DRE: Generating Recommendation Explanations by Aligning Large Language Models at Data-level

Anonymous ACL submission

Abstract

Recommendation systems play a crucial role in various domains, suggesting items based on user behavior. And the lack of transparency in presenting recommendations can lead to user confusion. Thus, recommendation explanation methods are proposed to generate natural language explanations for users, which usually require intermediary representations of the recommendation model or need to conduct latent alignment training to the recommendation model. However, this additional training step usually causes potential performance issues due to the different training objectives between the recommendation task and the explanation task.

In this paper, we introduce **D**ata-level **R**ecommendation **E**xplanation (DRE), a non-intrusive explanation framework for black-box recommendation models. We propose a data-level alignment method, leveraging large language models to reason relationships between user data and recommended items, without any additional training or intermediary representations for the recommendation model. Additionally, we also address the challenge of enriching the details of the explanation by introducing target-aware user preference distillation, utilizing item reviews. Experimental results on several benchmark datasets demonstrate the effectiveness of the DRE in providing accurate and user-centric explanations, enhancing user engagement with recommended items [1].

1 Introduction

Recommendation systems (RecSys) play a pivotal role in learning user preferences and interests by analyzing historical user behavior data (Cheng et al., 2016; Guo et al., 2017; He et al., 2017; Johnson et al., 2014). Subsequently, the RecSys recommends relevant items from extensive databases, which are widely used in diverse domains such

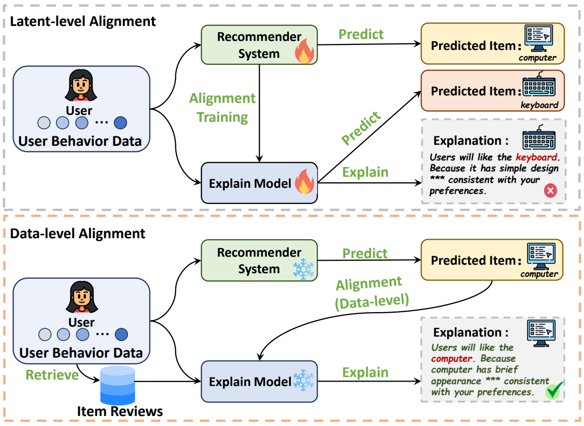


Figure 1: Comparison between existing latent-level alignment and our data-level alignment recommendation explanation method.

as e-commerce, news portals, and short video applications (Zhang et al., 2021; Koren et al., 2009; He and McAuley, 2016; Van den Oord et al., 2013). However, the direct presentation of recommended items may inadvertently confuse users, as they may not always comprehend the rationale behind a particular recommendation (Lei et al., 2023; Cheng et al., 2022, 2021). This lack of transparency impedes users' inclination to explore the recommended item further (Zhang et al., 2020a; Balog et al., 2019; Chen et al., 2020). Consequently, interpreting the recommendation results of a black-box recommender model logically has always been an important research direction (Bilgic and Mooney, 2005; Sharma and Cosley, 2013; Tintarev and Masthoff, 2010). Most of the existing methods (Xu et al., 2023; Wang et al., 2018b, 2024, 2023; Gao et al., 2023) usually focus on how to employ an additional explanation module to align with the recommendation system, subsequently generating natural language explanations.

However, there are two key challenges of these methods: (1) Existing methods (Lei et al., 2023; Xu et al., 2024; Chen et al., 2017, 2018) often involve

[1] Code is available at `https://anonymous.4open.science/r/DRE`

intrusion into the latent representations within the recommendation model, necessitating modifications to align the explanation and recommendation modules. Considering the different training objectives of these two modules, it could adversely affect the performance of both language generation and item recommendation. Moreover, although these methods aim to align two modules through training, they still cannot guarantee that the recommendation predictions of the two modules are consistent. Thus the discrepancies between the explained and recommended items may lead to user confusion. Additionally, in real-world applications, modifying the online serving recommendation model is very difficult. It also increases the overall system complexity, leading to a deep coupling between the recommendation and explanation modules. This does not align with the design principle of "low in coupling and high in cohesion" in software design. (2) The recommendation system based on ItemID models the co-occurrence relationships among items (Zhang et al., 2014, 2020b; Diao et al., 2014; Wang et al., 2018a), lacking an understanding of the specific semantic information about the items, such as the specific purposes of the products or the particular scenarios in which users use them. Thus, simply aligning the explanation module with the recommendation module cannot provide rich detailed semantic information about the item. However, to generate helpful explanations, the explanation module requires comprehensive and diverse information to avoid generating explanations with hallucination information.

In this paper, we propose the **D**ata-level **R**ecommendation **E**xplanation (DRE) which can be applied to any black-box recommendation model without accessing intermediate representations or modifying the model. To avoid modifying the recommendation system, we propose a *data-level alignment method* to align the explanation module and the recommendation model. Figure 1 shows the comparison between our proposed paradigm and existing methods. Since the large language models (LLMs) have shown strong reasoning capability in many tasks (Wei et al., 2022; Mann et al., 2020; Dong et al., 2019; Radford et al., 2018; Zhao et al., 2023; Xi et al., 2023), we propose to employ the LLM to reason the relationships between the user's historical data and recommended items. Specifically, we feed the input user historical behavior data used by the recommendation model and the recommended item to the LLM. And we leverage the internal knowledge of LLM to find a reasonable relationship between the user preference and the attributes of the recommended item. This data-level alignment method can align these two modules without requiring any internal representation or intermediate result of the recommendation model, and it can easily be plugged into any RecSys.

For the second challenge, due to the limited detailed information of item descriptions, relying solely on item descriptions for inferring relationships between items can sometimes be challenging in uncovering implicit relationship information. Therefore, we propose utilizing the reviews of the items purchased by users and the reviews of the target recommended items to enhance the explanation module's understanding of user preferences and the semantics of target items. Since there is a lengthy of reviews for items that users have purchased, extracting relevant information from these reviews and generating explanations that better align with user preferences is a challenge. Thus, we introduce the *target-aware user preference distillation* method, which leverages the understanding and reasoning capabilities of LLM, employing semantic matching to extract target-aware information from reviews on items previously purchased by users. Finally, by incorporating the extracted target-aware information, we generate explanations for the recommended target items. Experiments conducted on several benchmark datasets from recommendation systems demonstrate that our proposed DRE generates explanations accurately describing aspects that users care about, thereby enhancing user interest in recommended items.

Our contributions are as follows:

- We propose DRE, an LLM-based non-intrusive explanation framework for recommendation systems.
- We propose a data-level alignment method to align the explanation module and the recommendation model.
- We introduce a target-aware user preference distillation method to distill user-related information from item reviews.
- Experimental results on several benchmark datasets illustrate the advantage of DRE in terms of the accuracy of explanation.

2 Related Work

Explaining the black box of recommender systems has long been a prominent research direction in the field of recommender systems. Current research can be mainly divided into two categories. The first category focuses on identifying the most critical factors influencing recommendation results(Chen et al., 2016; Pan et al., 2020). Tan et al. (2021) formulate an optimization problem to generate minimal changes to item aspects, thereby altering the recommended result. These aspects can be viewed as the composition of an explanation detailing why the original item is recommended. Zilke et al. (2016); Lakkaraju et al. (2017); Shrikumar et al. (2017) define information-based measures to identify the attributes that the model utilizes from the input to generate explanations. The second category mainly focuses on training a surrogate model to explain the target model. For example, Wang et al. (2018b) propose a reinforcement learning framework that gets rewards from the environment and modifies recommendation explanation. Ma et al. (2019); Catherine et al. (2017) propose a framework for generating explanations based on the knowledge graph. Lei et al. (2023) employ LLMs as surrogate models, aiming to mimic and understand target recommender models by leveraging both natural language and latent spaces. After alignment, LLMs can generate target items and provide recommendation explanations. However, existing methods either rely solely on a few entity words or keywords as explanations or employ complex fine-tuning approaches to generate natural language explanations. It makes the explanations not natural or complex to use, which requires fine-tuning or modification of existing recommendation systems.

3 DRE Methodology

In this section, we detail the **D**ata-level **R**ecommendation **E**xplanation (DRE). An overview of DRE is shown in Figure 2.

3.1 Data-level Alignment

In order to generate precise explanations for recommended results, we propose a data-level alignment method to achieve behavioral consistency between the recommendation module and the explanation module. Given a list of items $I = \{I_1, I_2, \dots, I_N\}$ which is purchased by the user U, the recommendation model R predicts items I_p that the user U might find interesting. To achieve alignment between the recommendation module and the explanation module, previous methods typically fine-tune the explanation module to perform the recommendation prediction task as well, generating items I_p consistent with the predictions of the recommendation model R. However, this approach inevitably reduces the text generation capability of the explanation module due to changes in its model structure and parameters. In this paper, we propose leveraging the in-context learning and reasoning abilities of LLM to align the explanation module with the recommendation module. Given inputs I and outputs I_p that are consistent with the recommendation model R, LLM can learn this prediction pattern in the context and explore the associated relationships to generate natural language explanations.

3.2 Target-aware User Preference Distillation

Relying solely on item IDs and item descriptions for recommendation explanations may fail to capture the details or user actual experiences of the item, which are crucial for users. Therefore, we propose to incorporate the reviews of user-purchased items I and the target item I_p predicted by the recommendation model R to assist the explanation model in obtaining more item detail information. Given a purchased item I_i of user U, we retrieve M reviews $C^i = \{C_1^i, C_2^i, \dots, C_M^i\}$ of item I_i written by *other* users from the database, where each C_1^i represents a paragraph of natural language product review. Then, we can retrieve M user reviews for each purchased item I_i of user U, and then obtain a review set $C = \{C^1, C^2, \dots, C^N\}$ which contains $M \times N$ reviews of other users. Similarly, we can also retrieve M reviews for the target item I_p denoted as $C^p = \{C_1^p, C_2^p, \dots, C_M^p\}$ which is also written by other users. In this paper, we assume that the item characteristics described in the review set C are the key features that user U cares about, since the user U has bought these items. Therefore, we need to perform semantic matching between C and C^p to extract those item features that are both of interest to the user in the past purchased items and possessed by the target product I_p. We propose the *target-aware user preference distillation* method, which involves matching the target item reviews C^p with C to extract valuable information for generating recommendation explanations.

Since the description and reviews of items are usually quite long, and not all the information is helpful for generating recommendation explanations. For the target item I_p, we first construct an

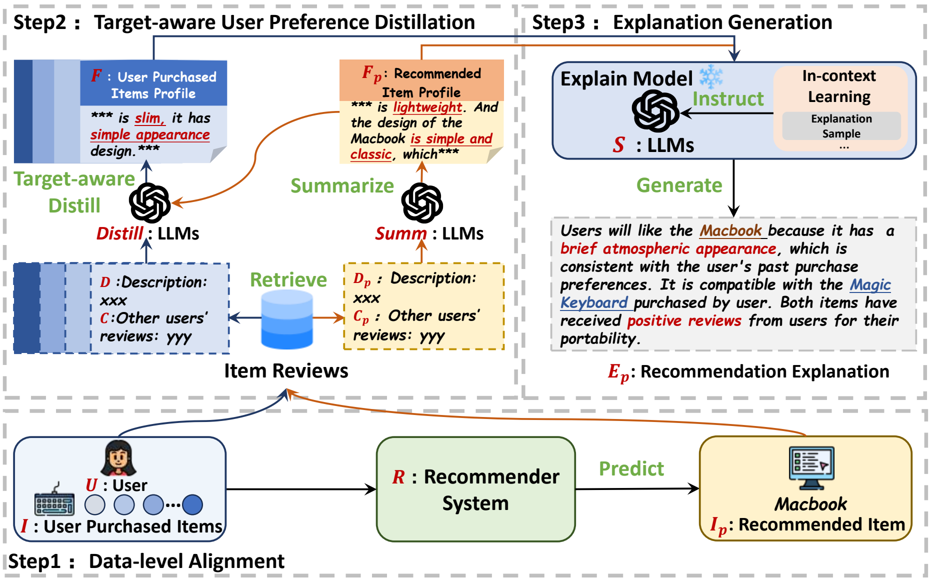


Figure 2: Overview of DRE, which firstly align the explanation module and recommender with **Data-level Alignment**, and then generate the explanation by incorporating details of target from **Target-aware User Preference Distillation**.

overview item profile F_p to distill the useful item features. We use the product description D_p and reviews information $C^p = \{C_1^p, C_2^p, \ldots, C_M^p\}$ of I_p as input and prompt the LLM to generate an item profile F_p:

$$F_p = \text{Summ}\left(\{C_1^p, C_2^p, \ldots, C_M^p\}, D_p\right), \qquad (1)$$

where F_p contains both the basic information of the target item and user usage experiences and Summ is an LLM-based module that is prompted by the following instructions:

> You are given item's description and reviews. Response item profile using the following format:
> item: {item name}
> description: {item description}
> other users' reviews: {item reviews}
> Extract key features from reviews.

However, not all the product features mentioned in F_p may be of concern to the user U. Therefore, we need to extract product features that user U care about from $C = \{C^1, C^2, \ldots, C^N\}$ associated with user behavior. Specifically, we use the item profile F_p of the target item to filter reviews in set C^i of item I_i:

$$F_i = \text{Distill}\left(F_p, \{C_1^i, C_2^i, \ldots, C_M^i\}, D_i\right), \qquad (2)$$

where D_i is the item description of item I_i, and Distill is an LLM-based module that is prompted by the following instructions:

> Finish history item profile using relevant features with recommended item, strictly adhere to the following format when responding:
> history item: {item name}
> genre: {item genre}
> relevant information: {item information}
> other users' reviews: {reviews}
> which relevant information mainly describes similarities between history item and recommended item, and summarize other users' reviews;

By integrating these two parts of information, we obtain the target-aware item profiles $F = \{F_1, F_2, \ldots, F_N\}$ for the items the user U has purchased.

3.3 Explanation Generation

Finally, we integrate the item profile F_p of the target item with the item profiles $F = \{F_1, F_2, \ldots, F_N\}$ of the items the user has purchased. We employ an in-context learning approach and instruct the LLM as follows to generate a logically coherent recommendation explanation that aligns with the recommendation system R and corresponds to user attention preferences:

$$E_p = S\left(F_p, \{F_1, F_2, \ldots, F_N\}\right), \qquad (3)$$

where S is an LLM-based module to generate the recommendation explanation which is instructed by the following instructions:

Now you are a recommendation assistant, combined with history relevant items, write an explanation of the recommended item. The format of response is as below:
item: {recommended item}
recommend reason: {reason}

4 Experimental Setup

4.1 Implementation Details

In our experiments, all DRE-C variants and the ChatGPT baseline use the gpt-3.5-turbo version, and the DRE-M variant and `Mistral` baseline use the Mistral $8 \times 7B$ version which is open-sourced. And we update the memory modules of agents in DRE after each turn, meaning that only the suggestions and experiences from the previous turn are retained.

4.2 Evaluation Metrics

To quantitatively measure the performance of DRE, we propose two evaluation metrics in our paper: (1)**Aspect Score**: We assume that the aspects mentioned in the review C_U^p of the target item I_p written by user U are crucial to the user. We use the review C_U^p as a reference of the explanation E_p. We first employ the LLM to extract aspects of the review C_U^p. Subsequently, we measure the alignment between recommendation explanations E_p and user preferences by calculating the extent of the aspect overlap between E_p and C_U^p:

$$\text{Aspect\_Score} = \frac{1}{N_a}\sum_{i=1}^{N_a} hit(i) \in [0,1], \tag{4}$$

where N_a is the number of aspects in the user review C_U^p. To capture the user's detailed intent, we set N_a=7. And when the aspect i in the explanation is semantically the same as the aspect in the recommendation explanations E_p then $hit(i) = 1$, otherwise, $hit(i) = 0$. (2)**Rating Score**: Following (Lei et al., 2023), to directly evaluate the quality of the generated explanation, we implement a three-level scoring criteria to quantitatively evaluate the explanation generated by models: (i) RATING-1: Poor Explanation, using chunks of original sentence from provided data. (ii) RATING-2: Acceptable Explanation, consider only one aspect of user history and reviews, explaining unrelated items together. (iii) RATING-3: Satisfactory Explanation. We employ the LLM to evaluate the generated explanation according to these criteria and calculate the average rating score over all the testset.

4.3 Dataset

In this paper, we employ two commonly used recommendation datasets in the experiments: Amazon (Ni et al., 2019) and Yelp [2]. In the Amazon dataset, we employ several categories, including Cell Phones & Accessories, Clothing Shoes & Jewelry, and Home & Kitchen. Intuitively, in order to better capture user preferences, we model user preferences only using positive user reviews. Cell Phones & Accessories contains 12,467 users, 6,977 items and 38,729 reviews. Home & Kitchen contains 16,102 users, 1,590 items, and 20,277 reviews. Clothing Shoes & Jewelry contains 19,310 users, 3,746 items and 24,712 reviews. To construct the user purchase history, we limit the items sequence to a minimum of 4 items on Clothing Shoes & Jewelry, Home & Kitchen, and a minimum of 3 items on Cell Phones & Accessories. The last item is then used as the prediction target item. We select 100 samples in each category as testset and each item has associated reviews. We filtered the data by removing the sample of items with fewer than 2 user-purchased items and no accompanying reviews from users.

In the Yelp dataset, we utilize attributes and categories associated with item as descriptions. The Yelp dataset consists of 12,377 users, 4,446 items, and 14,453 reviews. We also select 100 samples from the Yelp dataset as the test set and filter the data with a length of historical data of less than 3 or at least 1 review.

4.4 Comparison Methods

We compare DRE to a state-of-the-art LLM-based recommendation explanation method and several LLMs, including: (i) **`RecExplainer`** (Lei et al., 2023) introduces an explanation approach by leveraging LLM, which employs three methods - behavior alignment, intention alignment, and hybrid alignment - in the latent spaces. (ii) **`ChatGPT`** [3] is a closed-source LLM from OpenAI. We use the version gpt-3.5-turbo-0613. We conduct recommendation explanation as a prompt learning method that uses a single instruction with the same input data as our DRE. (iii) **`Mistral`** (Mix) is an open-source LLM and we use the mixture-of-experts version with 8×70 billion parameters, and use the same prompt as `ChatGPT`.

We also employ two variants of DRE: **DRE-C**

[2] `https://www.yelp.com/dataset`
[3] `https://chat.openai.com/`

Table 1: Recommendation explanation performance comparison. ‡ indicates significant improvement over `ChatGPT` with $p \leq 0.01$ according to a Student's t test.

Method	Home & Kitchen		Clothing Shoes & Jewelry		Cell Phones & Accessories		Yelp	
	Aspect (↑)	Rating (↑)	Aspect (↑)	Rating (↑)	Aspect (↑)	Rating (↑)	Aspect (↑)	Rating (↑)
`RecExplainer` (Lei et al., 2023)	0.6057	2.64	0.5628	2.68	0.6028	2.64	0.3238	2.86
`Mistral` (Mix)	0.7028	2.65	0.5757	2.79	0.6571	2.00	0.4642	2.65
`ChatGPT` [3]	0.6971	2.51	0.6362	2.86	0.6229	2.67	0.4200	2.79
DRE-M	0.7142	2.68	0.6485	2.89	0.6857	2.57	0.5542	2.82
DRE-C	**0.7714**‡	**2.88**†	**0.6728**‡	**2.94**‡	**0.7400**‡	**2.90**‡	**0.5600**‡	**2.91**‡
`DRE-C w/o Rev.`	0.6914	2.64	0.6400	2.65	0.6542	2.66	0.4242	2.83
`DRE-C w/o Dist.`	0.6278	2.79	0.5714	2.77	0.6057	2.89	0.5542	2.86
`DRE-C w/o Dist.`+F_p	0.5828	2.77	0.5671	2.82	0.5971	2.83	0.5028	2.83
`DRE-C w/` F_p	0.7385	1.64	0.5814	2.06	0.6585	2.03	0.4285	1.50

Table 2: Human evaluation results for two datasets.

	Clothing Shoes & Jewelry	Cell Phones & Accessories
`RexExplainer` (Lei et al., 2023)	1.80	1.80
`Mistral` (Mix)	1.60	1.87
`ChatGPT` [3]	1.87	1.60
DRE-M	2.60	2.53
DRE-C	2.67	2.73

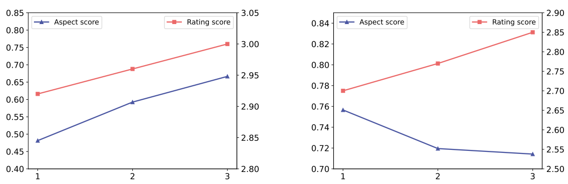
and **DRE-M** which use `ChatGPT` and `Mistral` as the LLM backbone respectively. To verify the effectiveness of each module in DRE, we also employ several ablation models: (i) **`DRE-C w/o Rev.`**: We remove all the reviews in our model and only use the description as input. (ii) **`DRE-C w/o Dist.`**: We directly summarize the description and reviews for the user-purchased item using Equation 1 without using the Distill method in Equation 2. (iii) **`DRE-C w/o Dist.`+F_p**: Based on `DRE w/o Dist.`, we also directly utilize the description and reviews of the target item without using the Summ method in Equation 1. (iv) **`DRE-C w/`** F_p: We directly generate the explanation by using the F_p as input to LLM, without using any information from user-purchased items. All the ablation studies are conducted based on **DRE-C**.

5 Experimental results

5.1 Main Results

Table 1 shows the performance of our proposed DRE and baselines in terms of two metrics. We can find that DRE shows superior performance in terms of all metrics compared to the state-of-the-art recommendation explanation method `RecExplainer`.This phenomenon indicates that compared to the latent-level alignment, our data-level alignment is capable of generating explanations of higher quality. Since we employ the data-level alignment method between the explanation model and the recommendation model, our DRE not only exhibits high quality, but also does not require any data for model training. This significantly enhances the applicability of the method, making it usable in scenarios without labeled data, and also reduces the issue of domain transfer caused by the labeled datasets.

We can also find our proposed DRE achieves superior performance compared with its LLM backbones respectively. Although the LLM backbones (*e.g.,* `Mistral` and `ChatGPT`) use the same input data as our proposed DRE, they cannot generate a high-quality recommendation explanation. Since LLMs can only reveal a limited relationship between user-purchased items and target item based solely on descriptions. This phenomenon demonstrates that our proposed target-aware user preference distillation method can assist the model in capturing more user preference information.



(a) Performance of using different numbers of user history.

(b) Performance of using different numbers of reviews.

Figure 3: Performance analysis of using different numbers of user history and reviews.

5.2 Ablation Study

To evaluate the effectiveness of each module in DRE, we also conduct ablation studies with model DRE-C, and the results are shown in Table 1. We found that the `DRE-C w/o Rev.` method achieves lower scores compared to other ablation models, indicating the effectiveness of integrating review information in our approach. Due to the complexity of information in reviews, generating meaningful explanations requires extracting target-aware information. Therefore, DRE-C w/o Dist. also exhibited lower performance after removal Distill module from DRE.

Additionally, since descriptions and reviews are usually quite long, extracting helpful information about recommended item requires distilling useful features from description and reviews. Therefore, DRE-C w/o Dist.+F_p method exhibited lower performance after removal Summ module from DRE-C w/o Dist.

5.3 Human Evaluation

In previous experiments, we used LLM to assess recommendation explanation quality. In this section, we employ two well-educated human annotators evaluate it directly. We use the same evaluation criteria as the rating score as shown in § 4.2. We conducted human evaluation on 60 randomly selected recommendation explanation samples from the Clothing Shoes & Jewelry and Cell Phones & Accessories dataset respectively. From Table 2, we can find that although the scores from the human evaluation and LLM scores (as shown in Table 1) do not fully align, the rankings among the baselines are consistent. To validate LLM-based evaluations, we assessed consistency with human evaluation using Cohen's kappa. The kappa value of 0.463 indicates moderate agreement, further supporting the consistency between LLM and human evaluations.

Additionally, to directly compare the differences in recommendation explanations generated by DRE-C and ChatGPT, we asked data annotators to directly compare the results from the two models. Specifically, we presented the data annotators with recommendation explanations generated by DRE-C and ChatGPT in random order and classified them according to the following criteria: (i) No significant difference between the two explanations; (ii) DRE-C better aligns with user preferences; (iii) ChatGPT better aligns with user preferences; The final results showed that 18.88% of samples are classified as category I, 61.11% as category II, and 20% as category III. This demonstrates that our proposed method offers significant advantages over directly prompting ChatGPT.

5.4 Case Study

Table 3 shows an example of recommendation explanations generated by ChatGPT, RecExplainer, and DRE based on information about user-

Table 3: Examples of the generated recommendation explanation of two baselines and DRE.

Item Information
Recommended Item: Kobwa (TM) Bumper Case for Apple iPhone 5, (Black & Clear) **User Historical Purchased Items:** [BlueAnt Supertooth Light Bluetooth Speakerphone (Black); XS Earbuds with Microphone and Volume Control.]
ChatGPT
Based on your purchased history ... I recommend the Kobwa (TM) Bumper Case for Apple iPhone 5 to you ... The black and clear color option matches your preference for sleek and minimalist designs (Hallucination for non-exist user preference). Additionally, this case is specifically designed for the iPhone 5 and 5G models, which you have shown interest in through your history of viewing electronic items (Not combining target-aware information).
RecExplainer
No, user will not like the item. Based on the user's purchase history ... shown an interest in technology-related items such as Bluetooth speakerphones, headphones ... user has shown an interest in accessories related to iPhones, such as earbuds and Bluetooth headphones (Not combining target-aware information.). From a recommendation model perspective ... the item matches their iPhone model and has features such as being made of hard plastic and soft rubber, it aligns with the user's preferences for durable and protective items. This consistency in compatibility may also appeal to the user. (Recommendation result not align with explanation result) ...
DRE
Based on your purchased history and other users' reviews, I recommend the Kobwa (TM) Bumper Case for Apple iPhone 5 to you ... While there may not be a direct relation between the Kobwa (TM) Bumper Case and the XS Earbuds in your viewing history, both items are accessories for Apple devices. Additionally, in the history item you viewed, the BlueAnt Supertooth Light Bluetooth Speakerphone (Black) is mentioned as being an electronic accessory with a black color option, similar to the Kobwa(TM) Bumper Case. Both items have also garnered positive feedback from users ... the Kobwa(TM) Bumper Case for Apple iPhone 5 would be a suitable recommendation for you.

purchased items and recommended item. The underlined text in the explanation indicates the recommended item and user-purchased items. We use the text in red to illustrate the shortcomings of the explanation, which is not generated by the model. The text in green shows target-aware information generated by the model. The text in blue represents the consistent information of reviews from user U for user-purchased items and recommended item. The target item profile and target-aware item profiles generated by DRE are shown in the Appendix 7.2. From this case, we can find that `ChatGPT` fails to establish convincing and reasonable relationships between recommended items and user preferences. Although `RecExplainer` employs the complicated alignment training step for the recommendation module, the generated explanation still fails to align with the recommendation result (as shown in the red text in the bracket). And DRE provides target-aware information that is persuasive and aligns with user preferences. This observation demonstrates that our proposed target-aware user preference distillation can effectively filter target-aware information from reviews and descriptions.

5.5 Analysis of Different Input

To verify the impact of the quantity of product reviews and the amount of user's historical purchase items on the model's performance, we measured the change in model performance under different input data settings. Figure 3(a) shows the effect of the amount of user's historical purchase items on the model's performance, From this figure, we can observe an upward trend in both aspect and rating scores, which demonstrates that incorporating more user historical purchase items into the model helps the model to more comprehensively understand user preferences.

Figure 3(b) shows the trend in model performance as the number of input reviews changes. As the number of item reviews a user has increased, the model pays more attention to these reviews, resulting in a focus on analyzing other user reviews of the item and a reduction in the description of item features. Since the aspect score focuses more on evaluating the description of the item features, this leads to a decrease in the score as shown in Figure 3(b). However, this decrease does not indicate a decline in the quality of the recommendation explanation. Therefore, the number of product reviews can be adjusted according to the user's preference to achieve the desired recommendation explanation.

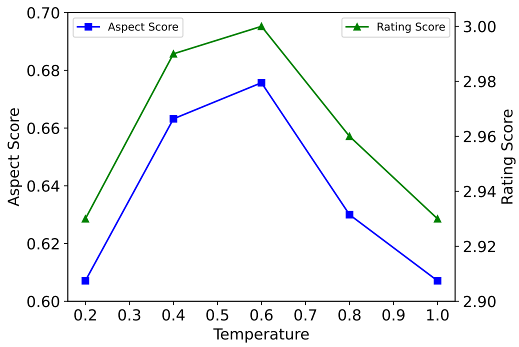


Figure 4: Performance of using different temperature settings in DRE.

5.6 Analysis of Different Hyper-parameters

The temperature parameter in the transformer-based language model controls the randomness and diversity of text generation, and higher temperature results in generating more diverse text [4]. To assess the influence of temperature setting on the DRE, we conducted experiments using different temperature configurations on the Home & Kitchen dataset. Since the recommendation explanation task requires both diverse explanations and fidelity to product attributes and user reviews, from Figure 4, we can find that both too high and too low temperature parameter can lead to a decrease in model performance.

6 Conclusion

In this paper, we introduced **D**ata-level **R**ecommendation **E**xplanation (DRE), a non-intrusive explanation framework for black-box recommendation models. We propose a data-level alignment method to align the explanation module and the recommendation model without additional parameter training or intermediate representations in recommendation model. Since the detailed information in the item description is limited, we propose the target-aware user preference distillation method to enhance semantic understanding by incorporating item reviews when generating recommendation explanations. Experimental results demonstrate the effectiveness of DRE in providing accurate and user-centric explanations, contributing to the improvement of recommendation system interpretability and user engagement.

[4] https://platform.openai.com/docs/guides/text-generation/completions-api

Limitations

In this paper, the gpt-3.5-0125 model we used can handle a maximum text length of 16k. In the real world, user historical interactions are often lengthy, leading to excessive text length that needs to be processed. Since existing long-context LLMs can easily handle large text inputs, our method can be readily adapted to these models for recommendation explanation. We plan to incorporate long-context LLMs into recommendation explanations in our future work.

Ethics Statement

While LLMs have the potential to generate hallucination information, our method leverages LLMs to distill target-aware information from ground truth data and generate explanations, ensuring that the explanations align as closely as possible with the user's information. As recommendation explanations are mostly applied in recommendation system, they are unlikely to raise significant ethical concerns.

References

Mixtral of experts: Mixtral-8x7b. https://mistral.ai/news/mixtral-of-experts/. Accessed: 2024-02-02.

Krisztian Balog, Filip Radlinski, and Shushan Arakelyan. 2019. Transparent, scrutable and explainable user models for personalized recommendation. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, pages 265–274.

Mustafa Bilgic and Raymond J Mooney. 2005. Explaining recommendations: Satisfaction vs. promotion. In *Beyond personalization workshop, IUI*, volume 5, page 153.

Rose Catherine, Kathryn Mazaitis, Maxine Eskenazi, and William Cohen. 2017. Explainable entity-based recommendations with knowledge graphs. *arXiv preprint arXiv:1707.05254*.

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*, pages 1583–1592.

Jingyuan Chen, Hanwang Zhang, Xiangnan He, Liqiang Nie, Wei Liu, and Tat-Seng Chua. 2017. Attentive collaborative filtering: Multimedia recommendation with item-and component-level attention. In *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 335–344.

Tong Chen, Hongzhi Yin, Guanhua Ye, Zi Huang, Yang Wang, and Meng Wang. 2020. Try this instead: Personalized and interpretable substitute recommendation. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pages 891–900.

Xu Chen, Zheng Qin, Yongfeng Zhang, and Tao Xu. 2016. Learning to rank features for recommendation over multiple categories. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 305–314.

Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, et al. 2016. Wide & deep learning for recommender systems. In *Proceedings of the 1st workshop on deep learning for recommender systems*, pages 7–10.

Mingyue Cheng, Zhiding Liu, Qi Liu, Shenyang Ge, and Enhong Chen. 2022. Towards automatic discovering of deep hybrid network architecture for sequential recommendation. In *Proceedings of the ACM Web Conference 2022*, pages 1923–1932.

Mingyue Cheng, Fajie Yuan, Qi Liu, Xin Xin, and Enhong Chen. 2021. Learning transferable user representations with sequential behaviors via contrastive pre-training. In *2021 IEEE International Conference on Data Mining (ICDM)*, pages 51–60. IEEE.

Qiming Diao, Minghui Qiu, Chao-Yuan Wu, Alexander J Smola, Jing Jiang, and Chong Wang. 2014. Jointly modeling aspects, ratings and sentiments for movie recommendation (jmars). In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 193–202.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. *Advances in neural information processing systems*, 32.

Yunfan Gao, Tao Sheng, Youlin Xiang, Yun Xiong, Haofen Wang, and Jiawei Zhang. 2023. Chat-rec: Towards interactive and explainable llms-augmented recommender system. *arXiv preprint arXiv:2303.14524*.

Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. Deepfm: a factorization-machine based neural network for ctr prediction. *arXiv preprint arXiv:1703.04247*.

Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*, pages 507–517.

Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*, pages 173–182.

Christopher C Johnson et al. 2014. Logistic matrix factorization for implicit feedback data. *Advances in Neural Information Processing Systems*, 27(78):1–9.

Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37.

Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. 2017. Interpretable & explorable approximations of black box models. *arXiv preprint arXiv:1707.01154*.

Yuxuan Lei, Jianxun Lian, Jing Yao, Xu Huang, Defu Lian, and Xing Xie. 2023. Recexplainer: Aligning large language models for recommendation model interpretability. *arXiv preprint arXiv:2311.10947*.

Weizhi Ma, Min Zhang, Yue Cao, Woojeong Jin, Chenyang Wang, Yiqun Liu, Shaoping Ma, and Xiang Ren. 2019. Jointly learning explainable rules for recommendation with knowledge graph. In *The world wide web conference*, pages 1210–1221.

Ben Mann, N Ryder, M Subbiah, J Kaplan, P Dhariwal, A Neelakantan, P Shyam, G Sastry, A Askell, S Agarwal, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Conference on Empirical Methods in Natural Language Processing*.

Deng Pan, Xiangrui Li, Xin Li, and Dongxiao Zhu. 2020. Explainable recommendation via interpretable feature mapping and evaluation of explainability. *arXiv preprint arXiv:2007.06133*.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Amit Sharma and Dan Cosley. 2013. Do social explanations work? studying and modeling the effects of social explanations in recommender systems. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1133–1144.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMLR.

Juntao Tan, Shuyuan Xu, Yingqiang Ge, Yunqi Li, Xu Chen, and Yongfeng Zhang. 2021. Counterfactual explainable recommendation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 1784–1793.

Nava Tintarev and Judith Masthoff. 2010. Designing and evaluating explanations for recommender systems. In *Recommender systems handbook*, pages 479–510. Springer.

Aaron Van den Oord, Sander Dieleman, and Benjamin Schrauwen. 2013. Deep content-based music recommendation. *Advances in neural information processing systems*, 26.

Lei Wang, Jingsen Zhang, Hao Yang, Zhiyuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Ruihua Song, Wayne Xin Zhao, Jun Xu, Zhicheng Dou, Jun Wang, and Ji-Rong Wen. 2024. User behavior simulation with large language model based agents. *Preprint*, arXiv:2306.02552.

Xiang Wang, Xiangnan He, Fuli Feng, Liqiang Nie, and Tat-Seng Chua. 2018a. Tem: Tree-enhanced embedding model for explainable recommendation. In *Proceedings of the 2018 world wide web conference*, pages 1543–1552.

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. IEEE.

Yan Wang, Zhixuan Chu, Xin Ouyang, Simeng Wang, Hongyan Hao, Yue Shen, Jinjie Gu, Siqiao Xue, James Y Zhang, Qing Cui, et al. 2023. Enhancing recommender systems with large language model reasoning graphs. *arXiv preprint arXiv:2308.10835*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. 2023. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*.

Lanling Xu, Junjie Zhang, Bingqian Li, Jinpeng Wang, Mingchen Cai, Wayne Xin Zhao, and Ji-Rong Wen. 2024. Prompting large language models for recommender systems: A comprehensive framework and empirical analysis. *arXiv preprint arXiv:2401.04997*.

Weiwen Xu, Deng Cai, Zhisong Zhang, Wai Lam, and Shuming Shi. 2023. Reasons to reject? aligning language models with judgments. *arXiv preprint arXiv:2312.14591*.

Qi Zhang, Jingjie Li, Qinglin Jia, Chuyuan Wang, Jieming Zhu, Zhaowei Wang, and Xiuqiang He. 2021. Unbert: User-news matching bert for news recommendation. In *IJCAI*, pages 3356–3362.

Yongfeng Zhang, Xu Chen, et al. 2020a. Explainable recommendation: A survey and new perspectives. *Foundations and Trends® in Information Retrieval*, 14(1):1–101.

Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. 2014. 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*, pages 83–92.

Yuan Zhang, Xiaoran Xu, Hanning Zhou, and Yan Zhang. 2020b. Distilling structured knowledge into embeddings for explainable and accurate recommendation. In *Proceedings of the 13th international conference on web search and data mining*, pages 735–743.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Jan Ruben Zilke, Eneldo Loza Mencía, and Frederik Janssen. 2016. Deepred–rule extraction from deep neural networks. In *Discovery Science: 19th International Conference, DS 2016, Bari, Italy, October 19–21, 2016, Proceedings 19*, pages 457–473. Springer.

7 Appendix

7.1 Computational Cost

Table 4: Statistics of token consumption for baselines. We show the token consumption of each module in DRE in the first three rows. The number in the bracket represents the percentage of tokens consumed by the module relative to the total token consumption of the model.

	Home & Kitchen	Clothing Shoes & Jewelry	Cell Phones & Accessories	Yelp
Sub-modules in DRE				
`Summ`	2059 (15.09%)	3138 (15.40%)	2530 (21.41%)	1438 (12.42%)
`Distill`	9046 (66.31%)	12752 (62.59%)	7055 (59.69%)	7847 (67.78%)
`Explain`	2536 (18.59%)	4484 (22.01%)	2234 (18.90%)	2293 (19.80%)
`DRE`	13641	20374	11819	11578
`ChatGPT`	3331	2227	3096	2850

Since our proposed DRE is a multi-module method based on prompting LLM, we provide statistics on the total token consumption of DRE and the token consumption of each module separately. Table 4 compares the token consumption of our proposed method with several baseline methods. Firstly, from the results, it can be seen that the Distill module in our proposed DRE consumes the most tokens compared to the other two modules. Since the Distill module is responsible for generating target-aware items profiles F_N, which requires using a large amount of item information as input and analyzing product associations, it consumes a significant number of tokens. Furthermore, as shown in the ablation study in Table 1, the Distill module contributes the most to the overall performance improvement in DRE (compared between DRE-C and `DRE-C w/o Dist.`).

The token consumption for the Summ module is mainly around 2k in three subsets in the Amazon dataset, while the token consumption for the Summ module in the Yelp dataset is lower than the other three datasets. Since the Yelp dataset treats categories and attributes as item descriptions, resulting in shorter item information compared to the other three datasets in Amazon, which have long item descriptions.

Since `ChatGPT` uses only simple instructions as prompts to directly generate recommendation explanations, its token consumption is lower than our method. However, the quality of the explanation generated by `ChatGPT` is significantly lower than those produced by our proposed DRE as shown in Table 1.

7.2 Case Study

The target item profile and target-aware item profiles generated by DRE.

Table 5: Details of the target item profile

Target Item Profile
item: Kobwa(TM) Bumper Case for Apple iPhone 5, 5G **description**: Kobwa(TM) Bumper Case is made of hard plastic and soft rubber, available in black and clear colors. It is compatible with the newest iPhone 5 5S. The package includes 1 case and 1 Kobwa's keyring. Only authorized Kobwa online retailers provide original packaging and keyring with printed logo. **other users' reviews**: Kobwa(TM) Bumper Case for Apple iPhone 5, 5G is commended for its affordable pricing and functionality. Some users noted slight stiffness in the volume button and the case's color not being entirely transparent. Despite the shipping delay and personal preference for covered back cases, the overall rating is positive due to the budget-friendly nature of the product.

Table 6: Details of the Target-aware Item Profile for BlueAnt Supertooth Light Bluetooth Speakerphone

Target-aware Item Profile: BlueAnt Supertooth Light Bluetooth Speakerphone
history item: BlueAnt Supertooth Light Bluetooth Speakerphone (Black) **genre**: electronics **relevant information**: Both the BlueAnt Supertooth Light Bluetooth Speakerphone and Kobwa(TM) Bumper Case focus on design and functionality. The BlueAnt speakerphone emphasizes hands-free technology with clear audio processing, while the Kobwa bumper case highlights a combination of hard plastic and soft rubber for iPhone protection. Both items aim to enhance user experience through innovative design and practical features. **other users' reviews**: Users appreciate the BlueAnt speakerphone for its outstanding audio quality, convenient design, and long-lasting battery life. They highlight the ease of use, clear communication, and smart features like the pop-out microphone and metallic visor clip. Despite minor issues like squishy volume buttons, the overall satisfaction is high.

Table 7: Details of the Target-aware Item Profile for XS Earbuds

Target-aware Item Profile: XS Earbuds
history item: XS Earbuds with Microphone and Volume Control, Bluetooth Headphones Noise Canceling **genre**: electronics **relevant information**: Both the XS Earbuds and Kobwa(TM) Bumper Case are designed for specific Apple devices - the XS Earbuds for iPhones and the Kobwa(TM) Bumper Case for iPhone 5 and 5G. They both provide secure mounting for Apple devices with different functionalities, with the XS Earbuds focusing on hands-free device usage while the Kobwa(TM) Bumper Case offers protection and style. **other users' reviews**: Users appreciate the secure grip and functionality of the iOttie Easy Flex 2, noting its strong suction cup and easy phone grip mechanism. Some users suggest improvements, like longer arms for better positioning or a more secure grip for larger phones. Overall, users find it durable, convenient for daily use, and suitable for various car models.