ELIXIR: Efficient and LIghtweight model for eXplaIning Recommendations

Ben Kabongo
Sorbonne University, CNRS, ISIR
Paris, France
ben.kabongo@sorbonne-universite.fr

Vincent Guigue
AgroParisTech, UMR MIA
Paris-Saclay
Palaiseau, France
vincent.guigue@agroparistech.fr

Pirmin Lemberger
Onepoint
Paris, France
p.lemberger@groupeonepoint.com

Abstract

Collaborative filtering drives many successful recommender systems but struggles with fine-grained user-item interactions and explainability. As users increasingly seek transparent recommendations, generating textual explanations through language models has become a critical research area. Existing methods employ either RNNs or Transformers. However, RNN-based approaches fail to leverage the capabilities of pre-trained Transformer models, whereas Transformer-based methods often suffer from suboptimal adaptation and neglect aspect modeling, which is crucial for personalized explanations. We propose ELIXIR (**E**fficient and **LI**ghtweight model for e**X**pla**I**ning **R**ecommendations), a multi-task model combining rating prediction with personalized review generation. ELIXIR jointly learns global and aspect-specific representations of users and items, optimizing overall rating, aspect-level ratings, and review generation, with personalized attention to emphasize aspect importance. Based on a T5-small (60M) model, we demonstrate the effectiveness of our aspect-based architecture in guiding text generation in a personalized context, where state-of-the-art approaches exploit much larger models but fail to match user preferences as well. Experimental results on TripAdvisor and RateBeer demonstrate that ELIXIR significantly outperforms strong baseline models, especially in review generation.

CCS Concepts

• **Information systems** → **Recommender systems**; • **Computing methodologies** → **Natural language generation**.

Keywords

Recommender Systems, Large Language Models, Explanation Generation, Aspect-based Recommendation, Neural Attention, Prompt tuning

ACM Reference Format:
Ben Kabongo, Vincent Guigue, and Pirmin Lemberger. 2025. ELIXIR: Efficient and LIghtweight model for eXplaIning Recommendations. In *Proceedings of (RecSys EARL '25)*. ACM, New York, NY, USA, 10 pages. https://doi.org/XXXXXXX.XXXXXXX

RecSys EARL '25, Prague, Czech Republic

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

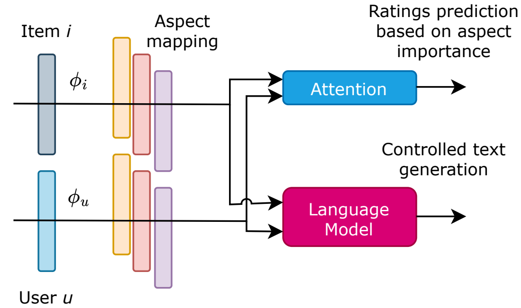


Figure 1: Schematic overview of our approach: starting from user (u) and item (i) inputs, aspect mapping extracts fine-grained features which are then weighted using a personalized attention mechanism. These refined representations drive both accurate ratings prediction—by emphasizing aspect importance—and controlled review generation via a language model.

Many successful recommendation systems rely on collaborative filtering (CF), which learns user preferences and item characteristics from interaction data [22, 24, 37]. Among CF methods, latent factor models based on matrix factorization (MF) [19, 23, 33] have shown strong performance. However, these models primarily rely on latent user and item representations, limiting their ability to capture fine-grained interactions and resulting in a lack of interpretability and explainability.

Beyond accuracy, explainability has become a central focus in recommender system research. Users increasingly expect not only relevant recommendations but also transparent justifications for these recommendations [13, 52]. This has led to a growing interest in leveraging language models to generate textual explanations [14, 27, 35], including tips [28] and reviews [10, 48]. The underlying assumption is that user-generated reviews encapsulate both overall sentiment and fine-grained opinions on specific aspects of an item, making them natural candidates for explainable recommendations. Several studies have explored explanation generation using recurrent neural networks (RNNs) [9, 10, 28, 36] and Transformers [5, 26, 27, 43]. However, RNN-based approaches are penalized by the lack of pre-training [3, 11, 41]. Meanwhile, existing Transformer-based methods often show limited performance gains due to suboptimal adaptation strategies. Furthermore, most of these methods often overlook aspect modeling, which is essential for capturing fine-grained user preferences and guiding text generation to combine personalization, relevance and fighting hallucinations. Aspect-based recommender systems (ABRS) already make it possible to model users' opinions on specific aspects of an item. Aspects are mainly extracted from reviews [7, 8, 15, 17] and processed by aspect-based sentiment analysis (ABSA) [30, 51] to build more detailed profiles [2, 4, 12, 18, 53]. More generally, aspect modeling enables us to refine a user's overall assessment of an item while detailing their opinions on specific aspects, which constitutes a first level of explanation. But this modeling also allows us to improve the textual explanations generated by incorporating fine-grained information specific to each aspect [35, 45].

In this paper, we introduce **ELIXIR** (**E**fficient and **LI**ghtweight model for e**X**pla**I**ning **R**ecommendations), a multi-task model, composed of two modules: the rating prediction module and the personalized review generation module. Our approach learns both global and aspect-based representations for each user and item from interaction data, optimizing three predictive objectives: overall rating prediction, aspect ratings prediction, and user review generation. We have built an original personalized attention mechanism to estimate the importance of different aspects for each user and each item, in addition to affinities. The review generation module derives a personalized continuous prompt from user and item representations, which is then fed into a language model to generate the review. Our approach is particularly parameter-efficient, starting from a mainly frozen pre-trained model where only the continuous prompt from user and item profiles is optimized. Detailed aspect profiles allow us to generate a complete, personalized explanation of the item recommendation through prompt tuning, a particularly parameter-efficient refining method [16]. An illustration of our approach is shown in Figure 1.

Experiments conducted on two real-world multi-aspect datasets, TripAdvisor and RateBeer with a T5-small language model (60M) [41], show that ELIXIR outperforms the state of the art yet based on LLMs several orders of magnitude larger. Ablation studies and empirical analyses further highlight the contributions of personalized attention and aspect modeling. Our contributions are as follows:

- We propose a unified architecture for aspect-based user and item profile modeling, for overall and aspect-specific ratings prediction, and review generation.
- We introduce a parameter-efficient generative architecture that outperforms the state-of-the-art in personalized review generation. Our prompt tuning approach on a small pre-trained language model demonstrates its effectiveness in controlling text generation.
- We conduct experiments and analyses on two multi-aspect real-world datasets. Our results demonstrate the superiority of our model in both rating prediction and review generation, as well as its potential for adaptation to datasets without explicit aspect annotations.

2 RELATED WORK

Recommender systems are among the early applications of representation learning, particularly through collaborative filtering (CF) [23, 37]. CF-based models capture user preferences by identifying similarities among users or items derived from interaction data. Various architectures have been proposed, ranging from matrix factorization [23, 33] to deep learning methods [19, 42]. However, these methods face challenges such as the sparsity of the interaction matrix and the cold start problem, where limited interactions hinder recommendations for new users or items. Content-based (CB) recommender models [20, 31] mitigate this issue by incorporating item characteristics, although they tend to recommend only similar items. Hybrid methods combine different recommendation techniques to overcome the limitations of individual approaches. Integrating additional information about users or items—such as item attributes [7, 8] or reviews [44, 45]—can significantly enhance the performance of CF-based models.

More specifically, aspect-based recommendation systems (ABRS) improve personalization by modeling user preferences and item characteristics at a finer level of granularity through specific attributes or aspects. A first category of ABRS methods extracts aspects from reviews in an unsupervised manner. For example, ALFM [7] uses an aspect-based topic model (ATM) to learn a multivariate distribution of topics from the reviews, while ANR [8] learns aspect representations for users and items by employing attention mechanisms to focus on the most relevant parts of the reviews. On the other hand, a second category of ABRS methods focuses on opinion analysis [2, 4], taking advantage of aspect-based sentiment analysis (ABSA) techniques [51] to extract opinions on various aspects from reviews. The key advantage of these methods lies in their greater interpretability and explicability compared to conventional recommendation approaches.

Explainability has become a central focus in research on recommender systems, with growing interest in the exploitation of language models to generate textual explanations [10, 26–28, 48]. The first methods were based on recurrent neural networks (RNNs), Att2Seq [10] and NRT [28] being notable examples. The latter, in particular, is a multi-task model that generates explanations while simultaneously predicting the overall rating. Other RNN-based methods integrate aspect modeling to better guide explanation generation [35, 36, 45]. Subsequently, several explanation generation models took advantage of Transformers. One of the first Transformers-based multi-task models is PETER [26], which integrates heterogeneous data, including word, user, and item embeddings, for overall rating prediction and explanation generation. PEPLER [27] is derived from PETER and also uses a pre-trained Transformer, GPT-2 [40]. This architecture has been extended [5, 39, 43]. Despite these advances, performance gains over RNN-based methods remain modest, particularly for generating long explanations. This reflects the limitations of previous Transformers adaptations for generating personalized explanations, as most of these methods neglect aspect modeling. Integrating aspect-specific information improves the quality of the generated explanations [35, 45]. To address this challenge, we propose an approach that combines prompt tuning [25] with the integration of aspect-aware information to better guide the generation process and improve the quality of personalized explanations.

3 ELIXIR

We propose **ELIXIR** (**E**fficient and **LI**ghtweight model for e**X**pla**I**ning **R**ecommendations), a multi-task model designed to predict the overall rating, aspect ratings, and personalized review for a given user-item pair. ELIXIR consists of two core components: the **rating prediction module** and the **personalized review generation module**. Figure 2 provides an overview of the complete ELIXIR architecture.

3.1 Notations and Problem Formulation

Let $\mathcal{U}$ be the set of users and $\mathcal{I}$ the set of items. We denote $|\mathcal{U}|$ and $|\mathcal{I}|$ the number of users and items, respectively. Let $\mathcal{R}$ be the set of interactions between users and items. For an interaction between a user u and an item i, let r_{ui} denote the overall rating and t_{ui} the review provided by user u for item i. The review t_{ui} is represented as a sequence of tokens, $t_{ui} = (y_1, y_2, \ldots, y_{|t_{ui}|})$, where each token y_k belongs to the vocabulary set $\mathcal{V}$. Given an application domain, let $\mathcal{A}$ denote the set of aspects of interest, and let $|\mathcal{A}|$ represent the number of aspects. Aspect ratings for an item i provided by a user u can either be explicitly available or extracted from the review t_{ui} using aspect-based sentiment analysis (ABSA) methods [30, 50, 51]. We denote r_{ui}^a as the rating assigned by user u to aspect a of item i. The interaction set is defined by:

$$\mathcal{R} = \left\{ \left(u, i, r_{ui}, t_{ui}, \{r_{ui}^a\}_{a\in\mathcal{A}}\right) \right\}. \tag{1}$$

In this article, we focus on an aspect-supervised framework to demonstrate the interest of our generated text control methodology. However, we assume that ABSA techniques have progressed sufficiently in recent years to reasonably consider moving to an unsupervised framework.

Our goal is to learn user and item representations from interaction data, capturing both global and aspect-level interactions to better model overall preferences and fine-grained interests. Our approach represents user preferences and item characteristics at both levels, using personalized attention to assess the importance of each aspect for individual users and items. We achieve this through three unified tasks: overall rating prediction, aspect-specific rating prediction, and personalized review generation, ensuring both high performance and explained recommendations.

3.2 Model

ELIXIR is composed of two key modules: one for rating prediction and the other for text generation. The first module relies on learned representations at two levels (global –$\mathbf{u}$, $\mathbf{i}$– and aspect –$\{\mathbf{a}_u\}_{a\in\mathcal{A}}$ and $\{\mathbf{a}_i\}_{a\in\mathcal{A}}$–) and an original personalized attention mechanism to link aspects in the rating prediction. These representations serve as a bridge to the generative module: they are used to build a continuous prompt $\mathbf{p}_{ui}$ to control the generation of personalized text. We denote θ as the set of model parameters.

3.2.1 Global and Aspect Levels Representations.

Global Level. We learn global representations, $\mathbf{u} \in \mathbb{R}^d$ for each user $u \in \mathcal{U}$, and $\mathbf{i} \in \mathbb{R}^d$ for each item $i \in \mathcal{I}$, which encode various information, including user preferences, item characteristics, and other relevant contextual factors.

Aspect Level. From the global representations, we aim to extract the preferences of user u and the characteristics of item i concerning aspect a. To achieve this, we define two functions, $\phi_{\mathcal{U}}^a, \phi_{\mathcal{I}}^a : \mathbb{R}^d \to \mathbb{R}^d$, which transform the global representations $\mathbf{u}$ and $\mathbf{i}$ into aspect-specific representations, denoted as $\mathbf{a}_u$ and $\mathbf{a}_i$, respectively. These aspect-aware representations are obtained as follows:

$$\mathbf{a}_u = \phi_{\mathcal{U}}^a(\mathbf{u}), \quad \mathbf{a}_i = \phi_{\mathcal{I}}^a(\mathbf{i}), \quad \mathbf{a}_u, \mathbf{a}_i \in \mathbb{R}^d. \tag{2}$$

Our experiments led us to choose non-linear projections (MLPs) for $\{\phi_{\mathcal{U}}^a, \phi_{\mathcal{I}}^a\}_{a\in\mathcal{A}}$. We denote $(\theta_{\mathcal{U}}, \theta_{\mathcal{I}}, \theta_{\mathcal{A}})$ as the set of model parameters corresponding to users, items, and aspects, respectively.

3.2.2 *Personalized Attention.*

In personalized recommendation systems, the importance of each aspect varies significantly depending on the user's preferences and the item's characteristics [6–8]. We model these variations in aspect importance using personalized attention, which estimates the weight of each aspect for each user and each item, inspired by attention mechanism from [46]. Given a user u, we estimate the relative importance weights of aspects, denoted by $\{\alpha_u^a\}_{a\in\mathcal{A}}$, using three linear mapping functions, $q_{\mathcal{U}}, k_{\mathcal{U}}, v_{\mathcal{U}} : \mathbb{R}^d \to \mathbb{R}^d$. Similarly, for an item i, we apply three linear mapping functions, $q_{\mathcal{I}}, k_{\mathcal{I}}, v_{\mathcal{I}} : \mathbb{R}^d \to \mathbb{R}^d$ to compute the aspect attention weights $\{\alpha_i^a\}_{a\in\mathcal{A}}$. The importance weights of aspect a for user u and item i, respectively denoted by α_u^a and α_i^a, are computed as follows:

$$\alpha_u^a = \frac{\exp(q_{\mathcal{U}}(\mathbf{u}) \cdot k_{\mathcal{U}}(\mathbf{a}_u))}{Z_u}, \quad \alpha_i^a = \frac{\exp(q_{\mathcal{I}}(\mathbf{i}) \cdot k_{\mathcal{I}}(\mathbf{a}_i))}{Z_i}, \tag{3}$$

where Z_u and Z_i are normalization terms. For a user, the aspect attention weights reflect the relative importance the user assigns to each aspect. For an item, these weights represent the average importance attributed to each aspect by users.

We then compute aggregated representations for the user u and the item i, denoted as $\tilde{\mathbf{u}}$ and $\tilde{\mathbf{i}}$, respectively. These representations dynamically capture the user's preferences and the item's characteristics across various aspects. They are obtained as follows:

$$\tilde{\mathbf{u}} = \sum_{a\in\mathcal{A}} \alpha_u^a v_{\mathcal{U}}(\mathbf{a}_u), \quad \tilde{\mathbf{i}} = \sum_{a\in\mathcal{A}} \alpha_i^a v_{\mathcal{I}}(\mathbf{a}_i), \quad \tilde{\mathbf{u}}, \tilde{\mathbf{i}} \in \mathbb{R}^d. \tag{4}$$

3.2.3 *Rating Prediction Module.*

The objective of the rating prediction module is to estimate both the overall rating r_{ui} and the aspect-specific ratings $\{r_{ui}^a\}_{a\in\mathcal{A}}$ for a user u and an item i.

Overall Rating Prediction. To predict the overall rating r_{ui} of user u for item i, we define a function $f : \mathbb{R}^{2\times d} \to \mathbb{R}$. The inputs to this function are the aggregated representations of the user and the item, denoted $\tilde{\mathbf{u}}$ and $\tilde{\mathbf{i}}$, respectively. These representations encapsulate the essential user preferences and item characteristics while dynamically integrating the relative importance of different aspects. The overall rating is predicted as follows:

$$\hat{r}_{ui} = f(\tilde{\mathbf{u}}, \tilde{\mathbf{i}}). \tag{5}$$

Aspect Rating Prediction. To predict the aspect-specific rating r_{ui}^a, which represents user u's evaluation of aspect a for item i, we define a function $g_a : \mathbb{R}^{2\times d} \to \mathbb{R}$. The inputs to this function are the representations of aspect a for the user and the item, denoted $\mathbf{a}_u$ and $\mathbf{a}_i$, respectively. These representations capture fine-grained user preferences and item characteristics for aspect a. The prediction of the aspect rating is given by:

$$\hat{r}_{ui}^a = g_a(\mathbf{a}_u, \mathbf{a}_i). \tag{6}$$

We denote θ_R as the set of parameters specific to the rating prediction module. Our experiments have led us to take non-linear functions (MLPs) for f and $\{g_a\}_{a\in\mathcal{A}}$.

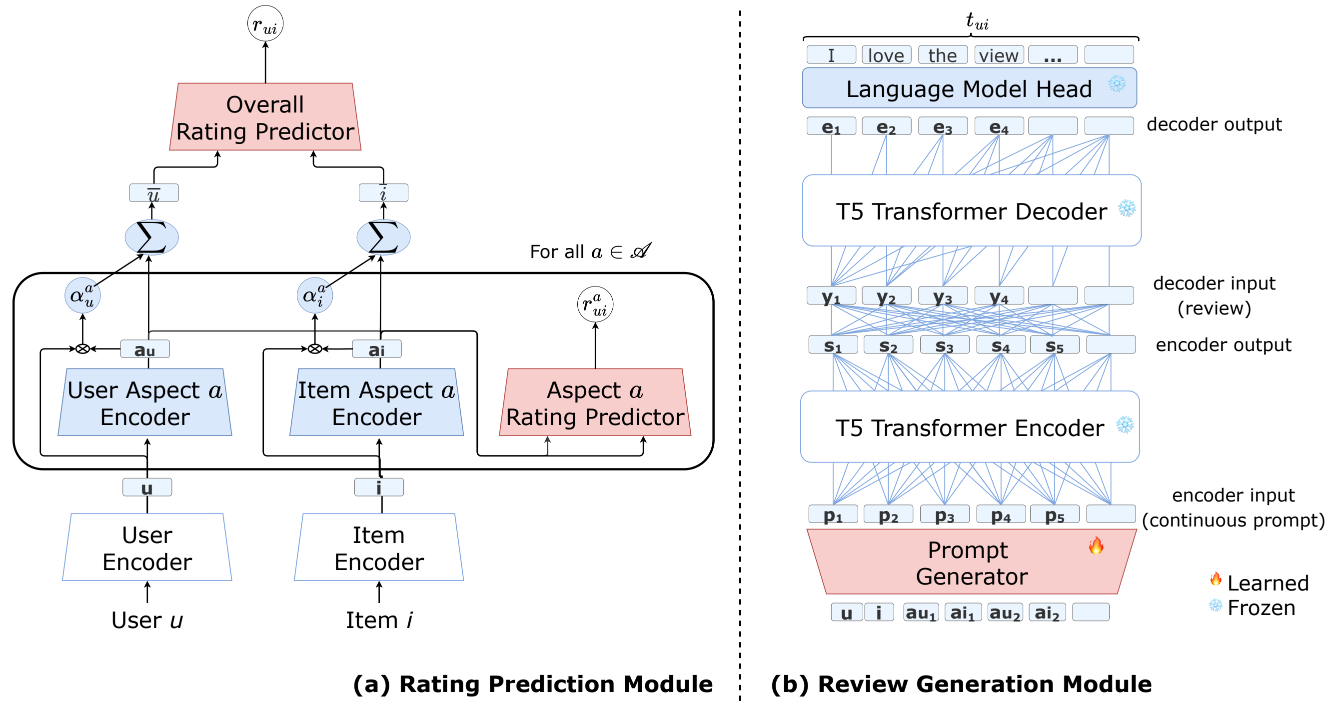


Figure 2: ELIXIR consists of two modules: (a) the rating prediction module on the left and (b) the personalized review generation module on the right. The model learns global representations for both the user and the item, and derives representations for each aspect $a \in \mathcal{A}$. The rating for aspect a, denoted r_{ui}^a, is obtained from the corresponding user and item aspect representations. The aspect representations are aggregated via personalized attention for both the user and the item to yield the overall rating r_{ui}. Finally, the complete set of global and aspect representations is used to generate a continuous prompt, which is fed to a pre-trained language model (T5) to produce the personalized review t_{ui}.

3.2.4 Personalized Review Generation. The generative module aims to construct a personalized text explaining recommendations, learned from users' past reviews. Our parameter-efficient approach operates in two steps: (1) generate $\mathbf{p}_{ui}$, a set of continuous representations from the aspect profiles (user and item); (2) optimize these token representations (prompt tuning) to match the target text t_{ui} while keeping the language model frozen. We denote θ_P as the set of model parameters for prompt generation.

Personalized Prompt Generation. From the global representations $\mathbf{u}$ and $\mathbf{i}$ of user u and item i, along with their aspect representations $\{\mathbf{a}_u, \mathbf{a}_i\}_{a\in\mathcal{A}}$, we employ a function $\psi : \mathbb{R}^{2(1+|\mathcal{A}|)\times d} \longrightarrow \mathbb{R}^{\eta\times d_w}$ to obtain the continuous personalized prompt. Here, η is a hyperparameter of the model that depends on the language model θ_{LM}. Our experiments with T5-Small have shown that 50 tokens effectively guide generation. Our hypothesis is that, in addition to the global representations, the user preferences and item characteristics encapsulated in the aspect-aware representations contribute to more effective guidance for the review generation process. The personalized prompt, denoted $\mathbf{p}_{ui} \in \mathbb{R}^{\eta\times d_w}$, consists of η tokens in the latent space of the language model θ_{LM} of dimension d_w. This prompt captures the same information as the global and aspect representations of the user and item, and is computed as follows:

$$\mathbf{p}_{ui} = \psi\big(\mathbf{u}, \mathbf{i}, \{\mathbf{a}_u, \mathbf{a}_i\}_{a\in\mathcal{A}}\big). \tag{7}$$

Review Generation. After obtaining the continuous personalized prompt $\mathbf{p}_{ui}$ for user u and item i, we employ a language model θ_{LM} to generate the review t_{ui}, conditioned on the prompt. In our experiments, we chose the T5 pre-trained model for θ_{LM}. The generation of the review t_{ui} for user u and item i, conditioned on the personalized prompt $\mathbf{p}_{ui}$, is formulated as follows:

$$\begin{aligned} P_{\theta_P,\theta_{LM}}(t_{ui}|\mathbf{p}_{ui}) &= P_{\theta_P,\theta_{LM}}\big((y_1, y_2, \ldots, y_{|t_{ui}|})|\mathbf{p}_{ui}\big) \\ &= \prod_{k=1}^{|t_{ui}|} P_{\theta_P,\theta_{LM}}(y_k|\mathbf{p}_{ui}, y_{<k}). \end{aligned} \tag{8}$$

3.3 Optimization

In this section, we discuss the optimization of the model, covering the loss functions and the training procedure.

3.3.1 Loss Functions. The set of model parameters, denoted as $\theta = (\theta_{\mathcal{U}}, \theta_{\mathcal{I}}, \theta_{\mathcal{A}}, \theta_R, \theta_P, \theta_{LM})$, is optimized by minimizing three loss functions corresponding to the three objectives: overall rating prediction, aspect ratings prediction, and review generation.

For overall rating prediction, we employ the mean squared error (MSE) loss, defined as:

$$\mathcal{L}_R = \frac{1}{|\mathcal{R}|} \sum_{(u,i)\in\mathcal{R}} \left(r_{ui} - \hat{r}_{ui}\right)^2. \tag{9}$$

For aspect ratings prediction, we use the average of the mean squared error (MSE) losses computed for each aspect:

$$\mathcal{L}_A = \frac{1}{|\mathcal{A}|} \sum_{a\in\mathcal{A}} \frac{1}{|\mathcal{R}|} \sum_{(u,i)\in\mathcal{R}} \left(r^a_{ui} - \hat{r}^a_{ui}\right)^2. \tag{10}$$

The total loss of the rating prediction module is a weighted combination of these two losses:

$$\mathcal{L}_{rating} = \alpha\mathcal{L}_R + (1-\alpha)\mathcal{L}_A, \tag{11}$$

where α is an hyperparameter that control the relative importance of the overall rating loss ($\mathcal{L}_R$) and the aspect ratings loss ($\mathcal{L}_A$).

For personalized review generation, we use the negative log-likelihood (NLL) loss, given by:

$$\mathcal{L}_{review} = -\frac{1}{|\mathcal{R}|} \sum_{(u,i)\in\mathcal{R}} \frac{1}{|t_{ui}|} \sum_{k=1}^{|t_{ui}|} \log P_{\theta_P,\theta_{LM}}\left(y_k \mid \mathbf{p}_{ui}, y_{<k}\right). \tag{12}$$

3.3.2 Training. Our method naturally aligns with prompt tuning [25], allowing us to effectively leverage the pre-trained knowledge. In this configuration, we freeze the language model parameters θ_{LM} during training and only optimize the prompt generation parameters θ_P. This strategy significantly reduces the number of trainable parameters, thus preserving the rich knowledge encoded within the pre-trained model. To maximize efficiency in this setting, we adopt a sequential training approach. Since the personalized prompt depends directly on global and aspect-aware user and item representations, we first optimize these representations by training the rating prediction module and minimizing the loss $\mathcal{L}_{rating}$. Once the learned representations are sufficiently informative, we then train only the prompt generation parameters in the personalized review generation module by minimizing the review generation loss $\mathcal{L}_{review}$, keeping the language model parameters frozen.

4 EXPERIMENTS

4.1 Implementation

We implement key functions in ELIXIR[1] using multi-layer perceptrons (MLPs). These include f for the overall rating prediction, $\{g_a\}_{a\in\mathcal{A}}$ for the aspect ratings prediction, $\{\phi^a_{\mathcal{U}}, \phi^a_{\mathcal{I}}\}_{a\in\mathcal{A}}$ for the aspect representations, and ψ for the personalised prompt generation. We conducted preliminary experiments by implementing the functions f and $\{g_a\}_{a\in\mathcal{A}}$ with a scalar product and the functions $\{\phi^a_{\mathcal{U}}, \phi^a_{\mathcal{I}}\}_{a\in\mathcal{A}}$ and ψ with linear projectors. However, the performance of the model is inferior compared to the performance obtained by implementing these functions with MLPs, suggesting that non-linearity is necessary to translate the different inputs into the different semantic spaces. For personalized user attention $(q_{\mathcal{U}}, k_{\mathcal{U}}, v_{\mathcal{U}})$ and item attention $(q_{\mathcal{I}}, k_{\mathcal{I}}, v_{\mathcal{I}})$ functions, we learn separate projection matrices. For example, for the function $q_{\mathcal{U}}$, we use a projection matrix $\mathbf{W}^q_{\mathcal{U}} \in \mathbb{R}^{d\times d}$. The generation of the review is given by $P_{\theta_P,\theta_{LM}}(t_{ui}|\mathbf{p}_{ui})$. In our implementation of ELIXIR, we utilize the pre-trained Transformer-based T5 language model [41], and our experiments follow the prompt tuning approach within a sequential training setup. The hyperparameter η, which specifies the number of prompt tokens, depends on the language model, as demonstrated in [25].

4.2 Ablations

We introduce various ablations of the ELIXIR model to assess the contribution of individual components, such as personalized attention and aspect modeling.

4.2.1 Personalized Attention. We denote *ELIXIR -Attention* as the ablation of the ELIXIR model in which personalized attention is replaced with max pooling. The aggregated aspect representations of user u and item i are computed using a max pooling operation over all their aspect representations, as follows:

$$(\tilde{\mathbf{u}})_j = \max_{a\in\mathcal{A}}(\mathbf{a}_u)_j, \quad (\tilde{\mathbf{i}})_j = \max_{a\in\mathcal{A}}(\mathbf{a}_i)_j, \quad \forall j \in \{1,\ldots,d\}. \tag{13}$$

4.2.2 Aspect Modeling. We denote *ELIXIR -Aspects* as the ablation of the ELIXIR model in which aspect modeling is removed. In this variant, the model does not learn aspect-aware representations for users and items, and we also omit aspect ratings prediction and personalized attention. For overall rating prediction, we redefine the function $f : \mathbb{R}^{2\times d} \to \mathbb{R}$, and for personalized prompt generation, we redefine the function $\psi : \mathbb{R}^{2\times d} \to \mathbb{R}^{\eta\times d_w}$. Since aspect modeling is omitted, these functions take as input only the global representations of the user and item, as follows:

$$\hat{r}_{ui} = f(\mathbf{u}, \mathbf{i}), \quad \mathbf{p}_{ui} = \psi(\mathbf{u}, \mathbf{i}). \tag{14}$$

4.2.3 Global Representations. We denote *ELIXIR -Global* as an ablation of the ELIXIR model where overall rating prediction, aspect ratings prediction, and personalized continuous prompt generation rely solely on the global representations of users and items. This ablation allows us to assess the combined importance of aspect modeling and personalized attention across all model tasks. In this model ablation, the functions $f : \mathbb{R}^{2\times d} \to \mathbb{R}$, $g_a : \mathbb{R}^{2\times d} \to \mathbb{R}$, and $\psi : \mathbb{R}^{2\times d} \to \mathbb{R}^{\eta\times d_w}$ for overall rating prediction, aspect a rating prediction, and prompt generation, are defined as follows:

$$\hat{r}_{ui} = f(\mathbf{u}, \mathbf{i}), \quad \hat{r}^a_{ui} = g_a(\mathbf{u}, \mathbf{i}), \quad \mathbf{p}_{ui} = \psi(\mathbf{u}, \mathbf{i}). \tag{15}$$

4.3 Experimental Setup

4.3.1 Datasets. Our experiments are conducted on two real-word multi-aspect datasets:

- **TripAdvisor** [2] [47]: includes reviews covering six aspects of hotels: cleanliness, location, room, service, sleep quality, and value.
- **RateBeer** [3] [32]: includes reviews focusing on four aspects of beers: appearance, aroma, palate, and taste.

For each dataset, we exclude interactions with missing information and retain only users and items that have at least 5 reviews. Table 1 provides a summary of the datasets before and after applying these filtering steps. Although the number of datasets is limited, we emphasize the diversity of the topics covered. We are confident

[1]Our implementation is available here https://github.com/BenKabongo25/aspect_explainable_recommender.

[2]https://www.cs.virginia.edu/~hw5x/Data/LARA/TripAdvisor/

[3]https://cseweb.ucsd.edu/~jmcauley/datasets.html#multi_aspect

Table 1: Datasets description. Statistics after filtering are shown in brackets.

Datasets	TripAdvisor	RateBeer
Aspects	6	4
Users	716 870 (8 830)	40 213 (8 384)
Items	10 008 (2 903)	110 419 (5 093)
Interactions	1.4M (62.7K)	2.8M (201.7K)

in the possibility of rapidly extending these experiments to aspect-unsupervised data using ABSA techniques.

4.3.2 Evaluation metrics. We use the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) metrics to evaluate the models on rating prediction. For review generation, we evaluate the models using text quality metrics, including METEOR [1], BLEU [38], ROUGE [29], and BERTScore [49].

4.3.3 Baselines. We compare ELIXIR against a diverse set of baselines, including traditional rating prediction methods, aspect-based recommendation models, explanation generation approaches, and multi-task models. To evaluate the contribution of the different components of ELIXIR, we also consider its various ablations, including ELIXIR -Attention, ELIXIR -Aspects, and ELIXIR -Global.

Ratings prediction. We employ the following baselines: Average, MF [23], MLP [19], and NeuMF [19]. The Average method predicts the mean rating across all user-item pairs. For aspect-based recommendation, we evaluate two baselines: ALFM [7] and ANR [8]. These methods learn representations of aspects from reviews in an unsupervised manner, making it challenging to align the learned aspects with those in the dataset. We consider the ablations ELIXIR -Global and ELIXIR -Attention as additional baselines specifically for aspect ratings prediction.

Review generation. We consider two categories of explanation generation models. The first category consists of RNN-based models, including Att2Seq [10] and the multi-task architecture NRT [28]. The second category includes multi-task Transformer-based models, such as PETER [26] and PEPLER [27]. PETER uses an unpretrained Transformer, while PEPLER employs the pre-trained GPT-2 [40] in a fine-tuning approach. Most of the text-based explanation generation methods considered rely on latent representations of users and items—typically encoded as just two tokens—to condition the entire explanation. We exclude models such as those proposed in [5, 39, 43], which are extensions of PETER and PEPLER, as their performance closely mirrors that of the original models.

4.3.4 Setup. Each dataset is split into training, validation, and test sets using an 80:10:10 ratio. All ratings, including overall and aspect ratings, are standardized to a scale of 1 to 5. Review lengths are capped at 128 words for all models. Models are trained on the training set, and hyperparameters are tuned using the validation set. The results reported correspond to the performance of the models on the test set. For most models, we use the hyperparameters specified in the original papers. Specifically, for PEPLER [27], we use GPT-2 (124M) [40] as detailed in the original paper. For ELIXIR, we use the T5-Small model (60M) [41], which is half the size of GPT-2. After preliminary experiments, we set η, the number of tokens in the personalized prompt, to 50. The model dimension, d, is set to 256, while the T5-Small model's dimension is 512 [41]. The number of layers in the MLPs is fixed at 2, and we use the Rectified Linear Unit (ReLU) [34] as the activation function. The dropout probability is set to 0.1 to prevent overfitting. We find that setting $\alpha = \frac{1}{|\mathcal{A}|+1}$, where $|\mathcal{A}|$ represents the number of aspects, strikes the best balance between performance on overall and aspect ratings. All models are trained for a maximum of 100 epochs. For ELIXIR, we employ sequential training and prompt tuning. The rating prediction module is trained for 50 epochs, followed by another 50 epochs for training the review generation module. Training is performed using the Adam [21] optimizer with a learning rate of 10^{-3}.

Table 2: Performance on overall rating prediction.

	TripAdvisor		RateBeer	
Model	RMSE ↓	MAE ↓	RMSE ↓	MAE ↓
Average	0.932	0.645	0.571	0.424
MF	0.840	0.646	**0.411**	**0.300**
MLP	<u>0.833</u>	<u>0.565</u>	0.464	0.324
NeuMF	0.840	0.570	0.473	0.329
ALFM	0.896	0.691	0.433	0.314
ANR	<u>0.847</u>	<u>0.607</u>	<u>0.423</u>	<u>0.308</u>
NRT	0.859	0.548	0.420	0.306
PETER	0.807	0.532	<u>0.415</u>	**0.300**
PEPLER	<u>0.779</u>	<u>0.478</u>	0.430	0.305
ELIXIR	**0.748**	**0.447**	<u>0.416</u>	<u>0.305</u>
-Attention	0.771	0.513	0.421	0.311
-Global	0.865	0.632	0.443	0.332

4.4 Ratings prediction

Overall rating. We compare ELIXIR to various baseline models for overall rating prediction, with results reported in Table 2. ELIXIR consistently outperforms all baselines across all metrics, particularly on the TripAdvisor dataset. On the RateBeer dataset, it remains among the top-performing models, with performance close to MF and PETER. The superiority of ELIXIR over traditional and multi-task approaches can be attributed to its integration of aspect modeling. Likewise, its advantage over aspect-based baselines highlights the benefits of supervised modeling, especially through directly extracting opinions on aspects from reviews. The ablations ELIXIR -Attention and ELIXIR -Global perform worse than the full version of ELIXIR, confirming the importance of personalized attention and the joint use of global and aspect-aware representations for achieving the highest accuracy.

Aspect ratings. We compare ELIXIR with the Average baseline and the ablations ELIXIR -Global and ELIXIR -Attention for aspect ratings prediction. For each model, we calculate the RMSE and MAE for all aspects and then report the mean and standard deviation of these metrics across all aspects. The results are presented in Table 3. ELIXIR consistently and significantly outperforms all other

Table 3: Performance on aspect ratings prediction. We report the mean and standard deviation of all aspects.

	TripAdvisor		RateBeer	
Model	RMSE ↓	MAE ↓	RMSE ↓	MAE ↓
Average	1.014 (0.087)	0.801 (0.057)	0.605 (0.011)	0.489 (0.023)
ELIXIR	**0.753 (0.081)**	**0.451 (0.053)**	**0.465 (0.034)**	**0.354 (0.030)**
-Attention	0.785 (0.073)	0.554 (0.051)	0.486 (0.031)	0.373 (0.030)
-Global	0.860 (0.076)	0.631 (0.056)	0.495 (0.035)	0.383 (0.031)

Table 4: Performance on review generation (%).

TripAdvisor	METEOR ↑	BLEU ↑	ROUGE-1 ↑	ROUGE-2 ↑	ROUGE-L ↑	BERT-P ↑	BERT-R ↑	BERT-F1 ↑
Att2Seq	18.611	04.690	28.783	06.473	18.523	85.348	83.676	84.490
NRT	17.219	03.405	25.833	05.194	17.539	82.828	81.533	82.161
PETER	17.955	03.943	27.974	05.906	18.252	85.037	83.823	84.406
$\text{PEPLER}_{\text{GPT-2}}$	24.340	11.400	33.831	11.679	22.452	82.635	84.945	83.726
$\text{ELIXIR}_{\text{T5-Small}}$	**42.752**	**33.544**	**53.285**	**37.878**	**44.053**	**90.686**	**88.478**	**89.554**
-Aspects	27.642	10.029	39.076	21.970	29.594	88.001	85.193	86.540
RateBeer	METEOR ↑	BLEU ↑	ROUGE-1 ↑	ROUGE-2 ↑	ROUGE-L ↑	BERT-P ↑	BERT-R ↑	BERT-F1 ↑
Att2Seq	18.611	04.690	28.783	06.473	18.523	85.348	83.676	84.490
NRT	24.963	08.737	32.589	11.472	26.629	85.046	82.992	83.985
PETER	28.818	11.518	35.504	13.620	29.668	87.340	85.621	86.448
$\text{PEPLER}_{\text{GPT-2}}$	28.266	10.143	32.444	11.182	26.248	84.020	86.063	84.990
$\text{ELIXIR}_{\text{T5-Small}}$	**40.763**	**24.160**	**46.371**	**25.818**	**39.461**	**90.483**	**89.135**	**89.792**
-Aspects	32.675	13.652	39.068	17.106	32.464	89.365	87.323	88.310

Table 5: Impact of the number of tokens in the personalized prompt (η) on review generation (TripAdvisor dataset).

η	METEOR ↑	BLEU ↑	ROUGE-2 ↑
PEPLER	24.340	11.400	11.679
2	12.159	01.282	04.436
5	16.730	03.546	06.046
10	21.160	07.692	10.330
20	<u>29.379</u>	<u>17.018</u>	<u>20.200</u>
50	**42.752**	**33.544**	**37.878**

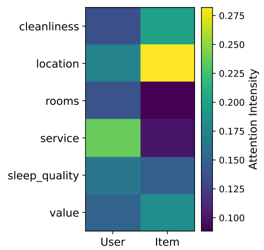
methods across all metrics and datasets. Both ablations surpass the Average baseline. We observe that the ELIXIR -Global ablation performs worse than ELIXIR -Attention. This performance gap arises because ELIXIR -Global relies solely on global representations, whereas ELIXIR -Attention retains aspect-specific representations. This clearly demonstrates the value of explicitly learning aspect-aware representations for accurate aspect ratings prediction. Moreover, the full ELIXIR model consistently outperforms the ELIXIR -Attention ablation, which replaces personalized attention with a max pooling operation. This further confirms that personalized attention provides a more effective aggregation strategy for user and item aspect representations than max pooling. Finally, an additional benefit of ELIXIR over traditional rating prediction or multi-task baselines is its capability to explicitly predict aspect ratings alongside overall rating and personalized review, providing richer and more informative recommendations.

4.5 Review generation

We evaluate ELIXIR against various baselines and the ELIXIR -Aspects ablation for review generation using text quality metrics, with the results reported in Table 4. ELIXIR and ELIXIR -Aspects consistently outperform all other models across all metrics and datasets, with ELIXIR achieving superior performance compared to its ELIXIR -Aspects ablation. These results highlight the limitations of RNN-based approaches in effectively capturing long-range context, and note that Transformer-based baselines (PETER and PEPLER) do not always significantly outperform RNN-based models, particularly on the TripAdvisor dataset. The underperformance of the ELIXIR -Aspects ablation confirms our hypothesis that incorporating aspect-aware information alongside global representations is crucial for enhancing the quality of generated reviews.

Furthermore, we analyze the impact of η, the number of tokens in the personalized prompt, on review generation quality for the TripAdvisor dataset. As shown in Table 5, starting from 20 tokens, ELIXIR outperforms the PEPLER baseline. The best performance is achieved with 50 tokens. All baseline methods condition their generation solely on learned user and item representations, typically represented by just two tokens. In contrast, our approach employs

Table 6: Visualization of attention on example from TripAdvisor dataset. For each aspect, we show the rating predicted by ELIXIR, along with the actual rating in brackets. The aspect ratings and their relative importance are contrasted with the user's actual review. An alignment is observed between the importance of the aspects and the content of the review.



Aspect	Rating	Ground truth review
Cleanliness	4.9 (5.0)	if we go back to paris, we are staying here again. the place is so charming and
Location	5.0 (5.0)	overlooks the beautiful luxembourg gardens. the staff were sooo hospitiable.
Rooms	5.0 (5.0)	always asking what they could do to help us. they arranged two tours for us,
Service	5.0 (5.0)	recommended places to eat and then made the reservations for us, arranged
Sleep	5.0 (5.0)	transportation from and to the airport, etc. royce and xavier, i can't thank you
Value	5.0 (5.0)	enough! also, so many places are in walking distance, like notre dame and the
Overall	4.9 (5.0)	louvre. you can't help but fall in love with this place!

a richer, personalized continuous prompt, with its length optimized as a hyperparameter based on the language model used. Notably, both ELIXIR -Aspects and the full ELIXIR model outperform the PEPLER model despite using T5-Small, which has roughly half the parameters of the GPT-2 model used by PEPLER. This further underscores the effectiveness and efficiency of our personalized prompt tuning approach.

4.6 Aspect modeling

Aspect modeling, by learning aspect-based representations for each user and item, enhances the performance of the model across all tasks. For overall and aspect-specific ratings prediction, as reported in Tables 2 and 3, both the ELIXIR model and its ELIXIR -Attention ablation, which replaces attention with max pooling while preserving aspect-aware representations, significantly outperform the ELIXIR -Global ablation, which relies solely on global representations. In review generation, the ELIXIR -Aspects ablation, which neither learns aspect-aware representations nor predicts aspect ratings, is also considerably outperformed by ELIXIR. These results highlight the crucial role of aspect-aware representations in encoding user preferences and item characteristics at the aspect level, leading to more accurate predictions and more informative review generation.

We conduct an empirical study to examine user representations based on the aspects learned by the model on the TripAdvisor dataset. These representations are projected into a two-dimensional space, and grouped by aspect. The results are presented in Figure 3. The representations based on the aspects learned by ELIXIR allow a coherent separation of users according to the aspects, thus demonstrating that the model effectively captures user preferences according to specific criteria. In particular, we observe an overlap between the aspects *rooms, service, sleep quality* and *cleanliness*, which suggests that these aspects are more similar to each other than to *location* and *value*. This highlights the advantage of aspect modeling, which allows learning user preferences and item characteristics at a fine-grained level, leading to better structured and more interpretable representations.

4.7 Personalized Attention

The personalized attention mechanism in ELIXIR infers the relative importance of aspects for each user and item. ELIXIR achieves significant performance gains over its ablations ELIXIR -Attention and ELIXIR -Global, particularly in overall rating prediction and aspect rating prediction (Tables 2 and 3). Notably, ELIXIR -Attention replaces personalized attention with max pooling, resulting in a performance drop, highlighting that personalized attention better aggregates aspect information than max pooling. The learned attention weights can be interpreted as proxies for the importance of each aspect to a given user or item. To validate this hypothesis, we analyzed attention weights for user-item pairs and examined their consistency with the actual review content. An example from TripAdvisor dataset is provided in Table 6. In this example, the most important aspect for the user, according to attention, is *service*, while for the item, it is *location*. Comparing these attention weights with the review content reveals a strong alignment with the user's preferences and the aspect importance inferred by ELIXIR. Notably, *service* is the most frequently mentioned aspect in the review, with a positive sentiment, while *location*, the most important aspect for the item, is also emphasized. These empirical analyses confirm that ELIXIR's personalized attention effectively captures the relative importance of aspects for users and items. Beyond predicting overall and aspect ratings and generating reviews, the inferred aspect importance can serve as an additional explanatory factor, enhancing the interpretability of recommendations.

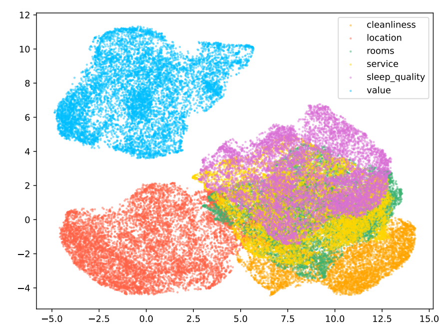


Figure 3: Projection and clustering of user aspect representations for the TripAdvisor dataset.

5 LIMITATIONS AND FUTURE WORKS

ELIXIR shows promising performance, but some limitations remain. Currently, the model relies on aspect annotations, which restricts its applicability in contexts where such annotations are unavailable or vary in quality. The experimental evaluation, limited to two specific domains, as well as the use of a small-scale model (T5-Small), may also constrain the generalizability and the quality of the generated explanations. To address these issues, future work will aim to develop unsupervised ABSA methods that automatically extract aspects from unannotated datasets, thereby making the approach more adaptable. Moreover, exploring hybrid architectures or employing larger language models that combine full fine-tuning with prompt tuning could improve the quality and fluency of the generated reviews. Integrating visual data and metadata represents another important avenue to enrich the system's contextual explainability. Finally, incorporating user feedback mechanisms could help refine personalization and allow the recommendations and their explanations to adapt in real time to users' evolving preferences.

6 CONCLUSION

In this paper, we introduce ELIXIR, a multi-task model composed of a rating prediction module —for overall and aspect ratings— and a personalized review generation module. Our approach jointly learns global and aspect-based representations for users and items, and derives a personalized continuous prompt in the semantic space of a language model. Personalized attention enables us to infer the relative importance of aspects for users and items, supporting our hypothesis that integrating aspect information enhances both recommendations and their explanations. We implement ELIXIR using the pre-trained T5 language model, employing prompt tuning and sequential training to leverage pre-training effectively. Our experiments on two real-world multi-aspect datasets, TripAdvisor and RateBeer, demonstrate the superiority of ELIXIR over strong baselines across all tasks, particularly in review generation. Empirical analyses and ablation studies confirm the significant contributions of each model component, especially aspect modeling and personalized attention.

References

[1] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*. 65–72.

[2] Konstantin Bauman, Bing Liu, and Alexander Tuzhilin. 2017. Aspect based recommendations: Recommending items with the most valuable aspects based on user reviews. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. 717–725.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

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

[5] 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. *arXiv preprint arXiv:2306.12657* (2023).

[6] Zhiyong Cheng, Ying Ding, Xiangnan He, Lei Zhu, Xuemeng Song, and Mohan S Kankanhalli. 2018. Aˆ 3NCF: An Adaptive Aspect Attention Model for Rating Prediction.. In *IJCAI*. 3748–3754.

[7] Zhiyong Cheng, Ying Ding, Lei Zhu, and Mohan Kankanhalli. 2018. Aspect-aware latent factor model: Rating prediction with ratings and reviews. In *Proceedings of the 2018 world wide web conference*. 639–648.

[8] Jin Yao Chin, Kaiqi Zhao, Shafiq Joty, and Gao Cong. 2018. ANR: Aspect-based neural recommender. In *Proceedings of the 27th ACM International conference on information and knowledge management*. 147–156.

[9] Felipe Costa, Sixun Ouyang, Peter Dolog, and Aonghus Lawlor. 2018. Automatic generation of natural language explanations. In *Companion Proceedings of the 23rd International Conference on Intelligent User Interfaces*. 1–2.

[10] 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*. 623–632.

[11] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

[12] Gayatree Ganu, Noemie Elhadad, and Amélie Marian. 2009. Beyond the stars: Improving rating predictions using review text content.. In *WebDB*, Vol. 9. 1–6.

[13] Yingqiang Ge, Shuchang Liu, Zuohui Fu, Juntao Tan, Zelong Li, Shuyuan Xu, Yunqi Li, Yikun Xian, and Yongfeng Zhang. 2024. A survey on trustworthy recommender systems. *ACM Transactions on Recommender Systems* 3, 2 (2024), 1–68.

[14] Shijiang Geng, Shuai Liu, Zhiyong Fu, Yong Ge, and Yu Zhang. 2023. Recommendation as Language Processing (RLP): A Unified Pretrain, Personalized Prompt & Predict Paradigm (P5). *arXiv preprint arXiv:2203.13366* (Jan 2023). doi:10.48550/arXiv.2203.13366

[15] Xinyu Guan, Zhiyong Cheng, Xiangnan He, Yongfeng Zhang, Zhibo Zhu, Qinke Peng, and Tat-Seng Chua. 2019. Attentive aspect modeling for review-aware recommendation. *ACM Transactions on Information Systems (TOIS)* 37, 3 (2019), 1–27.

[16] Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. 2024. Parameter-Efficient Fine-Tuning for Large Models: A Comprehensive Survey. *Transactions on Machine Learning Research* (2024).

[17] Emrul Hasan, Mizanur Rahman, Chen Ding, Jimmy Xiangji Huang, and Shaina Raza. 2024. Review-based Recommender Systems: A Survey of Approaches, Challenges and Future Perspectives. *arXiv preprint arXiv:2405.05562* (2024).

[18] Xiangnan He, Tao Chen, Min-Yen Kan, and Xiao Chen. 2015. Trirank: Review-aware explainable recommendation by modeling aspects. In *Proceedings of the 24th ACM international on conference on information and knowledge management*. 1661–1670.

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

[20] Umair Javed, Kamran Shaukat, Ibrahim A Hameed, Farhat Iqbal, Talha Mahboob Alam, and Suhuai Luo. 2021. A review of content-based and context-based recommendation systems. *International Journal of Emerging Technologies in Learning (iJET)* 16, 3 (2021), 274–306.

[21] Diederik P Kingma. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[22] Yehuda Koren. 2009. Collaborative filtering with temporal dynamics. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. 447–456.

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

[24] Yehuda Koren, Steffen Rendle, and Robert Bell. 2021. Advances in collaborative filtering. *Recommender systems handbook* (2021), 91–142.

[25] Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691* (2021).

[26] Lei Li, Yongfeng Zhang, and Li Chen. 2021. Personalized transformer for explainable recommendation. *arXiv preprint arXiv:2105.11601* (2021).

[27] Lei Li, Yongfeng Zhang, and Li Chen. 2023. Personalized prompt learning for explainable recommendation. *ACM Transactions on Information Systems* 41, 4 (2023), 1–26.

[28] 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*. 345–354.

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

[30] Bing Liu. 2022. *Sentiment analysis and opinion mining*. Springer Nature.

[31] Pasquale Lops, Marco De Gemmis, and Giovanni Semeraro. 2011. Content-based recommender systems: State of the art and trends. *Recommender systems handbook* (2011), 73–105.

[32] Julian McAuley, Jure Leskovec, and Dan Jurafsky. 2012. Learning attitudes and attributes from multi-aspect reviews. In *2012 IEEE 12th International Conference on Data Mining*. IEEE, 1020–1025.

[33] Andriy Mnih and Russ R Salakhutdinov. 2007. Probabilistic matrix factorization. *Advances in neural information processing systems* 20 (2007).

[34] Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*. 807–814.

[35] Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*. 188–197.

[36] 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)*. 706–711.

[37] Harris Papadakis, Antonis Papagrigoriou, Costas Panagiotakis, Eleftherios Kosmas, and Paraskevi Fragopoulou. 2022. Collaborative filtering recommender systems taxonomy. *Knowledge and Information Systems* 64, 1 (2022), 35–74.

[38] 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*. 311–318.

[39] Jakub Raczyński, Mateusz Lango, and Jerzy Stefanowski. 2023. The Problem of Coherence in Natural Language Explanations of Recommendations. In *ECAI 2023*. IOS Press, 1922–1929.

[40] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

[41] 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. *Journal of machine learning research* 21, 140 (2020), 1–67.

[42] Steffen Rendle, Walid Krichene, Li Zhang, and John Anderson. 2020. Neural collaborative filtering vs. matrix factorization revisited. In *Proceedings of the 14th ACM Conference on Recommender Systems*. 240–248.

[43] Ryotaro Shimizu, Takashi Wada, Yu Wang, Johannes Kruse, Sean O'Brien, Sai HtaungKham, Linxin Song, Yuya Yoshikawa, Yuki Saito, Fugee Tsung, et al. 2024. Disentangling Likes and Dislikes in Personalized Generative Explainable Recommendation. *arXiv preprint arXiv:2410.13248* (2024).

[44] Jie Shuai, Kun Zhang, Le Wu, Peijie Sun, Richang Hong, Meng Wang, and Yong Li. 2022. A review-aware graph contrastive learning framework for recommendation. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*. 1283–1293.

[45] Peijie Sun, Le Wu, Kun Zhang, Yu Su, and Meng Wang. 2021. An unsupervised aspect-aware recommendation model with explanation text generation. *ACM Transactions on Information Systems (TOIS)* 40, 3 (2021), 1–29.

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. arXiv:1706.03762 [cs.CL] https://arxiv.org/abs/1706.03762

[47] Hongning Wang, Yue Lu, and Chengxiang Zhai. 2010. Latent aspect rating analysis on review text data: a rating regression approach. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. 783–792.

[48] 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*, Vol. 37. 13816–13824.

[49] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675* (2019).

[50] Wenxuan Zhang, Yang Deng, Xin Li, Yifei Yuan, Lidong Bing, and Wai Lam. 2021. Aspect sentiment quad prediction as paraphrase generation. *arXiv preprint arXiv:2110.00796* (2021).

[51] Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2022. A survey on aspect-based sentiment analysis: Tasks, methods, and challenges. *IEEE Transactions on Knowledge and Data Engineering* 35, 11 (2022), 11019–11038.

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

[53] Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. 2014. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. 83–92.