

MAPLE: Enhancing Review Generation with Multi-Aspect Prompt Learning in Explainable Recommendation

Ching-Wen Yang¹, Zhi-Quan Feng¹, Ying-Jia Lin^{3,4}, Che-Wei Chen¹, Kun-Da Wu²,
Hao Xu², Jui-Feng Yao², Hung-Yu Kao^{4*}

¹Department of Computer Science and Information Engineering,
National Cheng Kung University, ²Google Pixel Software

³Department of Artificial Intelligence, Chang Gung University

⁴Department of Computer Science, National Tsing Hua University
P76114511@gs.ncku.edu.tw, hykao@cs.nthu.edu.tw

Abstract

The Explainable Recommendation task is designed to receive a pair of *user* and *item* and output explanations to justify why an item is recommended to a user. Many models approach review generation as a proxy for explainable recommendations. While these models can produce fluent and grammatically correct sentences, they often lack precision and fail to provide personalized, informative recommendations. To address this issue, we propose a personalized, aspect-controlled model called Multi-Aspect Prompt Learner (MAPLE), which integrates aspect category as another input dimension to facilitate memorizing fine-grained aspect terms. Experiments conducted on two real-world review datasets in the restaurant domain demonstrate that MAPLE significantly outperforms baseline review-generation models. MAPLE excels in both text and feature diversity, ensuring that the generated content covers a wide range of aspects. Additionally, MAPLE delivers good generation quality while maintaining strong coherence and factual relevance. The code and dataset used in this paper can be found at <https://github.com/Nana2929/MAPLE>.

1 Introduction

In the context of Natural Language Generation (NLG) explainable recommendation models, a good explanation is required to have the following characteristics: 1) *Diversity*: for the same item, a model should generate personalized rationales for different users. 2) *Factuality*: the recommended feature or content should be factually relevant to the item. 3) *Precision*: the recommended feature should be as precise (as opposed to general) as possible.

Within these criteria, we observe a "diversity-factuality dilemma", as illustrated in Figure 1. Models that generate generic aspects, such as

*Corresponding author.

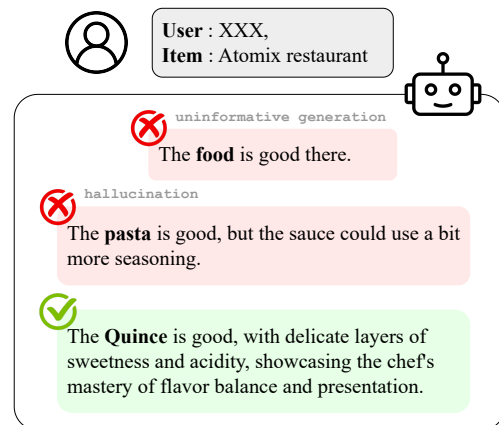


Figure 1: Overview of explainable recommendation: A simple illustration of the "diversity-factuality dilemma".

"food" in the context of restaurants, may provide safe and broadly applicable explanations but risk being uninformative and repetitive, which has been noted in prior studies (Li et al., 2017; Dong et al., 2017; Li et al., 2021, 2023). On the other hand, models creating diverse and creative aspects produce more engaging explanations but may compromise factual relevance by including irrelevant details, which is referred to as the hallucination problem. Few existing models effectively balance these traits.

To address this, we introduce the concept of "multi-aspect" from multi-aspect sentiment analysis (Lu et al., 2011; Xianghua et al., 2013), which assumes a limited inventory of multiple, distinct aspects like "food quality" or "service" in the restaurant domain, providing a more fine-grained understanding of user preferences. In this paper, we propose a simple yet effective two-stage tuning approach that integrates aspect as an auxiliary signal to improve the memorization of rich aspect terms, called MAPLE (Multi-Aspect Prompt Learner). For *diversity*, we utilize a Multilayer Perceptron (MLP) model to predict aspects and

employ a distribution-balanced loss function (Wu et al., 2020) to ensure diverse predictions. For *factuality*, we generate recommendation text using a retriever-reader framework, and for *precision*, we retrieve specific information to guide the LLM in generating user- and item-specific recommendation text. Our contributions are as follows:

1. MAPLE increases diversity, factuality, and precision of generated features, endorsed by the self-crafted explainability metrics from aspect-wise perspectives.
2. MAPLE’s generated explanations serve as good queries within the retriever-reader framework. By comparing with a latent personalized retriever model, we show that MAPLE more accurately predicts aspect relations, as evidenced by the results and case studies.
3. We renew the review-generation datasets in the restaurant domain (Li et al., 2020) to include higher quality aspect terms and additionally label associated aspect categories, enhancing the research in this field.

We treat MAPLE as a retriever in the retriever-reader framework with an LLM as the reader in Appendix D, demonstrating that MAPLE’s explanations combined with the LLM’s comprehension yield enriched, personalized results.

2 Related Works

2.1 Review Generation

Past works (Dong et al., 2017; Li et al., 2017, 2021, 2023) have proposed end-to-end frameworks for learning a short explanation (often part of the review) for a user-item pair. Experimental statistics and case studies show that these models generate repetitive, overly generic sentences that are still far from good rationales (Xie et al., 2023), and even suffer from hallucination issue (Xie et al., 2023; Maynez et al., 2020)¹. It is also observed that the uninformative sentences are often due to models’ being incapable of generating precise and informative aspect terms.

2.2 Aspect-aware Explanation Generation

According to Zhang et al. (2023) which studies into aspect-based sentiment analysis, aspect categories c refer to broad attributes (such as *food* or *service*

¹A model exhibiting hallucinations generates content that includes inaccuracies or irrelevant information.

in the restaurant domain) and aspect terms a refer to specific targets (such as *the beef tacos* in the review "The beef tacos here are amazing").

To address overly general aspect terms, Ni and McAuley (2018) proposed ExpansionNet, which expands short phrases into detailed explanations via a tri-encoder framework that predicts aspect-term importance using an aspect encoder. Aspect terms are obtained through ABAE (He et al., 2017), an unsupervised model that infers aspect categories and retrieves top-K aspect terms.

Building on ExpansionNet, UARM (Sun et al., 2021) enhances the use of inferred aspect embeddings from ABAE by converting them into aspect-aware user and item representations. Li et al. (2020) extracts review aspect terms with Sentires (Zhang et al., 2014), and its proposed model NETE does aspect-term condition generation in training. In the inference stage, NETE uses Point-wise Mutual Information (PMI) to rank and select an aspect term by user preference. However, aspect-term distribution might be too sparse to be captured by PMI. Therefore, there’s an emerging line of works that simply assume the ground-truth aspect terms are given: One of them is PETER+ (Li et al., 2021), which places aspect-term tokens behind user and item IDs to guide the generation of explanations. Another example is ERRA (Cheng et al., 2023), which leverages Sentence-BERT (Reimers and Gurevych, 2019) to encode aspect terms and the corpus. It searches for appropriate aspect and corpus vectors to support the text generation.

Our approach aligns more with the line of NETE, wherein in the training stage, MAPLE is trained by a given *aspect-category* signal, and in the inference stage, it decides which aspect signal it wants to use. Compared to ERRA, MAPLE fine-tunes the generation model and incorporates multi-aspect learning to predict aspects with greater diversity. This makes MAPLE a simple yet effective approach for controlling from a broad sense and generating fine-grained aspect terms.

2.3 Retrieval-Augmented Explanation Generation

To address the issue of factual relevance, a long-existing challenge in NLG task, Xie et al. (2023) proposes a retriever-reader two-stage framework called PRAG. In the first stage, a personalized retriever formulates a latent query based on the input user and item; in the second stage, a reader model generates the explanation, grounding it in the re-

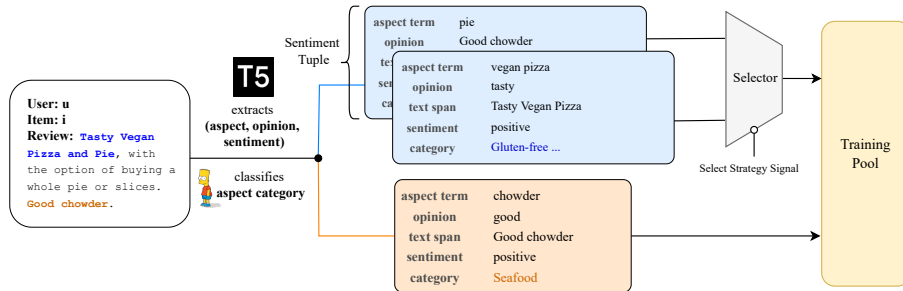


Figure 2: Automated sentiment-analysis pipeline for Multi-Aspect Review Segmentation.

trieved content, which consists of past reviews that are inherently factual. Given the current advancements in LLM (Rorseth et al., 2024), we argue that LLMs fit into the role of the reader very well as their reading comprehension ability can be leveraged not only to enhance factual relevance but also to refine the style of the generated reviews, making them more persuasive and closely aligned with real explanations. In our case studies in Appendix D, with the examples of employing MAPLE as a discrete retriever and LLM as a reader, we showcase the potential for improving both factual accuracy and the persuasiveness of explanations.

3 Methodology

In this work, we propose a two-stage training approach. We initially extract the aspect term from reviews in Section 3.2 to support the two-stage training. In Stage 1 (Section 3.3), we focus exclusively on the explanation generation task, optimizing it using the negative log-likelihood loss \mathcal{L}_T until convergence. In Stage 2 (Section 3.4), we introduce the recommendation loss, weighted by α , into the total loss to equip the ID embeddings with the selector ability. Finally, our inference process is presented in Section 3.5.

3.1 Problem Setup

Consider a set of users U and items I . For each user $u \in U$ and item $i \in I$ in the training set, there is an associated review r covering several aspects c_i , each explained in terms of that aspect. The task of joint aspect-recommendation-explanation is to learn a function $rec : (u, i) \rightarrow (\hat{c}_{u,i}, \hat{E}_{u,i})$, where $\hat{c}_{u,i}$ is the predicted aspect-category distribution $\in \mathcal{R}^{n_{\text{aspect}}}$ (n_{aspect} denotes the number of aspect categories) and $\hat{E}_{u,i}$ is the textual explanation given to justify the recommendation of item i to user u . In this context, $\hat{c}_{u,i}$ serves as a by-product that assists in generating $\hat{E}_{u,i}$ in the inference stage.

3.2 Multi-Aspect Review Segmentation

Instead of predicting aspect-term importance, as done by ExpansionNet and UARM, our model takes a broader perspective by learning aspect topics and extracting aspect terms from within these topics based on ID information. This approach prevents the model from overfitting to the characteristics of individual users or items, avoiding overly generic explanations that are conditioned on ID alone. By introducing aspect prompts c , the model focuses on learning the fine-grained aspect terms of the particular category, while the ID embeddings are learned as *selectors* for relevant aspects.

We employ an automated sentiment-analysis pipeline to extract sentiment tuples (*aspect term, opinion, sentiment, category*) from reviews. Concretely, we use a fine-tuned sentiment analysis model (Raffel et al., 2020) to extract sentiment tuples within the reviews and discard the reviews without tuples. We then use a zero-shot classifier (Lewis et al., 2020) to assign each extracted aspect term with an aspect category (for details on model and aspect inventory, see Appendix A.1 and A.3). Finally, for every user-item pair, we check if there are multiple mined aspect terms under the same aspect category; if yes, we choose the tuple with an arbitrary select strategy, e.g., choose the one that is the longest, and add the associated tuple into the training data. Figure 2 illustrates an example of a review and its extracted sentiment tuples, Since we take one tuple for each category, this review yields two tuples. For the "Gluten-Free, Vegan, Vegetarian" category, we take ("vegan pizza", ...) since it has a longer aspect term compared to "pie"; for the "Seafood" category, we take ("chowder", ...).

3.3 Stage 1: Explanation Generation

We adopt the continuous prompt learning approach proposed by Li et al. (2023). The input sequence can be represented as

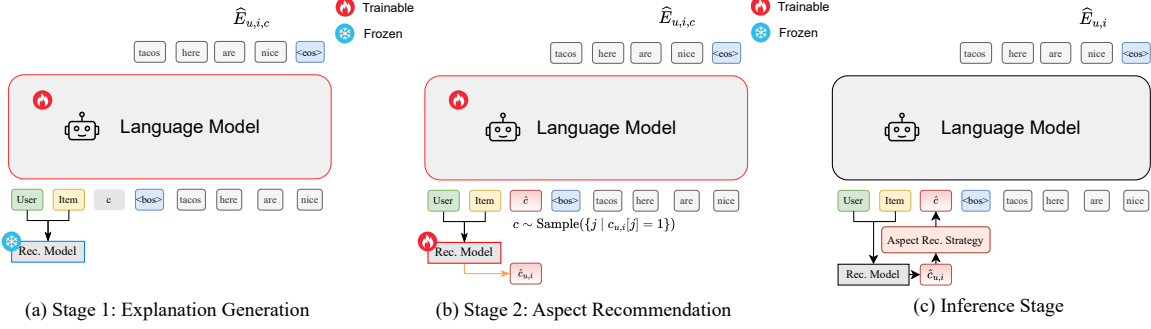


Figure 3: The proposed MAPLE architecture. MAPLE is trained on Stage 1: Explanation Generation and Stage 2: Aspect Recommendation as a unified architecture. During inference, MAPLE exploits the trained Aspect-Recommendation Model (Rec. Model) in Stage 2 to predict an aspect distribution $\hat{c}_{u,i}$ and obtain an aspect signal \hat{c} , and then feed it back to Stage 1 for generation.

$S = [u, i, c, \langle \text{bos} \rangle, e_1, \dots, e_{|E_{u,i,c}|}, \langle \text{eos} \rangle]$, where $\langle \text{bos} \rangle$ and $\langle \text{eos} \rangle$ are arbitrary tokens marking the beginning and end of a sentence, as shown in Figure 3 (a). Specifically, we prepare three sets of token embeddings $U \in \mathbb{R}^{|\mathcal{U}| \times d}$; $I \in \mathbb{R}^{|\mathcal{I}| \times d}$; $C \in \mathbb{R}^{|\mathcal{C}| \times d}$, where $\mathcal{U}, \mathcal{I}, \mathcal{C}$ represent the set of users, items and the predefined aspect categories of a dataset, respectively. We illustrate the details using aspect category because it is the same for obtaining ID representations. To obtain the representation of aspect category c , denoted \mathbf{c} , we index the token embeddings using the one-hot vector $g(c) \in \{0, 1\}^{|\mathcal{C}|}$.

We then train MAPLE to condition the user-item-aspect signals for text generation. We minimize the negative log-likelihood loss L_T , where \mathcal{T} denotes the training set; $E_{u,i,c}$ denotes the explanation segment for the user-item-aspect pair; $c_{3+t}^{e_t}$ is offset by 3 to accommodate the user, item, and aspect prompt tokens. As discussed in Section 3.2, rather than learning just one explanation for a user-item pair, a user-item can now pair with different aspect categories c_i depending on the categories labeled from the ground-truth review.

$$\mathcal{L}_T = \frac{1}{|\mathcal{T}|} \sum_{(u,i,c) \in \mathcal{T}} \frac{1}{|E_{u,i,c}|} \sum_{t=1}^{|E_{u,i,c}|} -\log c_{3+t}^{e_t} \quad (1)$$

3.4 Stage 2: Aspect Recommendation

During training, MAPLE takes a user-item pair and the auxiliary aspect category. However, in the inference stage, the aspect information may be absent. To close the gap between training and inference, we design an auxiliary task, called aspect recommendation, to recover the relationship between IDs

and their associated aspect categories. The input of the aspect-recommendation task is a pair of user, item IDs, the output is the predicted aspect category probabilities $c_{u,i}^{\hat{c}} \in \mathbb{R}^{n_{\text{aspect}}}$. We implement the model with an MLP architecture:

$$\hat{c}_{u,i} = \sigma(\mathbf{W}_c \delta(\mathbf{W}_h [\mathbf{u}; \mathbf{i}] + \mathbf{b}_h) + \mathbf{b}_c) \quad (2)$$

In equation 2, $\mathbf{W}_h \in \mathbb{R}^{h \times 2d}$ and $\mathbf{b}_h \in \mathbb{R}^h$ are the weights and biases of fully-connected layer(s), $\mathbf{W}_c \in \mathbb{R}^{n_{\text{aspect}} \times h}$ and $\mathbf{b}_c \in \mathbb{R}^{n_{\text{aspect}}}$ are the weights and biases of the output classifier layer, and δ and σ are the activation functions ReLU and Sigmoid, respectively.

However, the aspect-category label exhibits a very skewed, long-tail distribution (Figure 5), which poses a challenge in training the aspect-recommendation model, since a conventional classifier easily overfits the head classes. We employ the distribution-balanced loss function (Wu et al., 2020) to address this issue, effectively capturing aspect preferences from ID information rather than label distribution. It is formalized as:

$$\mathcal{L}_{DB}(x^k, y^k) = \frac{1}{n_{\text{aspect}}} \sum_{i=0}^{n_{\text{aspect}}} \hat{r}_i^k \left[y_i^k \log(1 + e^{-\Delta_i}) + \frac{1}{\lambda} (1 - y_i^k) \log(1 + e^{\lambda \Delta_i}) \right] \quad (3)$$

$$\Delta_i = z_i^k - \nu_i$$

In equation 3, $\mathcal{L}_{DB}(x^k, y^k)$ represents the distribution-balanced loss function for the k -th sample; \hat{r}_i^k represents the sampling weighting factor for the i -th aspect class of the k -th sample; y_i^k is the ground truth label for the i -th class of the k -th

sample; z_i^k is the logit for the i -th class of the k -th sample; ν_i is the class bias for i -th class, and λ is a hyperparameter regularizing sigmoid’s gradient issue.

In this stage, we add the recommendation loss weighted by α to the total loss, giving ID embeddings selector ability. Formally,

$$L = L_T + \alpha L_{DB}$$

We conducted preliminary tuning and found that $\lambda = 1$ yielded stable training, while larger values (e.g., $\lambda = 5$) led to underfitting. Thus, we use $\lambda = 1$ in all the reported experiments.

3.5 Inference Stage

Given user and item IDs, we first infer stage 2’s aspect-recommendation model and trim the predicted aspect distribution $\hat{c}_{u,i}$ to leave only the top 5, and then we sample K ($K \leq 5$) aspects with replacement from the trimmed distribution, using the predicted probabilities as sample weights. The K embeddings are then fused (by taking their mean) to form the aspect signal at position \hat{c} , as shown in Figure 3 (b). The trained transformer is thus conditioned on u, i, \hat{c} to generate a well-rounded explanation $\hat{E}_{u,i}$.

4 Experiment Setup

4.1 Dataset

In our experiments, we focus on the restaurant domain by benchmarking against two datasets from Yelp, which we refer to as **Yelp19** (Li et al., 2020) and **Yelp23** (Inc., 2023), distinguishing them based on the year the data was provided. Different restaurants offer a wide variety of unique items (e.g., specific dishes like *tonkatsu ramen* for a Japanese ramen store; *risotto* for an Italian restaurant). This distinctiveness in aspect terms allows for more precise and compelling explanations for a specific restaurant. We provide the dataset sources, statistics, and details of how we renew the datasets using the automated sentiment pipeline in Appendix A.

4.2 Automatic Evaluation Metrics

4.2.1 Diversity

We measure the diversity of generated sentences using token, sentence, and corpus-level evaluations. For token-based metrics, we use the **Distinct-N** metrics by Li et al. (2016). To assess the severity of repetitive sentences generated by the model, we utilize the **Unique-Sentence Ratio (USR)** Li et al.

(2020) and the derived metrics, **user-wise USR** and **item-wise USR**². For the corpus-level, we employ the entropy-based metric ENTR by Jhamtani et al. (2018).

4.2.2 Factuality

For factuality evaluation, Xie et al. (2023) introduces the *Entailment Ratio*, which measures how many statements are supported by existing reviews of the same item. However, the generated content can synthesize aspects and opinions from multiple reviews, which the *Entailment Ratio* metric may incorrectly classify as "unfactual." To address the above issue, we propose **Item-wise Feature Matching Ratio (iFMR)**. We treat the collection of extracted features³ in one restaurant’s reviews as its "menu"; our rationale is that if the generated content contains any feature from the menu, it is "factual." Formally,

$$\text{iFMR} = \frac{1}{N} \sum_{u,i} \delta \left(\exists f \in F_i : f \in \hat{E}_{u,i} \right)$$

$\hat{E}_{u,i}$ is the generated sentence for a user-item pair; F_i is the set of features associated with item i extracted from all training set reviews of that item; $\delta(x) = 1$ if x is true and $\delta(x) = 0$ otherwise.

4.2.3 Precision

To evaluate precision, which can also be interpreted as aspect-wise explainability, we adopt two metrics from Li et al. (2020): **Feature Coverage Ratio (FCR)** and **Ground-Truth Feature Matching Ratio (GT-FMR)**. FCR is computed as the number of distinct features contained in all the generated explanations, divided by the total number of the features in the training set⁴. **Ground-Truth Feature Matching Ratio (GT-FMR)** measures whether a generated explanation contains the feature in the ground-truth text.⁵ Additionally, it’s important to consider the perspective of restaurant owners, who invariably prefer that a greater number of their

²User-wise USR is calculated as the average USR ratio across a user’s generated explanations, with the same approach applied for item-wise USR.

³We use "features" interchangeably with "aspect terms" as Li et al. (2020) refer to aspect terms as features in their proposed metrics.

⁴Since we evaluate only a test subset, we count the features associated with its items and use the total as the denominator.

⁵It was originally proposed as a Feature-Matching Ratio (FMR). To distinguish it from the factuality metric item-wise FMR, we prefix it with "GT".

Yelp19											
Method	Factuality	Aspect-wise Exp.			Text Diversity					Gen. Quality	
	iFMR	FCR	iFCR	GT-FMR	USR	uUSR	iUSR	D-2	D-3	ENTR	MAUVE
Att2seq	0.693	0.055	0.063	0.082	0.348	<u>0.993</u>	0.864	0.802	0.858	<u>8.389</u>	<u>0.0404</u>
NRT	0.672	0.051	0.063	0.075	<u>0.384</u>	<u>0.993</u>	<u>0.894</u>	0.804	<u>0.867</u>	8.331	0.0402
PETER	0.704	<u>0.062</u>	0.057	<u>0.096</u>	0.263	0.991	0.733	<u>0.819</u>	0.858	8.173	0.0239
PEPLER	0.661	0.047	0.047	0.087	0.301	0.991	0.797	0.365	0.412	8.241	0.0061
ERRA	<u>0.775</u>	0.060	<u>0.064</u>	0.097	0.277	0.917	0.859	0.825	0.868	8.316	0.0267
MAPLE	0.807	0.185	0.108	0.087	0.951	0.999	0.997	0.684	0.808	11.015	0.0699
MAPLE-GT	0.684	0.148	0.086	0.167	0.475	0.994	0.890	0.801	0.867	9.362	0.0506

Yelp23											
Method	Factuality	Aspect-wise Exp.			Text Diversity					Gen. Quality	
	iFMR	FCR	iFCR	GT-FMR	USR	uUSR	iUSR	D-2	D-3	ENTR	MAUVE
Att2seq	0.618	0.012	0.022	0.230	0.253	<u>0.987</u>	<u>0.977</u>	<u>0.878</u>	0.851	8.024	<u>0.0634</u>
NRT	0.615	0.009	0.021	0.224	0.216	0.975	0.956	0.876	0.850	7.657	0.0486
PETER	0.678	0.015	0.024	<u>0.251</u>	0.272	0.962	0.934	0.867	0.847	8.022	0.0604
PEPLER	0.690	0.013	0.025	0.261	0.303	0.950	0.911	0.809	0.799	8.311	0.0426
ERRA	<u>0.775</u>	<u>0.033</u>	<u>0.03</u>	0.181	<u>0.388</u>	0.963	0.858	0.884	0.878	<u>8.992</u>	0.0267
MAPLE	0.794	0.063	0.041	0.206	0.855	0.999	0.997	0.816	<u>0.866</u>	11.485	0.2183
MAPLE-GT	0.691	0.039	0.029	0.333	0.620	0.998	0.994	0.892	0.886	10.316	0.1497

Table 1: Automatic evaluation results on 10,000 user-item pairs sampled from the test sets. The "i"-prefix is "item-wise" for short, and the "u"-prefix is "user-wise". The best and second-best performances are bold-faced and underlined, respectively. Since MAPLE-GT employs ground-truth aspects, we do not put this variant into comparisons. Unless otherwise stated, MAPLE in the paper defaults to the Supervised@3 aspect-recommendation strategy.

restaurant’s advantages be highlighted and recommended to users in the explanation texts. Therefore, we craft **item-wise Feature Coverage Ratio (iFCR)** to provide insights into how well a model covers the relevant features for each item based on the training data. For more implementation details and concrete formulas, refer to Appendix B.1.

4.2.4 Generation Quality

Unlike targeted generation tasks such as text summarization, explanation generation is a form of open-ended text generation. This nature makes the use of metrics like BLEU or ROUGE, which rely on exact token-by-token matches, less suitable. To address this, we employ **MAUVE** (Althubyani et al., 2024) to quantify the gap between generated explanations and human texts, providing a statistical measure of generation quality.

4.3 Baselines

We compare our model against two groups of baseline methods. The first group consists of classic end-to-end review generation models focused on NLG for explanations. We cover **Att2seq** (Dong et al., 2017), **NRT** (Li et al., 2017), **PETER** (Li et al., 2021), **PEPLER** (Li et al., 2023), and **ERRA** (Cheng et al., 2023). The second group is the retriever architecture in the retriever-reader framework, in which, to the best of our knowledge, only the personalized retriever in **PRAG** (Xie et al.,

2023) is available. For more details, see Appendix B.2.

5 Results and Analysis

5.1 Quantitative Analysis on Explanations

In our quantitative analysis, we compare MAPLE with baseline review-generation models that use item and user IDs but lack aspect input signals. Despite not multitasking on overall rating prediction, MAPLE significantly outperforms in factuality, feature coverage, and both sentence- and corpus-level diversity. MAPLE shows more than a 10% improvement in item-wise FMR, accurately recommending features in about 80% of explanations. While striving for precision, such as choosing specific features like "salmon sushi" over general "food," MAPLE experiences a trade-off between FCR and FMR, especially notable in its lower GT-FMR score on Yelp23. Nonetheless, it excels in textual diversity on Yelp19, as evidenced by high USR and ENTR scores, which indicate minimal sentence repetition and high creativity. Besides, MAPLE also achieves high MAUVE scores, demonstrating a generation distribution closely aligned with human texts. For qualitative analyses and case studies, please refer to Appendix D.

Yelp19									
Method	Factuality iFMR	Aspect-wise Exp.		Text Diversity			Gen. Qty. MAUVE	Text Sim. BLEU-4	Ranking HR
		iFCR	GT-FMR	USR	D-2	ENTR			
S@1	0.728	0.077	0.061	0.654	0.830	10.580	0.0560	0.297	0.350
S@2	0.789	0.099	0.081	0.840	0.768	11.094	0.0451	0.245	0.568
S@3	0.805	0.108	0.087	0.944	0.721	10.958	0.0699	0.238	0.716
S@4	0.823	0.125	<u>0.096</u>	0.971	0.666	11.497	0.0813	0.182	0.806
Heuristic@3	0.731	0.094	0.076	0.942	0.733	10.998	0.0504	0.230	0.419
GT@1	0.684	0.086	0.167	0.728	0.833	8.041	0.0506	0.585	-
Yelp23									
S@1	0.548	0.023	<u>0.232</u>	0.336	0.877	8.587	0.0548	1.528	0.710
S@2	0.725	0.035	0.200	0.661	0.856	11.487	0.2018	0.952	0.863
S@3	0.794	0.041	0.207	0.855	0.816	11.485	0.2183	0.947	0.916
S@4	0.834	0.054	0.227	0.921	0.734	12.182	0.1594	0.596	0.948
Heuristic@3	0.785	<u>0.042</u>	0.178	0.855	0.816	11.441	0.1796	0.861	0.811
GT@1	0.691	0.029	0.333	0.620	0.892	10.316	0.1497	2.616	-

Table 2: Comparison of MAPLE with different aspect selection strategies. S@k denotes Supervised strategies with different values of K . HR is short for Hit Ratio. For Yelp23, there might be multiple ground-truth review segments; in this case, we report its BLEU-4 with the multi-reference mechanism.

5.2 Ablation Study on Aspect Recommendation Strategies

The trained aspect-recommendation component aims to select the best-fit aspect categories based on user-item IDs. While its performance is not directly evaluated, it plays a crucial role in determining the textual quality and topic of the generated content. We conduct an ablation study on the aspect-recommendation strategies with or without the component and display selected statistics in Table 2.

1. **Supervised@K (K=1,2,3)**: Inferencing the trained aspect-recommendation component to select K aspects from a trimmed predicted aspect distribution (Appendix C.1).
2. **Heuristic@3**: Randomly sampling three categories from the intersection of item’s and user’s category histories.⁶
3. **GT@1**: Simulating the scenario where the user has either specified their interested aspect or the aspect-recommendation model is 100% accurate by directly using the ground-truth aspect.⁷

A proper mix of aspects enriches explanations.

Increasing the K value in the model enhances the hit ratio, which boosts corpus-level diversity and feature coverage. Conversely, it slightly reduces token-level diversity, as evidenced by lower Distinct-2 scores, and significantly diminishes text-similarity scores. This reduction in Distinct-2

⁶Default to item history if the intersection size is smaller than three.

⁷We do *not* inference the trained aspect-recommendation component for Heuristic and GT.

scores is attributed to the lengthening of sentences as K increases. Given that the BLEU (Papineni et al., 2002) is considered an indicator of textual relevance to the ground-truth reviews, we attempt to strike a balance between the diversity and textual relevance by monitoring BLEU-4 score. We employ the elbow method on BLEU-4 to identify the point where the slope sharply changes (when the component starts to produce overly generalized selection) and choose the point as our optimal K . We choose $K = 3$ as the default value. Additionally, while heuristic selection strategies yield suboptimal results in aspect-wise explainability, the supervised approach is preferred due to its higher aspect ranking scores.

Ground-truth aspects lead to precise feature prediction and higher textual relevance. When provided with the ground-truth aspect category, the model’s GT-FMR score significantly improves, demonstrating effective training of aspect prompts. This improvement is attributed to the accurate aspect signal, which narrows the feature pool and thereby enhances feature prediction accuracy. Furthermore, when conditioning on the ground-truth aspect, the BLEU-4 score tops all other strategies, showing that under the scenario where the user can provide the aspect category he/she is interested in, the explanation precision significantly improves. This underscores the value of precise user input in enhancing the relevance and accuracy of generated explanations.

5.3 Explainability of Aspect Prompts

Our hypothesis posits that MAPLE’s aspect prompts tokens better retain and utilize less frequent features, which we explored using t-

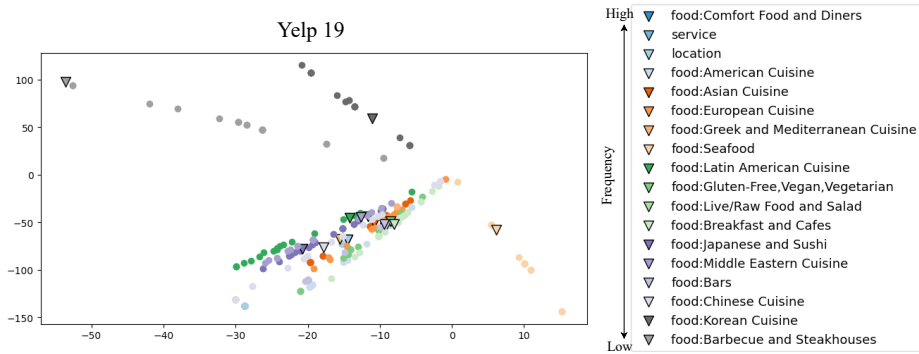


Figure 4: The t-SNE results on MAPLE trained on Yelp19 (fold 1) with S@3 strategy. The inverted triangle symbol indicates the aspect category prompt; the dots with the same color indicate the feature words belonging to the aspect category. We plot the 20 nearest feature words for each category.

Distributed Stochastic Neighbor Embedding (t-SNE) (van der Maaten and Hinton, 2008) to visualize the semantic clustering of feature words and aspect prompts. Figure 4 shows linear clusters for the aspects and their feature words, confirming MAPLE’s efficient learning of semantic associations. We further calculate the FCR scores by only considering the head 5 and the tail 5 aspect categories, denoted **head FCR** and **tail FCR**, and then present the statistics in Table 3. These results suggest that MAPLE’s aspect prompt tokens significantly improve its ability to remember and utilize even the less frequent features, maintaining competitive performance even for less common aspect categories.

Yelp19			
	head FCR	tail FCR	FCR
PETER	0.0635	0.0433	0.0616
PEPLER	0.0509	0.0237	0.0470
MAPLE	0.1993	0.1184	0.1846
Yelp23			
	head FCR	tail FCR	FCR
PETER	0.0158	0.0099	0.0147
PEPLER	0.0141	0.0065	0.0131
MAPLE	0.0698	0.0353	0.0629

Table 3: Head and tail FCR for MAPLE with sota baselines. FCR is attached for reference.

5.4 MAPLE as a Discrete Retriever

In this section, we examine the effectiveness of MAPLE as a discrete retriever within the retriever-reader framework, comparing it to the only known personalized retriever architecture from PRAG (Xie et al., 2023). We evaluate both models from two perspectives: the *latent* perspective and the

aspect perspective.

Latent Perspective. We project the generated explanations of MAPLE into the same latent space as PRAG’s queries by encoding it with all-mpnet-base-v2 to assess similarity against the ground-truth review, using Cosine Similarity and Mean Squared Error (MSE). Since PRAG’s objective during training is to minimize the MSE, it performs very well in this context, as the results shown in Table 4.

Aspect Perspective. We evaluate the model’s ability to identify relevant aspects and provide the results in Table 4. MAPLE uses an aspect recommendation module targeting relevant losses, whereas for PRAG, we utilize its retrieved reviews from the user and item history, and collect the top-3 reviews’ corresponding aspects as its prediction. Despite MAPLE’s queries showing only 30% similarity to ground-truth reviews, it more effectively identifies correct aspects, demonstrating stronger alignment with target goals. Overall, PRAG excels in generating queries similar to ground-truth reviews, while MAPLE outperforms in accurately targeting relevant aspects.

Model	Latent		Aspect	
	MSE	Cos Sim.	HR@3	F1
MAPLE	0.002	0.301	0.716	0.369
PRAG	0.001	0.439	0.389	0.228

Table 4: Comparison of Latent Metrics and Aspect Metrics for MAPLE and PRAG models

6 Conclusion

We introduce the Multi-Aspect Prompt LEarnner (MAPLE), a model that leverages user IDs and

multi-aspect signals to generate detailed and controllable explanatory texts. Our primary technical achievements include the simple yet effective integration of aspect information into representation learning, aspect recommendation, and the learning of review sentences. We also prove that MAPLE can serve as a good discrete retriever in a retriever-reader explainable pipeline.

7 Limitations

Despite its high features and textual diversity, the MAPLE model has several limitations. A notable challenge is the labeling of aspect categories. Although automated, this process still necessitates manual effort to define the aspect category inventory. The quality of these labels and their distribution across the dataset impact the training difficulty of the aspect recommendation component and subsequently the inference text quality and style. In cases where the label distribution is highly skewed, it might be necessary to optimize the aspect recommendation component separately. Additionally, our introduction of the item-wise Feature Matching Ratio marks a pioneering step towards enriching the aspect-wise factuality perspective of explainable recommendation model evaluations. While this metric adeptly identifies factual features in sentences, future research and methodological advancement are needed to improve its ability to detect non-factual elements.

Acknowledgments

We thank the Google Pixel Software Team for providing strong research support to this project. This work was partially supported by the National Science and Technology Council, Taiwan, under Grant NSTC 113-2223-E-007-019.

References

- Mohammed Althubayani, Zhijin Meng, Shengyuan Xie, Cha Seung, Imran Razzak, Eduardo Benitez Sandoval, Baki Kocaballi, Mahdi Bamdad, and Francisco Cruz Naranjo. 2024. [Percy: A multimodal dataset and conversational system for personalized and emotionally aware human-robot interaction](#). Preprint, arXiv:2412.04908.
- Hao Cheng, Shuo Wang, Wensheng Lu, Wei Zhang, Mingyang Zhou, Kezhong Lu, and Hao Liao. 2023. [Explainable recommendation with personalized review retrieval and aspect learning](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 51–64, Toronto, Canada. 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 *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 623–632, Valencia, Spain. Association for Computational Linguistics.
- Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2017. [An unsupervised neural attention model for aspect extraction](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 388–397, Vancouver, Canada. Association for Computational Linguistics.
- Yelp Inc. 2023. [Yelp open dataset](#). Accessed: 2023-10-15.
- 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 *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1661–1671, Melbourne, Australia. Association for Computational Linguistics.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. [BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, 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 *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.
- Lei Li, Yongfeng Zhang, and Li Chen. 2020. [Generate neural template explanations for recommendation](#). In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 755–764.
- Lei Li, Yongfeng Zhang, and Li Chen. 2021. [Personalized transformer for explainable recommendation](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4947–4957, Online. Association for Computational Linguistics.

- Lei Li, Yongfeng Zhang, and Li Chen. 2023. Personalized prompt learning for explainable recommendation. *ACM Transactions on Information Systems*, 41(4):1–26.
- Piji Li, Zihao Wang, Zhaochun Ren, Lidong Bing, and Wai Lam. 2017. Neural rating regression with abstractive tips generation for recommendation. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '17*, page 345–354, New York, NY, USA. Association for Computing Machinery.
- Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.
- Bin Lu, Myle Ott, Claire Cardie, and Benjamin K. Tsou. 2011. Multi-aspect sentiment analysis with topic models. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 81–88.
- Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2022. Unified structure generation for universal information extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5755–5772, Dublin, Ireland. Association for Computational Linguistics.
- Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.
- Jianmo Ni and Julian McAuley. 2018. Personalized review generation by expanding phrases and attending on aspect-aware representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 706–711, Melbourne, Australia. Association for Computational Linguistics.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Maria Pontiki, Dimitrios Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad Al-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *10th International Workshop on Semantic Evaluation (SemEval 2016)*.
- Maria Pontiki, Dimitrios Galanis, Harris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. Semeval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 486–495.
- Maria Pontiki, Dimitrios Galanis, John Pavlopoulos, Haris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.
- 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(1).
- Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *Preprint*, arXiv:1908.10084.
- Joel Rorseth, Parke Godfrey, Lukasz Golab, Divesh Srivastava, and Jaroslaw Szlichta. 2024. RAGE Against the Machine: Retrieval-Augmented LLM Explanations. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, pages 5469–5472, Los Alamitos, CA, USA. IEEE Computer Society.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.
- Peijie Sun, Le Wu, Kun Zhang, Yu Su, and Meng Wang. 2021. An unsupervised aspect-aware recommendation model with explanation text generation. *ACM Trans. Inf. Syst.*, 40(3).
- Laurens van der Maaten and Geoffrey E. Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.
- Tong Wu, Qingqiu Huang, Ziwei Liu, Yu Wang, and Dahua Lin. 2020. Distribution-balanced loss for multi-label classification in long-tailed datasets. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV*, page 162–178, Berlin, Heidelberg. Springer-Verlag.
- Fu Xianghua, Liu Guo, Guo Yanyan, and Wang Zhiqiang. 2013. Multi-aspect sentiment analysis for chinese online social reviews based on topic modeling and hownet lexicon. *Knowledge-Based Systems*, 37:186–195.
- Zhouhang Xie, Sameer Singh, Julian McAuley, and Bodhisattwa Prasad Majumder. 2023. Factual and informative review generation for explainable recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13816–13824.
- W. Zhang, X. Li, Y. Deng, L. Bing, and W. Lam. 2023. A survey on aspect-based sentiment analysis: Tasks, methods, and challenges. *IEEE Transactions on Knowledge & Data Engineering*, 35(11):11019–11038.

Yongfeng Zhang, Haochen Zhang, Min Zhang, Yiqun Liu, and Shaoping Ma. 2014. [Do users rate or review? boost phrase-level sentiment labeling with review-level sentiment classification](#). In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval, SIGIR '14*, page 1027–1030, New York, NY, USA. Association for Computing Machinery.

A Dataset Preparation

A.1 Additional Details for Multi-Aspect Review Segmentation

For the sentiment-analysis model, we use a T5-based fine-tuned information-extraction model `uie-large-en`⁸(Lu et al., 2022) on SemEval datasets (Pontiki et al., 2014, 2015, 2016). For the zero-shot aspect-category classifier, we use the off-the-shelf `facebook/bart-large-mnli` released by Meta. We formulate the classification task as multiple Natural-Language Inference problems: for one aspect term, we ask the question "`{aspect-term}`". For each aspect category, we use the prompt "This example is about `{category}`.", and select the category with the highest entailment logit as the final aspect category for the given aspect term.⁹ In this setup, we use T5 for precise span-level extraction, and BART for reliable category-level classification. We find this modular design to be more effective than relying on a single multi-task model.

A.2 Additional Dataset Details

	Yelp19	Yelp23
# of users	27,147	35,152
# of items	20,266	24,199
# of reviews	1,293,247	1,339,433
# of features	204,117	279,636
# of segments	1,293,247	3,079,123
# of categories	18	28
avg # of reviews/user	47.64	38.10
avg # of reviews/item	63.81	55.35
avg # of segments/review	1.0	2.298

Table 5: Dataset Statistics

For Yelp23, we downloaded all the reviews from the online Yelp-hosted dataset website¹⁰. Since the raw data contains a large number of reviews, we follow the guidelines of Li et al. (2020). That is, we recursively prune users and items with fewer than 20 interactions, and split the datasets randomly into training, validation, and testing at the ratio of 8:1:1 5 times, while ensuring a warm-start scenario. Note that we split in terms of user-item pairs, i.e., all of the associated $(u, i, c_1), \dots, (u, i, c_k)$ (k is the number of categories mined from the ground-truth

⁸<https://huggingface.co/luyaojie/uie-large-en>

⁹<https://huggingface.co/facebook/bart-large-mnli>

¹⁰<https://www.yelp.com/dataset>

review) are guaranteed to appear in one of the training, validation, or testing stages. For Yelp19, we utilize the data as processed by Li et al. (2020). Our sentiment analysis pipeline is applied to both Yelp19 and Yelp23. For Yelp23, we retain text spans illustrated in 2 as $E_{u,i,c}$. Since Yelp19 consists of only fragments of full reviews—which are not available to us—we treat the entire text associated with each user-item pair as $E_{u,i,c}$.

A.3 Aspect Category Inventory

We refer to the tags of items under Yelp website and SemEval workshop’s aspect category inventory to define our aspect category inventory. We also consider the quality of the zero-shot labeling result and slightly alter the label list accordingly, which results in different aspect inventory for different datasets.

B Experiment Setup

B.1 Automatic Evaluation Metrics

- **Feature Coverage Ratio (FCR):** Measures the proportion of distinct features captured in the generated explanations compared to the total feature set.

$$\text{FCR} = \frac{N_g}{|\mathcal{F}|}$$

where \mathcal{F} represents the aggregated collection of features belonging to the sampled test items, and N_g is the number of distinct features shown in the generated explanations.

- **Item-wise Feature Coverage Ratio (iFCR):** Calculates the average ratio of matched features to the total features for each item in the test set.

$$\text{iFCR} = \frac{1}{M} \sum_{i \in \text{test_set}} \frac{|f_i^{\text{matched}} \cap f_i^{\text{all}}|}{|f_i^{\text{all}}|}$$

where M is the number of items in the test set, f_i^{matched} are the features matched in the generated explanations for item i , and f_i^{all} are all the features associated with item i in the training set.

- **Ground-Truth Feature Matching Ratio (GT-FMR):** Assesses whether the features in the generated explanations match those in the ground-truth text.

$$\text{GT-FMR} = \frac{1}{N} \sum_{u,i} \delta \left(\exists f \in f_{u,i} : f \in \hat{E}_{u,i} \right)$$

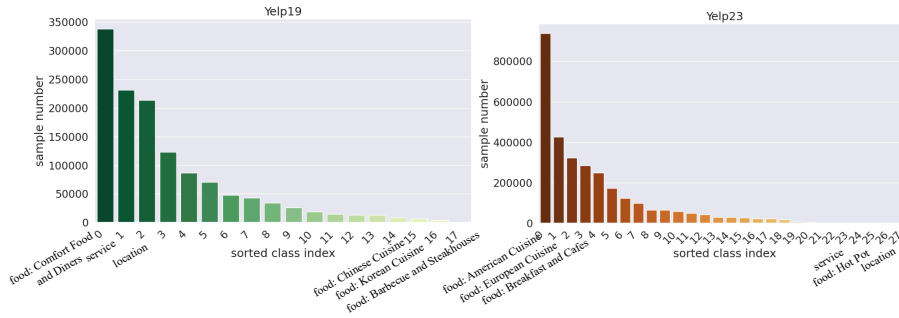


Figure 5: The long-tail dataset distribution of Yelp19 and Yelp23.

Dataset	Categories
Yelp19	American Cuisine, Asian Cuisine, Barbecue and Steakhouses, Bars, Breakfast and Cafes, Chinese Cuisine, Comfort Food and Diners, European Cuisine, "Gluten-Free, Vegan, Vegetarian", Greek and Mediterranean Cuisine, Japanese and Sushi, Korean Cuisine, Latin American Cuisine, Live/Raw Food and Salad, Middle Eastern Cuisine, Seafood, location, service
Yelp23	African Cuisine, American Cuisine, Asian Cuisine, Barbecue and Steakhouses, Breakfast and Cafes, Burmese and Mongolian Cuisine, Chinese Cuisine, Comfort Food and Diners, European Cuisine, Food Court and Stands, Gastropubs and Modern European, "Gluten-Free, Vegan, Vegetarian", Greek and Mediterranean Cuisine, Halal and Kosher, Hot Pot, Japanese, and Sushi, Korean Cuisine, Latin American Cuisine, Live/Raw Food and Salad, Middle Eastern Cuisine, Seafood, South Asian Cuisine, Southeast Asian Cuisine, Tapas Bars, ambiance, location, miscellaneous; service

Table 6: Aspect Inventories for Yelp19 & 23 Datasets, with 18 and 28 categories respectively.

where $\hat{E}_{u,i}$ is the generated sentence for a user-item pair, $f_{u,i}$ is the set of features extracted from the ground truth review, and $\delta(x) = 1$ if x is true, $\delta(x) = 0$ otherwise.

For informativeness (iFMR) and some of the aspect-wise explainability metrics (iFCR, FCR), with an eye to encouraging more fine-grained aspect terms and penalizing the overly generic terms, we filter out the keywords that are too short (less than 4 characters), too general or noisy (see below list of dummy_words).

dummy_words = ["and", "very", "the", "is", "a", "an", "it", "this", "that", "of", "in", "that", "are", "were", "was", "food"]

B.2 Additional Details for Baseline Models

The first group consists of classic end-to-end review generation models focused on NLG for explanations.

- **Att2Seq** (Dong et al., 2017): An LSTM-based model.

- **NRT** (Li et al., 2017): Another LSTM-based model designed for personalized review generation.
- **PETER** (Li et al., 2021): An unpretrained transformer-based model that integrates user and item embeddings into the generation process.
- **PEPLER** (Li et al., 2023): A model leveraging a pre-trained GPT-2 with two-stage tuning for generating explanations. We specifically use the *PEPLER-MF* variant, known for its superior text quality. Among these models, PEPLER’s architecture closely resembles ours, making it a crucial point of comparison.
- **ERRA** (Cheng et al., 2023): A model that combines both aspect-modeling and retrieval into the revised transformer architecture. While ERRA and MAPLE share similar core ideas, our model separates the retrieval process from itself and employs an LLM reader to comprehend the retrieved information from a broader perspective.

The second group of comparison methods involves models utilizing a retriever-reader framework.

- **PRAG** (Xie et al., 2023): A transformer-based model integrates user and item embeddings and review history. It employs a personalized attention mechanism to generate a latent query, which is optimized in an auto-encoder fashion against the embedding of the ground-truth review.

C Implementation for Experiments

C.1 Implementation Details

MAPLE is trained on the training set with hyperparameters fine-tuned on the validation set. Evaluation is performed on 10,000 test-set pairs, averaged over 5 splits for MAPLE and review-generation models, while PRAG is tested only on the first split due to its long training time. MAPLE is optimized using the AdamW optimizer (Loshchilov and Hutter, 2019), with a learning rate of 0.001 and a batch size of 196. The maximum sequence length is limited to 20 Byte Pair Encoding (BPE) (Sennrich et al., 2016) tokens. The training process is divided into two stages as discussed in Section 3.5: Stage 1 runs for up to 30 epochs with the tolerance times set to 5 (monitoring only the text loss); Stage 2 lasts for a maximum of 20 epochs with the tolerance times set to 2 (monitoring only the aspect-recommendation loss). We use different training datasets for each stage. In the first stage, each user-item pair is associated with various aspect categories c_i from the training set. In the second stage, focused on training the recommendation task, we trim the dataset so that each user-item pair appears only once. The aspect regularization coefficient α is set to 0.01. The embedding dimension for user ID, item ID, and aspect is set to 768. For the aspect-recommendation component, we use an MLP with two hidden layers of 256 and 128 dimensions. Regarding the Distribution Balance loss, we set the negative-tolerance coefficient λ to 1.0 and class bias ν_i to 0.05 for each class.

During inference, we apply greedy decoding for all models. For MAPLE, it sets K to 3 as the number of aspects to be mixed as the guiding signal. For the personalized retriever of PRAG, we employ its item-marginalization variant and train with the user, item ID dimension of 768; the rest follow the default settings. For all the other baseline models, we adopt the official implementation as well as the default hyperparameter settings.

C.2 Computational Resources and Cost Analysis

All experiments are run on a single NVIDIA GeForce RTX 3090 (24 GB) GPU.

Training and Inference Costs. It takes about 14 and 7 hours to train MAPLE on Yelp23 and Yelp19, respectively. At inference, MAPLE performs aspect recommendation and then generates explanations conditioned on the predicted aspect vector. This adds *only 0.02s latency per batch of 200 samples*, making it computationally comparable to baseline models (e.g., PETER, PEPLER, and ERRRA) under the same backbone.

+RAG Variant Resource Usage. In case studies (Appendix D), the MAPLE +RAG variant performs retrieval by encoding all reviews using a sentence transformer and computing similarity scores over a review subset. This process requires approximately 60 GB of CPU RAM. For language generation, we use the GPT-4 model via ChatGPT (accessed on June 10, 2024) with default settings and cleared history.

D Qualitative Case Studies

To address the issues of under-representation of user roles, we select several examples under the same item for the case study. Since the personalized retriever in Xie et al. (2023) produces a latent query for a user-item pair, we train the *embedding interpreter* as Xie et al. (2023) proposes to translate the query to a set of keywords: We input the latent query into the embedding interpreter and list the output keywords. In the implementation, it generates 5 keywords and then keeps only those appearing in the retrieved reviews, therefore, the displayed keywords may be fewer than five. In the enhanced +RAG model, we treat the prefix model as the retriever and utilize GPT-4 as the reader. We augment only MAPLE and PRAG with this setup. For MAPLE, we encode its explanation with `all-mpnet-base-v2`¹¹ and then treat them as queries to fetch the top-10 most similar reviews from the user and item history review pools, respectively. See 7 for illustration. For PRAG personalized retriever, we use its proposed item-marginalization variant (i.e., subtract the mean of latent queries generated for the same item, and use it as the query embedding so as to highlight the /textdistinctiveness of each item) to retrieve pertinent

¹¹<https://huggingface.co/sentence-transformers/all-mpnet-base-v2>

You are a restaurant recommendation explainer. Along with

1. a user,
2. a restaurant,

you are also given

3. a model predicted user's personal query to this restaurant, which may not be a true statement,
4. reviews about the restaurant and
5. reviews written by the users himself in the other restaurants.

With the above information, you should pinpoint a feature within the "4. restaurant reviews", A feature can be a dish or an aspect (eg. service, location, etc.) of the restaurant. The feature MUST be mentioned in the restaurant reviews. If the personal query mentions a feature, you can use that. The user's reviews on the other restaurants do not hold true for the current restaurant. You should explain why the user might like or dislike the feature. The explanation should be short and concise within 50 words. Try to summarize the opinions if there are many discussing the same feature. Begin your explanation with "You may be interested in".

```
User: {user} Restaurant: {item}
Personal query: {personal_query}
Restaurant reviews (where you find the feature to recommend the user):
{item_reviews}
User reviews (where you can refer to or identify user's preferences from):
{user_reviews}
```

Figure 6: Prompt template for the reader model. In terms of the personal query, for MAPLE+RAG, we use the MAPLE-generated explanation; for PRAG+RAG, we use the translated and filtered keyword set.

reviews from the user and item pool.

D.1 Case Study 1: Item Feature Precision

It can be observed that in Table 7, PRAG's retriever often produces queries focused narrowly on **service speed**, which lacks the breadth to capture the restaurant's primary attractions. Consequently, the **PRAG+RAG** mostly centers the discussions on service and wait times. Some more cases show that PRAG tends to prioritize service aspects. Conversely, MAPLE excels in identifying detailed features such as ice cream *flavors* (e.g., salted caramel, chocolate chips, strawberry), *styles* (e.g., twist), and its *complement* (e.g., the waffle cone), offering a variety of explanations tailored to different user preferences. In contrast, other review-generation models generally produce generic terms like "ice cream" or "toppings"; the best is probably PETER's "soft-served" in recommending user `xC-q_yh0XwcjRLImkS3RN`.

D.2 Case Study 2: Personalization

In exploring the personalization capabilities of recommendation systems, we focus on how effectively restaurant features are remembered and presented

in user-specific contexts, i.e., how item IDs impact the generated explanations. A pertinent question arises: How about the impacts of user IDs? We assume an imaginary baseline approach, which simply samples reviews from an item's review history. Could it possibly be as good as MAPLE or the other review-generation models? Or let's put it this way: what are the differences between it and the other review-generation models? We observe that user signals primarily influence the *style* and *tone* of the generated sentences. For instance, in Table 8, the user has a review history predominantly in French. Learning that the past reviews are primarily in French, the models generate sentences also in French. This level of personalization would be challenging to achieve with this random sampling baseline from the item's review history. Could such a baseline possibly rival in terms of personalization of MAPLE or other user ID-incorporated review-generation models? Moreover, MAPLE demonstrates an ability to retain and reflect personal details. For example, in Table 9, the user frequently mentions "Phoenix" in the review history, suggesting they might reside in the City of Phoenix. MAPLE captures this detail and

incorporates it into the generated explanation, highlighting the user's local preferences. This ability to integrate personal context into recommendations underscores the model's strength in tailoring content to individual users. These observations lead to the conclusion that the random baseline could not replicate the nuanced personalization achieved by models like MAPLE, which possess sophisticated mechanisms for personalizing generated sentences, adapting not only to users' linguistic preferences but also to subtle personal details, thereby enhancing the relevance and effectiveness of personalized recommendations.

User	Ground-Truth
Hci2c0qo98CO-Pv-VmV7gg	favorites are the bday cake one and smores
Method	Explanation
Att2seq NRT PETER PEPLER	it 's a good place to get a quick bite to eat the ice cream is good but the texture is a bit too sweet the ice cream is good the ice cream is a bit on the sweet side
PRAG PRAG (+RAG)	wait, parking, long You may be interested in the quick service despite the long lines at this restaurant. Reviews mention that even though the restaurant is small and the line can get long, the service is relatively quick.
MAPLE MAPLE (+RAG)	the ice cream is good but the waffle cone was a bit too soft for my liking You may be interested in the ice cream at Sweet Jesus. Although reviews indicate the ice cream itself is average, the unique toppings and creative presentation could be appealing to you, especially given your past enjoyment of special flavors like black sesame and matcha in waffle cones.
Ao-6FYE29-I8WwPg67806A	the ice cream was nice and creamy and not dripping all over the place like most other soft-serve ice cream I've had
Att2seq NRT PETER PEPLER	it was a good size the ice cream was good the ice cream was good the ice cream was good but the texture was a little off
PRAG PRAG (+RAG)	line, fast, worth You may be interested in the line management and atmosphere at this restaurant. Reviews frequently mention long lines and crowded conditions, which may not be worth the wait for some.
MAPLE MAPLE (+RAG)	i had the salted caramel and it was delicious You may be interested in the caramel offerings at Sweet Jesus. You may enjoy the salted caramel, which aligns with the restaurant's praised caramel and coconut soft serve. Reviews consistently highlight their delicious caramel-focused desserts like the sticky toffee pudding and caramel soft-serve , suggesting these might match your tastes.
xC-q_yh0XwcjRLimkS3RNng	raspberry puree and meringue crumble while the oh sweet baby jesus was caramel soft serve with peanut butter and caramel sauce
Att2seq NRT PETER PEPLER	it 's a good place to get a quick bite to eat the ice cream is good i 'm not a huge fan of the soft serve ice cream the ice cream was good but the texture was a little off
PRAG PRAG (+RAG)	wasn't, line, huge, nice You may be interested in the quick service and small seating area at Sweet Jesus. Although the restaurant is not large, reviews indicate that the line moves quickly and there is some outdoor seating, which might appeal to your preference for efficient service seen in your past reviews.
MAPLE MAPLE (+RAG)	i was craving something sweet so i ordered the chocolate chip pizookie which was a chocolate chip cookie You may be interested in the chocolate ice cream at Sweet Jesus. Reviews suggest it's exceptionally chocolatey and comes with unique toppings like marshmallows . This seems to align with your love for sweet but well-balanced desserts, as noted in your previous reviews of desserts.
flV6Cp3M2vHcqFBcCrGS1A	i would definitely recommend getting the cone in a cup as it does tend to be messier than your regular ice cream
Att2seq NRT PETER PEPLER	the ice cream is good but the ice cream is a bit too sweet for my liking the ice cream is a bit on the pricey side but the toppings are n't the ice cream was good but the ice cream wasn't that great the ice cream was good but the texture was a little off
PRAG PRAG (+RAG)	plus, wait, large, pretty You may be interested in the efficiency of the line at this restaurant. Despite the small size and often large crowds, reviews suggest that the line moves quickly and there's outdoor seating available. This might suit your preference for streamlined services, as seen in your past reviews.
MAPLE MAPLE (+RAG)	i've been here twice and i've tried the red velvet and the chocolate twist You may be interested in the " strawberry soft serve with caramel twist " at Sweet Jesus. This unique flavor combination has been noted for its deliciousness and distinctiveness, possibly aligning with your adventurous taste in ice cream flavors.

Table 7: Case studies on item id N_2yEZ41g9zDW_gWArFiHw (Sweet Jesus, an ice-cream shop) in Yelp19. There are a total of 35 reviews in the 10,000 test samples, 4 reviews are selected for demonstration. The recommended aspect terms mentioned in the explanations are boldfaced.

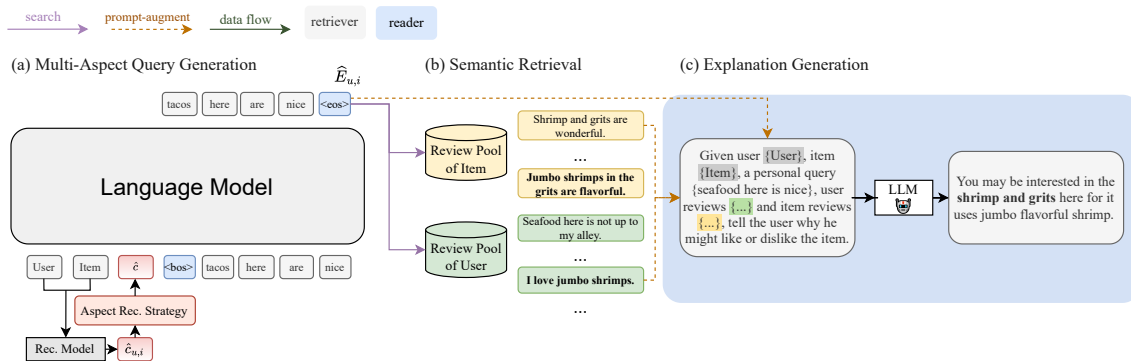


Figure 7: MAPLE and LLM in retriever-reader framework

User	Ground-Truth
m-2OHy6aP26oefPliiTocc	service is super quick but damn their food is good
Method	Explanation
Att2seq	le service est rapide et les serveurs sont
NRT	le service est rapide et les plats sont
PETER	le service est très agréable et le service est très agréable
PEPLER-MF	le service est the same .
PRAG	delish, duck, bbq, pork, sweet
PRAG (+RAG)	You may be interested in trying the roasted BBQ pork at this restaurant. Reviews mention it as "sweet and well roasted" and "delicious," aligning with your query interests in BBQ and delicious dishes. This choice could offer you a comforting and flavorful dining experience.
MAPLE	le menu est original , vous souhaitez manger un peu difficileux et authentiquiri avec la qualit et la qualit et les banqu simplement d'amoureux cocheureuse
MAPLE (+RAG)	You may be interested in the BBQ duck , praised in the reviews for being sweet and well-roasted. It aligns with your query's desire for unique, high-quality meals. The restaurant's ability to surprise with unexpectedly good dishes despite their simple descriptions might also appeal to your adventurous palate.

Table 8: Case studies on item id **OILK5_S-7svgSwbUfx1xYA** (King's Noodle Restaurant) in Yelp19. The recommended aspect terms mentioned in the explanations are boldfaced.

User	Ground-Truth
MG26Ips5arkOiFrvmmtcVg	the portions are small so you might want to order a side or appetizer with your entree
Method	Explanation
Att2seq	the food was great and the service was great
NRT	the food was good
PETER	the food is amazing
PEPLER-MF	the food is good
PRAG	spices, dish, amazing
PRAG (+RAG)	You may be interested in the food at this restaurant, particularly for its authentic Mexican flavors and well-loved dishes such as enchiladas and tacos . Reviews consistently highlight the deliciousness and unique preparation of their meals, aligning with your appreciation for varied and tasty dishes noted in your past reviews.
MAPLE	<p>this is a great place to relax and enjoy the beautiful Phoenix sun. (Restaurant Reviews)</p> <ul style="list-style-type: none"> - this is the place to go if you want some authentic Mexican food - a great local find - the food was excellent - pretty good tacos margaritas - and don't expect chips and salsa to arrive on your table when you sit down - it's not that kind of Mexican restaurant - it was fantastic; the enchiladas were nothing like traditional ones and were amazing. - unpretentious and positively delicious dishes prepared with love and taste like a practiced Mexican grandma labored in the kitchen to present you with an incredible meal - what a great find - what do you call two folded up pasty tortillas with unseasoned green sauce poured on top with an over abundance of crema and queso fresco - we paid the nice lady and the prices were decent <p>(User Reviews)</p> <ul style="list-style-type: none"> - this is going to be a popular spot in <u>Phoenix</u> - the food is the best in <u>Phoenix</u> and the service is always so friendly - very happy we found this hidden gem in Tempe - a perfect restaurant to bring the whole family - such a wonderful neighborhood restaurant - great place to try different dishes and they are all very good - a wonderful way to end your Sunday is with a campfire sundae! - the atmosphere however is a bit loud and their music makes it worse - make the trip and enjoy a wonderful breakfast with the family - the food was very good
MAPLE (+RAG)	You may be interested in the authentic Mexican food at Escobar Mexican Kitchen. Reviewers consistently praise the quality and traditional flavors, describing dishes as "unpretentious and positively delicious," reminiscent of a seasoned Mexican grandma's cooking. This aligns well with your preference for places to relax and enjoy good food under <u>the Phoenix sun</u> .

Table 9: Case studies on user id **MG26Ips5arkOiFrvmmtcVg** and item **Escobar Mexican Kitchen** in Yelp19. he recommended aspect terms mentioned in the explanations are boldfaced. The personalization details aligned in the explanation are underlined.