TASTE: A Context-Aware Explainable Recommendation Model

Anonymous ACL submission

Abstract

Travel recommendation involves complex, context-dependent decision-making that goes beyond static user preferences. While recent large language model (LLM)–based conversational recommender systems enable natural language interaction and explanation generation, most existing approaches treat user preferences as fixed and utilize user queries only for surface-level explanation, without adapting the underlying recommendation logic. This results in a mismatch between conversational intent and actual recommendation criteria.

In this paper, we propose TASTE (Query-Aware Travel Recommendation via Aspect-based Sentiment Profiling), a novel recommendation framework that dynamically adapts aspect-level preference weights according to natural language user queries while preserving long-term user preference profiles. TASTE constructs structured user preference representations using Aspect-Based Sentiment Analysis (ABSA) applied to review texts and combines learned aspect attention with query-derived aspect importance through a controllable weighting mechanism. Large language models are employed exclusively for query interpretation and explanation generation, ensuring that recommendation decisions remain deterministic, interpretable, and faithful to model-internal reasoning.

We conduct extensive experiments on three real-world datasets spanning restaurant, travel, and hotel domains (Yelp Restaurant, Yelp Travel-related, and TripAdvisor Hotel). Quantitative results demonstrate that TASTE achieves competitive rating prediction performance compared to strong rating- and review-based baselines, without sacrificing accuracy for interpretability. Qualitative and quantitative analyses further show that TASTE effectively captures query-aware preference shifts, producing meaningful changes in recommendation rankings for the same user under different queries. Finally, explanation quality evaluation using an LLM-as-a-Judge protocol confirms that TASTE generates coherent, relevant, and aspect-aligned explanations across domains and languages.

Overall, TASTE provides a unified framework for accurate, query-aware, and explainable travel recommendation, addressing key limitations of existing conversational recommender systems.

1 Introduction

Travel recommendation systems have been widely adopted across online platforms and the travel industry, and recent advances in large language models (LLMs) have accelerated the emergence of conversational, chatbot-based travel recommendation services. In practice, many platforms now employ natural language interfaces that accept free-form user queries and provide personalized recommendations, as exemplified by systems such as OpenAI (OpenAI, 2025), Expedia (Expedia, 2025), HAIJEKO (Jeju Air, 2025), and the Malaysia Airlines chatbot. These trends indicate a shift from conventional item recommendation toward interactive, dialogue-oriented recommendation systems.

This transition has been driven by three key technological advances. First, progress in natural language processing enables direct understanding of unstructured textual data such as reviews and conversational utterances, allowing recommendation models to move beyond rating prediction toward reasoning-based and explainable recommendations. Second, user modeling has evolved to extract personalized preferences from review text rather than relying solely on structured interaction data. Third, these advances have facilitated the rapid adoption of LLM-based conversational recommendation systems across domains including travel, e-commerce, and entertainment. Consequently, natural language–based preference modeling and explainable recommendation have emerged

as central research challenges in modern recommender systems.

Among recommendation domains, travel recommendation is characterized by particularly complex and context-dependent user preferences. Travel decision-making involves a sequence of interrelated choices—such as destination selection, accommodation, dining, and activity planning—rather than a single consumption decision (Sarkar et al., 2023). Throughout this process, users simultaneously consider functional utility and experiential factors, including travel purpose, companions, budget constraints, and preferred atmosphere.

Prior studies have shown that even the same user may apply substantially different preference criteria depending on situational context, reflecting differences between solo and group travel, leisure-oriented versus sightseeing-focused trips, and other contextual conditions (Renjith et al., 2020). As a result, travel recommendation has long been regarded as a domain in which personal context and subjective judgment play a more prominent role than in conventional content recommendation tasks such as music or movie recommendation (Lin and Liu, 2024). Moreover, fine-grained preferences over factors such as price, location, service quality, and atmosphere have been reported to exert a direct and significant impact on overall travel satisfaction (Payandenick and Othman, 2026). These findings highlight the limitations of static user profiles and fixed preference weights in modeling real-world travel decision-making.

Meanwhile, prior research on review-based and conversational recommendation has primarily focused on modeling long-term user preferences or generating explanations for recommended items. For example, studies have proposed constructing interpretable user profiles from review text to support explainable recommendations, while TEARS introduced a controllable recommendation framework that allows users to influence recommendation outcomes (Ramos et al., 2024). In addition, recent studies on ChatGPT-based travel recommendation systems have reported that LLM-generated explanations positively affect perceived relevance, credibility, and usefulness.

However, most existing approaches still assume that user preferences are static. Even when natural language queries are supported in conversational recommendation settings, queries are often used only for explanation generation rather than for actively adjusting recommendation criteria. Consequently, when a single user issues queries reflecting different travel purposes or contextual needs, the underlying preference weights remain unchanged, leading to a mismatch between conversational input and recommendation logic.

To address these limitations, we propose **TASTE (Query-Aware Travel Recommendation via Aspect-based Sentiment Profiling)**, a novel framework for conversational travel recommendation. TASTE constructs user preference profiles using Aspect-Based Sentiment Analysis (ABSA) over review texts and dynamically adjusts aspect-level importance weights according to the user's natural language query. By preserving long-term preferences while reweighting aspects based on query intent, TASTE enables context-adaptive recommendation behavior. Furthermore, aspect-level contribution scores are explicitly exposed, and large language models are used solely to generate natural language explanations grounded in these quantitative signals, thereby enhancing transparency and trustworthiness.

Based on this motivation, this study addresses the following research questions:

2 Related Works

2.1 LLM-based Recommendation

The rapid advancement of large language models (LLMs) has introduced a new paradigm in recommender systems research (Liu et al., 2024; Wu et al., 2024; Wang et al., 2024a). Early studies primarily explored the use of LLMs as text encoders or auxiliary knowledge sources to enhance existing neural recommendation models (Zhao et al., 2023). More recent work has moved toward deeper integration of LLMs into the recommendation pipeline, leveraging their capabilities in reasoning, summarization, and content generation.

Prior research broadly categorizes LLM-based recommendation approaches into LLM-enhanced recommender systems, generative recommendation, and hybrid frameworks. LLM-enhanced recommender systems include knowledge enhancement, interaction enhancement, and model enhancement strategies. Knowledge enhancement approaches enrich user and item representations through user history summarization, item attribute generation, and knowledge graph expansion (Devlin et al., 2019; Wang et al., 2024b; Tian et al., 2024; Cao et al., 2024; Ren et al., 2024; Sun et al., 2025). Interaction enhancement methods mitigate

data sparsity by generating pseudo interaction logs via prompt-based reasoning (Yue et al., 2023; Yang et al., 2025b). Model enhancement approaches distill semantic or reasoning signals derived from LLMs into lightweight recommendation models for efficient inference (Gheewala et al., 2024).

Generative recommendation adopts an alternative paradigm in which user histories and items are represented as natural language token sequences and recommendations are generated autoregressively. To support this paradigm, techniques such as ID tokenization, clustering-based indexing, and VAE-based encoding have been proposed. In addition, several studies employ LLMs in re-ranking stages or adopt retrieval-augmented generation (RAG)–based recommendation frameworks (Liu et al., 2025a; Zhou et al., 2025). Despite their strong reasoning capability, these approaches generally overlook the problem of dynamically adjusting internal preference structures according to user query context.

2.2 Review-based and Aspect-based Recommendation

User reviews provide a rich source of information for modeling user preferences, experiences, and emotions, and have been extensively studied in recommender systems. Early approaches enhanced user and item representations using review embeddings, while later studies significantly improved text understanding by incorporating pre-trained language models such as BERT (Liu et al., 2019). Models such as DAML jointly learn from ratings and review text, achieving complementary effects between numerical and textual signals (Liu et al., 2025b).

Graph-based approaches have further explored structural relationships within review data. APAHN models aspects as hyperedges to capture interactions among multiple aspects (Zhang et al., 2023), while the Multi-Aspect Enhanced GNN constructs a user–item–review tripartite graph to learn multi-relational representations (Hasan et al., 2025). The Multi-Factor Collaborative Prediction Model similarly improves recommendation performance by extracting multiple latent semantic factors from review text (Zhang et al., 2022). A comprehensive survey on review-based recommendation identifies text processing, sentiment analysis, aspect extraction, and multi-preference modeling as core components (Shen et al., 2025). However, many existing methods struggle to disentangle multiple coexisting aspects within a single review and to explicitly model both positive and negative preferences, motivating the adoption of ABSA-based approaches.

2.3 Aspect-based Sentiment Analysis

Aspect-Based Sentiment Analysis (ABSA) aims to identify aspects mentioned in text and determine the sentiment polarity associated with each aspect. ABSA has been widely adopted for fine-grained user preference modeling in recommendation systems. While early ABSA models relied on LSTM- and CNN-based architectures, recent approaches have achieved substantial performance gains using pre-trained language models.

Recent studies focus on more precise modeling of relationships between contextual semantics and aspects. DMAN dynamically computes token importance to reduce contextual noise (Xu et al., 2025), while DR-BERT enhances aspect-specific representations through context-aware token reweighting (Zhao et al., 2025). In addition, ZeroABSA and DS2-ABSA enable effective ABSA in low-resource and cross-domain settings through LLM-based data synthesis and domain generalization (Cheng et al., 2023; Yang et al., 2025a). These advances provide a robust foundation for extracting reliable aspect-level positive and negative preference signals from review text.

2.4 Explainable Recommendation and Rating Prediction Models

Explainability has become increasingly important for improving user trust and system transparency in recommender systems. Early studies proposed expressing aspect importance based on model-internal structures or learning interpretable user profiles from review text (Lei et al., 2024; Shimizu et al., 2025; Peng et al., 2024). With the emergence of LLMs, natural language–based explanation generation has been actively explored. Models such as MAPLE, RecExplainer, and Disentangling Likes and Dislikes align user preferences and model signals with natural language explanations (Bang and Song, 2025; Penaloza et al., 2024; Koren et al., 2009). More recent work further investigates uncertainty-aware explanations and LLM-friendly user profile representations (Mnih and Salakhutdinov, 2007; Ramos et al., 2024; Huang et al., 2013).

In parallel, traditional rating prediction models include MF, PMF, SVD++, and NMF (He et al., 2017; Li et al., 2021; Zheng et al., 2017), while

neural models such as NeuMF, DAML, and BERT-MEF integrate textual and semantic information to improve performance (Tay et al., 2018; Chen et al., 2018; Cheng et al., 2018). However, existing approaches remain limited in jointly leveraging aspect-level positive and negative preferences for both recommendation and explanation.

By integrating these research streams, this work distinguishes itself by constructing ABSA-based multi-aspect preference vectors from review text and dynamically adjusting aspect importance according to natural language query context. We empirically validate the effectiveness of the proposed TASTE framework on large-scale real-world datasets from the travel and dining domains.

3 Datasets

3.1 Data Sources and Domain Coverage

To analyze the complex and context-dependent decision-making process in travel recommendation, we employ three review-based datasets covering complementary travel-related domains. Travel experiences involve multiple interrelated decisions, including dining, sightseeing and activities, and accommodation, each governed by distinct evaluation criteria and preference structures. Accordingly, we select datasets that represent these major components of travel decision-making.

Specifically, we use the Yelp Restaurant dataset to model dining experiences, a Yelp Travel-related dataset to capture tourist attractions and activity-related evaluations, and the TripAdvisor Hotel (Hong Kong) dataset to analyze accommodation preferences. All datasets contain numerical ratings accompanied by relatively long and descriptive review texts, making them well-suited for aspect-level sentiment analysis and personalized preference modeling.

3.2 Data Description and Preprocessing

3.2.1 Dataset Characteristics

The Yelp Restaurant dataset consists of *user_id*, *business_id*, star ratings, and review text, and provides rich evaluations of core dining aspects such as food quality, service, ambiance, price, and location. The Yelp Travel-related dataset is constructed by filtering travel-, attraction-, and activity-related categories from the full Yelp corpus. It captures experiential and context-dependent factors, including sightseeing experiences, surrounding environments, and activity satisfaction. The TripAdvisor Hotel dataset contains *user_id*, *hotel_id*, star ratings, and review text, and includes fine-grained evaluations of accommodation attributes such as room quality, cleanliness, facilities, location, and staff service. Although the datasets differ in review length distributions, sparsity levels, and rating distributions, all provide sufficiently rich textual information required for ABSA-based multi-dimensional preference profiling.

3.2.2 Filtering and Cleaning

To ensure data quality and reliable preference estimation, we apply consistent preprocessing criteria across all datasets. We retain users with at least 10 reviews and items with at least 5 reviews, remove duplicate entries, and inspect rating distributions to exclude extreme outliers.

3.3 Aspect Definition and Data Splitting

Rather than relying on fully unsupervised aspect extraction, we predefine domain-specific aspect sets to enhance interpretability and enable consistent cross-domain comparison. For the restaurant domain, we define the aspect set $\mathcal{A} = \{food, service, ambience, price, location\}$. For the travel and hotel domains, we define $\mathcal{A} = \{room\_quality, service, facilities, location, value\}$. These aspects correspond to core evaluation factors that have been shown to strongly influence user satisfaction. Details of the ABSA model architecture and training procedure are described in Section 4.

All datasets are split into training, validation, and test sets using an 80/10/10 ratio with a fixed random seed to ensure reproducibility. The split is performed at the user–item interaction level, with the constraint that each user and item appears at least once in the training set to avoid cold-start issues during evaluation.

3.4 Summary Statistics

Table 1 summarizes the statistics of the final datasets. Due to computational constraints, we randomly sample approximately 5,000 users from the Yelp Restaurant and TripAdvisor Hotel datasets for experimental evaluation.

Dataset	#Users	#Items	#Reviews
Yelp Restaurant	5,000	31,321	129,489
Yelp Travel	948	2,942	14,658
TripAdvisor Hotel	5,000	6,280	177,726

Table 1: Summary statistics of the datasets used in this study.

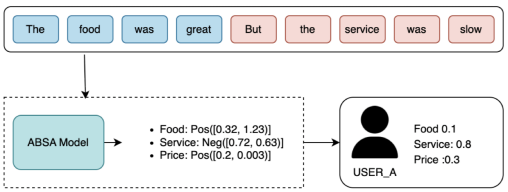

Figure 1: Illustration of the ABSA extraction process.

4 Method

4.1 Aspect-Based Sentiment Profiling

We propose **TASTE** (Query-Aware Travel Recommendation via Aspect-based Sentiment Profiling), a recommendation framework that dynamically adapts aspect importance based on natural language queries while preserving long-term personalized preferences.

4.2 Problem Formulation

Let $\mathcal{U} = \{u_1, \ldots, u_M\}$ denote a set of users and $\mathcal{I} = \{i_1, \ldots, i_N\}$ denote a set of items (restaurants or travel entities). Each user–item interaction (u, i) is associated with a rating r_{ui} and a review text t_{ui}. We define a set of aspects $\mathcal{A} = \{a_1, \ldots, a_K\}$ representing key preference dimensions (e.g., food, service, ambiance, price, location).

Given a user u, a natural language query q, and the user's historical reviews, our objectives are to: (1) generate a ranked list of top-K item recommendations, (2) provide natural language explanations grounded in aspect-level contributions, and (3) dynamically adjust recommendation criteria according to query context.

4.2.1 Aspect Sentiment Extraction

For each review text t_{ui}, we extract aspect-specific sentiment scores using a pre-trained DeBERTa-v3 model fine-tuned for Aspect-Based Sentiment Analysis (ABSA). The ABSA model outputs a probability distribution over sentiment polarities (negative, neutral, positive) for each aspect. We compute the sentiment score for aspect a as:

$$s_{ui}^{a} = p_{ui}^{a}(pos) - p_{ui}^{a}(neg), \tag{1}$$

where $p_{ui}^{a}(\cdot)$ denotes the predicted probability for aspect a. For each user–item pair, we obtain an aspect sentiment vector $\mathbf{s}_{ui} \in R^K$.

4.2.2 User Preference Normalization

To account for individual rating biases, we normalize ratings using user-specific statistics:

$$\tilde{r}_{ui} = \frac{r_{ui} - \mu_u}{\sigma_u}, \tag{2}$$

where μ_u and σ_u are the mean and standard deviation of user u's historical ratings. The normalized rating $\tilde{r}_{ui}$ is used as the training target.

4.3 TASTE Model Architecture

The TASTE model consists of four components: (1) embedding layers, (2) aspect attention mechanism, (3) query-aware dynamic weighting, and (4) rating prediction.

4.3.1 Embedding Layer

We learn latent representations for users and items, $\mathbf{e}_u, \mathbf{e}_i \in R^d$, and user-specific aspect bias vectors $\mathbf{b}_u \in R^K$. Aspect biases capture each user's inherent preference tendencies across aspects.

4.3.2 Aspect Attention Mechanism

Aspect sentiment scores are first adjusted by user biases:

$$\hat{\mathbf{s}}_{ui} = \mathbf{s}_{ui} + \mathbf{b}_u. \tag{3}$$

We compute aspect attention weights using a multilayer perceptron (MLP) with softmax normalization:

$$\alpha_{ui} = softmax(MLP([\mathbf{e}_u; \mathbf{e}_i; \hat{\mathbf{s}}_{ui}])), \tag{4}$$

where $[\cdot;\cdot]$ denotes vector concatenation. The attention vector α_{ui} reflects the relative importance of aspects learned from historical interactions.

4.3.3 Query-Aware Dynamic Weighting

To incorporate situational context, we extract query-specific aspect weights β_q from the user query q using an LLM, subject to $\sum_a \beta_q^a = 1$. We combine learned attention and query weights via linear interpolation:

$$\gamma_{ui} = \lambda \alpha_{ui} + (1 - \lambda)\beta_q, \tag{5}$$

where $\lambda \in [0, 1]$ controls the balance between long-term preferences and query intent. We set $\lambda = 0.5$ in all experiments.

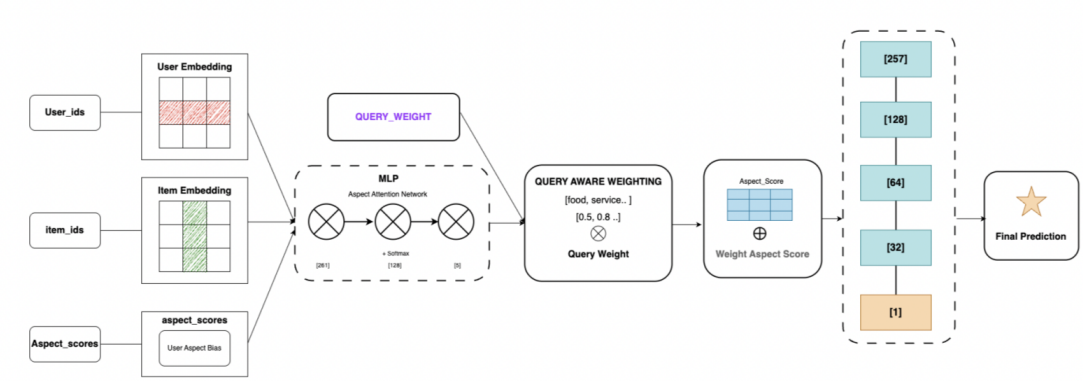

Figure 2: Overview of the TASTE model architecture.

4.3.4 Rating Prediction

We compute the query-aware weighted aspect score:

$$z_{ui} = \gamma_{ui}^{\top} \hat{\mathbf{s}}_{ui}. \tag{6}$$

The final rating prediction is given by:

$$\hat{r}_{ui} = f(z_{ui}, \mathbf{e}_u, \mathbf{e}_i) + b_i + b, \tag{7}$$

where $f(\cdot)$ is an MLP with ReLU activations and dropout, b_i is an item bias, and b is a global bias. Predictions are denormalized to obtain ratings on the original scale.

4.4 Training Objective

We train the model using Smooth L1 loss on normalized ratings:

$$\mathcal{L} = \sum_{(u,i)\in\mathcal{D}_{train}} SmoothL1(\tilde{r}_{ui}, \hat{r}_{ui}) + \eta\Theta_2^2, \tag{8}$$

where Θ denotes all trainable parameters and η is the L2 regularization coefficient.

4.5 Explainable Recommendation with LLM

For each recommended item, we generate natural language explanations grounded in the query-aware aspect weights γ_{ui}. The LLM takes as input the user query, predicted rating, and aspect weights, and produces a concise explanation focusing on the most influential aspects. Crucially, the LLM does not make recommendation decisions; it only translates quantitative model outputs into human-readable explanations, ensuring faithfulness and controllability.

4.6 Query-Aware Recommendation Pipeline

Given a user u and query q, we extract query-specific aspect weights, compute learned attention for candidate items, and fuse both signals to obtain query-aware weights. Items are ranked by predicted ratings, and explanations are generated for the top-K recommendations. Explanation quality is evaluated using an LLM-as-a-Judge protocol.

5 Experiments and Evaluation

We evaluate the proposed TASTE framework through a combination of quantitative and qualitative experiments. The evaluation focuses on three aspects: (i) rating prediction accuracy, (ii) query-aware adaptability, and (iii) explanation quality. Specifically, we address the following research questions.

RQ1 (Performance): Does TASTE achieve competitive rating prediction performance compared to existing baselines?

RQ2 (Query-awareness): Does TASTE dynamically adjust aspect importance and recommendation outcomes in response to different user queries?

RQ3 (Explanation Quality): Do aspect-grounded, LLM-based explanations improve explanation quality and faithfulness?

All experiments are conducted on three real-world datasets: Yelp Restaurant, Yelp Travel-related, and TripAdvisor Hotel. Dataset details and preprocessing procedures are described in Section 3.

5.1 Experimental Setup

5.1.1 Baselines

We compare TASTE with representative models from three categories.

Traditional Collaborative Filtering: MF (Zheng et al., 2017), NeuMF (Tay et al., 2018).

Review-based Models: DeepCoNN (**?**), MPCN (**?**), DAML (**?**), and NARRE (**?**).

Parameter	Value	Description
Embedding Dimension	128	Latent vector size
Number of Aspects	5	Domain-specific aspects
Batch Size	256	Mini-batch size
Learning Rate	1e-3	AdamW optimizer
Weight Decay	1e-3	L2 regularization
Max Epochs	30	Training iterations
Early Stopping	5	Validation patience
Query Weight λ	0.5	Attention-query balance
LLM Model	GPT-4	Query parsing and explanation

Table 2: Hyperparameters for TASTE.

Aspect-aware Models: ANCF (**?**), which incorporates aspect attention but does not support query-aware adaptation.

All models are trained on identical data splits with the same rating normalization strategy.

5.1.2 Training Configuration

Table 2 summarizes the hyperparameters used for TASTE.

5.2 Rating Prediction Performance (RQ1)

5.2.1 Evaluation Metrics

We evaluate rating prediction using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE):

$$MAE = \frac{1}{|\mathcal{D}_{test}|} \sum_{(u,i)\in\mathcal{D}_{test}} |r_{ui} - \hat{r}_{ui}|, \quad (9)$$

$$RMSE = \sqrt{\frac{1}{|\mathcal{D}_{test}|} \sum_{(u,i)\in\mathcal{D}_{test}} (r_{ui} - \hat{r}_{ui})^2}. \quad (10)$$

All metrics are computed on denormalized ratings (1–5 scale).

5.2.2 Results

Table 3 reports rating prediction performance across Yelp Restaurant, Yelp Travel-related, and TripAdvisor Hotel datasets. TASTE achieves the best or highly competitive MAE and RMSE on all three datasets, demonstrating that query-aware aspect modeling improves robustness without sacrificing predictive accuracy TASTE achieves consistently competitive performance across all datasets. While it does not always achieve the absolute best score, it outperforms most review-based models and maintains stable performance across domains. This suggests that explicit aspect modeling improves robustness without sacrificing accuracy.

5.3 Query-aware Recommendation Effects (RQ2)

We qualitatively evaluate query-aware behavior by issuing different queries for the same user and observing changes in aspect weights and top-K rankings. Table 4 illustrates representative query scenarios and resulting aspect weight shifts on Yelp Restaurant.

Aspect importance shifts align well with query intent and lead to meaningful ranking changes. Additional multilingual case studies are provided in Appendix A.

	Yelp Restaurant		Yelp Travel		TripAdvisor Hotel	
Model	MAE	RMSE	MAE	RMSE	MAE	RMSE
MF	1.0925	0.8398	0.9522	0.7318	0.8260	0.6286
NeuMF	1.0819	0.8261	0.9465	0.7110	0.8451	0.6373
DeepCoNN	0.9865	0.7486	0.8953	0.6803	0.5839	0.7843
MPCN	0.9936	0.7566	0.9013	0.6832	0.5856	0.7784
DAML	0.9865	0.7486	0.9040	0.6832	0.5866	0.7795
NARRE	0.7600	0.5789	0.7383	0.5590	**0.5029**	0.6703
ANCF	0.7644	0.5920	0.7714	0.5709	0.5161	0.6903
TASTE	**0.7577**	**0.5752**	**0.7524**	**0.5709**	0.5157	0.6940

Table 3: Rating prediction performance (MAE and RMSE) across three datasets. Lower values indicate better performance. Best results for each metric are highlighted in bold.

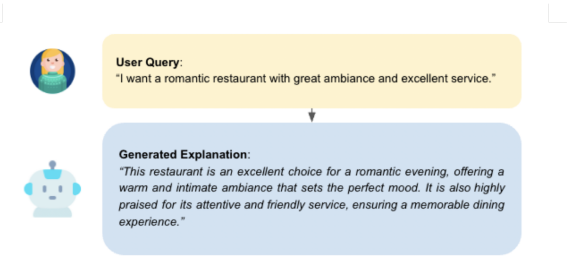

Figure 3: High-scoring example.

5.4 Explanation Quality (RQ3)

5.4.1 LLM-as-a-Judge Evaluation

We adopt the LLM-as-a-Judge protocol (**?**) to evaluate explanation quality. Explanations are scored along five dimensions: relevance, coherence, informativeness, aspect alignment, and naturalness.

5.4.2 Results

Table 5 reports average scores across datasets. Results indicate consistently high explanation quality, with strong coherence and relevance. Aspect alignment scores validate that explanations faithfully reflect model-internal reasoning. Representative examples are shown in Figure 3, with additional cases in the appendix.

Aspect	No Query	Romantic	Value	Convenience
Food	0.20	0.15	0.35	0.18
Service	0.20	0.30	0.15	0.20
Ambiance	0.20	0.40	0.10	0.12
Price	0.20	0.05	0.35	0.32
Location	0.20	0.10	0.05	0.18

Table 4: Aspect weight changes under different queries.

Dataset	Overall	Rel.	Coh.	Align.	Nat.
Yelp Restaurant	4.85	5.00	5.00	4.80	5.00
Yelp Travel	4.87	5.00	5.00	4.67	5.00
TripAdvisor	4.89	4.98	5.00	4.78	5.00

Table 5: LLM-as-a-Judge evaluation results.

6 Conclusion and Limitations

6.1 Conclusion

This paper proposed **TASTE**, a query-aware recommendation framework that integrates aspect-based sentiment profiling with dynamic weighting derived from natural language queries. Across three real-world datasets, TASTE achieved competitive rating prediction performance (MAE/RMSE) compared to strong rating- and review-based baselines, while additionally providing aspect-level interpretability and query-aware adaptation. Qualitative analyses further showed that different queries from the same user yield meaningful shifts in aspect importance and corresponding changes in top-K recommendations, indicating that TASTE can capture situational preference variations beyond static user profiles. Moreover, TASTE generates faithful and interpretable explanations grounded in quantitative aspect contributions. By restricting the role of LLMs to query interpretation and explanation generation (rather than decision-making), the framework maintains alignment between model reasoning and natural language explanations, resulting in consistently high explanation quality.

6.2 Limitations and Future Work

Despite its effectiveness, this work has several limitations. First, aspect granularity is constrained by a fixed, predefined aspect set (e.g., five aspects per domain), which may miss finer-grained or latent factors (e.g., quietness, hygiene, family-friendliness). Future work could explore automatic aspect discovery or hierarchical aspect representations.

Second, the fusion coefficient λ is fixed (we use $\lambda = 0.5$), which may not be optimal across users and query types. A natural extension is to learn λ dynamically from user characteristics, query semantics, or uncertainty estimates.

Third, TASTE prioritizes interpretability and query-aware controllability, and thus does not consistently achieve the lowest prediction error across all settings. Future work could investigate hybrid architectures that preserve aspect interpretability while improving predictive accuracy.

Fourth, explanation evaluation is limited by the lack of strong, directly comparable baselines for query-aware, aspect-grounded explanations. Future studies could develop controlled explanation baselines or compare against emerging LLM-based explanation frameworks.

Finally, we rely on LLM-as-a-Judge for scalable evaluation; however, human-centered studies are necessary to assess perceived usefulness, trust, and satisfaction. Future work should incorporate user studies (e.g., surveys or A/B tests) to validate the practical impact of query-aware explanations.

Overall, TASTE demonstrates that a lightweight, weight-based aspect modeling approach can balance accuracy, contextual adaptability, and explanation faithfulness, providing a foundation for human-centered conversational recommendation.

References

S. Bang and H. Song. 2025. Llm-based user profile management for recommender systems. *arXiv preprint arXiv:2502.14541*.

Y. Cao, N. Mehta, X. Yi, R. H. Keshavan, L. Heldt, L. Hong, and M. Sathiamoorthy. 2024. Aligning large language models with recommendation knowledge. In *Findings of NAACL*, pages 1051–1066.

C. Chen, M. Zhang, Y. Liu, and S. Ma. 2018. Neural attentional rating regression with review-level explanations. In *WWW*, pages 1583–1592.

H. Cheng, S. Wang, W. Lu, W. Zhang, M. Zhou, K. Lu, and H. Liao. 2023. Explainable recommendation with personalized review retrieval and aspect learning. *arXiv preprint arXiv:2306.12657*.

Z. Cheng, Y. Ding, X. He, L. Zhu, X. Song, and M. S. Kankanhalli. 2018. a^3ncf: An adaptive aspect attention model for rating prediction. In *IJCAI*, pages 3748–3754.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186.

Expedia. 2025. Expedia Korea. https://www.expedia.co.kr/. Accessed: Dec. 23, 2025.

S. Gheewala, S. Xu, S. Yeom, and S. Maqsood. 2024. Exploiting deep transformer models in textual review based recommender systems. *Expert Systems with Applications*, 235:121120.

E. Hasan, M. Rahman, C. Ding, J. X. Huang, and S. Raza. 2025. Review-based recommender systems: A survey of approaches, challenges and future perspectives. *ACM Computing Surveys*, 58(1).

X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua. 2017. Neural collaborative filtering. In *WWW*, pages 173–182.

K. Huang, N. D. Sidiropoulos, and A. Swami. 2013. Non-negative matrix factorization revisited. *IEEE Transactions on Signal Processing*, 62(1):211–224.

Jeju Air. 2025. Jeju Air AI Chatbot "Hi-Jeko". https://cschatbot-jwhiz.jejuair.net/chatbot?language=KO. Accessed: Dec. 23, 2025.

Y. Koren, R. Bell, and C. Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37.

Y. Lei, J. Lian, J. Yao, X. Huang, D. Lian, and X. Xie. 2024. Recexplainer: Aligning large language models for explaining recommendation models. In *KDD*, pages 1530–1541.

L. Li, Y. Zhang, and L. Chen. 2021. Personalized transformer for explainable recommendation. *arXiv preprint arXiv:2105.11601*.

J. Lin and X. Liu. 2024. Context-awareness-based intelligent recommendation method of tourism service resources. In *Proceedings of the 8th International Conference on Electronic Information Technology and Computer Engineering*, pages 5–10.

C.-C. Liu, H.-R. Yao, D.-C. Chang, and O. Frieder. 2025a. Treatrag: A framework for personalized treatment recommendation. In *Proceedings of the ACM Conference on Recommender Systems*, pages 690–695.

D. Liu, J. Li, B. Du, J. Chang, and R. Gao. 2019. Daml: Dual attention mutual learning between ratings and reviews for item recommendation. In *KDD*, pages 344–352.

J. Liu, T. Li, D. Wu, Z. Tang, Y. Fang, and Z. Yang. 2025b. An aspect performance-aware hypergraph neural network for review-based recommendation. In *WSDM*, pages 503–511.

Q. Liu, X. Zhao, Y. Wang, Y. Wang, Z. Zhang, Y. Sun, and F. Tian. 2024. Large language model enhanced recommender systems: A survey. *arXiv preprint arXiv:2412.13432*.

A. Mnih and R. R. Salakhutdinov. 2007. Probabilistic matrix factorization. *NeurIPS*, 20.

OpenAI. 2025. ChatGPT Image Generation. https://chatgpt.com/images/. Accessed: Dec. 23, 2025.

M. Payandenick and M. K. Othman. 2026. A multi-criteria recommendation system for personalised tourism experiences with user query analysis. *Information Technology & Tourism*, 28(1):3.

E. Penaloza, O. Gouvert, H. Wu, and L. Charlin. 2024. Tears: Textual representations for scrutable recommendations. *arXiv preprint arXiv:2410.19302*.

Y. Peng, H. Chen, C.-S. Lin, G. Huang, J. Hu, H. Guo, and X. Wang. 2024. Uncertainty-aware explainable recommendation with large language models. In *IJCNN*, pages 1–8.

J. Ramos, H. A. Rahmani, X. Wang, X. Fu, and A. Lipani. 2024. Transparent and scrutable recommendations using natural language user profiles. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Y. Ren, Z. Chen, X. Yang, L. Li, C. Jiang, L. Cheng, and J. Zhou. 2024. Enhancing sequential recommenders with augmented knowledge from aligned large language models. In *SIGIR*, pages 345–354.

S. Renjith, A. Sreekumar, and M. Jathavedan. 2020. An extensive study on the evolution of context-aware personalized travel recommender systems. *Information Processing & Management*, 57(1):102078.

J. L. Sarkar, A. Majumder, C. R. Panigrahi, S. Roy, and B. Pati. 2023. Tourism recommendation system: A survey and future research directions. *Multimedia Tools and Applications*, 82(6):8983–9027.

C. Shen, W. Wei, D. Wang, and Z. Wang. 2025. Zero-shot cross-domain aspect-based sentiment analysis via domain-contextualized chain-of-thought reasoning. In *Findings of EMNLP*, pages 4558–4573.

R. Shimizu, T. Wada, Y. Wang, J. Kruse, S. O'Brien, and J. McAuley. 2025. Disentangling likes and dislikes in personalized generative explainable recommendation. In *The Web Conference*, pages 4793–4809.

J. Sun, S. Qian, Z. Han, W. Li, Z. Qian, D. Yang, and G. Xue. 2025. Lkd-kgc: Domain-specific kg construction via llm-driven knowledge dependency parsing. *arXiv preprint arXiv:2505.24163*.

Y. Tay, A. T. Luu, and S. C. Hui. 2018. Multi-pointer co-attention networks for recommendation. In *KDD*, pages 2309–2318.

J. Tian, Z. Wang, J. Zhao, and Z. Ding. 2024. Mmrec: Llm-based multi-modal recommender system. In *SMAP*, pages 105–110.

Q. Wang, J. Li, S. Wang, Q. Xing, R. Niu, H. Kong, and C. Zhang. 2024a. Towards next-generation llm-based recommender systems: A survey and beyond. *arXiv preprint arXiv:2410.19744*.

Y. Wang, Y. Wang, Z. Fu, X. Li, W. Wang, Y. Ye, and R. Tang. 2024b. Llm4msr: An llm-enhanced paradigm for multi-scenario recommendation. In *CIKM*, pages 2472–2481.

L. Wu, Z. Zheng, Z. Qiu, H. Wang, H. Gu, T. Shen, and E. Chen. 2024. A survey on large language models for recommendation. *World Wide Web*, 27(5).

H. Xu, Y. Zhang, Q. Wang, and R. Xu. 2025. Ds2-absa: Dual-stream data synthesis with label refinement for few-shot aspect-based sentiment analysis. In *ACL*, pages 15460–15478.

C.-W. Yang, Z.-Q. Feng, Y.-J. Lin, C.-W. Chen, K.-D. Wu, H. Xu, and H.-Y. Kao. 2025a. Maple: Enhancing review generation with multi-aspect prompt learning in explainable recommendation. In *ACL*, pages 31803–31821.

W. Yang, W. Zhang, Y. Liu, Y. Han, Y. Wang, J. Lee, and P. S. Yu. 2025b. Cold-start recommendation with knowledge-guided retrieval-augmented generation. *arXiv preprint arXiv:2505.20773*.

Z. Yue, S. Rabhi, G. D. S. P. Moreira, D. Wang, and E. Oldridge. 2023. Llamarec: Two-stage recommendation using large language models for ranking. *arXiv preprint arXiv:2311.02089*.

C. Zhang, S. Xue, J. Li, J. Wu, B. Du, D. Liu, and J. Chang. 2023. Multi-aspect enhanced graph neural networks for recommendation. *Neural Networks*, 157:90–102.

K. Zhang, M. Zhang, H. Zhao, Q. Liu, W. Wu, and E. Chen. 2022. Incorporating dynamic semantics into pre-trained language model for aspect-based sentiment analysis. *arXiv preprint arXiv:2203.16369*.

W. Zhao, P. Zhang, H. Gu, D. Li, T. Lu, and N. Gu. 2025. Aotree: Aspect order tree-based model for explainable recommendation. In *AAAI*, pages 2273–2286.

X. Zhao, M. Wang, X. Zhao, J. Li, S. Zhou, D. Yin, and R. Guo. 2023. Embedding in recommender systems: A survey. *arXiv preprint arXiv:2310.18608*.

L. Zheng, V. Noroozi, and P. S. Yu. 2017. Joint deep modeling of users and items using reviews for recommendation. In *WSDM*, pages 425–434.

H. Zhou, H. Gu, Z. Zhan, X. Liu, K. Zhou, Y. Xiao, and T. Chen. 2025. The efficiency vs. accuracy trade-off: Optimizing rag-enhanced llm recommender systems using multi-head early exit. In *ACL*, pages 26443–26458.

A Additional Qualitative Examples

A.1 Appendix A: LLM-as-a-Judge Evaluation Example

Item ID: 7Cecref3HqLLpVLhMB1oTQ

Evaluation Scores:

- Overall Score: 4.80 / 5.0
- Relevance: 5.0 / 5.0
- Coherence: 5.0 / 5.0
- Informativeness: 5.0 / 5.0
- Aspect Alignment: 4.0 / 5.0
- Naturalness: 5.0 / 5.0

User Query: "I want a romantic restaurant with great ambiance and excellent service."

Generated Explanation:

> If you're looking for a romantic spot, this restaurant is an absolute gem. The ambiance is enchanting, with dim lighting and cozy seating that sets the perfect mood for a special evening. Paired with exceptional service, where the staff goes above and beyond to make guests feel welcomed and cared for, this place offers a truly memorable dining experience. The thoughtfully curated menu further enhances the overall experience, making it an ideal choice for a special night out.

Aspect Weights:

- Food: Query = 0.10, Attention = 0.18
- Service: Query = 0.25, Attention = 0.24
- Ambiance: Query = 0.50, Attention = 0.33
- Price: Query = 0.05, Attention = 0.15
- Location: Query = 0.10, Attention = 0.09

LLM-as-a-Judge Justification: The explanation directly addresses the user's intent by emphasizing romantic ambiance and high-quality service, which are the most important aspects highlighted in the query. The content is well-structured, logically coherent, and provides rich descriptive details that support decision-making. While the explanation could have mentioned food quality more explicitly, it remains highly engaging and natural overall.

Reasons for High Score:

- Strong relevance to the user query
- Natural and human-like language
- Clear and coherent structure
- Informative and decision-supportive content

A.2 Appendix B: Query-Aware Recommendation Case Study

This appendix presents a detailed case study demonstrating the query-aware recommendation mechanism of TASTE using a non-English (Korean) user query. The example illustrates the complete pipeline from query interpretation to final recommendations and explanations.

A.2.1 B.1 Query Information

User Query (English Translation): "I am looking for an affordable restaurant that is suitable for dining with my family."
User ID: `_BcWyKQL16ndpBdggh2kNA`

A.2.2 B.2 Query Parsing via LLM

The primary intent of this query is to find a family-friendly restaurant with good value for money. Accordingly, the highest weight is assigned to the *value* aspect. Since the query involves family dining, food quality and service are also emphasized. In addition, cleanliness and convenient location are considered important factors for family visits.

Derived Aspect Weights:

- Value: 0.30
- Food: 0.20
- Service: 0.20
- Cleanliness: 0.15
- Location: 0.15

A.2.3 B.3 Top-5 Recommendation Example

Top Recommendation (Rank #1): Item ID: 2o4GOsWEmcdwjgqnEw0w9A

Generated Explanation:

> This restaurant is highly rated for value for money, which directly aligns with your preference for an affordable family dining option. It also receives strong ratings across key aspects such as food quality, service, cleanliness, and location, making it an excellent choice for a comfortable and budget-friendly family meal.

Reason for Ranking #1:

- Highest predicted rating among candidate items (1.01)
- Strong value aspect score (0.895), matching query emphasis
- High combined weight on the value aspect (0.282)
- Consistently high scores across all aspects (0.776–0.89)