fig. 2 shows an overview of the proposed method, which is explained below.

1. Aspect Extraction
 - We input the reviews of both products at the same time into an LLM to extract common and representative aspects.
2. Evaluation Extraction
 - We input a review into the LLM to extract evaluations that match each aspect.
3. Sub-Aspect Extraction
 - For each aspect, we input the related evaluations into the LLM to extract more detailed sub-aspects.
4. Evaluation Summarization
 - For each sub-aspect, we input the evaluations into the LLM and summarize them into one sentence.
 - We then match summary sentences from two products using the same sub-aspects.

In the following sections, we explain each step in detail. To help explain, we use an example with two products in the earphone category and their reviews, as introduced in Section 4.2. The reviews, prompts, generated texts, and system screenshots included in this manuscript were originally created in Japanese; English translations are provided herein.

3.3 Aspect Extraction

First, we input all reviews for each product into an LLM to extract aspects from the reviews. We use 100 reviews for each product in this study. Fig. 3

You are a review analyst. Unless otherwise noted, always respond in Japanese.
 The following are 100 reviews each for Product A and Product B.
 Product A's review : ***Only the left earbud loses connection. I want a replacement with an improved version./ ...***
 Product B's review : ***I bought these wireless earphones for running, and they're amazing. No rubbing noise from cords, they don't fall out even while running, and they feel comfortable. / ...***
 Given these reviews, what aspects would be useful for comparing these two earphones?
 Please list 3 aspects that appear often in Product A's reviews, 3 that appear often in Product B's reviews, and 3 that appear in both.

Fig. 3. Prompt for extracting aspects.

shows an example of the prompt used when inputting reviews like: “*Only the left earbud loses connection.*” and “*I bought these wireless earphones for running, and they're amazing.*” In this example, only two reviews are shown, and the other reviews are left out to keep the paper simple. In the actual prompt, 100 reviews are used for each product.

After inputting the reviews for the two earphone products, the LLM output the following: Product A aspects: “Price,” “Sound quality,” and “Ease of use.” Product B aspects: “Connection stability,” “Battery life,” and “Design.” Common aspects: “Sound quality,” “Ease of use,” and “Bluetooth connection.” When there are overlapping aspects among the frequently mentioned aspects in Product A, those in Product B, and the common aspects in both, we remove the duplicates to create a list without overlaps. In this case, the seven aspects are: “Design,” “Bluetooth connection,” “Price,” “Sound quality,” “Ease of use,” “Connection stability,” and “Battery life.”

3.4 Evaluation Extraction

Next, in order to summarize reviews by aspect in the comparative table, we extract evaluations related to each aspect. As explained in Section 3.6, we also judge whether each evaluation is positive or negative so that we can create separate summary sentences for positive and negative opinions. Each review is input one at a time into an LLM, and evaluations related to the aspects found in Section 3.3 are extracted. Fig. 4 shows an example of the prompt used for the review: “*Only the left earbud loses connection. I want a replacement with an improved version.*” From this review, the LLM extracts an evaluation related to the aspect “Connection stability”: “*Only the left earbud loses connection,*” and judges it as negative. The number of evaluations extracted in this step is also used to narrow down the important aspects. Based on the extraction results, we narrow down the aspects to the top 5 with the highest number of mentions. For the earphone example, the top five aspects were: “Sound quality,” “Design,” “Ease of use,” “Bluetooth connection,” and “Connection stability.”

From the review given by the user, extract evaluations that clearly mention any of the aspects listed. Also, judge whether the evaluation is positive or negative, and output either “positive” or “negative.” Use the same wording for aspects as in the aspect list.
 Aspect list : [“Design,” “Bluetooth connection,” “Price,” “Sound quality,” “Ease of use,” “Connection stability,” “Battery life”]
 Review : *Only the left earbud loses connection. I want a replacement with an improved version.*

Fig. 4. Prompt for evaluation extraction.

You are a review analyst. Unless otherwise instructed, always respond in Japanese. Below are reviews from product A and B that mention the aspect of **sound quality**.
 Reviews: *The sound quality was pretty good.* / *The sound quality is good, so at this price, I think it's acceptable.* / ...
 When comparing these two earphones, what specific points should be used to compare sound quality
 Please list three sub-aspects that often appear in product A's reviews, three in product B's reviews, and three that appear in both.

Fig. 5. Prompt for extracting sub-aspects.

3.5 Sub-Aspect Extraction

After evaluation extraction, sub-aspects are extracted in order to match summaries in the comparative table. We input the evaluations grouped by aspect into an LLM and extract sub-aspects. Fig. 5 shows the prompt used when inputting evaluations related to the aspect “Sound quality,” such as “*The sound quality was pretty good.*” and “*The sound quality is good, so at this price, I think it's acceptable.*” The figure shows only two of the input evaluations, and the other evaluations are omitted. From the reviews about the aspect of “Sound quality,” the following nine sub-aspects were extracted: “Sound quality,” “Bass,” “Sound skipping,” “Clarity of treble,” “Overall satisfaction,” “Fit,” “Sound spread,” “Cost performance,” and “Connection stability.” After extracting the sub-aspects, we align each evaluation with the sub-aspects using an LLM. The prompt provided to the LLM includes the evaluation and the set of sub-aspects, with instructions to output which sub-aspects are contained in the evaluation.

3.6 Evaluation Summarization

The evaluations are summarized by sub-aspect obtained in Section 3.5. We input the evaluations that mention the same sub-aspect into an LLM and summarize them into one sentence. Fig. 6 shows an example of the prompt where we input positive evaluations related to the “Sound quality” aspect and the “Bass” sub-aspect, such as: “*The bass sounds deep and strong.*” and “*The sound is clear and the deep bass is solid, making it easy to listen to.*” Only two evaluations are shown in Fig. 6; the others are omitted.

In this example, the output summary was: “*The rich bass and clear sound quality stand out, providing a natural and pleasant listening experience.*”

Please summarize the following reviews in one sentence
 Reviews: *The bass sounds deep and strong./ The sound is clear and the deep bass is solid, making it easy to listen to./ ...*
 Return only the summary.

Fig. 6. Prompt for summarizing evaluations.

1.Top page

Aspect	Bose SoundSport Free	Anker Zolo Liberty
Usability	○ The wireless earphones are free from cable hassle and are rated highly for quick connection, ease of use, and light weight, making them comfortable for everyday use. (16) ✕ Some reviews mention the buttons on the earphones are hard to press, lack of explanation about the app download, and difficulty in setting up. (5)	○ These earphones are easy to set up and pair, and users appreciate the smooth charging and power on/off operation, making them stress-free and allowing focused music enjoyment. (17) ✕ While some adjustments to the earphones are necessary, there are reports of issues during pairing and initial setup, as well as warnings about ear tip compatibility. (5)
Sound Quality	○ Overall, the sound quality is rated as excellent, exceeding expectations, with good bass, immersive sound, and clarity. (54)	○ Many reviews express satisfaction with the sound quality, saying it's good for the price and sufficient for users who aren't picky about sound. (38)

2. Display of summarization for each aspect

Aspect : Usability

[-- Back to Top](#)

Bose SoundSport Free	Anker Zolo Liberty
○ These wireless earphones are praised for freeing users from cable hassle, offering quick connectivity, ease of use, and lightness, making them comfortable to use in various daily situations. (16) ✕ The stiffness of the earphone buttons, the lack of clear instructions on how to download the app, and the difficulty of the setup process have been pointed out. (5) ✕ Issues such as poor connection in the left ear and unstable Bluetooth connectivity have been	○ This earphone is praised for its easy setup and pairing, smooth and stress-free charging and power on/off functionality, and the ability to enjoy music with full concentration. (17) ✕ Although some physical adjustments to the device are necessary, issues have been reported during pairing and initial setup, and users should also be cautious about the compatibility of the ear tips. (5) ○ This radio has no sound interruptions and features a stable magnetic mounting mechanism.

3. Display of evaluations

Bose SoundSport Free

- These wireless earphones are praised for freeing users from cable hassle, offering quick connectivity, ease of use, and lightness, making them comfortable to use in various daily situations. (16)
- I expect the true wireless and water-resistant design to perform well.
- Wireless is especially convenient for those using it on trains. No cords to get caught and no need to wind them up and put them in your bag.

4. Display of original reviews

- I expect the true wireless and water-resistant design to perform well.
- It seems the release date was moved up, and I got the Midnight Blue model today. While there may be other products with better treble clarity, this one creates a great atmosphere for enjoying music overall. Since I run marathons, I expect the true wireless and water-resistant features to perform well during training (though the appropriateness of wearing earphones while jogging is debatable...). So far, the fit is excellent.
- ease of use

Fig. 7. Overview of the prototype system.

3.7 Prototype System of Comparative Table

Finally, we built a prototype system of a comparative table that links summary sentences, evaluations, and original reviews based on aspects. An example of user operations is shown in Fig. 7. This example is a part of the comparative table actually used in the user study described in Section 4.3.

On the main screen, the summary for the most frequently mentioned sub-aspect of each aspect is displayed. By clicking on an aspect shown on the main screen, the user can move to the detailed page for that aspect. Each aspect page shows the summaries for each sub-aspect and the number of evaluations included in each summary. When a summary is clicked, the evaluations used to generate that summary are displayed. The original review that includes that evaluation is shown by hovering the mouse over an evaluation.

4 Experiments

In this study, we conducted two experiments: one to evaluate the evaluation extraction and another with users. First, we explain the data used in both experiments, and then describe each experiment in detail.

4.1 Data

The following describes the data used in the experiment. For the product reviews, we used the Rakuten Ichiba review data contained in the Rakuten Dataset [11]. Rakuten Ichiba is a Japanese e-commerce site. This dataset includes information such as product name, product ID, product genre ID, review title, and review content. For review summarization, 100 reviews were used for each product. The products used in the comparative table were chosen based on the following conditions: they belong to the same product genre ID, each has at least 100 reviews, and their prices are similar. To check the usefulness of our method in different domains, we used two types of products: earphones and microwave ovens. In each domain, we selected two products. The selected products are referred to as Earphone A and B, and Microwave A and B. In total, the experiment was conducted using two product pairs from two different product categories.

4.2 Experiment for evaluation extraction

First, we checked how accurately the LLMs can automatically extract evaluations from reviews. To see if this extraction accuracy changes depending on the model, we did experiments using multiple LLMs.

Large Language Model In the experiment, we use Gemma-2-Llama Swallow 9B IT v0.1(Swallow)² and Gemma 3 12b (Gemma) [5] as LLMs. Gemma-2-Llama Swallow 9B IT is an LLM based on Gemma 2 9B, with continued pre-training and instruction tuning focused on improving its Japanese language ability. Gemma 3 12b is a model built based on the research and technologies used to create Google’s Gemini model. In the experiment, we used a 6-bit quantized version of Swallow and a 4-bit quantized version of Gemma to keep the model sizes similar. To ensure stable output, the temperature was set to 0 for all models.

Evaluation Method This section explains how we evaluated the evaluation extraction described in Section 3.4. To evaluate the extraction, we measured the accuracy of the match between the aspect and the extracted evaluation. For example, if the evaluation “*I’m very satisfied with the sound.*” is extracted for the aspect “Sound quality,” it is considered correct. On the other hand, if the evaluation “*The design is cool.*” is extracted for the aspect “Sound quality,” it is considered incorrect.

² <https://huggingface.co/tokyotech-llm/Gemma-2-Llama-Swallow-9b-it-v0.1>

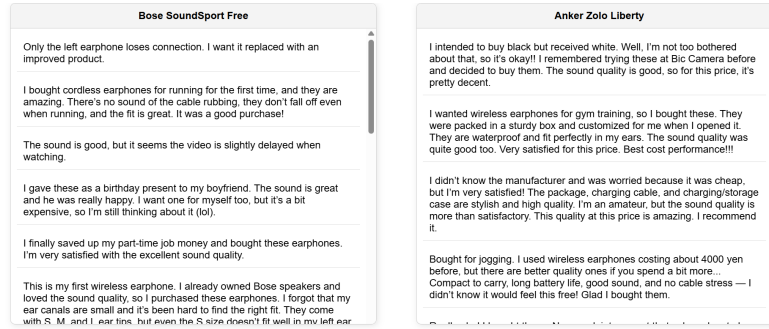


Fig. 8. Example of baseline system.

Results of Experiment for Extraction Evaluation For the top 5 aspects, Swallow extracted 314 evaluations for earphones and 249 for microwaves. Gemma extracted 391 evaluations for earphones and 308 for microwaves. The accuracy of evaluation extraction was 0.78 for Swallow and 0.85 for Gemma. Gemma outperformed Swallow in this task. The impact of accuracy on system usability is discussed in Section 5.

4.3 User Study

In the user study, we tested whether the proposed comparative table helps users compare products more effectively.

Large Language Model In the user study, we used the LLM Gemma-3-12b (Gemma) [5], which showed good results in the LLM comparison experiment in Section 4.2.

Participants The participants in the user study were 16 students from the University of Hyogo (8 male, 8 female). The user study lasted 1.5 hours per participant, and each participant was compensated approximately JPY 2,000. The experiment was conducted between May 29 and June 23, 2025.

Baseline Method As a baseline method, we used a system that displays the same 100 reviews for each product used in our proposed method, showing the reviews of the two products side by side. Fig. 8 shows an example of baseline system. The system displayed the reviews of each product next to each other, allowing users to scroll through the reviews for each product separately.

Task In the user study, participants first completed a training task to learn how to use the two systems. Then, they did a task which they compared two

Table 1. Tasks completed by each participant.

Participant	Task 1	Task 2
1, 5, 9, 13	Baseline method / Earphones	Proposed method / Microwave
2, 6, 10, 14	Proposed method / Earphones	Baseline method / Microwave
3, 7, 11, 15	Baseline method / Microwave	Proposed method / Earphones
4, 8, 12, 16	Proposed method / Microwave	Baseline method / Earphones

products using each system and answered a questionnaire afterward. Each participant completed two tasks, one for each system, in a within-subjects design. As shown in Table 1, the order in which participants performed each task with each system was counterbalanced. After completing all tasks, participants answered a final questionnaire. In the user study, each participant first compared either the earphones or the microwave ovens using either the system with the comparative table or the baseline system, and decided which product they would buy. Then, in the second task, they compared the other product category using the system they did not use in the first task. To make the task feel like real online shopping, we gave them a specific situation for using the product, along with product images, prices, and specifications. The two situations were: (1) Earphones to use at home while studying or attending online meetings (2) A microwave oven for someone starting to live alone. The following is an example of the task description for choosing “Earphones to use at home while studying or attending online meetings”

— A task example —

You are planning to buy new earphones for studying at home and joining online meetings. You have narrowed your choices down to two earphones. Using the system, compare the two products based on the reviews and decide which one to buy. You have 15 minutes for this task. If you make your decision before the time is up, you may stop the comparison. Also, please refer to the product prices and specifications shown below while comparing.

The time limit for each task was set to 15 minutes. If participants were able to decide which product to buy before the time was up, they were allowed to finish early.

Questionnaire After each task, a post-task questionnaire was conducted. In the post-task questionnaire, participants were asked which of the two products they chose, why they made that choice, what were the similarities and differences between the two products, and how much time they spent making their decision. In addition, participants answered three questions using a 5-point Likert scale (1: Strongly disagree, 2: Disagree, 3: Neither agree nor disagree, 4: Agree, 5: Strongly agree). The three questions were: “I was able to get enough information within the time limit.” “The system was helpful for comparing the products from

various aspects.” “The system helped me to understand which aspects to use for comparison.” These questions were asked to measure information satisfaction, variety of aspects, and discovery of new aspects. Furthermore, the usability of the system was evaluated using SUS [3]. SUS is a standardized usability score based on 10 questions, with a maximum score of 100 points. For SUS evaluation, participants answered 10 questions using a 5-point Likert scale.

After all tasks were completed, a final questionnaire was conducted. In the final questionnaire, participants were asked which system was more suitable for comparison and why. They were also asked to provide positive and negative points regarding the proposed method.

Study Results We report the results of the user study in this section. To compare the proposed method and the baseline method, we used Welch’s t -test with a 5% significance level.

First, the average time taken was 13.21 minutes for the baseline method and 11.58 minutes for the proposed method. There was no significant difference.

Next, to objectively evaluate whether the comparative table helped with the comparison, we analyzed the free-text answers from the questionnaires given after each task. Specifically, we analyzed the question that asked participants to list as many similarities and differences between the two products as they could think of. We counted one similarity or difference per aspect as one point and compared the baseline method and the proposed method. The average number of similarities mentioned for the two products was 3.56 for the baseline method and 4.56 for the proposed method, with the proposed method being higher. Similarly, for differences between the two products, the baseline method had an average of 2.25 points and the proposed method had 2.50 points, again with the proposed method being higher. There was no significant difference in either case. These answers also included information from specification sheets. To analyze how much information came from the comparative table, we excluded the similarities and differences that could be found only in the specification sheets. For similarities, the baseline method had 2.31 and the proposed method had 2.94, with the proposed method being higher. For different points, the baseline method had 1.56 and the proposed method had 2.25, again with the proposed method being higher. There was no significant difference in both cases.

The average and standard deviation of questionnaire scores by method and item are shown in Table 2. From the results, the proposed method scored higher than the comparative table method on all items. Also, there were significant differences in all cases.

The SUS scores were 62.19 for the baseline method and 75.63 for the proposed method. It is said that the average SUS score is 68, so the proposed method was higher than the average ³.

We describe the results of the final questionnaire. When asked “Which system was more suitable for comparison?”, all 16 participants chose the proposed method. Some of the reasons given were: “*It was easier to compare because the*

³ <https://measuringu.com/>

Table 2. Questionnaire Results.

Questionnaire item	Baseline method		Proposed method	
	Mean	Std.	Mean	Std.
Q. I was able to get enough information within the time limit.	3.75	1.18	4.94	0.25
Q. The system was helpful for comparing the products from various aspects.	2.69	1.30	4.69	0.48
Q. The system helped me to understand which aspects to use for comparison.	2.69	1.20	4.93	0.25

key points were summarized,” and “Since the information was organized by aspect, it was easy to compare the part I care most about.” On the other hand, some points for improvement were also mentioned. For example, *“Some positive summaries included negative evaluations in the details,”* and *“Some reviews were not related to the aspect, such as reviews about microwave sounds being included under the heating function.”* These comments pointed out that sometimes the summaries did not match the evaluations, and the aspects did not match the reviews. There were also suggestions to improve usability, such as: *“It would be more helpful if the positive or negative tone was easier to see at a glance,”* and *“I would prefer if the full review only appeared when I clicked on it.”*

5 Discussion

In the evaluation experiment, we checked the accuracy of evaluation extraction for two models. As a result, both models had about 0.8 accuracy. This shows that the models could extract evaluations related to each aspect with fairly high accuracy. However, in the final questionnaire asking for points to improve the comparative table, there were comments such as *“There was misclassification.”* and *“There were unrelated opinions included, and I was worried about that.”* When there are misclassifications, users of the comparative table can easily notice them, which may affect the trust in the table. Since the extracted evaluations are used later for sub-aspect extraction, classification, and summarization, we think that improving extraction accuracy in the future will lead to better matching, easier comparisons, and more accurate comparative tables.

In the user study, we automatically created a comparative table using an LLM for two product domains and compared it with a baseline method. For the earphones, aspects such as “Sound quality,” “Design,” “Ease of use,” “Bluetooth connection,” and “Connection stability” were obtained, and for the microwave ovens, aspects such as “Ease of use,” “Low price,” “Heating function,” “Compact size,” and “Quick response” were obtained. We were able to create the comparative table on demand. As a result of the user study, the proposed method was rated higher than the baseline method in subjective evaluations. On the other hand, there was no significant difference between the proposed method and the

baseline method in the results of listing similarities and differences. When extracting aspects and sub-aspects, the prompt instructed the LLM to extract aspects that appeared frequently in one product and aspects that were common to both products. This was intended to extract aspects (sub-aspects) that represent similarities and differences between the two products. However, this approach may not have been very effective. In the comparative table, positive and negative evaluations for each sub-aspect are shown side by side. With this style, the positive and negative evaluations for the same aspect do not always represent similarities. For example, for the aspect “Price,” one product may have a positive evaluation like *“The price is high, but I’m happy I bought it,”* while the other product may have *“I’m happy I could buy it for under 10,000 yen.”* In this case, both summary sentences are positive opinions about the same aspect, but they do not share the same opinion. In the future, it will be necessary to match not only by aspects or sub-aspects but also by the actual content of the evaluations. Also, in the final questionnaire asking about improvements to the comparative table, there was a comment that said, *“It would be better if the correspondence between the items was easier to understand.”* Therefore, we need to not only think about how to match the evaluations better but also improve the UI so that users can find similarities and differences more easily.

6 Conclusion and Future Work

In this study, we used an LLM to automatically summarize product reviews and generate a comparative table that shows the summaries by aspect.

To test the usefulness of the proposed method, we conducted a user study. Participants answered a questionnaire on three points: “I was able to get enough information within the time limit.”, “The system was helpful for comparing the products from various aspects.”, “The system helped me to understand which aspects to use for comparison.” The average scores for the proposed method were higher than the baseline method in all three items. In all three items, the difference was statistically significant. In the final questionnaire after all tasks were completed, all participants answered that the proposed method was more suitable for comparison. Based on these results, we can say that the proposed method is useful for product comparison. From the questionnaire about the proposed method, users pointed out positive points such as the summaries being organized by aspect and evaluation, and the number of opinions being clearly shown. On the other hand, some users mentioned that the content of the summary did not always match the actual evaluations. This shows that improvements are needed in the accuracy of evaluation extraction and summary generation.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Numbers JP25K03229, JP24K03228, and JP25K03228. We also used “Rakuten Dataset” (<https://rit>).

rakuten.com/data_release/) provided by Rakuten Group, Inc. via IDR Dataset Service of National Institute of Informatics.

References

1. Amplayo, R.K., Angelidis, S., Lapata, M.: Aspect-Controllable Opinion Summarization. In: Proc. of EMNLP’21. pp. 6578–6593 (2021)
2. Atri, Y.K., Iyer, A., Chakraborty, T., Goyal, V.: Promoting Topic Coherence and Inter-Document Consorts in Multi-Document Summarization via Simplicial Complex and Sheaf Graph. In: Proc. of EMNLP’23. pp. 2154–2166 (2023)
3. Brooke, J.: SUS:A ‘Quick and Dirty’ Usability Scale. Usability Evaluation In Industry **189**(194), 4–7 (1996)
4. Dave, K., Lawrence, S., Pennock, D.M.: Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In: Proc. of WWW’03. pp. 519–528 (2003)
5. Gemma Team: Gemma 3 (2025), <https://goo.gle/Gemma3Report>
6. Gunel, B., Tata, S., Najork, M.: STRUM: Extractive Aspect-Based Contrastive Summarization. In: Companion Proc. of the ACM Web Conference 2023. pp. 28–31 (2023)
7. Gunel, B., Wendt, J.B., Xie, J., Zhou, Y., Vo, N., Fisher, Z., Tata, S.: STRUM-LLM: Attributed and Structured Contrastive Summarization. arXiv preprint arXiv:2403.19710 (2024)
8. Iso, H., Wang, X., Angelidis, S., Suhara, Y.: Comparative Opinion Summarization via Collaborative Decoding. In: Findings of the Association for Computational Linguistics. pp. 3307–3324 (2022)
9. Lerman, K., McDonald, R.: Contrastive Summarization: An Experiment with Consumer Reviews. In: Proc. of HLT-NAACL’09 (Companion Volume: Short Papers). pp. 113–116 (2009)
10. Mehta, P., Majumder, P.: Content Based Weighted Consensus Summarization. In: Proc. of ECIR’18. pp. 787–793 (2018)
11. Rakuten Group, Inc: Rakuten Ichiba data. Informatics Research Data Repository, National Institute of Informatics. (dataset). (2020), <https://doi.org/10.32130/idr.2.1>
12. Tripathy, A., Agrawal, A., Rath, S.K.: Classification of sentiment reviews using n-gram machine learning approach. Expert Systems with Applications **57**, 117–126 (2016)
13. Yan, J.N., Liu, T., Chiu, J., Shen, J., Qin, Z., Yu, Y., Lakshmanan, C., Kurzion, Y., Rush, A., Liu, J., Bendersky, M.: Predicting Text Preference Via Structured Comparative Reasoning. In: Proc. of ACL’24 (Volume 1: Long Papers). pp. 10040–10060 (2024)
14. Zhang, X., Wei, Q., Song, Q., Zhang, P.: TOMDS (Topic-Oriented Multi-Document Summarization): Enabling Personalized Customization of Multi-Document Summaries. Applied Sciences **14**(5) (2024)
15. Zhang, Y., Zhang, N., Liu, Y., Fabbri, A., Liu, J., Kamoi, R., Lu, X., Xiong, C., Zhao, J., Radev, D., McKeown, K., Zhang, R.: Fair Abstractive Summarization of Diverse Perspectives. In: Proc. of NAACL-HLT’24 (Volume 1: Long Papers). pp. 3404–3426 (2024)