Appendix for Factual and Informative Review Generation for Explainable Recommendation

A Dataset Statistics

	Movie	TripAdivor	Yelp
#user	7506	9765	27,147
#item	7360	6280	20,266
#interaction	441,783	320,023	1,293,247

Table 1: Overview of the datasets used in our experiments.

We provide an overview of dataset statistics in Table 1. Note that this is the exact dataset used in previous works in the literature, such as in Li, Zhang, and Chen and in Geng et al..

B Implementation Details

We implement all of the experiments in PyTorch[1]. For PRAG, we implement the model using Hugging Face Transformers library (Wolf et al. 2020). Unless otherwise specified, the hyper-parameter setting of models follow the recommendation by Hugging Face library's default model trainer. For hyper-parameter tuning, we always select model with fewer parameters in case of tied performance.

For PRAG's personalized retriever, we experiment with 2 to 4 layers of transformer blocks on TripAdvisor dataset, and use the best performing hyper-parameter across the remaining datasets. Similarly, we tune the SVD-based rating prediction head using 5, 10, and 15 latent factors on TripAdvisor dataset, and use it across other datasets.

For PRAG's embedding estimator, we experiment with keeping 2, 4, and 6 (all) layers of the full pre-trained GPT model, and select the best performing model on TripAdvisor dataest, and use it across the remaining datasets.

We adopt publicaly available implemantation of Att2Seq, NRT, PETER, and PEPLER[2], and follow their recommended hyper-parameter settings. For PRAG-Optimus, we adopt implementation by Iso et al., and follow their hyper-parameter setting. Specifically, we set latent dimension to 512, with 2 free bits for VAE's training. We fine-tune the base Optimus model for 1 epoch, following recommendation by its original authors (Li et al. 2020). We follow Iso et al.'s implementation for other hyper-parameters such as learning rate and batch size.

For the T5-based summarizer (SUM), we run our experiment using Huggingface Transformer, and follow their default training hyper-parameters. We adopt early stopping and stop model training after there is no improvement after 3 epochs. When constructing parallel data for summarization, we retrieve for each target review a set of reviews that has the smallest cosine distance in MPNet encoder's sentence embedding space. Meanwhile, we set the minimum cosine distance threshold to 0.65, and discard reviews that has greater cosine distance to the target review. However, due to reviews in Amazon Reviews dataset being extremely diverse, we losen such constraint to 0.55 for this dataset specifically to ensure there is enough training instance for tuning a summarization model.

We note that the hyper-parameter search is not comprehensive for PRAG and SUM, and there could be better settings that yields superior performance than we reported. We will release our implementation of PRAG, Optimus and SUM upon acceptance.

C Computational Resource

Experiments are conducted on Titan RTX (24GB) and Titan X (12GB) GPUs.

D User Study

We use Amazon Mechanical Turk for user study[3]. We randomly sample 150 instance from each selected model's generated explanation, and present output of PRAG and an opponent model in pairs to the user at survey time. This resulted in 1350 surveys across all model and dataset combinations. For each survey, we ask the human evaluator to read a group of 8 reviews generated by PRAG and the opponent model. Then, we ask the evaluator to select the better model given the question "Considering English language fluency only, compare R1 and R2" for **fluency**. For **informativeness**, we ask the evaluator to select the better model given the question "Which response do you think is more informative? (e.g. contains less repetetive content and gives you more

[1] https://pytorch.org/

[2] https://github.com/lileipisces/NLG4RS

[3] https://www.mturk.com/

	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**
PRAG$_{alt}$	**89.3**	79.6	85.6	44.8	39.3	46.6	84.2	81.8	84.2	**12.0**	**12.0**	**12.0**	74.4	79.9	72.5	**24.4**	**44.0**	20.0

Table 2: **Automatic evaluation results** with alternative prompt.

`insight)`”. To ensure a wider range of human evaluators, we publish each survey in stand-alone manner, so that we could obtain feedbacks from a wider audience.

E Additional Qualitative Results of Personalized Retriever

We provide additional un-cropped examples of the retriever's retrieval result with and without marginalization. As shown in Table 4, 5, and 6, the retriever with marginalization could retrieve reviews with more consistent topics across the three datasets. Note that Table 4 is the full version for the presented result in the main content.

F Measuring Informativeness via Token Recall

As a sanity check, we additionally provide analysis for informativeness via recall of tokens from the product's previous reviews. As shown in Table 3, models augmented by our personalized retriever could consistently achieve higher token recall.

G Experiment with Alternative Prompt

To ensure PRAG's explanation generation performance is not sensitive to prompt choice, we report performance with alternative prompt. In this case, the model was asked “`what was good?`” or “`what was not so great?`”, depending on the sign of the predicted rating adjustment score. We report the alternative model PRAG$_{alt}$'s performance in Table 2. As shown, there is no significant performance variation given the alternative prompt. We note that there exists gradient-based methods for automatically construction prompts, such as in Shin et al., which could potentially improve PRAG's performance in future works.

H Additional Details for Baseline Models

Att2Seq (Dong et al. 2017) Uses LSTM as a decoder for review generation; we adopt the implementation of Li, Zhang, and Chen, removing the attention layer that harms the readability of generated content.

NRT (Li et al. 2017) Was designed to generate a tip and a rating estimation based on user and item ids. We directly use the ‘tip’ generation module to generate reviews for fair comparison with other models.

	Movie	TripAdivor	Yelp
Att2Seq	0.87	1.27	0.66
NRT	0.83	1.38	0.73
PETER	0.86	1.32	0.68
PEPLER	0.89	1.36	0.44
Optimus	**0.98**	1.28	0.71
SUM	1.02	**1.70**	**0.85**
PETER+	0.74	1.00	0.58
PRAG	0.95	**1.70**	0.74

Table 3: **Average token-level recall (percentage)** of the models. Recall here measures how many tokens from previous reviews of a product is mentioned in the generated explanation.

PEPLER (Li, Zhang, and Chen 2022) Is a prompt-finetuned GPT-2 model that could also conduct rating estimation by dot product between the learned user and item latent representations.

PETER (Li, Zhang, and Chen 2021) Is a transformer-based model that is modified to conduct rating regression and review generation. The model could also take in an arbitrary number of categorical features. In our work, we denote the base model as PETER, and an additional conditional model that has access to the ground truth aspect as PETER+.

I Additional Details for UnifiedQA

UnifedQA (Khashabi et al. 2020) is a pre-trained question-answering model that is optimized to produce answers given a wide variety of quesion-context formats and has strong out-of-domain generalization ability. We adopt it as our question-answering component for its good performance on unseen datasets.

J Potential Alternatives to Question-answering Component

We note that in additional to using UnifiedQA, it is possible to adopt other pre-trained models for aggregating retrieved reviews. To this end, we have also attempted using GPT-2 in place of UnifiedQA. Specifically, we concatenate the retrieved reviews as context, and append a single sentence

“Question: What was great? Answer:” to the end of the given context to prompt GPT-2 for generating the corresponding answer. However, we found that GPT-2 cannot generate coherent answers, and often produce nonsensical tokens such as “xa0”.

K Qualitative Analysis

We provide results generated from PRAG (in comparison to baseline models) from different domains in Table 7, Table 8, and Table 9. As shown in Table 7, PRAG is able to mention specific aspects of a movie, such as plots (the car-chasing scene), whereas the summarizer favors generic comments such as the movie is “good” and “worth watching”. Similarly, as shown in Table 9, PRAG could produce specific dishes such as “house salad” when commenting on a restaurant. We hypothesize this is the reason why PRAG is rated more informative in many cases.

Further, using a pre-trained component makes PRAG less prone to grammar errors. For example, NRT generated an extra letter “a” at the end of example 6 in Table 8, possibly due to RNNs are not as powerful for modeling long-term dependency as pre-trained transformers. Such an advantage could explain why PRAG is rated more fluent compared to the baseline models.

Retrieved Reviews
1. the decor of the hotel is greatly refined. ...
2. the building and decoration are very nice and very tastefully decorated.
3. a four seasons in every respect it is an architecturally interesting and esthetically pleasing property in a great location.
4. the rooms are stunning.
5. the hotel concierge also made excellent restaurant recommendations.
6. one night we dined in the restaurant and the staff and management were exceedingly attentive to us.
7. with some of the best hotels and restaurants in the world. i stayed in this beautiful hotel during my birthday week.
8. on arrival my wife and i did a quick walk around the property and were enticed by the seaside grill
1.) - excellent spa and small but nice pool on the top floor- good breakfast buffet- fast wifi.
2. the hotels location is outstanding ca n’t be better. the room service is good.
3. the rooms are stunning.
4. indeed everything about this hotel is quality.
5. this is a wonderful hotel (i visited several other hotels and they
6. all looked bad compared with the four seasons) in a great location in an exciting city.
7. and the lobby tea bar is lovely with wonderful food and great service. the service and concierge was great and helped us fall in love with budapest. you will not be disappointed in this hotel.
8. the hotel concierge also made excellent restaurant recommendations

Table 4: Un-cropped results of retrieved reviews, with (up) and without (down) marginalization. (TripAdvisor)

References

Dong, L.; Huang, S.; Wei, F.; Lapata, M.; Zhou, M.; and Xu, K. 2017. Learning to Generate Product Reviews from Attributes. In *EACL*, 623–632. Valencia, Spain: Association for Computational Linguistics.

Geng, S.; Fu, Z.; Ge, Y.; Li, L.; de Melo, G.; and Zhang, Y. 2022. Improving Personalized Explanation Generation through Visualization. In *ACL*, 244–255. Dublin, Ireland: Association for Computational Linguistics.

Iso, H.; Wang, X.; Suhara, Y.; Angelidis, S.; and Tan, W.-C. 2021. Convex Aggregation for Opinion Summarization. In *Findings of EMNLP*, 3885–3903. Punta Cana, Dominican Republic: Association for Computational Linguistics.

Khashabi, D.; Min, S.; Khot, T.; Sabharwal, A.; Tafjord, O.; Clark, P.; and Hajishirzi, H. 2020. UNIFIEDQA: Crossing Format Boundaries with a Single QA System. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 1896–1907. Online: Association for Computational Linguistics.

Li, C.; Gao, X.; Li, Y.; Peng, B.; Li, X.; Zhang, Y.; and Gao, J. 2020. Optimus: Organizing Sentences via Pre-trained Modeling of a Latent Space. In *EMNLP*, 4678–4699. Online: Association for Computational Linguistics.

Li, L.; Zhang, Y.; and Chen, L. 2021. Personalized Transformer for Explainable Recommendation. In *ACL-IJCNLP*, 4947–4957. Online: Association for Computational Linguistics.

Li, L.; Zhang, Y.; and Chen, L. 2022. Personalized Prompt Learning for Explainable Recommendation. *arXiv preprint arXiv:2202.07371*.

Li, P.; Wang, Z.; Ren, Z.; Bing, L.; and Lam, W. 2017. Neural Rating Regression with Abstractive Tips Generation for Recommendation. *SIGIR*.

Shin, T.; Razeghi, Y.; Logan IV, R. L.; Wallace, E.; and Singh, S. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4222–4235. Online: Association for Computational Linguistics.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; Davison, J.; Shleifer, S.; von Platen, P.; Ma, C.; Jernite, Y.; Plu, J.; Xu, C.; Le Scao, T.; Gugger, S.; Drame, M.; Lhoest, Q.; and Rush, A. 2020. Transformers: State-of-the-Art Natural Language Processing. In *EMNLP: System Demonstrations*, 38–45. Online: Association for Computational Linguistics.

Retrieved Reviews
1. crowded seating and always busy.
2. king 's is completely packed but without a lineup- we got seated quickly.
3. the place was rammed and we even had a random person eat at our table because there was n't enough room.
4. if you are alone (or less than 8 people) then they will ask you to sit next to complete strangers but what annoys me is that it was n't packed and i still have to sit next to strangers even though several tables are still open (and not reserved).
5. service was average ..
6. i visited on friday evening and the place was packed. horrible service ..
7. but super fast and the restaurant is clean.
8. and combined with the waitresses coming in and out and the mob of people waiting to be seated.
9. sometime members of certain organized groups get in loud arguments , but just mind your own business and chalk it up to ambiance i used to come here with my friend shrimp boy in the 90 's .
1. the food is quite good for downtown standards.
2. it 's a great resto for a quick bite of good. food was amazing .. they also have other nice dishes.
3. they provide copious amounts side dishes.
4. king noodle food is top notch and prices were reasonable (total was under 13) - that 's saying a lot as i have been spoiled coming from markham scarborough.
5. i also love getting their freshly made shrimp rice rolls and wontons.
6. the vegetables and various fish meat balls were very fresh and i absolutely loved their mini buffet bar and bottomless drinks station.
7. the bbq pork and duck are sweet and well roasted.
8. the wait times here for food might test your patience but the dish that ultimately arrives at your table is always worth it

Table 5: Un-cropped results of retrieved reviews, with (up) and without (down) marginalization. (Yelp)

Retrieved Reviews
1. with a terrific performance as sport.
2. the you should check out this excellent package that gives the show its due.
3. and his superb performance is simply devastating.
4. i watched this show since it was first aired and it is such a great balance of drama and humor and a great story line.
5. my girlfriend got the first series on dvd and we 'd been watching it in mini-mararathons now and then.
6. i think that i started watching it when there were reruns on abc family since i started on the 1st season.
7. witty observations and fascinating characters are apparent on both shows.
8. and his performance is typically excellent. and great ending) -ps i lo. the mostly dark cinematography of the movie is extremely fitting and effective
1. but the way things end up coming together is uncanny–it 's the kind of movie you could watch 5 times and still notice a detail that previously went over your head.
2. the mostly dark cinematography of the movie is extremely fitting and effective. and great ending) -ps i lo. both masterpieces in their own right–william foster 's wildly erratic and tempermental behavior strongly brings to mind the former 's robert dupea (jack nicholson).
3. despite the fine performances and excellent production values.
4. the you should check out this excellent package that gives the show its due.
5. pumpkin seems torn between wanting to be a touching unconventional love story and a spoof on what they call 'teen sex comedies'.
6. the show pulls you in to the love stories only to let you down as the heroines turn on their lovers and of course blame their wealthy backgrounds.
7. think 1940s screwball comedy mixed with a planned community drama.
8. ellen burstyn really was alice hyatt–her performance is brilliant and flawlessly convincing

Table 6: Un-cropped results of retrieved reviews, with (up) and without (down) marginalization. (Movies and TV)

Retrieved Reviews
1. what was not good ? the good guys get into wild cars chases which causes damage to public property.
2. what was not good ? the ending was a little too predictable
3. what was great ? everyone playing their parts in this film are superb
4. what was great ? great character development.
5. what was great ? twists and turns
6. what was great ? this movie is extremely entertaining and extremely inventive
7. what was great ? the cast was really great with the acting
8. what was not good ? boring and slow.
1. if you are looking for a good slasher flick
2. a movie about a girl who is stranded in a gloom
3. this movie is very good and worth watching
4. the movie is very well acted and has a great cast
5. the movie is very well acted and has a very good plot
6. this is a great movie and i recommend it to anyone
7. this is a very good movie
8. slasher flick i'm not sure what to expect from this sequel

Table 7: Examples of generated text on movie reviews corpus. PRAG wins in both informativeness and fluency on this sample. (Up: PRAG; Down: Summarizer)

Retrieved Reviews
1. what was not good ? the pool is almost too cold to use. average service
2. what was great ? laundry room on premises
3. what was not good ? the pool is almost too cold to use. average service. captive market means big prices. before entering bathroom
4. what was great ? laundry room on premises
5. what was great ? shuttle bus is free and runs every 15 or 20 minutes
6. what was great ? laundry room on premises
7. what was great ? laundry room on premises
8. what was not good ? it took a long time to get served a drink
1. the pool area is a bit small but the pool area is very nice
2. the hotel is very good value for money
3. we had a great stay at the hotel in the past
4. the pool area is very nice and the pool area is very nice to
5. and the buffet breakfast was excellent
6. rooms are spacious and well appointed with a large bathroom a
7. the hotel is a little dated but the rooms are very well appointed
8. we had a great stay at the hotel in the past

Table 8: Examples of generated text on hotel reviews corpus. PRAG wins in both informativeness and fluency on this sample. (Up: PRAG; Down: NRT)

Retrieved Reviews
1. what was great ? parking spot.
2. what was not good ? rudeness of the staff
3. what was great ? the bar service was great
4. what was great ? the food being served around me looked and smelled great
5. what was great ? the food and drinks are always great.
6. what was great ? the bar plate is excellent and the house salad was the best we have ever had
7. what was great ? the food being served around me looked and smelled great
8. what was great ? the salads and appetizers are decent
1. the pizza was delicious
2. the drinks are good
3. i 'm not a big fan of the UNK
4. the atmosphere is great
5. the atmosphere is great
6. the atmosphere is great
7. the drinks are cheap and the bartenders are friendly
8. the drinks are strong and the bartenders are friendly

Table 9: Examples of generated text on hotel reviews corpus. PRAG wins in both informativeness and fluency on this sample. (Up: PRAG; Down: Peter)