Generating Experiential Descriptions and Estimating Evidence Using Generative Language Model and User Products Reviews

Shen Chenfu
Graduate School of Science and Engineering
Aoyama Gakuin University
Kanagawa, Japan
shen@sw.it.aoyama.ac.jp

Yoshiyuki Shoji
Faculty of Informatics
Shizuoka University
Shizuoka, Japan
shojiy@inf.shizuoka.ac.jp

Takehiro Yamamoto
School of Social Information Science
University of Hyogo
Kobe, Japan
t.yamamoto@sis.u-hyogo.ac.jp

Katsumi Tanaka
Faculty of Informatics
The University of Fukuchiyama
Kyoto, Japan
tanaka-katsumi@fukuchiyama.ac.jp

Martin J. Dürst
College of Science and Engineering
Aoyama Gakuin University
Kanagawa, Japan
duerst@it.aoyama.ac.jp

***Abstract*—This paper introduces a method to transform technical product descriptions into user-friendly experiential descriptions, while also highlighting relevant parts of the original description. Product descriptions often are hard to understand without prior knowledge. For example, a beginner with a camera cannot understand technical descriptions like "ISO sensitivity 51,200". Our method translated this description to more relatable phrases such as "captures clear faces even at night." Our method adopts a generative language model to enable such experiential description generation and evidence estimation. Our method first trains a model with pairs of product descriptions and reviews. The trained model generates many candidate experiential descriptions when given product descriptions. After training, our method uses an ablation-based approach to estimate the evident description of the generated candidates. It checks for the frequency of words in the generated narrative when a portion of the description is removed. For example, terms like "night" or "clear" became less prevalent in reviews when "ISO sensitivity" was removed from the input description. Subject experiments with the actual review dataset verified our method's effectiveness in generating accurate narratives highlighting product features.**

***Index Terms*—Generative Language Model, Catchphrase, Slogan**

I. Introduction

In recent years, people who have never shopped online might be in the minority. The proliferation of online shopping sites and the rise of social commerce platforms like eBay have facilitated casual Internet trading. People of all ages and genders increasingly select and purchase online products. In this context, the importance of finding suitable products from the vast array available on online shopping sites has grown.

Typically, when seeking a product for a particular purpose, many people read the product description. However, product descriptions on online shopping sites often require domain-specific knowledge for comprehension. For instance, a novice who wishes to purchase a camera from an online site that can beautifully capture birds. Generally, camera product pages feature specification information such as "Image Sensor: APS-C" or "ISO 25,600". For people familiar with cameras, these specifications serve as valuable decision-making references. Yet, for novices without prior knowledge, understanding the implications of these specifications proves challenging, making it hard to determine if the camera suits the purpose of capturing birds.

In such cases, user-submitted product reviews act as additional sources of information to aid product selection. If a review mentions that the camera can capture birds beautifully, even novices can discern that the camera might be suitable for capturing distant, small, and moving subjects. Such experiential descriptions by purchasers assist many novices in judging the product's utility and making purchase decisions.

However, not all cameras suitable for bird photography necessarily feature a review stating their capability to capture birds beautifully. And if multiple cameras bear such reviews, deciding on the best option becomes complex. One must understand experiential descriptions and product descriptions to make informed decisions. Specifically, customers must sift through numerous reviews and product descriptions to deduce relationships like "Products often reviewed as capable of beautifully capturing birds tend to have high zoom capabilities and fast shutter speeds." Grasping the relation between product features and their results significantly burdens novices.

This research introduces an algorithm that, when fed a product description, outputs potential experiential descriptions from using the product and indicates the parts of the product description that serve as the basis for these narratives. Figure 1 illustrates a sample input-output scenario for the proposed application. The algorithm accepts any product's description.

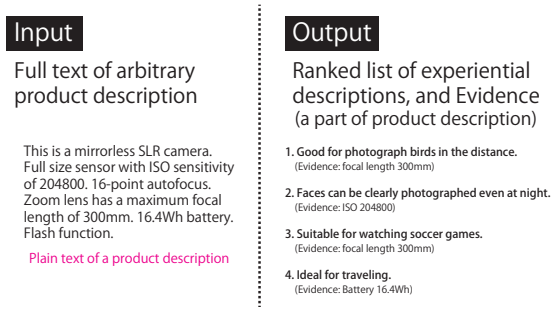

Fig. 1. Expected Inputs and Outputs. When a product description is entered, an experiential description tailored to its characteristics is output. Evidence is assigned to each description.

It then auto-generates multiple reviews likely to associate with the product's features. For each generated review, the system determines and indicates the specific feature in the product description that led to that narrative. For instance, a specialized term like "ISO sensitivity 409,600" can be translated into an experiential description such as "Can capture clear faces even at night," followed by the justification "Based on: ISO sensitivity 409,600". Even novices can expect a more straightforward product selection process by juxtaposing experiential descriptions and their justifications.

This study employed generative language models pre-trained on extensive corpora to convert product descriptions into experiential descriptions. This approach was adopted to:

- Understand the vague and diverse expressions found in user reviews,
- Facilitate natural language inference, and
- Cater to products that lack reviews.

The structure of this paper is as follows: The paper consists of a total of six sections. In this section, the background and motivations leading to this research have been discussed. In section II, related works about this research are introduced, and the positioning of this study is clarified. Section III describes the specific methodology proposed in this research. Section IV elaborates on the evaluation of the proposed methodology. In Section V, the results of the experiments are discussed. Section VI concludes this method and evaluation result, and shows future work.

II. Related Work

This research relates to three areas: studies utilizing user reviews, studies on generating recommendation phrases, and studies presenting reasons for recommendations. In sections II-A, II-B, and II-C, relevant studies in each area are introduced and discussed, respectively.

A. Studies Utilizing User Reviews

Recently, methods recommending items based on user reviews are becoming more common. For example, Zheng *et al.* [1] propose a deep learning model that jointly learns product or service attributes and user behaviors from user reviews to make recommendations.

B. Studies on Generating Recommendation Phrases

In this research, we generate experiential descriptions from product explanations. Studies have been done on automatically generating recommendation phrases in fields such as information recommendation. For example, Zhang *et al.* [2] report the results of introducing an Automatic Product Copywriting Generation system (APCG) using a machine learning model on the JD.com product recommendation platform, generating 2.53 million product descriptions over seven months. This research shows the feasibility and effectiveness of recommendation phrases using machine learning.

Additionally, Zhang *et al.* [3] generated recommendation phrases by inputting product descriptions based on a Self-labeling Conditional Variational Autoencoder (SLCVAE). Li *et al.* [4] proposed a Neural Template (NETE) description generation framework to balance sentence quality and expressiveness in recommendation systems using a generative approach. Chan *et al.* [5] proposed a machine learning model called S-MG Net for creating product descriptions and advertisements spanning multiple products. Deng *et al.* [6] proposed SGS-PAC, a system for automatically personalizing advertisement content according to consumer needs. While these studies aim to generate product advertisements or catchphrases, our study aims to make it easier for beginners to select products by presenting both the phrase and its rationale.

C. Studies on Presenting Reasons for Recommendations

This section discusses studies on presenting rationales for recommendations and explainability in machine learning models. Recently, the research area known as "Explainable AI," which explains the reasoning of machine learning models, is rapidly growing [7]. Especially in fields like product descriptions, as in this research, studies aiming to clarify reasons for recommendations and ensure trustworthiness are becoming active [8].

Zhang *et al.* [9] proposed an Explicit Factor Model (EFM) that presents both recommended and non-recommended products according to user interests based on product features and user tendencies and provides reasons for why a product is or isn't recommended.

Sinha *et al.* [4] investigated user perceptions across five music recommendation systems and found that users have a favorable view of recommendations when they feel there is high transparency in their reasons. This study indicates the effectiveness of presenting reasons during recommendations.

III. Method

Our method accepts product descriptions as input and outputs several experiential descriptions highlighting the product's features. Simultaneously, it provides the portion of the product description that serves as the basis for each description. To generate these experiential descriptions, a method for fine-tuning the pre-trained language model, GPT-2, is

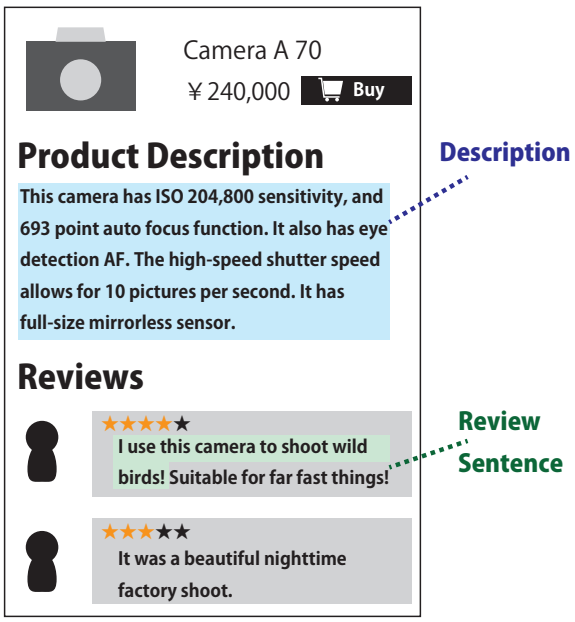

Fig. 2. An example of training data consists of a product description and reviews.

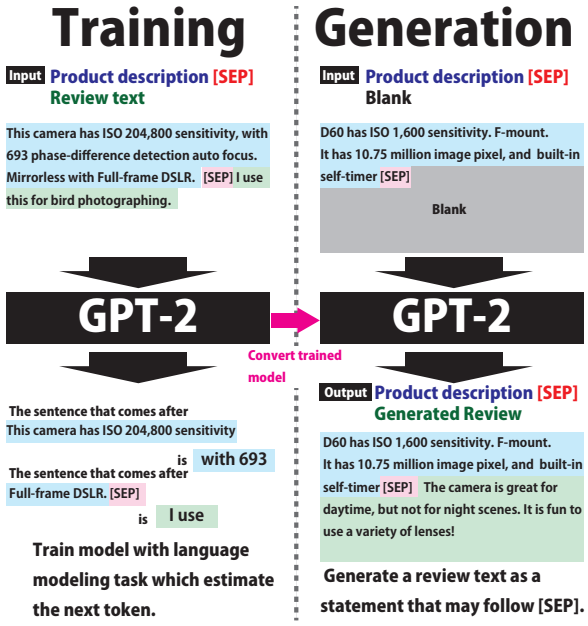

Fig. 3. Example of actual input/output text for the training and generation phases. Given a product description, the model generates a review likely written for it.

proposed. Furthermore, a method based on ablation is also introduced to estimate the foundation for the generated descriptions.

A. Cleansing and Preprocessing of Product and Review Data

This method utilizes product descriptions and their associated user reviews as datasets (illustrated in Figure 2). Since the review data consists of general user submissions, the content varies in quality; some might not be in proper Japanese, while others may be irrelevant to the product. Thus, data cleansing was necessary for preparing the learning data. Specifically, product descriptions and user reviews underwent rule-based data removal. A blacklist of words was created to eliminate sentences containing specific terms. Product descriptions often include information irrelevant to the product, such as promotions and shipping costs. Consequently, a manual list of terms frequently used in sales or shop information was created, encompassing words like "special price," "shipping," and "shipping fee." Reviews containing these terms were removed from product descriptions.

B. Fine-tuning using Product Descriptions and User Reviews

A language model that generates multiple potential reviews for a given product description was developed using GPT-2. Although GPT-2 is a general-purpose large-scale generative language model, publicly available models typically predict and produce sentence continuations. Adjustments were made to input product descriptions and output reviews by fine-tuning such models.

The data structure and tasks during actual training are illustrated on the left side of Figure 2. For learning purposes, pairs consisting of product descriptions and corresponding written reviews were extracted. In cases where multiple reviews were written for a single product, they were kept separate, resulting in numerous pairs with the same product description but different reviews. Subsequently, the product description and its review were connected using a unique token. Specifically, the commonly used [SEP] separator token was inserted between the product description and the review, combining them into one sentence. By training with such data, GPT-2 learns that pre-[SEP] content tends to be product-description-like, while post-[SEP] content leans towards review-like sentences. In this manner, a language model was created that generates a review as a continuation of the input text when provided with a product description and [SEP].

C. Generating Experiential Descriptions Using Fine-Tuned GPT-2

Experiential descriptions are generated using the fine-tuned GPT-2 language model (as shown on the right side of Figure 3). Providing the model with any product description followed by [SEP] produces a review as a continuation of the given product description. The generated reviews vary in length, from one-line or one-sentence descriptions to several lines. For this study, the generated reviews were split into individual sentences, and those ranging from 15 to 70 characters were extracted for experiential descriptions.

D. Estimation and Ranking of the Basis in the Product Description

Our method ascertains which portion of the product description each generated experiential description pertains to. For this purpose, a method based on the Ablation Study,

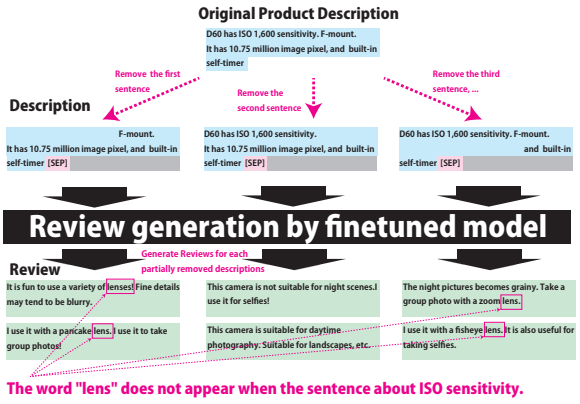

Fig. 4. Overview of ablation-based method

widely employed in machine learning, was utilized. In ablation, certain features are removed during training to determine their contribution to the machine learning output, compared with when these features are not removed. Following this methodology, a portion of the description was concealed, and the frequency of words in the generated review was used to determine which description influences which experiential statement. The experiential descriptions were then ranked based on the confidence of their bases.

The outline of the rationale estimation method based on ablation is depicted in Figure 4. First, one sentence is removed from the product description, which consists of multiple sentences. Then, the trained model is prompted to generate 1,000 experiential descriptions using the remaining part of the product description as input. Next, only the nouns that appear within the set of the 1,000 generated experiential descriptions are extracted. By doing this, words that no longer appear in the generated reviews when a specific sentence is removed from the input are aggregated based on their appearance probability. For instance, when a description related to "shutter speed" is omitted from the product description, there's a tendency for reviews containing terms like "sports" or "without blur" to be generated less frequently. In this manner, the word appearance probability for each sentence in the description is calculated by shifting which section of the product description is removed.

Since the data used in the experiment is in Japanese, words are not separated by spaces. Therefore, a morphological analyzer was used to split the sentences into words and estimate their parts of speech. For this purpose, MeCab was used as the morphological analyzer. Additionally, because reviews contain many colloquial expressions and proper nouns, MeCab-ipadic-NEologd, a morphological dictionary with many new words and specialized terms, was used.

In the next step, the experiential descriptions created in Section III-C are ranked using their probability of occurrence. This study aims to produce experiential descriptions with high accuracy in their basis estimation. Thus, regardless of the likelihood of the description being generated (*i.e.*, its plausibility) or whether it reflects the product's merits, the ranking is based solely on the credibility of its basis.

We compile the differences in occurrence probabilities of words in a generated review to determine which sentence, when hidden, makes it less likely to produce a given word. From this data, irrespective of the group of experiential descriptions generated from different product descriptions, the top five words with the most significant differences in occurrence probabilities are extracted. Note which product description sentences were removed when these words were extracted. Then, those containing these words are extracted from the experiential descriptions generated using the full product description (with no sentences removed). Ultimately, these extracted experiential descriptions can be obtained as having their basis in the noted sentences of the product description. An explanation is provided on why the final extracted experiential descriptions can be considered to have their basis in the noted sentences of the product description.

For instance, when generating an experiential description from a product description by removing only one sentence l_i, assume the occurrence probability of a word w decreases compared to when another sentence l_j is removed. This implies that the experiential description containing the word w correlates with the sentence l_i in the product description, and thus, it can be considered as its basis.

E. Presentation of Experiential Descriptions and the Corresponding Basis within Product Descriptions

This section presents the experiential descriptions and their corresponding basis in the product descriptions, as identified in Section III-D. The number of experiential descriptions within the group obtained in Section III-D is not fixed. Therefore, in this study, a single description is randomly selected from this group and presented along with the portion of the product description that serves as its basis. Specifically, five sets of words have the most significant difference in occurrence probability and their corresponding basis in the product descriptions. For each set, one complementary experiential description is presented. Thus, five experiential descriptions and their corresponding portions of product descriptions are presented.

IV. Evaluation

In this section, we discuss the dataset used to demonstrate the usefulness of the proposed method, details of the evaluation experiment conducted using this dataset, and its results. An experiment involving human participants was conducted in this study to evaluate the efficacy of the proposed method.

A. Dataset

To create training data for fine-tuning the GPT-2 model, product information and reviews from Rakuten Ichiba, a primary Japanese online shopping site, were used. In the experiments, we evaluated four product categories: earphones and cameras, where catalog specifications are essential; and T-shirts and bread, where product descriptions from the catalog may not be considered reliable.

Product descriptions and reviews of products with at least one review were extracted for data extraction and preprocessing. Given the token limitation of 512 tokens, product descriptions were truncated to 320 tokens, and review texts were truncated to 192 tokens. These were then concatenated using the [SEP] token to form the training data. Regarding dataset size, there were approximately 13,000 sets of product descriptions and reviews for earphones and around 3,500 sets for cameras.

B. Implementation

The GPT-2 model used in this study is a large-scale Japanese language model, which Rinna made open-source. This model was trained over about a month using 70 gigabytes of Japanese text from the CC-100 dataset, ensuring its versatility. The GPT-2 model was then further trained using the dataset described in Section IV-A.

C. Comparison Methods

In this section, we describe the methods compared with the proposed method to investigate its effectiveness. To evaluate the proposed method, we compare it with methods that use different approaches during fine-tuning: an approach based on co-occurrence frequency, a method that does not present any justification from the product description, and a method that only shows the original product description as a reference. Specifically:

- **Proposed Method**: The method proposed in this study generates experiential descriptions using GPT-2 fine-tuned with both product descriptions and review texts.
- **Review-only Training**: A method using GPT-2 fine-tuned exclusively with product review texts.
- **No Justification Presentation**: A method that presents the experiential description generated by GPT-2 without estimating the justification from the product description, displaying the description with the top five most frequent words from the generated experiential descriptions.
- **Co-occurrence Frequency (Baseline)**: A method that computes word sets co-occurring in product descriptions and corresponding reviews. The method then ranks them by the frequency of co-occurrence and extracts actual review sentences from the dataset containing words co-occurring with words in the given product description. The extracted sentence from the product becomes the justification.
- **Only Product Description Presentation**: A method that only displays the unaltered product description without any experiential description.

The baseline method based on co-occurrence frequency extracts potential justifications from the product description without any generation, resulting in one sentence from the reviews. An experiment also included GPT-2 trained only on review texts. This is because, in product reviews, specifications often mentioned in product descriptions can also appear. Therefore, even without introducing on product descriptions, if a product description is given to GPT-2 trained to continue reviews, it might generate experiential descriptions corresponding to the specifications. These methods were then compared with the proposed method.

D. Experimental Method

This section presents the experimental method used to evaluate the experiential descriptions and their justification from the product descriptions generated by the proposed method.

An evaluation was conducted using a human subject experiment. Three participants evaluated the experiential descriptions in a survey format. These descriptions were listed in advance for the participants to read. The actual product and experiential descriptions generated from the evaluation were provided to the participants during the evaluation.

The participants were evaluated based on the following six criteria:

1) **Naturalness**: Whether the experiential description is in natural Japanese,
2) **Experientialness**: Whether the experiential description seems genuinely experiential,
3) **Correctness**: Whether the content of the experiential description is correct,
4) **Interestingness**: Whether reading the experiential description and its justification in the product description sparked interest in the product,
5) **Helpfulness**: Whether understanding of the product deepened after seeing the experiential description and its justification in the product description,
6) **Evidentness**: Whether the correct part of the product description was presented as the justification for the experiential description.

Each criterion was evaluated using a 5-point Likert scale. Among these criteria, the sixth one, concerning whether the correct section of the product description was given as justification, might be challenging to assess without deep knowledge of the product domain. Therefore, for this particular criterion, an additional expert from each domain (a total of two experts) was recruited to evaluate.

E. Experimental Setup

For the experiment, 15 product descriptions were prepared for each product domain, totaling 60 product descriptions. For every product description, five experiential descriptions and their respective justifications from the product description were presented for each method. It is worth noting the differences in presentation depending on the method:

- For the "Product Description Only" method, neither the experiential description nor the justification from the product description was presented.
- For the "Without Justification" method, only the experiential description was shown.
- For all other methods, pairs of experiential descriptions and their justifications from the product descriptions were presented.

To ensure a blind evaluation, the presentations were randomly shuffled so the participants could not determine which

TABLE I
AVERAGE RATINGS FOR BOTH EARPHONES AND CAMERA PRODUCT CATEGORIES (OUT OF 5; ** $p < 0.01$, * $p < 0.05$)

Evaluation Item	Proposed Method	Review Learning	Co-occurrence	Without Justification	Product Description
Naturalness	****4.36**	4.25	4.18	*4.01	
Experientialness	4.20	**4.04	**4.32**	**3.96	
Correctness	**3.66	****3.86**	3.45	3.40	
Interestingness	****3.90**	**3.87	3.58	**3.23	3.80
Helpfulness	**3.80	****3.86**	3.61	**3.16	3.77
Evidentness	**3.39	****3.69**	3.00		

TABLE II
AVERAGE RATINGS FOR EARPHONES PRODUCT CATEGORY (OUT OF 5; COMPARED TO CO-OCCURRENCE ** $p < 0.01$, * $p < 0.05$)

Evaluation Item	Proposed Method	Review Learning	Co-occurrence	Without Justification	Product Description
Naturalness	***4.35**	4.18	4.16	**3.81	
Experientialness	**4.35**	*4.12	4.34	**3.82	
Correctness	*3.59	****3.83**	3.40	3.33	
Interestingness	****3.93**	**3.88	3.52	**3.11	3.78
Helpfulness	*3.77	****3.88**	3.56	**3.09	3.73
Evidentness	3.23	****3.70**	3.04		

TABLE III
AVERAGE RATINGS FOR THE CAMERA PRODUCT CATEGORY (OUT OF 5, COMPARED TO CO-OCCURRENCE ** $p < 0.01$, * $p < 0.05$)

Evaluation Item	Proposed Method	Review Learning	Co-occurrence	Without Justification	Product Description
Naturalness	**4.36**	4.32	4.20	4.20	
Experientialness	*4.06	**3.96	**4.30**	4.11	
Correctness	*3.73	****3.88**	3.51	3.47	
Interestingness	***3.88**	*3.87	3.64	*3.36	3.82
Helpfulness	3.82	**3.84**	3.65	**3.24	3.80
Evidentness	**3.54	****3.67**	2.96		

TABLE IV
AVERAGE RATINGS FOR T-SHIRT PRODUCT CATEGORY (COMPARED TO CO-OCCURRENCE ** $p < 0.01$, * $p < 0.05$)

Evaluation Item	Proposed Method	Review Learning	Co-occurrence	Without Justification	Product Description
Naturalness	4.56	**4.32	**4.67**	**4.46	
Experientialness	4.29	**4.13	**4.44**	*4.24	
Correctness	****3.44**	**3.43	2.91	2.95	
Interestingness	**3.72	*3.64	3.40	3.31	***3.80**
Helpfulness	3.50	3.62	3.62	**3.28	**3.87**
Evidentness	**3.42	****3.54**	2.71		

method produced which output. Each participant read the experiential descriptions and justifications for every product and evaluated them based on the six criteria. In cases where a method did not present either the experiential or product description, the corresponding evaluation criteria were omitted.

This experimental setup was meticulously crafted to comprehensively evaluate each method's outputs without introducing biases. The design allows for a clear comparison and an understanding of the strengths and weaknesses of each approach.

F. Experimental Results

This section presents the results of the subject experiments described in Section IV-D. A t-test was performed to compare the proposed method with the co-occurrence-based method. Items with a statistically significant difference are denoted with an asterisk (*).

The results, which compile data from the earphone and camera product categories, are shown in Table I. The results exclusive to the earphone and camera categories are is depicted in Table II and III, respectively. In these domains, where catalog specifications are crucial, the proposed method demonstrated a significantly higher accuracy concerning the naturalness of the Japanese language, correctness, understanding of the product, and purchase desire than the method that merely extracts relevant sections.

Similarly, the T-shirt product category results, where catalog information isn't as vital, are shown in Table IV. The comparison concerning bread is provided in Table V. Based on the experimental outcomes, significant differences were observed across all product domains in multiple evaluation metrics, including the level of interest in the product, when comparing the proposed method to the baseline. Moreover, the approach utilizing GPT-2 trained solely on review texts also demonstrated significantly higher accuracy across all product domains in multiple metrics compared to simple extraction.

As an example of generated experiential description, table VI shows translated sections for product description, experiential description, and justification, specifically for earphones. The generated description certainly describes the noise cancellation performance of the earphones. The content is specific enough to understand the comfort of use.

TABLE V
AVERAGE RATINGS FOR THE BREAD PRODUCT CATEGORY (COMPARED TO CO-OCCURRENCE ** $p < 0.01$, * $p < 0.05$)

Evaluation Item	Proposed Method	Review Learning	Co-occurrence	Without Justification	Product Description
Naturalness	4.28	**4.47	4.27	4.27	
Experientialness	**3.65	3.91	**4.00**	**3.45	
Correctness	**3.34	**3.51	2.81	2.66	
Interestingness	**3.54	**3.77	3.22	*2.91	**3.82
Helpfulness	**3.48	**3.64	3.05	*2.82	**3.84
Evidentness	**3.29	**3.66	2.84		

TABLE VI
HIGH-SCORING EXAMPLE OF EARPHONES IN THE PROPOSED METHOD

Product Description	Delicate and rich sound (unique driver × high-res playback). Equipped with Anker's proprietary driver "A.C.A.A 3.0," it realizes delicate and rich sound quality where even subtle sounds are audible, thanks to two dynamic drivers. Supporting the high-quality codec LDAC, it transmits three times more information than regular codecs (Bluetooth A2DP's SBC, 328kbps, 44.1kHz), allowing for faithful music reproduction to the original sound. 360° audio experience with 3D audio. Anker's unique algorithm processes the sound source in real-time, providing an acoustic experience like being at a live venue or cinema. The gyroscope sensor detects head movements, constantly offering an immersive musical experience. You can choose between Music Mode and Movie Mode. Ultra Noise Cancelling 2.0: Anker's proprietary Ultra Noise Cancelling 2.0 automatically adjusts the strength of noise canceling according to the surrounding noise level, maximizing immersion in music without being affected by the environment. Health monitoring includes heart rate, stress checks, posture reminders, and workout functions.
Experiential Description	I use them regularly during my commute on the subway and airplanes. Even in crowded trains, the noise canceling doesn't cut off, and I have no complaints.
Evidence	Ultra Noise Cancelling 2.0: Anker's proprietary Ultra Noise Cancelling 2.0 automatically adjusts the strength of noise canceling according to the surrounding noise level, maximizing immersion in music without being affected by the environment.

TABLE VII
HIGH-SCORING EXAMPLE OF CAMERAS IN THE PROPOSED METHOD

Product Description	With technology developed for full-frame cameras, the accuracy, speed, and tracking of eye detection have been greatly improved from this camera's Eye AF. It's now easy and reliable to continuously focus on the eyes, even in moving portrait shots. Eye AF activates simply by half-pressing the shutter, and you can switch between the left and right eyes of the participant. Furthermore, 'Real-time Eye AF' is also compatible with certain animals. This allows for high-speed and precise detection and tracking of the eyes of pets and wildlife. A densely packed 425-point phase-detection AF sensor covers about 84 % of the imaging area. Additionally, contrast AF points increased from 169 in this camera to 425. The conventional Lock-on AF has been revamped with the 'Real-time Tracking' feature. You can designate the subject, half-press the shutter, or touch it on the monitor, and the camera will automatically track it with high accuracy. Silent shooting is now available during high-speed continuous shooting. It allows shooting in quiet scenes without worrying about shutter noise, ensuring you don't miss the decisive moment. It's also compatible with AF-C, Real-time Eye AF, and Real-time Tracking.
Experiential Description	I photograph animals, so when I take close-up shots, the expressions of the moving subjects are captured exactly as they are.
Evidence	You can shoot quietly even in scenes where you want to be silent, without worrying about shutter noise, ensuring you don't miss the decisive moment.

Other generated result is shown in the table VII as an additional example regarding cameras. In this output example, the specific information that the shutter is quiet was rewritten as the experience of being able to approach an animal and take a picture without being noticed. Therefore, many participants rated this review description as correct.

V. DISCUSSION

In this section, based on the results obtained from the evaluation experiments, we discuss the effectiveness and characteristics of the proposed method. Overall, the experimental results revealed a significant difference across all product domains in multiple evaluation metrics, such as the level of interest in the product, compared to the baseline method. This indicates that generative approaches can offer a more engaging experiential description than simple extraction algorithms. We will further examine the effectiveness of the proposed method from several perspectives.

First, we consider the features that appear in each product domain. For the earphones and camera domains, the proposed method and the method that presents only the product description received equivalent scores in two metrics: interest in the product and the depth of product understanding. Conversely, for the T-shirt and bread domains, the method that presented only the product description received higher scores for these two metrics.

These products possess multiple specification items in the domains of earphones and cameras. For instance, earphones have various specifications like being wired or wireless, having noise-canceling features, and the type of drivers used. A product description stating "Equipped with noise-canceling functionality" could correspond experientially as "Train noises are not bothersome", thus specifications can serve as the basis for experiential descriptions. However, in the T-shirt and bread domains, the number of specifications written for the product is limited. Subjective elements such as appearance or taste are challenging to infer from product descriptions. Therefore, products that do not have detailed specifications in their descriptions may not be suitable for our method.

Considering the earphone and camera domains where our method performed particularly well, we reflect on the effectiveness of presenting experiential descriptions. When comparing the proposed method to the one that only presents the product description, no significant difference was observed in both metrics: interest in the product and depth of understanding. Given that some of the experiential descriptions generated by our method clearly contradicted the original product descriptions, improving the accuracy of these narratives could further demonstrate the effectiveness of presenting such descriptions.

The experiential descriptions generated in our method are based on only one sentence from the product description, discussing only one specification. However, the entire product description naturally encompasses all specifications of the product. Thus, there's an inherent difference in the amount of information between experiential descriptions and product descriptions, which might have influenced the evaluations.

Next, we discuss the effectiveness of presenting the basis for the product description. Comparing our method to the method

that does not present this basis revealed significant differences in all five metrics. This confirms that presenting the basis is more effective than not doing so.

We also reflect on the characteristics of the algorithm used to estimate the basis. Although the ablation-based estimation received relatively high evaluations (3.39 out of five points), many examples were generated where the basis was incorrect. One reason could be redundant statements in the product descriptions or multiple product features influencing a single experience. Our method estimated the basis by masking one sentence from the product description at a time. This approach faced challenges when one specification presupposed the presence of another. Strategies like masking similar sections at the word level for training could be considered in the future.

Finally, we discuss using Different Fine-tuning Tasks for GPT-2. A method that employed GPT-2, fine-tuned using only review texts, achieved high ratings comparable to the proposed method, despite being fed product descriptions not present in the training data. This can be attributed to the numerous instances where product descriptions and reviews contained overlapping information. Reviews often included statements such as "this specification makes it suitable for this application," which would typically be found in product descriptions. Vendors can draft product descriptions in the dataset from Rakuten Ichiba used in this study. As a result, in addition to catalog specifications and objective product details, some descriptions integrated excerpts from reviews as "customer voices" or included experiential statements as "salesperson recommendations." Such content possibly contributed to the model's high performance trained solely on reviews when generated from product descriptions.

VI. Conclusion and Future Challenges

This research proposed an algorithm that, when provided with a product description, outputs an experiential statement suited for it, and the corresponding basis within the product description. The proposed method utilized GPT-2, fine-tuned with product descriptions and review texts, to generate these experiential descriptions. By creating review texts while obscuring parts of the product description, the study inferred how specific contents in a product description would likely influence certain review narratives.

A comparative evaluation of the outputs from the proposed method and other methods across six evaluation metrics was conducted through experimental evaluations. The results reaffirmed the efficacy of generative approaches like the proposed method and methods showcasing the underlying basis. However, numerous instances where the presented basis within the product description was incorrect were observed. As detailed in section III-D, the current approach extracts the top five terms based on simple occurrence probability and then randomly selects and presents an experiential statement containing any of these terms. Enhancing this process is anticipated to improve the quality of the presented experiential descriptions and the accuracy of their corresponding basis. Addressing these challenges is deemed as the future direction for this research.

Acknowledgment

This work was supported by JSPS KAKENHI Grants Number 21H03775, 21H03774, and 22H03905. We used "Rakuten Dataset [1]" provided by Rakuten Group, Inc. via IDR Dataset Service of National Institute of Informatics [10].

References

[1] L. Zheng, V. Noroozi, and P. S. Yu, "Joint deep modeling of users and items using reviews for recommendation," in *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, 2017, pp. 425–434.

[2] X. Zhang, Y. Zou, H. Zhang, J. Zhou, S. Diao, J. Chen *et al.*, "Automatic product copywriting for e-commerce," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 11, 2022, pp. 12 423–12 431.

[3] Y. Zhang, Y. Wang, L. Zhang, Z. Zhang, and K. Gai, "Improve diverse text generation by self labeling conditional variational auto encoder," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2767–2771.

[4] R. Sinha and K. Swearingen, "The role of transparency in recommender systems," in *CHI '02 Extended Abstracts on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 2002, pp. 830–831. [Online]. Available: https://doi.org/10.1145/506443.506619

[5] Z. Chan, Y. Zhang, X. Chen, S. Gao, Z. Zhang, D. Zhao, and R. Yan, "Selection and generation: Learning towards multi-product advertisement post generation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 3818–3829. [Online]. Available: https://aclanthology.org/2020.emnlp-main.313

[6] S. Deng, C.-W. Tan, W. Wang, and Y. Pan, "Smart generation system of personalized advertising copy and its application to advertising practice and research," *Journal of Advertising*, vol. 48, no. 4, pp. 356–365, 2019.

[7] D. Gunning and D. Aha, "DARPA's explainable artificial intelligence (XAI) program," *AI magazine*, vol. 40, no. 2, pp. 44–58, 2019.

[8] A. Vultureanu-Albişi and C. Bădică, "Recommender systems: An explainable AI perspective," in *2021 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, 2021, pp. 1–6.

[9] Y. Zhang, G. Lai, M. Zhang, Y. Zhang, Y. Liu, and S. Ma, "Explicit factor models for explainable recommendation based on phrase-level sentiment analysis," in *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, ser. SIGIR '14. New York, NY, USA: Association for Computing Machinery, 2014, p. 83–92. [Online]. Available: https://doi.org/10.1145/2600428.2609579

[10] Rakuten Group, Inc., "Rakuten dataset," 2020, https://doi.org/10.32130/idr.2.1, https://rit.rakuten.com/data_release/.

[1] Rakuten Dataset (in Japanese): https://rit.rakuten.com/data_release/