UNIVERSITY OF CALIFORNIA SAN DIEGO

Towards Trustworthy Natural Language Explanations for Recommender Systems

A thesis submitted in partial satisfaction of the
requirements for the degree Master of Science

in

Computer Science

by

Zhouhang Xie

Committee in charge:

Professor Julian McAuley, Chair
Professor Zhiting Hu
Professor Jingbo Shang

2023

The Thesis of Zhouhang Xie is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2023

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor, Prof. Julian McAuley, for his guidance and support over the past two years. Thank you for all the discussions and suggestions. Thank you for being there and anchoring my research when life hits me from the most unexpected angle. I would also like to thank my committee members, Prof. Zhiting Hu and Prof. Jingbo Shang, for serving on my committee and offering their valuable time and precious comments.

To Prof. Sameer Singh, for introducing me to the world of Natural Language Processing. To Dr. Bodhisattwa Prasad Majumder, I wish you the best in Seattle. To Zhankui, thank you for all the insightful discussions. To other members at Lab McAuley, I feel incredibly grateful to be able to meet you all. To Emily, Nadyn, Wendy, and other CSE department members and ISPO members for their kindness through that unsettling summer.

Chapter 1, in full, is a reprint of the material presented at the Southern California Natural Language Processing Symposium, 2022 ("On Faithfulness and Coherence of Language Explanations for Recommendation Systems", Zhouhang Xie, Julian McAuley, and Bodhisattwa Prasad Majumder). The thesis author was the primary investigator and author of this paper.

Chapter 2, in full, is a reprint of the material as it appears in Proceedings of the AAAI Conference on Artificial Intelligence, 2023 ("Factual and Informative Review Generation for Explainable Recommendation", Zhouhang Xie, Sameer Singh, Julian McAuley, and Bodhisattwa Prasad Majumder). The thesis author was the primary investigator and author of this paper.

ABSTRACT OF THE THESIS

Towards Trustworthy Natural Language Explanations for Recommender Systems

by

Zhouhang Xie

Master of Science in Computer Science

University of California San Diego, 2023

Professor Julian McAuley, Chair

Product reviews are a form of user feedback that provide richer information than traditional signals in recommender systems, such as star ratings and implicit feedback. Meanwhile, a review is also a justification for the user's rating of a product. Previous works show that recommendation models can predict user ratings more accurately by jointly learning to generate reviews. The generated reviews can also serve as recommendation explanations, making the recommender system more interpretable. However, existing works evaluate these generated explanations using traditional natural language generation metrics only, overlooking trustworthiness, an important aspect of model explanations. In this thesis, we focus on two properties of trustworthy recommendation explanations: faithfulness, how truthfully do explanations reflect

the decision process of the model, and factuality, whether the generated content accurately reflects the characteristics of the corresponding product. Specifically, this thesis includes two directions: (1) we propose a set of methods for evaluating the faithfulness and semantic coherency of recommendation explanations, and (2) we develop a personalized retrieval-augmented model that can generate factual and informative reviews to explain its recommendation predictions.

Chapter 1

On Faithfulness and Coherence of Language Explanations for Recommendation Systems

Reviews contain rich information about product characteristics and user interests and thus are commonly used to boost recommender system performance. Specifically, previous work show that jointly learning to perform review generation improves rating prediction performance. Meanwhile, these model-produced reviews serve as recommendation explanations, providing the user with insights on predicted ratings. However, while existing models could generate fluent, human-like reviews, it is unclear to what degree the reviews fully uncover the rationale behind the jointly predicted rating. In this work, we perform a series of evaluations that probes state-of-the-art models and their review generation component. We show that the generated explanations are brittle and need further evaluation before being taken as literal rationales for the estimated ratings.

1.1 Introduction

Product reviews capture rich information about user preferences and thus improve recommender system performance McAuley et al. (2012); McAuley and Leskovec (2013a); Zheng et al. (2017); Tay et al. (2018); Chen et al. (2018); Pugoy and Kao (2020, 2021). Meanwhile, advancements in text generation enable generating realistic synthetic reviews conditioning on

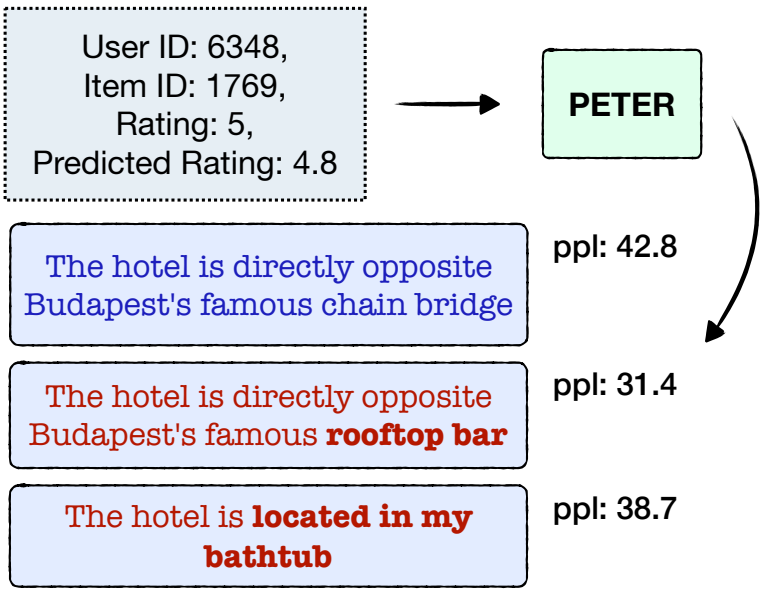

Figure 1.1. PETER, a state-of-the-art model assigns lower perplexity to factually incoherent reviews.

user and item identifiers, as well as additional features such as historical reviews Li and Tuzhilin (2019), product metadata Ni and McAuley (2018); Dong et al. (2017), knowledge graph embedding Li et al. (2021a), and sometimes the rating itself Chen et al. (2021). Recently, there has been increasing interest in coupling rating estimation and review generation, treating generated reviews as *explanations* for model recommendations Ni et al. (2017); Sun et al. (2020); Li et al. (2020b, 2021b); Hada et al. (2021).

In the current literature, the quality of the generated explanations are usually measured by perplexity and overlapping-based metrics such as Distinct-N Li et al. (2016), Rouge score Lin (2004), and BLEU score Papineni et al. (2002) with respect to the ground truth reviews. However, while these evaluations measure fluency and word-overlapping, they do not warrant the the generated reviews' quality as explanations.

Specifically, overlapping metrics overlook two core aspect of natural language explanations (NLEs): (1) *faithfulness*, how truthfully do the generated explanations reflect the decision process for the models rating prediction, and (2) *semantic coherence*, how well the model capture the users' true interest towards the product. To highlight the potential issue associated with

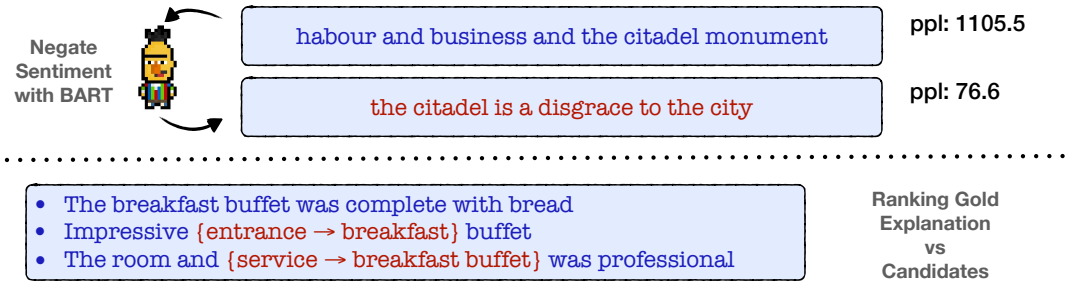

Figure 1.2. Up: adversarial invariance ratio. Down: robustness.

current evaluation, consider the review text for a restaurant "*I love this hotel because it has great service*" with a rating of 5, where the explanation generated is "*I love this hotel because it has great cookies*" with the correct predicted rating. The generated explanation deviates from the ground truth sentence by only one word, yet completely changes the rationale for the rating. However, the currently widely used automatic metrics will still assign a high score to the generated review. Further, there is no guarantee that even *cookie* is truly accountable for the predicted rating.

To address these discrepancies, we argue that NLEs for recommendation systems should be evaluated as explanations, similar to NLEs in NLP tasks. **In this work**, we probe review-as-explanation models in explainable recommendation literature. Our results show a concerning trend that current models struggle to produce reviews that are semantically coherent with the ground truth reviews, and are inconsistent with the explanation they produce. We encourage researchers and practitioners take beyond-overlapping evaluations into account when training review generation models for explanations. Better evaluations could lead to deeper understanding of capabilities of these generated rationales, and foster more trustworthy explainable recommendation systems.

1.2 Problem Definition and Models

This section is organized as follows: we first cover the task of joint review-rating generation, then introduce the models used in our experiments.

1.2.1 Problem Setup

Given a user u and an item i, the task of joint review-rating generation aims at predicting an associated rating $\hat{r}$ as well as a natural language explanation $\hat{e}$ [1]. During training, the model jointly minimizes the negative log likelihood (NLL) of ground truth reviews in the corpus, as well as the mean squared error (MSE) of their associated rating.

1.2.2 Models

We compare four recent models in the literature that covers a variety of commonly used architectures in natural language generation: Att2Seq Dong et al. (2017), NRT Li et al. (2017), PEPLER Li et al. (2022), and PETER Li et al. (2021b). Among the models, Att2Seq and NRT are based on Long Short-Term Memory (LSTM) Hochreiter and Schmidhuber (1997), while PEPLER combines a pre-trained GPT-2 model Radford et al. (2019) with prompt tuning. Finally, PETER adopts non-auto-regressive transformer architecture.

Meanwhile, since Li et al. noted incorporating specific content words significantly improves generation quality, we follow the original paper and condition the model on a content word, denoted by PETER$_{cond}$. Note that following the previous implementation, the aspect word in the dataset is extracted from *ground truth* review, giving PETER$_{cond}$ an unfair advantage. We thus use PETER$_{cond}$ as an upper-bound baseline.

1.2.3 Datasets

We conduct our experiments on Yelp[2] (Y.), TripAdvisor[3] (T.), and Movies and TV category from Amazon dataset He and McAuley (2016) (M.). These are standard dataset commonly used to benchmark joint review-rating estimation models Li et al. (2020b).

[1] which is commonly the associated review
[2] https://www.yelp.com/dataset/challenge
[3] https://www.tripadvisor.com

1.3 Evaluating Faithfulness and Semantic Coherence

We generate 10,000 explanations for each model on each dataset, and perform a set of evaluations as described in this section.

1.3.1 Faithfulness

When reviews are treated as natural langauage explanations, joint review-rating prediction models could be categorized into self-rationalizing models. Jacovi and Goldberg argues that the quality of NLE [4] should be evaluated by both their plausibility, how convincing the explanations are to humans, and faithfulness, how truthful they reflect the models' decision process. We focus on model faithfulness in this section.

By definition, a faithful explanation will truthfully represent the decision process of a model. However, directly measuring faithfulness is infeasible due to the black-box nature of deep neural networks. We instead design a set of proxy tasks that test *unfaithful* behavior of joint review-rating estimation models.

Adversarial In-variance Ratio (AIR).

Since the explanation generation by the model is representative of the model's belief of the reasons behind the rating prediction, we argue that such belief must be robust to sentiment perturbations. In other words, assume a model generates a sequence $\hat{e}_{u,i}$ as an explanation, the sentiment-negated counter explanation $\neg\hat{e}_{u,i}$ should not receive a higher likelihood (lower perplexity) than the original review. Illustration of selected sub-experiments are as shown in Figure 1.2.

Concretely, we take 4 or 5-star (positive sentiment) ratings from the test set and rewrite their sentiment to negative using a pretrained BART model[5], and let the target model rank the ground truth and rewritten review with perplexity. We mark the models' decision as flipped if

[4] NLR in Jacovi and Goldberg's work

[5] dapang/yelp_pos2neg_lm_bart_large from huggingface.

it assigns lower perplexity to the rewritten review with the negated sentiment. In this case, the model's explanation is thus unfaithful. Note that this means a random baseline would achieve 50 percent in AIR.

Mean Reciprocal Rank against Alternative Explanations (MRR-AE).

As pointed out in Jacovi and Goldberg, a model is unfaithful if it provides a different interpretation for the same decision by the same model. That is, the model should be able to differentiate its generated review from other candidate explanations. Following this intuition, we argue that the model should have the ability to pick out its generated review from other reviews, such as random reviews drawn from the dataset or adversarially constructed ones, as shown in Figure 1.2.

To measure this, we sample 100 reviews randomly from the test dataset for each gold review, and replace the aspect in the sampled sentences with the aspect covered by the ground truth. We then let the target model rank the 100 sentences along with the gold review with perplexity score and measure its performance with mean reciprocal rank (MRR). The random baseline for MRR-AE is thus around 5 percent.

Text-label Agreement Error (TLAE).

As faithful explanations, the generated reviews should strongly correlate with predicted ratings. To measure this, we train a BERT Vaswani et al. (2017) based auxiliary rating regressor based on *only* user reviews on the training set of the models being evaluated. At test time, we measure the Mean Squared Error of the auxiliary predictor on *generated* reviews and regressor-predicted ratings.

1.3.2 Semantic Coherence

Traditional evaluation metrics use in the literature focuses on word-overlapping, and thus would be insensitive to mismatched content words. To address this issue, we argue that generated explanations should be evaluated by its semantic coherence. Concretely, we adopt

Table 1.1. Evaluation results on the datasets.

	Faithfulness									Semantic Coherence									Rec.		
Metric	**AIR↑**			**MRR-AE↑**			**TLAE↓**			**Entail↑**			**BERTS.↑**			**BARTS.↓**			**RMSE↓**		
Model	**M.**	**T.**	**Y.**	**M.**	**T.**	**Y.**	**M.**	**T.**	**Y.**	**M.**	**T.**	**Y.**	**M.**	**T.**	**Y.**	**M.**	**T.**	**Y.**	**M.**	**T.**	**Y.**
Att2Seq	14.6	43.5	47.9	23.5	19.2	23.5	n/a	n/a	n/a	6.6	2.7	5.6	0.08	0.16	0.10	5.95	5.97	5.97	n/a	n/a	n/a
NRT	55.8	43.0	**55.8**	15.1	18.2	22.6	**1.39**	0.84	**1.18**	3.2	1.5	1.4	-0.21	-0.15	-0.19	6.54	6.72	6.6	**0.95**	**0.79**	**1.01**
PETER	60.3	47.3	49.4	18.1	**22.0**	26.3	**1.39**	0.84	**1.18**	8.2	3.5	7.9	0.11	0.17	0.12	5.95	5.96	5.88	**0.95**	0.81	**1.01**
PEPLER	**70.0**	**63.3**	21.0	16.6	16.3	7.3	**1.39**	0.84	**1.18**	4.4	3.4	4.6	0.13	0.19	0.23	5.95	5.93	6.99	1.25	1.71	1.69
PETER$_{cond}$	19.1	52.0	55.7	**27.9**	20.2	**27.9**	**1.39**	**0.76**	**1.18**	**27.7**	**24.2**	**24.3**	**0.18**	**0.30**	**0.25**	**5.47**	**5.93**	**5.17**	**0.95**	1.81	1.02

two recent, state-of-the-art semantic evaluation metrics: BERTScore Zhang et al. (2020) and BARTScore Yuan et al. (2021)[6]. Further, we use a pre-trained entailment model to check whether the generated content entails the ground truth review. We report the percentage of entailment (Entail), where a good model should have high ratio.

1.4 Empirical Results

Faithfulness.

Our main evaluation results are as shown in table 1.1. While model-generated explanation generally matches the predicted rating (TLAE), most models have near random performance against sentiment perturbations (AIR). Meanwhile, although PEPLER is the most robust to sentiment perturbation, it is not as competitive as other models in terms of recommendation performance (RMSE). This illustrates the potential risk of powerful language models giving a false sense of explainability simply due to their strong language modeling ability. In other words, the explanations are plausible but not faithful under Wiegreffe et al. (2021)'s framework.

Semantic Coherence.

From coherence evaluations, we could see the model generally struggle to capture the exact aspect that the user cares about, resulting in a low entailment ratio (Entail) compared to PETER$_{cond}$. This can be corroborated by BERTScore and BARTscore, highlighting the importance of conducting semantic evaluations for explanation generation.

[6]BERTScore is cosine similarity-based (larger means better) and BARTScore is NLL based (smaller means better).

1.5 Discussion and Analysis

Non-robust correlation between generated review and estimated rating.

Based on TLAE, we could observe that the generated review is indeed correlated to the predicted rating. However, the models being evaluated all demonstrate near-random AIR scores, showing such correlations are brittle, and the language model's belief is entangled with other rating irrelevant factors.

Weak correlation between generated item aspect and the item.

From the MRR-AE score, we could see that models generally perform poorly in ranking the generated review against synthetic alternative explanations, where the description of a random item is used in place of the generated one. This behavior shows the model fails to establish robust connections between generated reviews and the corresponding items.

How much can reviews explain rating?

While reviews boost recommender system performance, they cannot *fully* explain the corresponding rating. In particular, human-written explanations are inherently limited in *discovery* tasks, where a machine learning model needs to demonstrate beyond-human performance Tan (2022). As a result, human-written explanations are not the complete reason for inferring the label. Consequently, maximizing the likelihood of explanations does not guarantee the best task performance Carton et al. (2022), which could be corroborated by findings from the recommender system community Sachdeva and McAuley (2020) and PEPLER model from our evaluation.

1.6 Implications and Recommendations

Calibrate user expectations.

End users often trust that algorithm explanations are faithful Jin et al. (2022)[7] and hope they could receive reliable explanations Lakkaraju et al. (2022), we recommend practitioners inform the users about the probablistic nature of generated reviews as explanations. For example,

[7]Jin et al. study was based on medical image

if a system generates an explanation related to 'bbq pork ribs', it would be more of an indication of user's interest in smokehouse cuisines rather than the dish itself.

Develop hybrid systems.

While PETER$_{cond}$ acts as an "unfair" baseline in our experiment section, the model would be a great tool in a larger pipeline, where users actively provide feedback. Similarly, recent works starts to explore natural language as an *interface* in pipeline systems for non-language based explanations Slack et al. (2022). We encourage the community to consider conditional and pipeline systems in addition to end-to-end models.

Better evaluations.

We note that although evaluations generally depend on the use-cases of the model, and a powerful model does not necessarily need to satisfy faithfulness, plausibility, and semantic coherence simultaneously, it is advisable to perform beyond-overlapping evaluations before assuming the literal validity of generated recommendation NLRs.

1.6.1 Related Works

Faithfulness of Natural Language Explanations.

Jacovi and Goldberg argues that the quality of NLE from this class of models should be evaluated by their faithfulness, how truthful (and thus consistent) do they reflect the models' decision process. Under this setup, our work's evaluation differs from existing evaluations in the literature in that we clearly distinguishes faithfulness from general language quality. Wiegreffe et al. approach this problem by measuring the connection between labels and explanations, yet their evaluation do not take the semantics of the generated explanation itself into account.

Analyzing Model Decision in NLP.

Another related line of work is analysis of model decision boundaries. Common strategies usually involves adversarially probing the model, such as using counterfactual data Wu et al. (2021), constrast sets Gardner et al. (2020) and semantically preserving modifications of sentence

characteristics Ribeiro et al. (2018, 2020); Longpre et al. (2021). Our work deviates from prior works as we establish connections of adversarial evaluation directly with model faithfulness.

1.7 Conclusions

Joint review-rating prediction models could generate high-quality reviews while producing accurate rating estimations. However, it is unclear whether the generated reviews could be leveraged as precise recommendation rationales. We conduct a set of evaluation that benchmark faithfulness and semantic coherence of state-of-the-art models. We show more careful evaluations are needed before generated reviews could be taken as fully accountable explanations.

1.8 Acknowledgement

This chapter, in full, is a reprint of the material presented at the Southern California Natural Language Processing Symposium, 2022 ("On Faithfulness and Coherence of Language Explanations for Recommendation Systems", Zhouhang Xie, Julian McAuley, and Bodhisattwa Prasad Majumder). The thesis author was the primary investigator and author of this paper.

Chapter 2

Factual and Informative Review Generation for Explainable Recommendation

Recent models can generate fluent and grammatical synthetic reviews while accurately predicting user ratings. The generated reviews, expressing users' estimated opinions towards related products, are often viewed as natural language 'rationales' for the jointly predicted rating. However, previous studies found that existing models often generate repetitive, universally applicable, and generic explanations, resulting in uninformative rationales. Further, our analysis shows that previous models' generated content often contain factual hallucinations. These issues call for novel solutions that could generate both *informative* and *factually grounded* explanations. Inspired by recent success in using retrieved content in addition to parametric knowledge for generation, we propose to augment the generator with a personalized retriever, where the retriever's output serves as external knowledge for enhancing the generator. Experiments on Yelp, TripAdvisor, and Amazon Movie Reviews dataset show our model could generate explanations that more reliably entail existing reviews, are more diverse, and are rated more informative by human evaluators.

2.1 Introduction

Recently, there has been increasing interest in treating review generation as a proxy for explainable recommendation, where generated reviews serve as rationales for the models'

recommendationsLi et al. (2016, 2021b); Ni et al. (2017); Ni and McAuley (2018). However, existing models commonly generate repetitive and generic content, resulting in uninformative explanations Geng et al. (2022). Further, when evaluating the factuality of generated reviews using pre-trained entailment model, our analysis shows that existing models are also susceptible to factual hallucination, a long-existing challenge in many natural language generation (NLG) tasks Pagnoni et al. (2021); Maynez et al. (2020). Specifically, the models often generate statements that are not supported by information about the corresponding product in the training set. Both nonfactual and uninformative explanations are undesirable, as end users would look for recommendation rationales that truthfully reflect the characteristics of the product without being overly generic. Thus, these problems limit the usability of natural language explanations (NLE) produced by existing explainable recommendation models.

In order to address the issue that models commonly generate univiersally correct explanations, previous works experimented with diversifying generated reviews using distantly retrieved images as additional signals Geng et al. (2022). However, recommender system datasets do not always have associated images, and Geng et al. proposed to retrieve images from the web using available textual data. While this method indeed significantly diversifies the generated natural language explanations (NLE), there is no guarantee that the retrieved content will truthfully represent the quality of the corresponding product. Thus, the generator needs to condition on a given feature or aspect the user cares about at inference time, commonly extracted from the *ground-truth* review. This limits the usability of models as such user input might not be available at inference time. Another line of work attempted to incorporate a pre-trained language model for better generation quality Li et al. (2021b). However, the same study shows that pre-trained language models such as GPT-2 struggle to produce diverse reviews while maintaining competitive recommendation accuracy. Thus, generating informative and factual reviews without having access to information in ground truth reviews remains an open problem.

Recent advances in knowledge-grounded NLG show that retrieving unstructured text as supplements for the model's parametric knowledge significantly improves factuality and

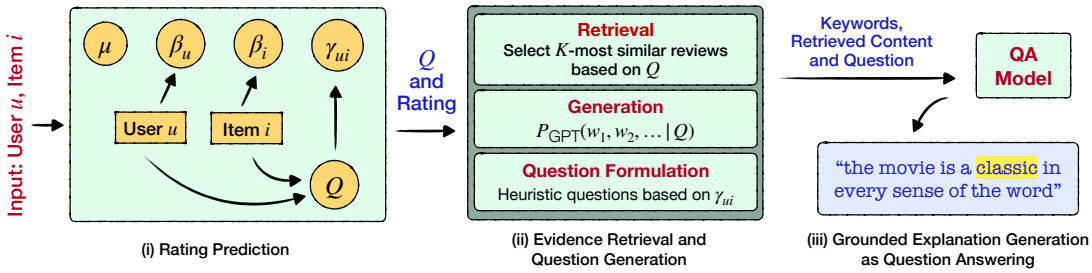

Figure 2.1. The proposed framework PRAG.

diversity of generated content Lewis et al. (2020); Guu et al. (2020). Inspired by such success, in this work, we propose to leverage existing reviews as additional context for generating recommendation explanations. Specifically, we propose **P**ersonalized **R**etriever **A**ugmented **G**enerator[1] (**PRAG**), a model loosely based on the retriever-reader framework for explainable recommendation. PRAG consists of a personalized retriever that can accurately give a rating estimation and generate a latent query for review prediction given the input user and item. To encourage *factual* explanations, we formulate the task of NLE generation as question-answering, where a reader model produces the explanation grounded on retrieved content. Meanwhile, to ensure the *informativeness* of the explanations, we estimate a set of personalized high-tf-idf keywords using the latent query, and use these keywords to guide the answer of the reader model. In this way, the reader model produces the final explanation by abstracting input text snippets and keywords, which yields superior quality compared to previous work in both automatic and human evaluation. Our contributions are as follows:

- To our best knowledge, we are the first work that evaluates and highlights the importance of factuality in natural language explanation for recommender systems.
- We develop a novel personalized retriever-reader model for generating factual and informative recommendation explanations.
- We personalize question-answering for generating recommendation explanations.

[1] https://github.com/zhouhanxie/PRAG

2.2 Related Work

Explainable recommendation aims at providing users insight for a recommender systems' decision. Following earlier works that provide topic words and product features as explanations McAuley and Leskovec (2013b); Zhang et al. (2014), recent works are increasingly focusing on generating reviews as explanations Ni et al. (2017); Ni and McAuley (2018); Hada et al. (2021); Dong et al. (2017). However, these works are not focused on the factuality of the generated content. Meanwhile, existing works require training a language model from scratch, while previous attempts at leveraging pre-trained language models either require decoding-time search Hada et al. (2021), or are not as performant as other recommendation models Li et al. (2022).

It is to be noted that there are commonly used existing metrics such as Distinct-N Liu et al. (2022) and unique sentence ratio (USR) Li et al. (2020b) that focus on diversity of generated explanations. However, a common measure used to increase diversity is to generate a sentence based on words and phrases from the *ground-truth* review, such as in Geng et al. and Ni and McAuley. Our work differs from previous works in that we consider specifically the case where no information from the ground truth review is given, which is common in recommender systems.

Outside of recommender systems, there is a general trend for using natural language as explanations for various tasks, such as text classification Hancock et al. (2018); Zhou et al. (2020), image captioning Majumder et al. (2021b); Marasović et al. (2020) and question answering Dalvi et al. (2021); Tafjord and Clark (2021). However, expert annotated explanations are usually unavailable in recommender system datasets, and reviews are usually noisy by nature and require further processing Bražinskas et al. (2020, 2021). This leaves learning to generate explanations from noisy supervision an open challenge, which we seek to address in this work.

Another related existing problem in natural language generation tasks is safe and generic outputs. For example, this problem is well studied in dialogue generation, where previous work shows that exposure bias and maximum likelihood objective lead the model to produce universally

applicable responses Zhou et al. (2021). In another parallel line of work on evaluating natural language explanations, Wiegreffe et al. pointed out that universally correct explanations are undesirable, calling for diverse and informative explanations. However, these works commonly focus on traditional NLP tasks and well-formulated datasets.

2.3 PRAG: Setup and Overview

We first introduce the joint review-rating prediction task, then cover our model's general architecture.

2.3.1 Problem Setup

Consider a set of users U and items I, where each user $u \in U$ and item $i \in I$ in the training set is associated with a real numbered rating $r_{u,i} \in \mathscr{T}$ and a review $e_{u,i} \in E$ that serves as explanations for the users rating decision. The task of joint rating-explanation estimation learns a function rec : $u, i \rightarrow \hat{r}_{u,i}, \hat{e}_{u,i}$, where $\hat{e}_{u,i}$ is a textual explanation that informs the user of the reasoning behind $\hat{r}_{u,i}$. Note that broadly speaking, explanations can be in various forms, such as topic words and feature importance, here, we use $e_{u,i}$ to denote natural language explanation specifically.

2.3.2 Model Overview

The architecture of PRAG is depicted in Figure 2.1. Given the input user u and item i, we first obtain semantic embeddings for all related existing reviews. Here, "related reviews" are reviews written by the user or written about the item. Then, given the historical reviews, a personalized retriever model produces a latent query $Q_{u,i}$ that is close to the ground truth explanation given the input user and item. A rating prediction module then uses such a latent query to produce a rating estimation.

To generate an explanation, we exploit the rich semantic information in $Q_{u,i}$. Specifically, we (1) retrieve a set of existing reviews G based on $Q_{u,i}$, where $G \subseteq E$, and (2) generate a set

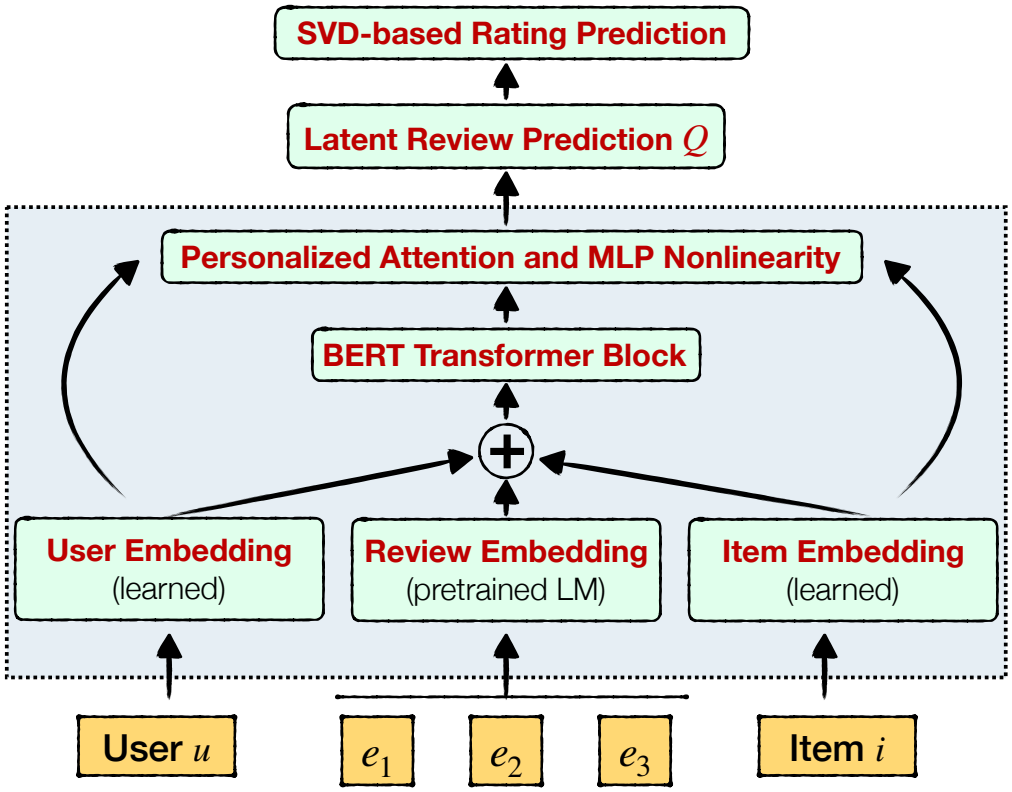

Figure 2.2. Overview for the retriever architecture.

of high tf-idf scored keywords that the user would care about based on $Q_{u,i}$ and G. Finally, a keyword-guided question answering model produces the final explanation based on the keywords, retrieved reviews, and the question. We now provide descriptions for each component.

2.4 Personalized Retriever

The architecture of our retriever model is as shown in Figure 2.2. The review, item, and user embedding go through a stack of transformer blocks and are then pooled to generate the estimated review embedding personalized for each user.

2.4.1 Embedding Reviews

Pre-trained language model representations contain rich semantic information about sentences and supports arithmetic operations such as addition and dot product. Following

previous success in pre-computing embeddings for efficient retrieval, such as in RAG Lewis et al. (2020), we obtain sentence embeddings for all reviews using a pre-trained sentence encoder Φ that outputs d dimensional sentence representations. In practice, we adopt the MPNET Song et al. (2020) model from huggingface transformers Wolf et al. (2020) for balanced efficiency and embedding quality. We note any state-of-the-art general-purpose sentence encoding model can be used for PRAG, and provide additional analysis with T5 Raffel et al. (2020) sentence encoders in our experiment section.

2.4.2 Review Aggregation

Our base personalized retriever model is based on BERT architecture Devlin et al. (2019). Specifically, we treat each historical review in the training corpus as a token. When predicting the review a user would give to an item, the input to the model is a set of historical reviews written by the user and written for the item. Since these tokens do not represent consecutive words in a sentence as in the original BERT model, we do not use any position embedding. However, historical reviews play different roles when they are related to the item versus being related to the user, thus we choose to maintain learnable embedding $c \in \mathbb{R}^d$ to represent these two distinct scenarios, which are added to the original review embedding instead of position embeddings.

Finally, we maintain a learnable embedding $v_u \in \mathbb{R}^d$ and $v_i \in \mathbb{R}^d$ for each user and item to model user preferences, following common practice in recommender systems. For the input user and item, we look up the corresponding embedding and add these embeddings to each of the input review embedding. The final input review embedding to the model is thus:

$$v_u + v_i + c.$$

Such a final embedding is then passed through a stack of 2-layer transformers to process the cross-review relationship in the input.

2.4.3 Personalized Attention

Not every review in the input is important. Thus, it is crucial to select reviews that are helpful to the recommendation model. To achieve this, we develop a personalized attention module for weighing review representations from the base transformer model. Specifically, we obtain an attention score with respect to each piece of input reviews using a standard linear layer with a Rectified Linear Unit (ReLU) activation function using the concatenation of the review embedding, user embedding, and item embedding. Following previous works' insight that recommendation models benefit from un-smoothed attention scores' ability to discard irrelevant items Zhou et al. (2018), we conduct weighted pooling directly using the normalized attention scores.

2.4.4 Review Embedding Prediction

Finally, the weighted sum of review embeddings is passed through a multi-layer-perception (MLP) layer to produce the final latent query $Q_{u,i} \in \mathbb{R}^d$. During training, we minimize the L2 distance between the produced query and the embedding of the corresponding ground truth review, $\mathscr{L}_{\text{retrieve}}$, following previous success in using vectorized reviews to regularize recommender models Hada et al. (2021). Such embedding thus represents the predicted semantics of the input user and item, which could then be used to retrieve relevant reviews that are semantically similar to the predicted review from the existing review corpus.

2.4.5 Rating Prediction

To perform rating prediction, we combine HFT McAuley and Leskovec (2013b), a strong matrix-factorization based explainable recommendation model with a modified wide-and-deep Cheng et al. (2016) architecture. Specifically, the original HFT model makes predictions by modifying the following equation:

$$\text{rec}(u,i) = \gamma_u \times \gamma_i + \beta_u + \beta_i + \mu,$$

where μ is set to the global mean value of all ratings, and β_u, β_i are the learned bias. The model further ties either γ_u or γ_u to topic models, and learns product or user-specific topics by jointly minimizing the rating regression loss and the negative likelihood of the corresponding review.

We extend the HFT model by using latent query $Q_{u,i}$. Specifically, we adopt the estimated latent review as a new source of semantic information in place of topic models. For example, when associating item features to semantic information, we use a simple multi-layer-perception (MLP) to map $Q_{u,i}$ to γ_i (or γ_u). Meanwhile, previous work has shown that using a shallow linear (wide) layer for memorizing simple patterns could increase the performance of recommendation models Cheng et al. (2016). Following this intuition, we add an additional linear (wide) layer to the original model

$$\text{rec}(u,i) = \text{MLP}(Q) \times \gamma_i + \text{wide}(Q) + \beta_u + \beta_i + \mu.$$

We learn to predict rating using the standard squared loss $\mathscr{L}_{\text{rating}}$. The final joint loss for training the personalized retriever is then

$$\mathscr{L}_{\text{retrieve}} + \mathscr{L}_{\text{rating}}.$$

2.5 Explanation Generation as Keyword-guided Question Answering

To generate an explanation, we source information from retrieved reviews and treat the task of explanation generation as question answering. Specifically, the reader model should be able to answer why a higher (or lower) rating *adjustment score* γ_{ui} is being produced, as this is the only inter-user-item factor in rating prediction. To achieve this, we first train an embedding estimator that generates *informative* keywords using the mean embedding of latent query $Q_{u,i}$ as well as its corresponding set of retrieved reviews. After this, we probe a question-answering model trained to *factually* reflect content from the input with natural language prompts while

incorporating the generated keywords. The schema of the explanation generation pipeline is as shown in Figure 2.3.

To facilitate factual behavior for both the embedding estimator and the question-answering model, we design an aggregated task where the ground-truth is *guaranteed* to have a strong correlation to model input at training time. Concretely, we train a model to recover informative keywords in a review from the latent vector encoded from that specific review. Now we discuss each of the components in detail.

2.5.1 Retrieving Reviews

Given that the latent query is optimized to be similar to the ground truth review, a natural scheme for review retrieval is to rank existing reviews' semantic embedding with respect to $Q_{u,i}$ using similarity metrics. However, in practice, we found the model often produces overly-generic retrieval results, i.e., the reviews for multiple users tend to be similar, as observed in previous work Geng et al. (2022). To address this issue, we propose to retrieve existing reviews using characteristics that are *specific* to the user. In other words, we want to de-emphasize explanations that a model will produce for *every* user based on the item. We do so by examining the explanation for other users, estimating the latent explanation that are produced for all users, and marginalizing out such a universal explanation.

Concretely, we sample a batch of users at inference time, and obtain each sampled user's corresponding latent query $Q_{u_n,i}$. Then, we estimate the explanation that would be produced for every user using the mean embedding of the batch of predicted queries. Finally, we subtract such mean embedding from the original $Q_{u,i}$, effectively marginalizing out the user-agnostic aspect of the latent query. Since this changes the magnitude of $Q_{u,i}$, we rank existing reviews using cosine distance instead of L2 distance, and select top reviews as the retrieved content. Note that if $Q_{u,i}$ is tied to γ_i, we marginalize out the universal explanation for a batch of items instead, as we need to de-emphasize the explanation being produced for every item. We provide qualitative and quantitative analysis of such marginalization in Section 2.7.3.

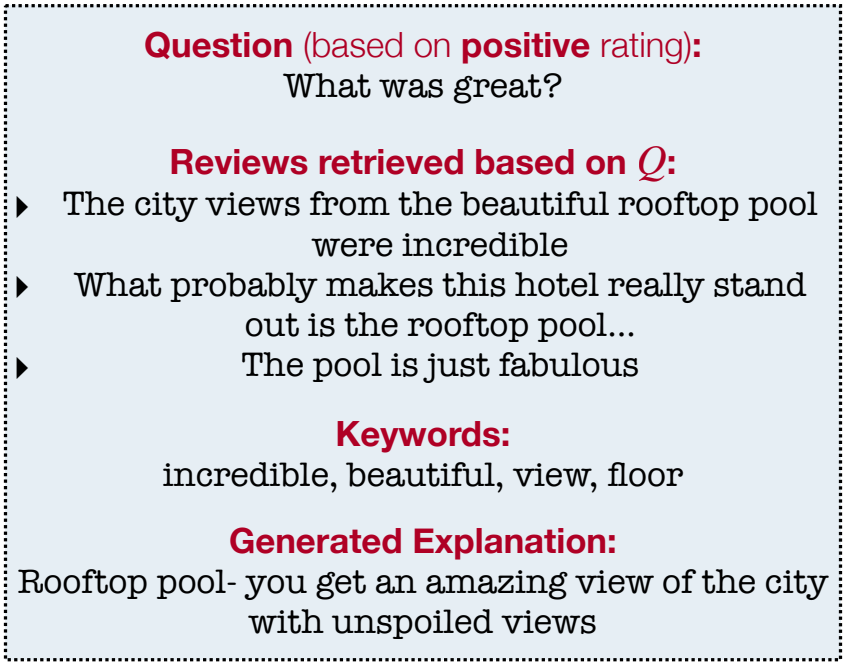

Figure 2.3. PRAG use retrieved historical reviews as source text to generate explanations.

2.5.2 Informative Keyword Generation

Retrieved reviews are by nature noisy, and multiple aspects in the reviews could simultaneously contribute to the model's rating decision. However, to generate realistic reviews, the generator must select a few aspects to focus on, as natural reviews typically focus on a few central topic words. To ensure the informativeness of the generated review, we design an embedding estimator that generates high tf-idf scored keywords given the personalized latent query. We then use the keywords in the explanation generation phase for improved informativeness.

To achieve this, we adopt a pre-trained GPT-2 model[2], and formulate the optimization objective as the task of embedding-to-keyword prediction. Specifically, given a latent embedding $\Phi(e)$ encoded from an original review e in the training set, we train the embedding estimator on recovering the 5-highest tf-idf scored words from the original review. In practice, we use the concatenated target keywords as desired model output and fine-tune the language model using the standard MLE objective. At inference time, we could then condition the embedding estimator

[2]https://huggingface.co/distilgpt2

on the corresponding latent query since the query is optimized to be in the semantic space of the sentence encoder Φ.

2.5.3 Explanation Generation

For explanation generation, we adopt an abstractive question-answering model that generates an answer given a question and relevant information[3], and exploit the in-context learning ability of large-scale, pre-trained language models Brown et al. (2020). Specifically, we ask the model to answer the question "`what was great?`" or "`what was not good?`", depending on the sign of the predicted rating adjustment score. Meanwhile, we note that the choice of prompt does not significantly influence the generation quality, and provide experiment with alternative prompts in the appendix.

To guide the model to cover the aspect that the user cares about, we want to encourage the model's output to contain at least one of the keywords produced by the embedding estimator. To achieve this, we first apply constrained decoding Hokamp and Liu (2017) using the original, unguided question answering model, forcing the output to include at least one of the input keywords. However, this could result in ungrammatical text. To address this issue, we manually rephrase 100 outputs for each dataset, which is easily achievable. This results in a high-quality dataset where each input text is paired with a set of keywords, with a ground-truth answer that contains at least one of the keywords. We then fine-tune the original question answering model on this dataset for a maximum of 10 epochs, following previous work on few-shot question answering Ram et al. (2021). The fine-tuned model could then be applied for efficient explanation generation *without* any decoding-time constraints.

[3]https://github.com/allenai/unifiedqa

2.6 Experiments

2.6.1 Datasets

We conduct our experiments on three publicly available dataset and splits from various domains Li et al. (2020b): Yelp[4] (restaurant reviews), TripAdvisor[5] (hotel) and Amazon Movies and TV He and McAuley (2016). Note that the data splits guarantee that products in the test set always appear in the training set.

Table 2.1. Automatic evaluation results on test sets.

	Entail			**D-1**			**D-2**			**ENTR**			**USR**			**MAUVE**		
Method	**Movie**	**Trip**	**Yelp**	**Movie**	**Trip**	**Yelp**	**Movie**	**Trip**	**Yelp**	**Movie**	**Trip**	**Yelp**	**Movie**	**Trip**	**Yelp**	**Movie**	**Trip**	**Yelp**
Att2Seq	25.6	12.2	35.9	39.9	34.6	43.1	75.9	75.4	78.1	9.56	8.11	8.44	41.7	21.0	39.9	3.0	1.4	3.9
NRT	36.1	10.0	31.4	44.0	32.4	41.0	77.8	72.8	76.6	7.5	7.5	8.3	36.1	46.3	44.4	3.0	3.0	4.2
PETER	29.0	17.5	44.5	27.7	26.8	29.5	58.6	60.7	60.4	10.5	10.1	10.7	60.7	57.2	58.2	3.7	2.3	2.2
PEPLER	17.9	11.0	16.0	23.2	23	25.5	51.5	52.2	52.5	11.1	10.0	11.0	52.6	41.7	49.1	1.1	0.4	0.4
OPTIMUS	25.1	22.8	11.5	31.9	32.8	33.2	77.3	77	79.3	10.3	8.5	10.7	**98.5**	92.1	**96.1**	3.5	3.3	4.5
SUM	49	29.5	30.8	22.1	18.7	20	67.1	61	63.7	11.2	10.4	11.5	95.3	**94.7**	94.8	5.8	4.7	5.4
PETER+	40.0	32.6	59.4	43.9	**42.6**	47.0	78.4	81.9	83.1	9.48	8.53	9.85	60.6	31.5	52.8	12.9	5.3	10.4
PRAG	**88.8**	**80.1**	**86.2**	**45.6**	39.9	**47.1**	**84.3**	**82.2**	**84.7**	**12.0**	**12.0**	**11.9**	71.8	76.5	70.4	**23.1**	**42.8**	**20.3**

2.6.2 Baselines

We compare four commonly used models in the literature: Att2Seq Dong et al. (2017), NRT Li et al. (2017), PETER Li et al. (2021b), and PEPLER Li et al. (2022). Among the models, Att2Seq and NRT are LSTM-based models, while PETER is a transformer-based model. We additionally incorporate a variant of PETER that conditions on a topic word from ground-truth review at inference time, denoted by PETER+. Finally, PEPLER adopts a pre-trained GPT-2 model with prompt tuning for explanation generation. We provide more details for each of the models in the appendix.

Further, we show that when augmented with our personalized retriever, opinion aggregation approaches such as summarization models could seamlessly integrate into the retriever-reader

[4]https://www.yelp.com/dataset/challenge
[5]https://www.tripadvisor.com

framework. We demonstrate this by proposing two *novel* baselines, where summarization models are used to aggregate retrieved reviews.

PRAG-Optimus (OPTIMUS)

Adopts a state-of-the-art pre-trained VAE-based language model based on BERT and GPT Li et al. (2020a). Recent studies find VAE-based language models could be used for unsupervised opinion summarization Iso et al. (2021). However, Iso et al.'s approach requires searching over large amounts of potential sequences at inference time, particularly when there are multiple inputs. Thus, we use the original Optimus model without inference-time searching as our baseline. Specifically, we fine-tune a pre-trained Optimus model on each dataset for 1 epoch on language modeling, per recommendation by the original authors, and condition the GPT-based generator on the mean embedding of retrieved reviews at inference time.

PRAG-SUM (SUM)

Following previous success in training summarizers by learning to recover the target review from a set of distantly retrieved similar reviews Amplayo and Lapata (2020), we train our summarizer in a leave-one-out fashion over sets of similar reviews. Specifically, for each review in the training set, we retrieve a set of most similar reviews using cosine distance in sentence encoder Φ's semantic space. These retrieved reviews then serve as the input for the model. Then, the summarization module is trained to recover the original review that is used for retrieval from the retrieved reviews. For training, we use the same pre-trained T5 model as PRAG as our base model for a fair comparison and train the model using maximum likelihood estimation (MLE) objective. We provide more details for data processing and training of SUM in the appendix.

2.6.3 Automatic Evaluation

For automatic evaluation, we generate 10,000 samples from each baseline model, and measure the performance in terms of factuality, informativeness, and generation quality. We provide additional recommendation performance analysis on all eligible models on the whole

test set.

Factuality.

Factuality constraints that the models' generated explanations are factually correct. Intuitively, the statement in a generated explanation is correct if it could be supported by any existing reviews in the training data. Otherwise, the model is exhibiting **hallucinations** Maynez et al. (2020). As reported by the same work, entailment models could better measure the factual consistency between reference text and generated content. We follow their findings and evaluate the entailment relationship between generated reviews with reviews of the same product in the training set. Specifically, for each generated explanation, we check whether the explanation entails *any* reviews from the training set for the same product using a pre-trained entailment model [6]. We note that a generated explanation is factual if it entails any piece of existing reviews for the product, and report the entailment ratio. Specifically, we report the percentage of entailment as Entail.

Informativeness.

We measure the informativeness of generated explanations using token, sentence, and corpus level evaluations. Concretely, We evaluate the models using Distinct-1 and Distinct-2 (D-1, D-2) scores Li et al. (2016), Unique Sentence Ratio (USR) as proposed by Li et al. (2020b) and ENTR Jhamtani et al. (2018) following previous work on diversifying generated content Majumder et al. (2021a).

Generation quality.

To measure the generated explanation quality, we opt to measure how human-like the generated explanations are. Specifically, we adopt MAUVE Pillutla et al. (2021), a distribution-based evaluation metric that measures how close the generated contents are to the ground truth corpus.

[6] https://huggingface.co/prajjwal1/roberta-large-mnli

Recommendation performance.

We evaluate the recommendation performance using Rooted Mean Squared Error (RMSE) score following previous works done using the same dataset.

2.7 Results

2.7.1 Quantitative Evaluation

We report PRAG's performance compared to baseline models as in Table 2.1. PRAG consistently outperforms the baseline models in terms of both diversity and informativeness, and can generate high-quality sentences compared to human-written reviews.

Retrieval component improves generation quality.

As shown in Table 2.1, conditioning on additional information improves both model diversity and factuality. Specifically, OPTIMUS, SUM, and PRAG consistently outperform other models that only maintain vector representation for each user and item. The generated content has both higher sentence-wise diversity (from D-1 and D-2 scores) and higher corpus-level diversity (from ENTR and USR scores), as well as being closer to human written reviews (from MAUVE scores). This shows that the personalized retrieval component could improve generation quality.

Training data affects generator factuality.

While retrieved historical reviews encourage more factual output in general, this does not guarantee strong factuality. In particular, the T5-based summarizer and OPTIMUS model have limited improvement in terms of factuality compared to other baseline models. This highlights the major cause of hallucination is directly training generators on noisy data. Specifically, Maynez et al. reported that ground-truth sequences in the training data that contain hallucinated content would trigger the model to be less factual at inference time in summarization. In parallel, Longpre et al. reported that noisy retrieved content would cause the generator to hallucinate for question answering.

Table 2.2. RMSE scores for recommendation performance.

	Movie	**TripAdivor**	**Yelp**
NRT	0.79	0.95	1.01
PETER	0.80	0.95	1.01
PEPLER	1.71	1.25	1.69
PRAG_u	0.80	0.95	1.01
PRAG_{T5}	0.80	0.96	1.00
PRAG	0.79	0.95	1.01

We note that a similar case applies to review generation as well. In particular, there is no guarantee that the ground truth review will entail *any* of other reviews for the same product in the training set. While this behavior is natural for human Maynez et al. (2020), the model ended up learning to hallucinate during optimization as a result. By leveraging a reader model trained to faithfully present content in the input text, PRAG sidesteps this issue, and thus has the best factuality performance across models.

2.7.2 Recommendation Performance

To verify PRAG's ability for recommendation, we report RMSE scores on the three datasets as shown in Table 2.2. Since SUM and OPTIMUS are based on the same retriever model as PRAG, we report PRAG's performance only in this section. Further, Att2Seq cannot produce rating estimations, and thus we omit the model in the table. We validate that the performance of baseline models is consistent with previous works on the same dataset Geng et al. (2022). As shown in Table 2.2, PRAG achieved state-of-the-art rating estimation performance.

2.7.3 Human Evaluation

We perform human evaluation using 150 test samples with several of the strongest baselines. Specifically, we compare PRAG's performance against NRT, PETER and SUM. We omit Att2Seq and PEPLER since the models are not as performant as other models in recommendation performance, and leave PETER+ out since ground-truth aspects are not always available. Finally, we pick the summarizer as a stronger retriever-augmented baseline (compared

Table 2.3. Human evaluation results of PRAG versus baseline models.

	Fluency			Informativeness		
PRAG vs.	**Movie**	**Trip**	**Yelp**	**Movie**	**Trip**	**Yelp**
NRT	**37**	**16**	**13**	**28**	**1**	**3**
PETER	**20**	**26**	**15**	-2	**12**	**27**
SUM	**8**	**5**	**3**	0	-6	**5**

to OPTIMUS) based on automatic evaluation results. We compare the generated reviews in terms of (1) **fluency** compared to other models (2) **informativeness** to the user. We report the results in Table 2.3. Similar to automatic evaluation results, the retriever could reliably boost generation performance: PRAG is consistently ranked as the most fluent model by human evaluators, while SUM has almost on-par performance due to having access to the personalized retriever. Meanwhile, PRAG and SUM almost invariably win in terms of informativeness compared to previous works.

Analysis and Discussion

Analysis of marginalization and retrieval reliability.

To validate whether marginalizing the predicted embedding before performing retrieval truly uncovers user or item characteristics instead of resulting in random vectors, we perform a simple test by comparing the retrieval result between a pair of mirroring retrievers. Specifically, we hypothesize that given a user and an item, a retriever that ties $Q_{u,i}$ to γ_u and a retriever that ties $Q_{u,i}$ to γ_i should achieve agreement after marginalization. That is, the two retrievers should be able to retrieve similar reviews. To verify this, we report the average exact-match between two retrievers using agreement at 5. As shown in Table 2.5, the retrievers consistently achieved significantly higher agreement compared to a random baseline across three datasets. Further, we report examples of retrieved reviews with and without marginalization. As shown in Table 2.4, after marginalization, there is a clear trend that reviews related to aesthetic aspects are being retrieved, as opposed to a set of generic reviews without marginalization.

Table 2.4. Retrieved reviews (cropped, from top-5 results).

Retrieved Reviews
the decor of the hotel is greatly refined. ... the building and decoration are very nice and very tastefully decorated. a four seasons in every respect it is an architecturally interesting and esthetically pleasing property in a great location. the rooms are stunning.

Table 2.5. Average agreement-at-5 of mirroring retrievers.

	Movie	**TripAdivor**	**Yelp**
Retriever	2.87	2.44	2.50
Random	0.54	0.61	0.27

Analysis of sentence embedding model.

We hypothesize that any strong sentence embedding model could be used for PRAG. To validate such a hypothesis, we train PRAG model on all three datasets using a pre-trained T5-based sentence encoder. We report the performance as in Table 2.2. As shown, there is no significant performance variation across different types of sentence embeddings.

Effect of tying user or item Factors to latent query.

Similar to the HFT McAuley and Leskovec (2013b) model, the rating estimation component in PRAG could either tie the user or item factor to $Q_{u,i}$. We conduct experiments using both types of architectures, and also report our results in Table 2.2. Similar to findings reported for the HFT model McAuley and Leskovec (2013b), the performance is generally dataset-dependent, and the design could be viewed as a hyper-parameter.

2.8 Summary and Outlook

In this work, we propose PRAG, a retriever-reader model that can generate factual and diverse explanations for recommendation. Experiments on three real-world datasets show

PRAG can generate both factually grounded and informative explanations. We also investigate the cause of hallucinated content in review generation, and demonstrate the benefit of training text generation models on hallucination-free tasks and datasets. Meanwhile, although we adopted a question-answering model for explanation generation, PRAG's retrieval component could provide support for personalizing a wider range of knowledge-based tasks, such as personalized conversational recommendation, summarization, and product description generation.

2.9 Acknowledgement

This chapter, in full, is a reprint of the material as it appears in Proceedings of the AAAI Conference on Artificial Intelligence, 2023 ("Factual and Informative Review Generation for Explainable Recommendation", Zhouhang Xie, Sameer Singh, Julian McAuley, and Bodhisattwa Prasad Majumder). The thesis author was the primary investigator and author of this paper.

Bibliography

Reinald Kim Amplayo and Mirella Lapata. 2020. Unsupervised opinion summarization with noising and denoising. In *ACL*, pages 1934–1945, Online. Association for Computational Linguistics.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020. Unsupervised opinion summarization as copycat-review generation. In *ACL*, pages 5151–5169, Online. Association for Computational Linguistics.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2021. Learning opinion summarizers by selecting informative reviews. In *EMNLP*, pages 9424–9442, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Samuel Carton, Surya Kanoria, and Chenhao Tan. 2022. What to learn, and how: Toward effective learning from rationales. In *Findings of ACL*.

Chong Chen, Min Zhang, Yiqun Liu, and Shaoping Ma. 2018. Neural attentional rating regression with review-level explanations. *WWW*.

Huimin Chen, Yankai Lin, Fanchao Qi, Jinyi Hu, Peng Li, Jie Zhou, and Maosong Sun. 2021. Aspect-level sentiment-controllable review generation with mutual learning framework. In *AAAI*.

Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, Rohan Anil, Zakaria Haque, Lichan Hong, Vihan Jain, Xiaobing Liu, and Hemal Shah. 2016. Wide & deep learning for recommender systems. In *1st Workshop on Deep Learning for Recommender Systems*, DLRS 2016,

page 7–10, New York, NY, USA. Association for Computing Machinery.

Bhavana Dalvi, Peter Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. 2021. Explaining answers with entailment trees. In *EMNLP*, pages 7358–7370, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Li Dong, Shaohan Huang, Furu Wei, Mirella Lapata, Ming Zhou, and Ke Xu. 2017. Learning to generate product reviews from attributes. In *EACL*.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of EMNLP*.

Shijie Geng, Zuohui Fu, Yingqiang Ge, Lei Li, Gerard de Melo, and Yongfeng Zhang. 2022. Improving personalized explanation generation through visualization. In *ACL*, pages 244–255, Dublin, Ireland. Association for Computational Linguistics.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: Retrieval-augmented language model pre-training. *arXiv preprint arXiv:2002.08909*.

Deepesh V. Hada, M Vijaikumar, and Shirish K. Shevade. 2021. Rexplug: Explainable recommendation using plug-and-play language model. *SIGIR*.

Braden Hancock, Paroma Varma, Stephanie Wang, Martin Bringmann, Percy Liang, and Christopher Ré. 2018. Training classifiers with natural language explanations. In *ACL*, pages 1884–1895, Melbourne, Australia. Association for Computational Linguistics.

Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. *WWW*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9:1735–1780.

Chris Hokamp and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. In *ACL*, pages 1535–1546, Vancouver, Canada. Association for Computational Linguistics.

Hayate Iso, Xiaolan Wang, Yoshihiko Suhara, Stefanos Angelidis, and Wang-Chiew Tan. 2021. Convex Aggregation for Opinion Summarization. In *Findings of EMNLP*, pages 3885–3903, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *ACL*.

Harsh Jhamtani, Varun Gangal, Eduard Hovy, Graham Neubig, and Taylor Berg-Kirkpatrick. 2018. Learning to generate move-by-move commentary for chess games from large-scale social forum data. In *ACL*, pages 1661–1671, Melbourne, Australia. Association for Computational Linguistics.

Weina Jin, Xiaoxiao Li, and Ghassan Hamarneh. 2022. Evaluating explainable AI on a multi-modal medical imaging task: Can existing algorithms fulfill clinical requirements? In *AAAI*. AAAI Press.

Himabindu Lakkaraju, Dylan Slack, Yuxin Chen, Chenhao Tan, and Sameer Singh. 2022. Rethinking explainability as a dialogue: A practitioner's perspective. *arXiv preprint arXiv:2202.01875*.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *NeurIPS*.

Chunyuan Li, Xiang Gao, Yuan Li, Baolin Peng, Xiujun Li, Yizhe Zhang, and Jianfeng Gao. 2020a. Optimus: Organizing sentences via pre-trained modeling of a latent space. In *EMNLP*, pages 4678–4699, Online. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *NAACL-HLT*, San Diego, California.

Junyi Li, Wayne Xin Zhao, Zhicheng Wei, Nicholas Jing Yuan, and Ji-Rong Wen. 2021a. Knowledge-based review generation by coherence enhanced text planning. *SIGIR*.

Lei Li, Yongfeng Zhang, and Li Chen. 2020b. Generate neural template explanations for recommendation. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 755–764, New York, NY, USA. Association for Computing Machinery.

Lei Li, Yongfeng Zhang, and Li Chen. 2021b. Personalized transformer for explainable recommendation. In *ACL-IJCNLP*.

Lei Li, Yongfeng Zhang, and Li Chen. 2022. Personalized prompt learning for explainable recommendation. *arXiv preprint arXiv:2202.07371*.

Pan Li and Alexander Tuzhilin. 2019. Towards controllable and personalized review generation. In *EMNLP-IJCNLP*, Hong Kong, China.

Piji Li, Zihao Wang, Zhaochun Ren, Lidong Bing, and Wai Lam. 2017. Neural rating regression with abstractive tips generation for recommendation. *SIGIR*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*.

Siyang Liu, Sahand Sabour, Yinhe Zheng, Pei Ke, Xiaoyan Zhu, and Minlie Huang. 2022. Rethinking and refining the distinct metric. In *ACL*, pages 762–770. Association for Computational Linguistics.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *EMNLP*.

Bodhisattwa Prasad Majumder, Taylor Berg-Kirkpatrick, Julian McAuley, and Harsh Jhamtani. 2021a. Unsupervised enrichment of persona-grounded dialog with background stories. In *ACL-IJCNLP*, pages 585–592, Online. Association for Computational Linguistics.

Bodhisattwa Prasad Majumder, Oana-Maria Camburu, Thomas Lukasiewicz, and Julian J. McAuley. 2021b. Rationale-inspired natural language explanations with commonsense. *ICML*, abs/2106.13876.

Ana Marasović, Chandra Bhagavatula, Jae sung Park, Ronan Le Bras, Noah A. Smith, and Yejin Choi. 2020. Natural language rationales with full-stack visual reasoning: From pixels to semantic frames to commonsense graphs. In *Findings of EMNLP*, pages 2810–2829, Online. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *ACL*.

Julian McAuley and Jure Leskovec. 2013a. Hidden factors and hidden topics: understanding rating dimensions with review text. *RecSys*.

Julian McAuley and Jure Leskovec. 2013b. Hidden factors and hidden topics: Understanding rating dimensions with review text. In *ACM RecSys*, RecSys '13, page 165–172, New York, NY, USA. Association for Computing Machinery.

Julian McAuley, Jure Leskovec, and Dan Jurafsky. 2012. Learning attitudes and attributes from multi-aspect reviews. *ICDM*, pages 1020–1025.

Jianmo Ni, Zachary C. Lipton, Sharad Vikram, and Julian McAuley. 2017. Estimating reactions and recommending products with generative models of reviews. In *IJCNLP*, Taipei, Taiwan.

Jianmo Ni and Julian McAuley. 2018. Personalized review generation by expanding phrases and attending on aspect-aware representations. In *ACL*.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *NAACL-HLT*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers. In *NeurIPS*.

Reinald Adrian Pugoy and Hung-Yu Kao. 2020. Bert-based neural collaborative filtering and fixed-length contiguous tokens explanation. In *AACL*.

Reinald Adrian Pugoy and Hung-Yu Kao. 2021. Unsupervised extractive summarization-based representations for accurate and explainable collaborative filtering. In *ACL-IJCNLP*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Ori Ram, Yuval Kirstain, Jonathan Berant, Amir Globerson, and Omer Levy. 2021. Few-shot question answering by pretraining span selection. In *ACL-IJCNLP*, pages 3066–3079, Online. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically equivalent adversarial rules for debugging NLP models. In *ACL*, Melbourne, Australia.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist. *arXiv preprint arXiv:2005.04118*.

Noveen Sachdeva and Julian J. McAuley. 2020. How useful are reviews for recommendation? A critical review and potential improvements. In *SIGIR*. ACM.

Dylan Slack, Satyapriya Krishna, Himabindu Lakkaraju, and Sameer Singh. 2022. Talktomodel: Understanding machine learning models with open ended dialogues. *ArXiv*, abs/2207.04154.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. In *NeurIPS*.

Peijie Sun, Le Wu, Kun Zhang, Yanjie Fu, Richang Hong, and Meng Wang. 2020. Dual learning for explainable recommendation: Towards unifying user preference prediction and review generation. *WWW*.

Oyvind Tafjord and Peter Clark. 2021. General-purpose question-answering with macaw. *CoRR*, abs/2109.02593.

Chenhao Tan. 2022. On the diversity and limits of human explanations. In *NAACL*.

Yi Tay, Anh Tuan Luu, and Siu Cheung Hui. 2018. Multi-pointer co-attention networks for recommendation. *CIKM*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *NeurIPS*, 30.

Sarah Wiegreffe, Ana Marasović, and Noah A. Smith. 2021. Measuring association between labels and free-text rationales. In *EMNLP*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *EMNLP: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S. Weld. 2021. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. In *ACL*.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *ICLR*.

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 *SIGIR*, pages 83–92. ACM.

Lei Zheng, Vahid Noroozi, and Philip S. Yu. 2017. Joint deep modeling of users and items using reviews for recommendation. *WSDM*.

Guorui Zhou, Xiaoqiang Zhu, Chenru Song, Ying Fan, Han Zhu, Xiao Ma, Yanghui Yan, Junqi Jin, Han Li, and Kun Gai. 2018. Deep interest network for click-through rate prediction. In *SIGKDD*, KDD '18, page 1059–1068, New York, NY, USA. Association for Computing Machinery.

Wangchunshu Zhou, Jinyi Hu, Hanlin Zhang, Xiaodan Liang, Maosong Sun, Chenyan Xiong, and Jian Tang. 2020. Towards interpretable natural language understanding with explanations as latent variables. In *NeurIPS*.

Wangchunshu Zhou, Qifei Li, and Chenle Li. 2021. Learning from perturbations: Diverse and informative dialogue generation with inverse adversarial training. In *ACL-IJCNLP*, Online.