# Unlocking the 'Why' of Buying: Introducing a New Dataset and Benchmark for Purchase Reason and Post-Purchase Experience

**Anonymous ACL submission**

## Abstract

Explanations are crucial for enhancing user trust and understanding within modern recommendation systems. To build truly explainable systems, we need high-quality datasets that elucidate why users make choices. While previous efforts have focused on extracting users' post-purchase sentiment in reviews, they ignore the reasons behind the decision to buy.

In our work, we propose a novel *purchase reason explanation* task. To this end, we introduce an LLM-based approach to generate a dataset that consists of textual explanations of why real users make certain purchase decisions. We induce LLMs to explicitly distinguish between the reasons behind purchasing a product and the experience after the purchase in a user review. An automated, LLM-driven evaluation, as well as a small scale human evaluation, confirms the effectiveness of our approach to obtaining high-quality, personalized explanations. We benchmark this dataset on two personalized explanation generation tasks. We release the code and prompts to spur further research[1].

## 1 Introduction

Providing user-understandable explanations to justify recommendations could enhance the effectiveness, persuasiveness, and user satisfaction of recommendation (Zhang and Chen, 2020). This has been confirmed by both user studies (Zanker, 2012) and many real products such as Microsoft Office 365 (Xu et al., 2020), JD.com (an e-commerce website) (Zhang et al., 2014a) and Spotify (McInerney et al., 2018). Explanations can be presented in various styles, such as a piece of text, a relevant user or item, a radar chart, an image or a set of reasoning rule (Tintarev and Masthoff, 2015). With the advancement of natural language generation techniques, a short narrative (e.g., a sentence) becomes one of the most popular explanation style, which is the focus of our work.

Table 1: An example compares existing tasks with our proposed purchase reason task and the post-purchase experience task.

| |
|---|
| <u>Product</u>:<br>Google Pixel 8 - Unlocked Android Smartphone<br><u>Review</u>:<br>I bought this phone as a birthday gift for my teenage daughter who is a fan of AI features. My daughter loves the AI photo editor, with which she successfully removed a stranger from our recent family reunion photo. I highly recommend this phone. |
| (***proposed new task***) Purchase reason: Birthday gift for a teenage daughter who likes AI features.<br>(***proposed improved task***) Post-purchase experience: The daughter loves the AI photo editor and found it a useful tool. Highly recommend.<br>(*existing task*) Common snippet based experience: highly recommend.<br>(*existing task*) Feature based experience: AI photo editor. |

To pave the way of building a more effective recommendation system, we need datasets that explain the reasoning behind user choices. Existing works leverage user reviews to mine explanations. One work (Li et al., 2021a) extracts the most commonly occurred (near) duplicate sentences across reviews, resulting in short and generic comments about the items. For instance, "*Excellent movie*" is the top 1 extracted explanation in Amazon Movies & TV review dataset. Others extract review sentences (Geng et al., 2022) or segments (Ni et al., 2019) that mention one or more pre-generated item features/aspects. An example explanation is "*The quality of the material is great*" where "quality" and "material" are two features of the item (Ni et al., 2019).

Since reviews are written after a purchase, user's sentiments towards the item are primarily based on post-purchase user experience. Existing explanation datasets, therefore, focus on how the item was commented on, rather than the reasons behind the initial decision to buy. In other words, these

[1]The URL will be provided after the anonymity period.

explanations are good for understanding whether a customer is satisfied by an item after buying it, rather than why they have purchased it in the first place. We argue that understanding the purchase reasons is crucial for personalized recommender systems, especially because recommendations are made *before* the user purchases an item. They reveal the user's personal information needs and motives, often beyond the particular reviewed item, which can help us develop more persuasive explanations and build more comprehensive user profiles (see Table 1 for an example).

In this work, we introduce a novel *purchase reason* explanation task. Our approach leverages a language model (LLM) to generate a dataset from user reviews, capturing purchase reasons and in the meantime, generating a highly relatable, personalized post-purchase experience. We propose four dimensions to measure the data quality and validate the effectiveness of our approach through automated LLM evaluator and small-scale human feedback. The resulting dataset is a high-quality, personalized set of explanations. With this dataset, we can develop models to generate explanations from two aspects, the relevance to the user need (i.e., purchase reason) and the preference for the particular item (i.e., post-purchase experience). We benchmark the dataset against these two generation tasks with an LLM and explore different user and item representations.

To summarize, the main contributions of this work are as follows:

- To the best of our knowledge, we are the first to propose the task of purchase reason explanation.
- We propose a simple yet effective LLM-powered approach to generate a high-quality, personalized explanation dataset consisting of both purchase reasons and post-purchase experience.
- We benchmark the tasks of purchase reason and post-purchase experience generation in the context of recommender systems.
- To spur further research, we release the code and all the prompts for generating the dataset and benchmarking.

## 2 Related Work

In this section, we first review prior work on dataset construction for personalized explanations and explanation generation methods in the context of explainable recommender systems. We then discuss the use of large language models (LLMs) for text generation and evaluation.

### 2.1 Personalized Explanation Dataset

Constructing a high quality personalized explanation dataset is the key to build an explainable model for personalized recommendation. Prior work extracts various information from user reviews, such as using the entire reviews (Chen et al., 2018; Li et al., 2017), sentences (Chen et al., 2019; Wang et al., 2018b), aspect-specific sentences (Geng et al., 2022), or elementary discourse units (Ni et al., 2019). They discard segments that are too personal (containing first-person or third-person pronouns) or too short/long. Li et al. (2021a) extract commonly occurred near-duplicate sentences as explanations, which are often short and generic (e.g., "*Excellent movie*"). The above approaches capture customer satisfaction rather than purchase motives because they do not distinguish pre- from post-purchase experiences.

Li et al. (2017) used Tips[2], a concise form of reviews, as justifications for choosing certain businesses. But Tips have low availability (e.g., Yelp mobile). To our knowledge, we are the first to extract purchase reasons directly from regular reviews for creating personalized explanation datasets.

### 2.2 Personalized Explanation Generation

Explainable recommendation (Zhang and Chen, 2020) has drawn considerable research attention. We limit our discussion to personalized, textual explanation generation, which aims to justify why the recommended item might match a user's interest. Ranking based generation (Li et al., 2021a, 2023a) selects explanations from a pre-generated candidate pool, where explanations could be represented as IDs. Template-based explanation (Zhang et al., 2014b; Wang et al., 2018a) selects candidate words to fill into a template sentence. Both approaches tend to generate generic explanations, lacking language flexibility and personalization.

Natural language generation (NLG) is widely used to generate free-text explanations. Early studies fine-tuned seq2seq models like LSTM (Costa et al., 2018), GRU (Li et al., 2017; Ni et al., 2019), Transformer (Li et al., 2021b) and T5 (Liu et al., 2023) or GPT-2 (Li et al., 2023b). Limited by the

[2] https://www.yelp-support.com/article/What-are-tips?l=en_US

model capability, however, the generated explanations are often generic and not fluent (Liu et al., 2023). Recent work with powerful LLMs (e.g., ChatGPT (Liu et al., 2023)) in zero or few-shot setups significantly improved quality.

A common challenge for NLG based methods is representing users and items. Most studies use textual descriptions of users (e.g., reviews) and items (e.g., title). User and item IDs are also considered (Liu et al., 2023; Li et al., 2023b), which are helpful for capturing interactions while limiting generalization to new IDs.

A shared challenge of all these approaches is how to systematically evaluate the generated personalized explanation, which we address by introducing a new dataset and benchmark.

### 2.3 Text Generation and Evaluation With LLMs

LLMs have shown impressive abilities in generation tasks (Zhao et al., 2023), matching commercial translation tools (Jiao et al., 2023) and human writers (Zhang et al., 2024). However, they can produce hallucinations (Huang et al., 2023). Readers can refer to a few recent survey papers for comprehensive discussions (Zhang et al., 2023; Huang et al., 2023).

LLMs have been widely used to evaluate the quality of generated text (Zhao et al., 2023) across various domains such as summarization (Wang et al., 2023a), translation (Kocmi and Federmann, 2023) and personalized text generation Wang et al. (2023b), often achieving state-of-the-art or competitive correlation with human judgments.

Our work is related to this general line of research as we use LLMs to generate purchase reasons from user reviews and to evaluate the quality of the extracted explanation.

## 3 Generating the Explanation Dataset

As we discussed in Section 2.1, prior work that extracts potential explanations from user reviews has several key limitations. Firstly, most of existing work heavily relies on the user's sentiment on the item rather than their initial motives. As the reviews are generally written after purchasing the item, it is more common that these sentiments stem from user experience that is post-purchase, rather than the original reasons that motivate the users to purchase. Consequentially, the extracted information may well explain the user's rating of the item, but it might not be a good explanation of their purchase decision. Secondly, the prior generated explanations tend to be generic and non-personal. For instance, Li et al. (2021a) extracted the most frequently occurring near-duplicate sentences as explanations, and Ni et al. (2019) discarded text segments that are personal (containing first-person or third-person pronouns). This is not hard to understand: while the assessments of an item are often shared across a majority of users, the reasons for making a purchase can be much more personal.

To address these limitations, we explicitly distinguish the purchase reasons from post-purchase experience. These two concepts have strong roots in the business and marketing literature, where purchase reason is related to purchase intent (or intention) (Chang and Wildt, 1994) and purchase motive (or motivation) (Sirgy, 1985), while post-purchase experience is related to customer (or product) satisfaction (Oliver and Swan, 1989; Richins and Bloch, 1991). In particular, purchase intent refers to the likelihood that a user will buy a product, which is influenced by a number of factors such as personal needs, customer preferences, brand equity, trust, perceived value of the product, etc. (Cobb-Walgren et al., 1995; Bleize and Antheunis, 2019) In our work, we do not further distinguish these factors and in general refer to them as *purchase reasons*. On the other hand, *post-purchase experience* encodes a variety of factors that influences post-purchase customer satisfaction, which is highly related to the user's attitude towards the product or business and may further influences their future purchasing behavior (Anderson et al., 1979). In our work, the practical difference between purchase reason and post-purchase experience lies in whether the conveyed information is before the user purchases the item or after. The former explains the user's decision to purchase the item, and the latter explains the decision of rating the item.

As LLMs (eg, ChatGPT[3], GPT-4 (OpenAI, 2023), Gemini (Gemini Team, 2023)) have demonstrated strong performance in text generation, we propose to utilize LLMs to generate both types of explanations based on user reviews, rather than adopting extractive methods by the prior works. We carefully devise strategies to improve generation quality and combat hallucinations. Moreover, we use LLMs as an auto-rater to assess the generated explanations from multiple aspects and further

[3] https://openai.com/chatgpt

improve data quality. To be specific, we demonstrate our solution with Gemini Ultra and Amazon product review 5-core dataset [4](Ni et al., 2019). We detail our end-to-end solution in the following.

### 3.1 Generating Explanations using LLM

We formally define the goal: given a product and an associated user review, asking an LLM (Gemini Ultra in our case) to generate the user's purchase reason and post-purchase experience for this product. We start with generating the two explanations as two separate tasks and identify a few issues with the initial exploration. Firstly, we find an LLM does not have a clear distinction between purchase reasons and post-purchase experience. The two generated explanations have overlaps or appear in the wrong explanation group. Secondly, we notice that users do not always explicitly express their purchase reasons in the review. In those cases, the model either infers the potential purchase reasons from the product information (e.g., some highlighted features by sellers) or generates them by hallucination. We devise the following strategies to address the issues and list the final prompt in Appendix (Table 12).

- We ask the model to perform the task of purchase reason and post-purchase experience generation simultaneously in one prompt. This forces the model to draw a clear boundary between the two types of explanations.

- To alleviate hallucination, we break down purchase reasons into *explicit* (reasons the user explicitly mentioned in the review) and *implicit* (reasons that can be inferred from the review and the product information). We allow the model to leave any of them as empty when no enough information is provided.

- We require the model to provide supporting evidence for its generated purchase reasons. This helps combat hallucination and promotes more accurate and concise explanations.

We also experiment with including few-shot examples in in-context learning for the explanation generation task. We do not find significant performance gains by including few-shot examples.

[4] https://nijianmo.github.io/amazon/index.html

### 3.2 Rating Explanations using LLM

Prior work has demonstrated a great success of automatic evaluation of text generation with LLMs (detailed in Section 2.3). Inspired by this, we explore the usage of LLMs in assessing the quality of our explanation dataset. In particular, we propose to measure the quality of purchase reasons and post-purchase experience from the following four dimensions. We conduct two separate evaluations for the two types of explanations (see Table 13 and Table 14 in Appendix for the specific prompts). We further leverage the four evaluation results to filter out noises, aiming at improving the dataset quality.

- Hallucination: if the explanation contains any completely irrelevant information that are not described or implied in the product information or the user review.

- Correctness: if the model correctly identifies the explanation type. We observe that the model sometimes confuse post-purchase experience as purchase reason, but rarely the opposite.

- Completeness: if the explanation covers all relevant aspects present in the product and the review.

- Personalization: if the explanation contains adequate personal context and is not too generic. It is common that some reviews are generic. This is not a quality metric but a tool to characterize the personalization level.

## 4 Dataset Evaluation and Analysis

As mentioned, we demonstrate our LLM based solution with Gemini Ultra and Amazon product review 5-core dataset (Ni et al., 2019). The full 5-core dataset consists of reviews from all users and items that have at least 5 reviews, resulting in 75 million reviews in total. Considering the cost of using LLMs, we randomly sample 10K reviews for experiments. We apply the LLM generator to construct the explanation dataset and utilize the LLM auto-rater to judge the data quality. In the following, we first validate the effectiveness of our LLM auto-rater and then describe the characteristics of the generated explanation dataset.

### 4.1 Effectiveness of LLM Auto-rater

To evaluate the effectiveness of the LLM auto-rater, we conduct a small scale human annotation with

Table 2: The percentage of examples that LLM auto-rater and human annotator agree. " Hall."/"Compl." denote hallucination and completeness, respectively.

| | Hall. | Correct | Compl. | Personal |
|---|---|---|---|---|
| Reason | 96% | 92% | 84% | 73% |
| Experience | 98% | N.A. | 80% | 80% |

four annotators by following the guidelines as the auto-rater uses. As a pilot study, we ask human annotators to label explanations for 20 reviews. We find there is a high agreement among annotators: they have perfect consensus on hallucination and personalization dimension, and pick up a common answer for 90% of explanations in terms of correctness and completeness. We further ask annotators to discuss the conflicting answers to reach a consistent interpretation of the guidelines. In the later formal annotation, we work on a larger set of 100 randomly sampled reviews, and each explanation is annotated by one person.

As detailed in Table 2, we find LLM auto-rater achieves a strong correlation with human judgement in hallucination (agree on 96% of reviews for purchase reason and 98% for post-purchase experience) and correctness (92% of reviews). However, they disagree more on completeness measurement, only reaching an agreement on 84% and 80% of reviews for purchase reason and post-purchase experience, respectively. The main reason is that the LLM auto-rater is good at capturing subtle information presented in the review and the product, and thus is more strict than human. For instance, human label purchase reasons for 91% of reviews as complete while auto-rater is 77% only. The personalization annotation, aiming at characterizing the explanations, is more challenging due to its subjective nature. The agreement between the LLM auto-rater and human rater is 73% and 80% for purchase reason and post-purchase experience, respectively. The LLM auto-rater labels more explanations as personal than human, indicating human are more conservative than an LLM in judging personalization.

## 4.2 Quality of Generated Explanations

We show the evaluation results of the generated explanations by the LLM auto-rater in Table 3. To understand how the quality varies across products, we show the break down results for top five product categories. Overall, the LLM generator performs very well in combating hallucination with a very low hallucination rate of 0.49% for purchase reason and 0.21% for post-purchase experience. Moreover, the LLM generator has a clear distinction about the two explanations, only confusing post-purchase experience as purchase reason in 0.61% of cases.

As expected, the explanation complete rate is low (77.23% for purchase reason and 77.23% for post-purchase experience). As discussed in Section 4.1, the LLM auto-rater is very strict in completeness evaluation, aiming at capturing all the subtle explanations. This suggests that the actual completeness rate could be higher than the LLM auto-rater detected.

Within the five product categories, we find that "Electronics" has the highest purchase reason complete rate (81.85%), while "Clothing, Shoes Jewelry" is the lowest (71.49%). We hypothesize the discrepancy might due to the different purchase and review behaviors for the two categories. Unlike electronics buyers, who often explicitly explain their purchase motivations in reviews, those buying clothes, shoes, or jewelry are less likely to do so. This means an LLM must infer these motivations from more subtle cues within the reviews and is more prone to neglect some reasons. For post-purchase experience, we also observe that "Electronics" has the highest complete rate (72.51%) than other product categories. The possible reason might be that "Electronics" reviews tend to focus on the product features, and use more direct languages, which is easy for an LLM to capture all post-purchase experience.

## 4.3 Characteristics of the Dataset

Our LLM generator successfully identifies purchase reason in 96.4% of reviews. This includes 61.8% of reviews with explicitly stated reasons and 70.9% where reasons are inferred from the review and the product information. Our LLM identifies post-purchase experience in 88.2% of reviews, slightly lower than purchase reason. This is because we ask the LLM only to identify explicitly described experience, without making any inference. Table 4 shows explanations for two reviews in our dataset.

Overall 70.1% of purchase reasons and 71.18% of post-purchase experience are rated as personal. Among these, reviews for "Books" have the highest proportion of personalized purchase reason (78.89%) but the lowest proportion of personalized post-purchase experience. This is likely because buyers of books often start with a personal connec-

Table 3: The auto-rater results on purchase reason.

| | All | Books | Clothing, Shoes & Jewelry | Home & Kitchen | Electronics | Sports & Outdoors |
|---|---|---|---|---|---|---|
| Purchase reason | | | | | | |
| Hallucination Rate | 0.49% | 0.38% | 0.62% | 0.37% | 0.44% | 0.31% |
| Correct Rate | 99.39% | 99.46% | 99.52% | 99.08% | 99.56% | 99.07% |
| Complete Rate | 77.23% | 73.30% | 71.49% | 77.29% | 81.85% | 80.00% |
| Personal Rate | 70.01% | 78.89% | 68.32% | 67.12% | 61.50% | 66.05% |
| Post-purchase experience | | | | | | |
| Hallucination Rate | 0.21% | 0.22% | 0.07% | 0.19% | 0.11% | 0.16% |
| Complete Rate | 77.23% | 72.66% | 78.21% | 80.47% | 82.04% | 79.78% |
| Personal Rate | 71.18% | 63.97% | 73.77% | 73.99% | 75.98% | 73.46% |

tion – why they chose the book, and then the focus shifts to an objective description of the book itself after reading, making the post-purchase experience less personalized.

We further characterize our dataset from linguistic aspects (detailed in Table 5). On average, purchase reasons are more concise (11.39 words and 10.16 words for explicit and implicit reason) than post-purchase experience (22.19 words). Both of them are much shorter than the original reviews (60.6 words). Since explanations are short, we use a very simple type to token ratio method (i.e., unique word count/total word count) to measure its lexical diversity (Templin, 1957). We find that our generated explanations have a very high lexical diversity (close to 1) and rarely use duplicate tokens, while product information is the most likely to use duplicate tokens to advocate product features. Finally, we assess the dataset's readability using the Flesch-Kincaid Grade Level metric (Flesch, 1948), a standard measurement in NLP research. Higher scores indicate greater readability. Our results align with expectations: product information and implicit purchase reasons (inferred from product information), are the most difficult to read. Conversely, user reviews and explicit purchase reasons are generally easier to understand.

Table 4: Example explanations in our dataset.

**Example 1**
Product: SVINZ Digital Calendar Alarm Day Clock with 3 Alarm Options, Extra Large Non-Abbreviated Day; Month SDC008-2 Color Display Settings
Review: I bought this for my aging parents who love it! The numbers are easy to read and also serves as a dim night light. Setting the alarm is not intuitive which may be difficult for those who are older/less tech savvy but there are many alarm features which make it very useful (i.e. Different options for everyday, weekday only, weekend alarm,s etc).
Explicit purchase reason: To help aging parents who need a clock with easy-to-read numbers and a dim night light.
Implicit purchase reason: Multiple alarm options (everyday, weekday only, weekend only)
Post-purchase experience: The customer's parents love the clock, indicating that it met their expectations for readability and functionality.

**Example 2**
Product: ViewHD 2 Port 1x2 Powered HDMI Mini Splitter for 1080P 3D | Model: VHD-1X2MN3D
Review: I bought this because my fire tv was causing directv to complain about hdcp compliance. I haven't had any problems since I've installed it.
Explicit purchase reason: To resolve HDCP compliance issues between the customer's Fire TV and DirecTV.
Implicit purchase reason: None.
Post-purchase experience: The splitter successfully resolved the HDCP compliance issues, as the customer reported no further problems after installing it.

Table 5: The linguistic characteristics of our dataset. Higher Fleshch score indicates greater readability. Large lexical diversity score indicates higher diversity.

| | Word count | Lex. diversity | Fleshch |
|---|---|---|---|
| Product | 184.84 | 0.68 | 49.71 |
| User review | 60.6 | 0.83 | 79.21 |
| Explicit reason | 11.39 | 0.97 | 64.63 |
| Implict reason | 10.16 | 0.96 | 48.75 |
| Experience | 22.19 | 0.93 | 51.16 |

## 5 A Benchmark of Recommendation Explaination

We benchmark our newly developed dataset in the novel explanation generation tasks – purchase reason and post-purchase experience, in the context of recommender systems.

### 5.1 Task Definition

Given certain user information and the item information, a model is tasked to explain 1) why the user would purchase the item (i.e., purchase reason) and 2) what would the user's experience be after using the item (i.e., post-purchase experience). Furthermore, we investigate the potential benefits of including auxiliary information, such as the user's actual rating for the given item, to the explainability of the tasked model. We are interested in whether models can perform better with the predicted rating as a hint. Additionally, we explore the impact of adding an auxiliary task – rating prediction.

To summarize, we consider the following tasks.

- **Task 1:** Given information about a user and an item, generate an explanation for recommending this item to the user.
- **Task 2:** Given information about a user and an item and the user's ground-truth rating for the item, generate an explanation for why the item was recommended to the user.
- **Task 3:** Given information about a user and an item, generate an explanation for recommending the item as well as a prediction of the user's rating on the item.

### 5.2 Experiment setup

We again choose large language models (LLMs) as the tasked models for benchmarking. It has been shown that explanations generated by LLMs are significantly favored by human raters over small models (Liu et al., 2023). Specifically, we employ Gemini Ultra (Gemini Team, 2023) as the LLM for various tasks in a zero-shot setup (see Table 16 in Appendix for the basic prompt). We evaluate model generated explanations against the "ground-truth" explanations provided in our dataset. We use a variety of metrics to measure the generation performance: BLEU (Papineni et al., 2002), Rouge1, Rouge2, and RougeLsum (Lin, 2004).

### 5.3 User and Item Representation

We are interested in learning the impact of different representation methods for the user and the item. To this end, we enhance our explanation dataset with past reviews for each user and each item. We base our analysis on our main task, task 1, defined in Section 5.1.

For user representation, we consider two options: 1) **UserReview**, composed of the user's reviews written before purchasing the given item, and 2) **UserProfile**, where an LLM summarizes the same past reviews as the user profile (see Table 15 in Appendix for the prompt). For product representation, we explore three options: 1) **Item**, including the item's title and description, 2) **ItemReview**, which additionally incorporates past reviews written by other users before the given user's purchase, and 3) **ItemProfile**, where replaces the raw past reviews as an LLM generated summary. For past reviews, we select up to 10 of the most recent past reviews, with a maximum limit of 8k tokens.

As presented in Table 6, we see that representing users by their raw past reviews and items by their metadata is most effective in purchase reason generation. Post-purchase experience generation follows similar trend, but have a slight performance improvement by incorporating raw past reviews (Table 7). For users, summarizing their past reviews has the risk of losing personal specific information, while for items, including all reviews of other users may bring irrelevant information. Further research is needed to extract useful information from such noisy data. In the following experiments, we report the results using UserReview+Item combination.

Table 6: Performance on purchase reason generation using different user and item representations.

| Method | BLEU | R1 | R2 | RLsum |
|---|---|---|---|---|
| UserReview-ItemReview | 5.91 | 22.13 | 7.71 | 20.21 |
| **UserReview-Item** | 6.46 | 22.14 | 8.45 | 20.38 |
| ItemReview | 4.15 | 17.68 | 6.00 | 16.42 |
| UserProfile-ItemProfile | 4.64 | 17.95 | 5.71 | 16.66 |
| UserReview-ItemProfile | 3.88 | 15.44 | 5.05 | 14.48 |
| UserProfile-ItemReview | 5.25 | 8.67 | 2.94 | 7.99 |
| UserProfile | 5.24 | 18.71 | 6.51 | 17.32 |

Table 7: Performance on post-purchase experience generation using different user and item representations.

| Method | BLEU | R1 | R2 | RLsum |
|---|---|---|---|---|
| **UserReview-ItemReview** | 3.99 | 21.59 | 5.96 | 16.53 |
| UserReview-Item | 3.66 | 21.35 | 5.34 | 16.21 |
| ItemReview | 3.39 | 21.06 | 5.00 | 15.77 |
| UserProfile-ItemProfile | 3.70 | 21.25 | 5.18 | 15.94 |
| UserReview-ItemProfile | 3.51 | 20.35 | 4.60 | 15.17 |
| UserProfile-ItemReview | 3.89 | 9.36 | 2.54 | 7.2 |
| UserProfile | 3.85 | 20.78 | 4.92 | 15.94 |

### 5.4 Impact of task formulations

We study the impact of the auxiliary information and an auxiliary task. They respectively correspond to task 2 and task 3 defined in Section 5.1. The results, presented in Table 8, revealed no significant performance differences between the various task formulations. This finding indicates that LLMs exhibit robustness to variations in auxiliary information and tasks when operated in zero-shot settings.

### 5.5 Variation across Product Categories

We further break down the model performance by product category. As shown in Table 9 for purchase reason generation, the model excels in the "Electronic" and "Home and kitchen" categories, while it underperforms in "Clothing, Shoes & Jewelry" and "Books". This aligns with our earlier findings in the auto-rater evaluation results (Section 4.2), where

Table 8: Performance of explanation generation based on different task formulations.

| Task | BLEU | R1 | R2 | RLsum |
|---|---|---|---|---|
| Purchase reason | | | | |
| Task 1 | 6.46 | 22.14 | 8.45 | 20.38 |
| Task 2 | 6.51 | 21.96 | 8.35 | 20.12 |
| Task 3 | 6.44 | 21.77 | 8.34 | 19.97 |
| Post-purchase Experience | | | | |
| Task 1 | 3.66 | 21.35 | 5.34 | 16.21 |
| Task 2 | 3.74 | 21.75 | 5.55 | 16.36 |
| Task 3 | 4.49 | 22.41 | 6.00 | 17.62 |

Table 9: Purchase reason generation performance per product category. Fashion denotes Clothing, Shoes and Jewelry.

| Category | BLEU | R1 | R2 | RLsum |
|---|---|---|---|---|
| Overall | 6.46 | 22.14 | 8.45 | 20.38 |
| Books | 5.36 | 19.44 | 6.68 | 17.50 |
| Electronics | 8.74 | 27.10 | 12.31 | 25.32 |
| Home, Kitchen | 7.21 | 24.17 | 9.35 | 22.54 |
| Sports, Outdoors | 6.15 | 22.04 | 8.02 | 20.27 |
| Fashion | 4.87 | 18.80 | 5.70 | 17.05 |

generated text for “Books” and “Clothing, Shoes & Jewelry” exhibited lower generality, as reflected in their "Non-personal" ratings. Conversely, the “Electronic” and “Home and kitchen” categories show higher generality. Consequently the explanation generation task is easier in these domains. Table 11 (in Appendix) shows the results for post-purchase experience generation with similar trend.

### 5.6 Discussions

Table 10 shows two example purchase reasons generated by our model. We observe that the model performance is contingent on the availability of purchase reason information within product metadata and the level of personalization in user reviews. When product metadata provides comprehensive insights into purchase reasons and user reviews are less individualized, the model exhibits strong predictive capabilities. However, the model’s performance degrades when attempting to generate purchase reasons that are more personalized, which may not be simply available from item metadata. This limitation calls for the development of models capable of constructing more comprehensive personal profiles to better predict such individualized purchase reasons.

## 6 Conclusion

We introduce a novel task of purchase reason explanation, aiming at better capturing what affects a user’s decision to purchase a product. We propose

Table 10: Case studies on purchase reason generation.

| **Example where the model generates a good explanation.** |
|---|
| Product title: DEDC 1999-2007 Right Passengers Side Power Towing Mirrors Fit Ford Super Duty F250 F350 F450 1999 2000 2001 2002 2003 2004 2005 2006 2007<br>Ground-truth purchase reason: Needed a replacement right-side mirror for their Ford Super Duty truck (year 1999-2007)<br>Predicted purchase reason: User Anonymous may need replacement mirrors for their Ford Super Duty truck. |
| **Example where the model generates a poor explanation.** |
| Product title: The Whole Bible Story: Everything That Happens in the Bible in Plain English<br>Ground-truth purchase reason: To read the Bible in a condensed format, allowing the reader to easily navigate to specific chapters.<br>Predicted purchase reason: Interest in religion and biblical studies |

an LLM-based approach to generating a high quality, personalized dataset that consists of textual explanations of purchase reasons and post-purchase experiences based on user reviews. As the first of its kind, we demonstrate the dataset’s value by benchmarking it against purchase reason and post-purchase experience generation tasks. To empower further research, we release all scripts and prompts used for dataset creation and benchmarking.

With this new dataset, it is interesting to benchmark more explainable recommendation or explanation generation models, especially by refining user and item representations. Our results indicate that relying solely on all past reviews can introduce noise. It will be interesting to explore whether limiting the past reviews that are more similar to the item or the user could be beneficial, aligning with the ideas in retrieval augmented generation (Gao et al., 2024). We leave this as future work.

## 7 Limitation

We propose a LLM-based solution for explanation dataset generation and auto evaluation. We validate this approach with Gemini Ultra and Amazon review dataset. Our approach is general, however, the performance with smaller models (e.g., Gemini Nano) and other model families (e.g., GPT-4, Claude, LLaMA) is still an open question. Similarly, we have not explored the generalization of our approach in reviews from other domains such Yelp and TripAdvisor.

## References

Ronald D Anderson, Jack L Engledow, and Helmut Becker. 1979. Evaluating the relationships among attitude toward business, product satisfaction, experience, and search effort. *Journal of Marketing Research*, 16(3):394–400.

Daniëlle NM Bleize and Marjolijn L Antheunis. 2019. Factors influencing purchase intent in virtual worlds: a review of the literature. *Journal of Marketing Communications*, 25(4):403–420.

Tung-Zong Chang and Albert R Wildt. 1994. Price, product information, and purchase intention: An empirical study. *Journal of the Academy of Marketing science*, 22:16–27.

Chong Chen, Min Zhang, Yiqun Liu, and Shaoping Ma. 2018. Neural attentional rating regression with review-level explanations. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, page 1583–1592.

Xu Chen, Yongfeng Zhang, and Zheng Qin. 2019. Dynamic explainable recommendation based on neural attentive models. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Cathy J Cobb-Walgren, Cynthia A Ruble, and Naveen Donthu. 1995. Brand equity, brand preference, and purchase intent. *Journal of advertising*, 24(3):25–40.

Felipe Costa, Sixun Ouyang, Peter Dolog, and Aonghus Lawlor. 2018. Automatic generation of natural language explanations. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion*, IUI '18 Companion, New York, NY, USA. Association for Computing Machinery.

R Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey.

Gemini Team. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2022. Recommendation as language processing (rlp): A unified pretrain, personalized prompt & predict paradigm (p5). In *Proceedings of the 16th ACM Conference on Recommender Systems*, RecSys '22, page 299–315, New York, NY, USA. Association for Computing Machinery.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine. *arXiv preprint arXiv:2301.08745*.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Lei Li, Yongfeng Zhang, and Li Chen. 2021a. Extra: Explanation ranking datasets for explainable recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2463–2469, New York, NY, USA. Association for Computing Machinery.

Lei Li, Yongfeng Zhang, and Li Chen. 2021b. Personalized transformer for explainable recommendation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4947–4957, Online. Association for Computational Linguistics.

Lei Li, Yongfeng Zhang, and Li Chen. 2023a. On the relationship between explanation and recommendation: Learning to rank explanations for improved performance. *ACM Trans. Intell. Syst. Technol.*, 14(2).

Lei Li, Yongfeng Zhang, and Li Chen. 2023b. Personalized prompt learning for explainable recommendation. *ACM Trans. Inf. Syst.*, 41(4).

Piji Li, Zihao Wang, Zhaochun Ren, Lidong Bing, and Wai Lam. 2017. Neural rating regression with abstractive tips generation for recommendation. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, page 345–354, New York, NY, USA. Association for Computing Machinery.

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

Junling Liu, Chao Liu, Peilin Zhou, Renjie Lv, Kang Zhou, and Yan Zhang. 2023. Is chatgpt a good recommender? a preliminary study.

James McInerney, Benjamin Lacker, Samantha Hansen, Karl Higley, Hugues Bouchard, Alois Gruson, and Rishabh Mehrotra. 2018. Explore, exploit, and explain: personalizing explainable recommendations with bandits. In *Proceedings of the 12th ACM Conference on Recommender Systems*, RecSys '18, page 31–39, New York, NY, USA. Association for Computing Machinery.

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.

Richard L Oliver and John E Swan. 1989. Equity and disconfirmation perceptions as influences on merchant and product satisfaction. *Journal of consumer research*, 16(3):372–383.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Marsha L Richins and Peter H Bloch. 1991. Post-purchase product satisfaction: Incorporating the effects of involvement and time. *Journal of Business Research*, 23(2):145–158.

M Joseph Sirgy. 1985. Using self-congruity and ideal congruity to predict purchase motivation. *Journal of business Research*, 13(3):195–206.

Mildred C Templin. 1957. Certain language skills in children; their development and interrelationships. *Child Welfare Monograph*, 26.

Nava Tintarev and Judith Masthoff. 2015. *Explaining Recommendations: Design and Evaluation*, pages 353–382. Springer US, Boston, MA.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023a. Is ChatGPT a good NLG evaluator? a preliminary study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11, Singapore. Association for Computational Linguistics.

Nan Wang, Hongning Wang, Yiling Jia, and Yue Yin. 2018a. Explainable recommendation via multi-task learning in opinionated text data. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, page 165–174, New York, NY, USA. Association for Computing Machinery.

Xiting Wang, Yiru Chen, Jie Yang, Le Wu, Zhengtao Wu, and Xing Xie. 2018b. A reinforcement learning framework for explainable recommendation. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 587–596.

Yaqing Wang, Jiepu Jiang, Mingyang Zhang, Cheng Li, Yi Liang, Qiaozhu Mei, and Michael Bendersky. 2023b. Automated evaluation of personalized text generation using large language models. *arXiv preprint arXiv:2310.11593*.

Xuhai Xu, Ahmed Hassan Awadallah, Susan T. Dumais, Farheen Omar, Bogdan Popp, Robert Rounthwaite, and Farnaz Jahanbakhsh. 2020. Understanding user behavior for document recommendation. In *Proceedings of The Web Conference 2020*, WWW '20, page 3012–3018, New York, NY, USA. Association for Computing Machinery.

Markus Zanker. 2012. The influence of knowledgeable explanations on users' perception of a recommender system. In *Proceedings of the Sixth ACM Conference on Recommender Systems*, RecSys '12, page 269–272, New York, NY, USA. Association for Computing Machinery.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024. Benchmarking Large Language Models for News Summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Yongfeng Zhang and Xu Chen. 2020. Explainable recommendation: A survey and new perspectives. *Found. Trends Inf. Retr.*, 14(1):1–101.

Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. 2014a. 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*, SIGIR '14, page 83–92, New York, NY, USA. Association for Computing Machinery.

Yongfeng Zhang, Haochen Zhang, Min Zhang, Yiqun Liu, and Shaoping Ma. 2014b. Do users rate or review? boost phrase-level sentiment labeling with review-level sentiment classification. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, page 1027–1030, New York, NY, USA. Association for Computing Machinery.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen

Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

# A Appendix

Table 11: Post-purchase experience generation performance per product category. Fashion denotes Clothing, Shoes and Jewelry.

| Category | BLEU | R1 | R2 | RLsum |
|---|---|---|---|---|
| Overall | 3.66 | 21.35 | 5.34 | 16.21 |
| Books | 3.05 | 19.71 | 3.94 | 14.85 |
| Electronics | 4.17 | 23.10 | 6.55 | 17.60 |
| Home, Kitchen | 4.16 | 22.52 | 6.22 | 17.29 |
| Sports, Outdoors | 3.79 | 21.89 | 5.73 | 16.51 |
| Fashion | 3.71 | 21.33 | 5.50 | 16.22 |

Table 12: Prompt used to generate purchase reason and post-purchase experience based on a product and an associated user review.

You are a customer engagement specialist at Amazon, please analyze:
1. explicit_purchase_reason: why this customer purchased a product based on their reviews on Amazon. Describe in detail the thought processes before the purchase. Leave null if not mentioned.
2. implicit_purchase_reason: why this customer purchased this product, not mentioned in the reviews, can be from the product description.
3. purchase_reason_explanation: why do you think this is the purchase reason.
4. post_purchase_experience: how did the product meet this user's expectation, describe it in 2 to 3 lines.

Be as specific and relating to personal context as much as possible.

Analyze purchase reason and experience based on this product review and product information.

Product:
*Actual product information including title and description.*

Customer Review:
*Actual review content.*

Please answer in json format, for example:
```
{
"explicit_purchase_reason": "........"
"implicit_purchase_reason": "........"
"purchase_reason_explanation": "........"
"post_purchase_experience": "........"
}
```

Answer:

Table 13: Prompt used to evaluate purchase reasons based on a product and an associated user review.

You are a customer engagement specialist manager at Amazon and your specialist is trying to write a summary of why a specific purchase happened. Please assess the summary using the following criteria:

personal: Answer "Good", "Average" or "Bad", if the summary contains adequate personal context and is not too generic.
completeness: Answer "Yes" or "No" if most of the purchase reasons are covered in the summary answer "Yes".
completeness_reason: leave null if most of the reasons are covered.
personal: Answer "Good" or "Bad", if the answer contains adequate personal context and is not too generic.
personal_evidence: show evidence that this contains adequate personal context and is not too generic.
hallucination: Answer "Hallucination" or "Factual", are there any completely irrelevant information introduced, that are not implied in the product information or user review
hallucination_reason: if there are irrelevant information not implied in the product information or user review, show evidence
incorrect: Answer "Yes" if the summary contains information from after the purchase, else "No"
incorrect_reason: what information is from after the purchase?

Please answer in json.
Example assessment
{
"completeness": "Yes",
"completeness_reason": "...",
"personal": "Good",
"personal_evidence": "...",
"hallucination": "Factual",
"hallucination_reason": "...",
"incorrect": "Yes",
"incorrect_reason": "..."
}

Now let's take a look at the following specialist's task and summary and provide the assessment:

Product:
*Actual product information.*

Customer Review:
*Actual review content.*

Specialist summary of purchase reason:
*Actual purchase reasons.*

Assessment:

Table 14: Prompt used to evaluate post-purchase experience based on a product and an associated user review.

You are a customer engagement specialist manager at Amazon, your specialist is trying to write a summary of user post purchase sentiment.
Please assess the summary using the following criteria:
personal: Answer "Good" or "Bad", if the answer contains adequate personal context and is not too generic.

personal_evidence: show evidence that this contains adequate personal context and is not too generic.
hallucination: Answer "Hallucination" or "Factual", are there any completely irrelevant information introduced, that are not implied in the product information or user review
hallucination_reason: if there are irrelevant information not implied in the product information or user review, show evidence
completeness: Answer "Yes" if all the post purchase sentiments are covered in the summary and "No" otherwise.
completeness_reason: what are the missing sentiments?

Please answer in json.
Example assessment
```
{
"personal": "Good",
"personal_evidence": "..."
"hallucination": "Hallucination"
"hallucination_reason": "..."
"completeness": "Yes"
"completeness_reason": "..."
}
```

Example assessment
```
{
"personal": "Average",
"personal_evidence": "..."
"hallucination": "Factual"
"hallucination_reason": "..."
"completeness": "No"
"completeness_reason": "..."
}
```

Now let's take a look at the following specialist's task and summary and provide the assessment:
Product:
*Actual product information.*

Customer Review:
*Actual review content.*

Specialist summary of post-purchase use experience:
*Actual post-purchase use experience.*

Assessment:

Table 15: Prompt used to rewrite user history

This is user Anonymous who left some past reviews on Amazon, please take a look at this user's past review history on other products

Past reviews from user Anonymous on other products:
*User's past reviews*

Let's identify a few past purchases from this user and predict user's purchase reason for past products and post purchase sentiments
1. explicit_purchase_reason: why user Anonymous could purchase this past product, as inferred from user's past reviews.
2. implicit_purchase_reason: why user Anonymous could purchase this past product, not mentioned in user user's past reviews.
3. purchase_reason_explanation: why do you think this could be the purchase reason
4. post_purchase_experience: how did this past product meet user Anonymous' expectation, based user user's past reviews, describe it in 2 to 3 lines.

For example:
Past item 1: ....
{
"explicit_purchase_reason": "........"
"implicit_purchase_reason": "........"
"purchase_reason_explanation": "........"
"post_purchase_experience": "........"
}

Past item 2: ....
{
"explicit_purchase_reason": "........"
"implicit_purchase_reason": "........"
"purchase_reason_explanation": "........"
"post_purchase_experience": "........"
}

Past item 3: ....
{
"explicit_purchase_reason": "........"
"implicit_purchase_reason": "........"
"purchase_reason_explanation": "........"
"post_purchase_experience": "........"
}

Table 16: Prompt used to conduct task 1, purchase reason generation

This is a user who left some past reviews on Amazon, please take a look at this user's past review history on other products and predict this user's purchase reason for this product and post purchase sentiment:
1. explicit_purchase_reason: why this user could purchase this product, as inferred from this user's past reviews.
2. implicit_purchase_reason: why this user could purchase this product, not mentioned in this user's past reviews, can be from the product description.
3. purchase_reason_explanation: why do you think could be the purchase reason.
4. post_purchase_experience: how could this product meet this user's expectation based this user's past reviews, describe it in 2 to 3 lines.

Product:
*Actual product information including title and description.*

Past reviews from this user on other products:
*Past review content with past product metadata.*

Now let's predict this user's purchase reason for this product and post purchase sentiment: Please answer in json format, for example:
{
"explicit_purchase_reason": "........"
"implicit_purchase_reason": "........"
"purchase_reason_explanation": "........"
"post_purchase_experience": "........"
}

Answer: