Snippet-based Conversational Recommender System

Anonymous ACL submission

Abstract

Conversational Recommender Systems engage users in interactive dialogues to gather preferences and provide personalized recommendations. While existing studies have advanced conversational strategies, they often rely on predefined attributes or expensive, domain-specific annotated datasets, which limits their flexibility in handling diverse user preferences and adaptability across domains. We propose SNIPREC, a novel resource-efficient approach that leverages user-generated content, such as customer reviews, to capture a broader range of user expressions. By employing large language models to map reviews and user responses into concise *snippets*, SNIPREC represents user preferences and retrieves relevant items without the need for intensive manual data collection or fine-tuning. Experiments across the restaurant, book, and clothing domains show that snippet-based representations outperform document- and sentence-based representations, achieving Hits@10 of 0.25-0.55 with 3,000 to 10,000 candidate items while successfully handling free-form user responses.

1 Introduction

Conversational Recommender Systems (CRS) aim to gather user preferences through conversation and provide personalized recommendations based on user responses. To achieve this, a CRS must effectively: a) organize information about target items, b) interpret user responses to identify relevant items, and c) pinpoint significant aspects to further solicit user preferences. This paper focuses on tasks (a) and (b), which have received less attention than (c) but are especially important for handling vague preference descriptions like "I'm looking for *local* restaurants." We propose a method that uses large language models (LLMs) to discover and leverage the wide variety of domain-specific topics naturally arising in conversations.

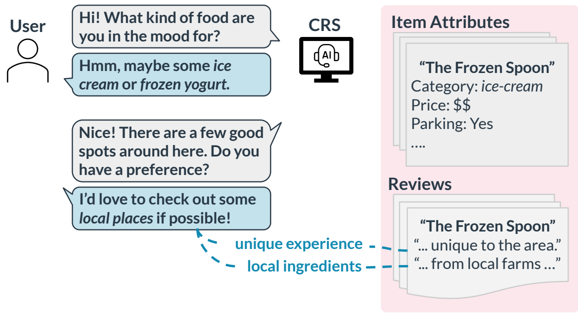

Figure 1: Example interactions with a conversational recommender system (CRS) driven by customer reviews. CRS informed by customer reviews can manage diverse user expressions that cannot be handled solely by item attributes and more effectively retrieve relevant items.

Data-driven strategies for (c) conversation strategies have seen significant progress through extensive research (Sun and Zhang, 2018; Lei et al., 2020a,b; Deng et al., 2021). The majority of these studies depend on predefined item attributes, such as restaurant categories or movie genres, assuming that (a) all items fall into these categories and (b) user responses can be mapped onto these attributes and their values. Another major line of work has developed CRS systems that mimic human conversation strategies (Li et al., 2018; Kang et al., 2019). In this approach, flexible conversations result from learning implicit representations of (a) items and (b) user responses based on large conversation datasets.

Yet, collecting conversation data that covers a wide variety of items and user preferences remains challenging, and adapting to new domains is difficult. In this paper, we aim to advance retrieval-based CRS (Gupta et al., 2023) by embracing a wide range of user expressions and preferences and enabling easy adaptation to new domains without additional in-domain data collection or fine-tuning. We focus on the comprehensive acquisition of item information and the flexible interpretation of user utterances, a direction orthogonal to conversation strategies. Specifically, we utilize user-generated content (UGC) like customer reviews to capture domain-specific and diverse user expressions and preferences. Systems supported by rich UGC data can potentially manage any information users mention and connect them to relevant items more effectively (illustrated in Figure 1). We leverage LLMs for their strengths in language understanding and information extraction (Wei et al., 2024; Li et al., 2023b) to derive high-quality insights from UGC, improving conversational recommendations in a resource-efficient and adaptable manner.

We introduce SnipRec, a system that mines and uses **snippets** from UGC (§3). A snippet is an atomic unit of information that conveys either an objective fact or a subjective opinion and serves to represent items and capture user preferences. The translation to snippets is executed by LLMs in a few-shot manner without the need for domain-specific data annotation or fine-tuning. In this study, we offload the conversation strategy to LLMs, which have proven effective in eliciting information from users (Li et al., 2023a; He et al., 2023).[1]

We evaluate SnipRec by an LLM-based user simulation–an emerging paradigm for CRS evaluation (Yoon et al., 2024; Wang et al., 2023). The simulator emulates diverse users based on customer reviews, supporting dialogues on a wide range of topics. Experiments on three datasets from distinct domains (restaurants, books, and clothing) show that snippet-based representations consistently improve item retrieval performance compared to document and sentence-based representations in CRS (§4). In particular, SnipRec using both GPT-4o-mini and LLaMA-3.3-70B models improved Hits@10 by 0.1-0.25. We also observed that the LLM-based snippet extraction methods performed reliably in most cases (>97% of faithfulness to the context) through automatic and manual evaluations.

This study advances previous CRS work in several ways. (1) We leverage UGC to capture a wider range of diverse expressions and long-tail information beyond predefined attributes. (2) We introduce SnipRec, an LLM-driven CRS approach that represents item information and user preferences using snippets. While enabling rich representations, our method reduces the need for domain-specific annotation or training, allowing straightforward domain adaptation. (3) We empirically demonstrate the benefits of snippet-based item representations for retrieval-based CRS.

[1]The conversation module can be replaced with other advanced methods, but that is beyond the scope of this paper.

2 Task and Problem Setting

Given a set of candidate items (e.g., restaurants, hotels) with user-generated content (e.g., reviews), the goal is to recommend the most suitable item through a multi-turn conversation between a *seeker* (user) and a *recommender* (system). The recommender focuses on two primary tasks: search and conversation strategy.

- **Search** aims to retrieve relevant items based on the implicit and explicit preferences of the seeker. At each turn, the seeker responds to a question from the recommender. The recommender then uses this information to retrieve candidate items.
- **Conversation strategy** aims to ask clarifying questions for multiple turns to elicit specific preference information from the seeker, which helps to refine the search space effectively.

These two tasks work together, with the search refining the recommendation pool and the conversation strategy uncovering valuable preferences. This iterative approach ensures a dynamic and user-centric dialogue experience that adapts to the user's needs. We assume that the seeker has a specific target item in mind and only provides relevant information to the open questions asked by the system without offering extraneous information. The overall objective is to minimize the number of turns needed to identify the seeker's target item.

3 Methodology

Figure 2 illustrates the workflow for the search step. First, *item* snippets are extracted from reviews to represent item knowledge (§3.1). During the conversation, the seeker's responses are parsed into *query* snippets (§3.2). The recommender leverages these query snippets to retrieve relevant item snippets and rank items (§3.3). Separately, we delegate the conversation strategy to an LLM (§3.4), as this aspect is beyond the main scope of our work.

3.1 Snippets from Item Reviews

People often express their needs in various ways (Lyu et al., 2021), such as "I am looking for a restaurant for a family gathering." and "I'd prefer a place with a good view." These free-form expressions can go beyond few predetermined item attributes which many existing CRS studies rely on.

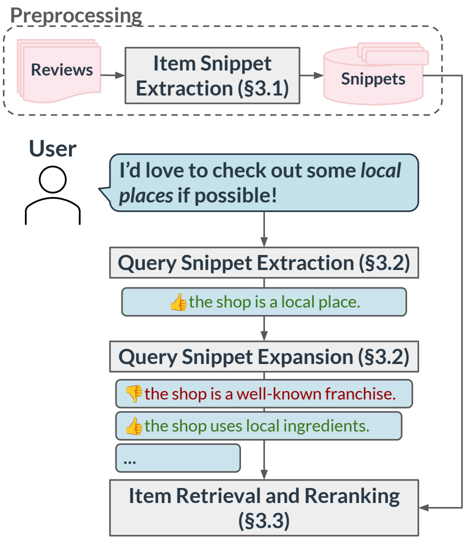

Figure 2: The search workflow of SnipRec.

In contrast, UGC like customer reviews captures any information that matters to users. Reviews offer opinions and facts about various aspects of items, making them a rich source for mining item knowledge.

Reviews may be used as a complete document (Wang et al., 2023) or segmented by sentence (Lu et al., 2021), but its nature poses challenges. Reviews can be verbose or omit details, which calls for more flexible information extraction. To obtain useful insights, we propose decomposing reviews into short snippets, each conveying atomic meaning (Nenkova and Passonneau, 2004), inspired by recent advances in fact-checking (Wanner et al., 2024; Min et al., 2023).

Motivated by the strong language understanding and information extraction capabilities of LLMs, we use LLMs with a few-shot learning method to decompose item reviews into **item** snippets, focusing on factual descriptions or subjective opinions related to a specific aspect of the item (Prompt included in Appendix C.1). Given an item i and its associated reviews R^i, we extract item snippets from each review and merge them into a snippet set, i.e., $S^i = \bigcup_{r \in R^i} \mathsf{DECOMP}(r)$, as shown below:

- item i: The Kebab Kitchen (restaurant)

⇒ reviews R^i = {"Perfect for a family gathering party...", "We ordered classic hummus, chicken kebab.", ...}

⇒ snippets S^i = {"This restaurant is perfect for a family gathering party.", "This restaurant serves classic hummus.", "This restaurant serves chicken kebabs.", ...}

This process often necessitates flexible filtering, completion, and rephrasing, a task at which LLMs excel.[2] Additionally, we transform the predefined attributes in the dataset for i into snippets using an LLM as described in Appendix A.1.1.

3.2 Snippets from Seeker Responses

At each turn t, the seeker articulates preferences in their response a_t. We apply a similar process to extract snippets from a_t (details in Appendix C.2). Specifically, we decompose each user response a_t into several **query** snippets by $\mathsf{DECOMP}(a_t)$, using a weakly canonicalized format similar to item snippets. For example, the utterance "I'm looking for a cozy cafe" is converted to "this cafe has a cozy atmosphere." The conversion reduces semantic differences between query and item snippets to support accurate retrieval.

Every query snippet also includes a *sentiment* indicator. This indicator denotes the user's preferences or dislikes for a particular aspect. For instance, if the user says "I'd love some pizza, but I want to avoid noisy places," the decomposition yields two snippets: "this restaurant serves pizza" and "this restaurant is noisy." These snippets are labeled as "prefer" and "dislike", respectively.

Snippet Expansion Query snippets often express indirect, long-tail subjective preferences with diverse lexical choices. To alleviate retrieval inaccuracies caused by this, we use LLMs (Appendix C.3) and expand the query snippets through three transformations: *paraphrase*, *support*, and *opposite* inspired by prior work (Xu and Xu, 2024; Lyu et al., 2024). After expansion, the query snippets at turn t are defined as:

$$S_t^u = \{s,\ \mathsf{paraphrase}(s),\ \mathsf{support}(s),\ \mathsf{opposite}(s) : \forall s \in \mathsf{DECOMP}(a_t)\}$$

These transformations are exemplified below:

(1) $\mathsf{paraphrase}$("...warm service...") $\xrightarrow{\text{expand}}$ "...friendly server..."

(2) $\mathsf{support}$("...comfortable seating...") $\xrightarrow{\text{expand}}$ "...cozy atmosphere..."

(3) $\mathsf{opposite}$("service is prompt" (prefer)) $\xrightarrow{\text{expand}}$ "wait time is long" (dislike)

[2] While LLMs may introduce hallucination (instances where extracted snippets contain non-existent information from the original review text), our evaluation shows that such errors are quite rare (§4.4). In general, the snippets are finer-grained than the original text.

3.3 Item Retrieval and Reranking

Given the query snippets at each turn S_t^u, we use dense retrieval (Karpukhin et al., 2020) to identify relevant item snippets from the entire set S^I $(= \cup_{i \in I} S^i)$. To ensure the retrieved snippets contain supporting information for each query snippet, we re-rank them using a natural language inference (NLI) model. Then, we score items based on the retrieved and re-ranked snippets.

Retrieval of Item Snippets Concretely, we use an off-the-shelf pretrained encoder to embed item snippets S^I and query snippets S_t^u. For each query snippet in S_t^u, we retrieve top-k similar item snippets based on cosine similarity, denoted as:

$$S_t^{I,u} = \bigcup_{s_t^u \in S_t^u} \underset{s^i \in S^I}{\text{arg top } k} \left(\mathsf{similarity}(\varepsilon(s^i), \varepsilon(s_t^u))\right),$$

where ε is an embedding function and $|S_t^{I,u}| = |S_t^u| \times k$.[3]

Next, we rank the item snippets within each group, where the groups are organized based on the query snippets used as the retrieval keys. Since vector search returns semantically similar content but does not guarantee relevance of retrieved item snippets, we use an NLI model to ensure these snippets truly satisfy the query snippets.

The ranking of a specific item snippet is based on the entailment score between each item snippet and its corresponding query snippet, as provided by the NLI model, and is performed independently within each set of item snippets retrieved by each query snippets, rather than across all retrieved item snippets in one turn. The higher entailment score, the higher their rank is (thus lower value in ranking position). We discard item snippets with scores below a threshold $t_{\text{entailment}}$. To summarize, for each query snippet $s_{t,k}^u \in S_t^u$ (the query snippets extracted from the utterance in the turn t), we obtain the rank of an item snippet $s^i \in S_t^{I,u}$ as follows:

$$\mathsf{rank}_{s_{t,k}^u}(s^i) \propto 1/\mathsf{NLI}(s^i \xrightarrow{entail} s_{t,k}^u)$$

$$\text{s.t.} \quad \mathsf{NLI}(s^i \xrightarrow{entail} s_{t,k}^u) \geq t_{\text{entailment}}$$

This process results in a smaller set of item snippets for each query snippet, which we denote as $S_t^{\prime I,u}$.

Item Re-ranking For each item i, we now update its score using the retrieved item snippets associated with it. Their within-group ranks are aggregated to calculate the item score. We employ Reciprocal Rank Fusion (Cormack et al., 2009), weighing each item snippet by $1/(\kappa + \mathsf{rank})$.[4] If multiple snippets for item i are included in the same ranking, we only consider the one with the highest rank, ignoring lower-ranked snippets. The score for item i is calculated as follows:

$$\mathsf{SCORE}_t(i) = \mathsf{SCORE}_{t-1}(i) + \sum_{s_{t,k}^u \in S_t^u} \sum_{s^i \in S_t^{\prime I,u} \cap S^i} \frac{\text{sgn}(s_{t,k}^u)}{\kappa + \mathsf{rank}_{s_{t,k}^u}(s^i)},$$

where sgn indicates the sentiment polarity that each query snippet $s_t^u \in S_{t,k}^u$ has:

$$\text{sgn}(s_{t,k}^u) = \begin{cases} 1, & \text{if sentiment is } \textit{prefer} \\ -1, & \text{if sentiment is } \textit{dislike} \end{cases}$$

This determines whether the item snippets it retrieves contribute positively or negatively to $\mathsf{SCORE}_t(i)$. A higher positive score results in a higher ranking for the item. Items that do not have any retrieved item snippets are disregarded in the item reranking process.

3.4 Clarification Questions

At each step, the recommender asks a clarification question to capture the seeker's preferences and narrow the search space. As the primary focus of this study is not on conversation strategies, we delegate this step to an LLM, which has shown effectiveness in generating context-specific questions in dialogue settings in recent studies (Li et al., 2023a; He et al., 2023). Although other conversation strategies exist (Lei et al., 2020a,b; Xu et al., 2021; *inter alia*), this LLM-based approach meets our need to elicit free-form utterances from users in a resource-efficient way.

We prompt the LLM to produce a relevant question based on the conversation history (Appendix C.4). This approach avoids extensive data annotation and fine-tuning, facilitating domain adaptation. Note that, similar to existing work (Wang et al., 2023), our question generator depends solely on the LLM's internal knowledge without incorporating signals from snippet extraction or item retrieval. Nevertheless, linking snippet retrieval and question generation is an interesting direction for future research.

[3] k is a hyperparameter.

[4] κ is a hyperparameter. Following common practices, we set $\kappa = 60$.

Dataset	Items	Users	Reviews	Snippets
Restaurant	3,007	129,503	170,996	986,819
Book	10,000	232,379	249,596	999,931
Clothing	10,000	417,501	442,295	1,725,964

Table 1: Dataset statistics.

4 Experiments

To validate the effectiveness of snippet-level representations for conversational recommendation, we evaluate SNIPREC on three datasets. We also validate the role of LLMs as conversation managers, evaluate the quality of extracted snippets, and test the reliability of our LLM-based user simulator via manual and automatic analyses.

4.1 Experimental Setup

Dataset We use the Yelp dataset (restaurant) and Amazon Reviews dataset (book and clothing) (Hou et al., 2024) which includes reviews and corresponding item information.[5] For Yelp dataset, we extract businesses located in Philadelphia, the city with the most registered businesses in the dataset, and retain only those with "food" in their Yelp categories and at least 10 reviews. For Amazon Reviews dataset, we picked the Books and Clothing categories and sampled 10,000 items with at least 10 reviews, each of which is a verified purchase and has at least 1 helpful vote. Table 1 shows the statistics of our datasets.

Baselines To test the effectiveness of the snippet-based representation and the proposed techniques, we compare SNIPREC with LLM-based systems without query snippet extraction or expansion (§3.2), using three different representations of UGC: (1) a **document-based** baseline that directly uses raw reviews without decomposition,[6] (2) a **sentence-based** baseline that splits review documents into individual sentences using spaCy (Honnibal et al., 2020)[7], and (3) a **snippet-based** baseline that uses the extracted item snippets (§3.1). While sentences are similar to snippets, they tend to be more diverse in content and expression.

Implementation Details We used GPT-4o-mini[8] for snippet extraction (§3.1) and experimented with two LLMs as the *recommender* system (§3.2, 3.4): GPT-4o-mini and LLaMa-3.3 (70B parameters). We tuned the hyperparameters and prompts of SNIPREC and baselines based on Hits@10 scores from the validation set. We tuned the number of retrieved snippets $k \in \{100, 500, 1000\}$ on validation set to find the best performing k for different experiment setting. We used BGE (Xiao et al., 2024)[9] for dense retrieval and NLI model by (Nie et al., 2020) with $t_{\text{entailment}} = 0.2$ for post-checking. These components were not fine-tuned during the experiments and are interchangeable with similar models. Experiments were run on a machine with 8 NVIDIA A100 GPUs, mainly for the NLI model and vector search engine. Each turn took around 5–10 seconds with a single GPU, using approximately 10GB of GPU memory. Appendix B.3 provides a detailed breakdown of runtime.

Evaluation Protocol We conduct simulated conversations up to five turns with an LLM-based user simulator (§4.2). To ensure that the user has an initial prompt and that the first step effectively narrows down the scope of reranking, the recommender starts with a fixed question: "`Hello, what category of restaurant are you looking for?`", or "`Hello, what category of books/clothing items are you looking for?`". The seeker is limited to responding to the provided questions. Based on the seeker's response, the recommender evaluates candidate items and calculates ranking metrics. At each subsequent turn, the recommender asks a brief clarification question. We compute Hits@k ($k = 1, 5$, and 10) and Mean Reciprocal Rank (MRR) at each turn. In cases of tied ranking, a random rank is assigned within the tied group, following (Sun et al., 2020).

4.2 User Simulator

A user simulator for CRS mimics real user behavior and preferences during interactions, enabling testing and evaluation without actual user involvement. Recently, LLM-based user simulators have emerged as a promising approach for CRS evaluation (Yoon et al., 2024; Liang et al., 2024; Kim et al., 2024), allowing scalable simulation of human-like interactions.

Following existing work, we provide the simulator with the following information: a) Yelp categories and attributes of the target item, or Amazon dataset's features, description, authors, categories

[5] See Appendix A.3 for details about the data use.

[6] Like our baseline, some studies use review documents without decomposition as part of retrieval targets (Gupta et al., 2023; Wang et al., 2023; Kook et al., 2025)

[7] `en_core_web_sm` was used.

[8] We use `gpt-4o-mini-2024-07-18` for all experiments involving GPT-4o-mini.

[9] `https://huggingface.co/BAAI/bge-base-en-v1.5`

			GPT-4o-mini		LLaMA-3.3 70B	
Dataset		**Method**	**Hits@10**	**MRR**	**Hits@10**	**MRR**
Yelp	Baseline	Document	0.283 $_{[0.26-0.31]}$	0.148 $_{[0.13-0.17]}$	0.257 $_{[0.23-0.28]}$	0.131 $_{[0.11-0.15]}$
		Sentence	0.386 $_{[0.36-0.42]}$	0.231 $_{[0.21-0.25]}$	0.396 $_{[0.37-0.43]}$	0.227 $_{[0.21-0.25]}$
		Snippet	0.414 $_{[0.38-0.44]}$	0.257 $_{[0.23-0.28]}$	0.415 $_{[0.38-0.45]}$	0.255 $_{[0.23-0.28]}$
	SNIPREC	Snippet	0.454 $_{[0.42-0.48]}$	**0.277** $_{[0.25-0.30]}$	0.438 $_{[0.41-0.47]}$	0.263 $_{[0.25-0.30]}$
		+ Expansion (§3.2)	**0.464** $_{[0.43-0.49]}$	**0.277** $_{[0.25-0.30]}$	**0.459** $_{[0.43-0.49]}$	**0.278** $_{[0.24-0.29]}$
Books	Baseline	Document	0.422 $_{[0.39-0.45]}$	0.227 $_{[0.21-0.25]}$	0.360 $_{[0.33-0.39]}$	0.209 $_{[0.19-0.23]}$
		Sentence	0.479 $_{[0.45-0.51]}$	0.364 $_{[0.34-0.39]}$	0.422 $_{[0.39-0.45]}$	0.325 $_{[0.30-0.35]}$
		Snippet	0.504 $_{[0.47-0.54]}$	0.385 $_{[0.36-0.41]}$	0.453 $_{[0.42-0.48]}$	0.349 $_{[0.32-0.38]}$
	SNIPREC	Snippet	**0.552** $_{[0.52-0.58]}$	**0.429** $_{[0.39-0.44]}$	0.495 $_{[0.46-0.53]}$	0.376 $_{[0.40-0.46]}$
		+ Expansion (§3.2)	0.545 $_{[0.51-0.58]}$	0.414 $_{[0.36-0.42]}$	**0.513** $_{[0.48-0.54]}$	**0.389** $_{[0.35-0.40]}$
Clothing	Baseline	Document	0.167 $_{[0.14-0.19]}$	0.098 $_{[0.08-0.11]}$	0.170 $_{[0.15-0.19]}$	0.099 $_{[0.08-0.11]}$
		Sentence	0.192 $_{[0.17-0.22]}$	0.125 $_{[0.11-0.14]}$	0.191 $_{[0.17-0.22]}$	0.123 $_{[0.10-0.14]}$
		Snippet	0.212 $_{[0.19-0.24]}$	0.142 $_{[0.12-0.16]}$	0.208 $_{[0.18-0.23]}$	0.141 $_{[0.12-0.16]}$
	SNIPREC	Snippet	**0.253** $_{[0.23-0.28]}$	0.146 $_{[0.13-0.16]}$	0.249 $_{[0.22-0.28]}$	0.143 $_{[0.12-0.16]}$
		+ Expansion (§3.2)	0.232 $_{[0.21-0.26]}$	**0.152** $_{[0.13-0.17]}$	**0.254** $_{[0.23-0.28]}$	**0.154** $_{[0.14-0.17]}$

Table 2: **Main results.** Hits@10 and MRR after 5 turns are reported with 95% confidence intervals in brackets.

and prices; b) a summary of positive reviews about the target and c) a review representing seeker's preference, along with general instructions and conversation history. GPT-4o-mini serve as the simulator's backbone throughout all experiments.

We use positive reviews as seed data, selecting reliable users and high-quality reviews based on the following criteria: a) review ratings of four or five, b) at least one useful (Yelp) / helpful (Amazon) vote for the review, and c) users who have written between 10 and 99 reviews. We then identify (user, item) pairs through maximum bipartite matching (Hopcroft and Karp, 1971) to maximize pair extraction without overlap, sampling 500 pairs for validation and 1,000 for testing for each dataset. We refer to these selected users as *seed users*. Next, we use GPT-4o-mini for post-processing. Following (Kim et al., 2024), we summarize general opinions from non-seed users. For each item, we provide GPT with Yelp/Amazon categories, Yelp attributes or Amazon features & descriptions, and five positive and five negative reviews (prioritizing those with higher usefulness scores) to generate summaries of five sentences each for positive and negative aspects. Finally, we anonymize proper nouns in both seed reviews and summaries to prevent answer leakage. Note that the seed users' reviews are not used by recommender systems for preventing shortcuts in item retrieval.

The simulator is prompted to generate natural utterances by answering the questions from the recommender based on the information of the target item and the review text. The prompt is designed to provide responses that are relevant, on-topic, and faithful to the context. Section 4.5 presents our manual evaluation results of the reliability of the LLM-based user simulator. The implementation details can be found in Appendix A.2.1.

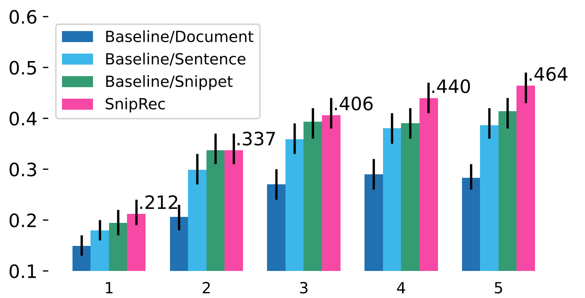

Figure 3: Hits@10 (vertical axis) over five turns (horizontal axis) in the restaurant dataset (GPT-4o-mini). The numbers represent SNIPREC scores. Similar trends were observed in other datasets for both GPT and LLaMA. Appendix B.1 presents full results.

4.3 SNIPREC Performance

Main Results: Table 2 presents the main results, demonstrating the effectiveness of snippet-based representations. We observe that more fine-grained representations (from document-level to sentence-level, and to snippet-level) lead to improved retrieval performance in the baseline systems for both GPT and LLaMA. For example, the Hits@10 scores improved from 0.283 to 0.414 on the restaurant dataset, from 0.422 to 0.504 on the book dataset, and from 0.167 to 0.212 on the clothing dataset for GPT. SNIPREC outperformed the baseline by a clear margin with both models due to the advanced processing of query snippets (§3.2).

Analysis: While query expansion (§3.2) improved performance consistently for LLaMA-based systems, it had mixed effects on GPT-based ones. Based on comparison of some examples, we conjecture that this may stem from differences in content diversity and structure. GPT tends to generate more formulaic item attributes (e.g., "The book is a children's book" $\xrightarrow{\text{expand}}$ "The book features engaging activities for young readers"). LLaMA, though structurally similar, creates more unique and concrete descriptions, e.g., "The story features talking animals." The ideal level of creativity may differ across domains, as indicated by the restaurant domain results. This suggests future research opportunities to optimize snippet expansion.

Progression Over Turns: We observed that item retrieval performance clearly improved over five conversation turns on the restaurant dataset Figure 3. Hits@10 for SnipRec increased by 0.252 from the first turn to the fifth. Similar trends were noted in other datasets with both GPT and LLaMA.[10] This finding corroborates recent studies reporting that LLMs can function as robust conversation modules.

Comparisons with Existing CRS Methods SnipRec operates in cold-start settings without the need for conversational training data, carving out a distinct niche from most existing CRS systems (§5). Unlike traditional approaches, we do not restrict interactions to predefined structured item attributes or assume access to in-domain training data. In contrast, existing CRS methods, such as attribute-based systems (e.g., Zhang et al., 2018) and those limited to specific dialog datasets (e.g., Li et al., 2018), fundamentally depend on these data types. As a result, direct empirical comparisons with existing baselines are not feasible. Our experiments demonstrate that SnipRec effectively handles aspects of recommended items that are conventionally underutilized in CRS, with ablation studies validating the effectiveness of our snippet-based approach.

4.4 Analysis: Quality of Snippets

Item snippets We evaluated item snippets based on three criteria: *faithfulness*—whether the extracted snippets contain factual errors; *atomicity*—whether the snippets represent minimal meaning units; and *decomposition completeness*—whether the decomposition covers all key information from the original text. See Appendix A.1.2 for details and examples. Three of the authors reviewed the same 100 snippets, resolving unclear cases through discussion, and found that only 3% contained hallucinations.[11] Regarding atomicity, 97% of snippets were atomic with respect to their aspect (e.g., a snippet focusing on a restaurant's interior). However, 43% could be further decomposed into multiple propositions within the same aspect. Table 3 shows some examples. For completeness, we evaluated snippets extracted from 30 reviews and found that decomposition was complete for 90% of Amazon book reviews, and 70% of reviews from both the Amazon clothes and Yelp datasets. In summary, the item snippets were mostly faithful, consistent with findings from previous studies. Although there were occasional limitations in atomicity and completeness, our end-to-end evaluation demonstrates that the snippets are effective for retrieval.

Atomic and Faithful Snippet (94/100)	
R	[...] a great cookbook [...] you'll actually learn more about why recipes are constructed like they are.
s	The book provides explanations for recipe components.
Hallucination (3/100)	
R	[...] I ended up buying their last four bottles, so they hooked me up with a 10% discount.
s	Discount is offered on bulk purchases.
Atomic Snippets at the Aspect-level (97/100)	
R	[...] Yeah the area is sketchy but their lot is kind of fenced in and the folks there seem fine [...]
s	The area is sketchy, but the lot is fenced in and feels safe.

Table 3: Examples of evaluated snippets. R and s denote a review and a snippet, respectively.

Extracted and Expanded Query Snippets Our experimental results show mixed outcomes regarding the effectiveness of query expansion. To analyze this further, two of the authors evaluated the extracted and expanded queries across 30 conversation turns for both GPT-4o-mini and LLaMA (See Appendix A.1.3 for details). We found that 100% of the extracted query snippets were faithful to the original user query. Expanded queries were evaluated based on whether they could be plausibly inferred from the extracted query and context. Averaging results from both LLMs, we observed expan-

[10] Full results are available in Appendix B.1. Conversation examples are provided in Appendix D.2.

[11] We further automatically evaluated 1,500 snippets using an LLM-as-judge approach and observed a similar result.

sion accuracy of 97% for opposite, 82% for paraphrase, and 70% for support. Our ablation study confirms that removing opposite and paraphrase expansions leads to greater performance degradation than removing support expansions (Appendix B.2). If training data is available, fine-tuning the scoring components may improve robustness to noisy expansions (Kook et al., 2025).

4.5 Analysis: Reliability of User Simulator

Although LLM-based user simulation has been successfully demonstrated in prior work, we further validated its reliability within our problem setting. Details are provided in Appendix A.2.2. Five of the authors evaluated the user simulator's responses over 50 turns based on three binary criteria: (1) *relevant*—whether the response fully addresses the posed question; (2) *on-topic*—whether the response remains focused on the asked topic; and (3) *faithful*—whether the response is free from hallucination. The aggregated scores indicate that the responses are highly relevant (98.37) and on-topic (97.48). Additionally, the responses were mostly factually accurate (84.55), and even when they were not, the errors were generally minor (e.g., requesting quick services when it was not mentioned in the seed review).

5 Related Work

CRS Many studies focus on data-driven conversation strategies (Jannach et al., 2021). These methods effectively elicit user preferences but are often limited by fixed predefined attributes (Zhang et al., 2018; Lei et al., 2020a) or rely on expensive conversation data (Li et al., 2018; Hayati et al., 2020). In contrast, our work addresses a different challenge by representing item information and user preferences through snippets extracted from UGC. This approach can capture a broader range of user expressions across various domains.

UGC Previous work has incorporated customer reviews in recommender systems through fine-tuned black-box embeddings, which often obscure the interpretability of the extracted information (Sachdeva and McAuley, 2020; Lu et al., 2021). Recent work incorporates review data directly into LLM-based systems to produce (non-conversational) recommendations (Xu and Xu, 2024; Lyu et al., 2024). Inspired by advances outside recommendation tasks (Min et al., 2023; Wanner et al., 2024), we extract explicit snippets from UGC to improve item retrieval in CRS.

LLMs in CRS Recent work demonstrates the strength of LLMs as zero-shot agents for managing dialogues in CRS (He et al., 2023; Wang and Lim, 2023; Friedman et al., 2023; Wang et al., 2023). Other studies highlight the strong question-asking capabilities of LLMs in different scenarios (Li et al., 2023a; Zhang et al., 2024). LLMs have also enabled a new paradigm for simulating realistic conversations at scale (Wang et al., 2023; Yoon et al., 2024; Liang et al., 2024; Kim et al., 2024). These findings motivate us to delegate both the conversation strategy and user simulation to LLMs, allowing our work to focus on representation through few-shot snippet extraction.

6 Conclusion

In this paper, we propose leveraging UGC to capture a wide range of user preferences and effectively address the challenges faced by systems based on predefined attributes. SnipRec utilizes snippets mined from customer reviews along with the implicit knowledge of LLMs to accurately identify relevant items for free-form user queries. Our approach is resource-efficient and adaptable to various domains, as demonstrated by our experiments. The results also confirm the reliability of LLM-based snippet extraction and the user simulator for evaluation.

This work opens several avenues for future research. For instance, as our analysis suggests, the processes of snippet extraction, expansion, and retrieval could be refined to better model user preferences (e.g., by considering the certainty associated with each snippet and how snippet expansion and retrieval can be adapted to conversational contexts, or what kinds of snippet expansion contribute better to conversational recommendation performance). Given the reliability of the user simulator demonstrated in this study, these modules could also be fine-tuned using user simulation as the reward model. Furthermore, our problem setting could be extended to capture more complex user behaviors. Currently, our user simulator assumes that users have a single target item in mind. Potential alternative setting could involve scenarios where users are interested in multiple target items that share some characteristics, or that the user has more diverse intents other than seeking for particular items.

Limitations

Interaction Type: Our study focuses on the backend recommender system, which reranks items based on the user intents provided at each turn, rather than on the frontend handling of diverse interaction types. While some interactions, such as providing feedback or explanation, fall beyond our problem setting where the recommender solely asks questions, they are common in early-stage human-to-human dialogue (Lyu et al., 2021). In reality, users may also present contradictory requests, ask questions in reverse, or exhibit other complex behaviors. While SNIPREC does not directly handle these patterns, they can be addressed by an upper-layer module that abstracts them into user intents, which can then be processed by our recommendation backend. Exploring how snippet-based representations can be extended to support these richer interactions, such as social explanations (Pecune et al., 2019), remains an interesting direction for future work.

Reliance on UGC: The proposed method relies on the availability and quality of user-generated content. Many domains such as restaurants, e-commerce, and movies have abundant UGC. However, in domains or languages where UGC is scarce or of low quality, the applicability and performance of SNIPREC may be limited.

Reliance on LLMs: Our approach uses LLMs, which have limitations including high computational cost, response latency, hallucination, and potential biases. We provide empirical results to quantify these issues and demonstrate the method's utility (§4). Although recent studies highlight the utility of LLMs as proxies for human users (§5), future research should explore how well they align with real users in depth. Addressing these concerns presents a potential avenue for future research.

Ethical Considerations

Several ethical factors should be considered when using the proposed system, particularly regarding the generation of clarification questions. Since the system relies on LLMs, its output may reflect biases in the training data. For example, in response to a general query like "I'm looking for Asian food," the system might suggest stereotypical dishes such as dumplings or sushi due to skewed data distributions. Additionally, there is a risk of the system generating harmful questions. While safeguard techniques can mitigate these issues to some extent (Wang et al., 2024), caution is still required.

References

Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 758–759, New York, NY, USA. Association for Computing Machinery.

Yang Deng, Yaliang Li, Fei Sun, Bolin Ding, and Wai Lam. 2021. Unified Conversational Recommendation Policy Learning via Graph-based Reinforcement Learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 1431–1441, New York, NY, USA. Association for Computing Machinery.

Luke Friedman, Sameer Ahuja, David Allen, Zhenning Tan, Hakim Sidahmed, Changbo Long, Jun Xie, Gabriel Schubiner, Ajay Patel, Harsh Lara, Brian Chu, Zexi Chen, and Manoj Tiwari. 2023. Leveraging Large Language Models in Conversational Recommender Systems. *arXiv*.

Raghav Gupta, Renat Aksitov, Samrat Phatale, Simral Chaudhary, Harrison Lee, and Abhinav Rastogi. 2023. Conversational recommendation as retrieval: A simple, strong baseline. In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, pages 155–160, Toronto, Canada. Association for Computational Linguistics.

Shirley Anugrah Hayati, Dongyeop Kang, Qingxiaoyang Zhu, Weiyan Shi, and Zhou Yu. 2020. INSPIRED: Toward Sociable Recommendation Dialog Systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 8142–8152, Online. Association for Computational Linguistics.

Zhankui He, Zhouhang Xie, Rahul Jha, Harald Steck, Dawen Liang, Yesu Feng, Bodhisattwa Prasad Majumder, Nathan Kallus, and Julian Mcauley. 2023. Large Language Models as Zero-Shot Conversational Recommenders. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, page 720–730, New York, NY, USA. Association for Computing Machinery.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

John E. Hopcroft and Richard M. Karp. 1971. A n5/2 algorithm for maximum matchings in bipartite. In *12th Annual Symposium on Switching and Automata Theory (swat 1971)*, pages 122–125.

Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. 2024. Bridging Language and Items for Retrieval and Recommendation. *arXiv*.

Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. 2021. A Survey on Conversational Recommender Systems. *ACM Computing Surveys*, 54(5):105:1–105:36.

Dongyeop Kang, Anusha Balakrishnan, Pararth Shah, Paul Crook, Y-Lan Boureau, and Jason Weston. 2019. Recommendation as a Communication Game: Self-Supervised Bot-Play for Goal-oriented Dialogue. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 1951–1961, Hong Kong, China. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 6769–6781, Online. Association for Computational Linguistics.

Minjin Kim, Minju Kim, Hana Kim, Beong-woo Kwak, SeongKu Kang, Youngjae Yu, Jinyoung Yeo, and Dongha Lee. 2024. Pearl: A Review-driven Persona-Knowledge Grounded Conversational Recommendation Dataset. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 1105–1120, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Heejin Kook, Junyoung Kim, Seongmin Park, and Jongwuk Lee. 2025. Empowering retrieval-based conversational recommendation with contrasting user preferences. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 7692–7707, Albuquerque, New Mexico. Association for Computational Linguistics.

Wenqiang Lei, Xiangnan He, Yisong Miao, Qingyun Wu, Richang Hong, Min-Yen Kan, and Tat-Seng Chua. 2020a. Estimation-Action-Reflection: Towards Deep Interaction Between Conversational and Recommender Systems. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 304–312, New York, NY, USA. Association for Computing Machinery.

Wenqiang Lei, Gangyi Zhang, Xiangnan He, Yisong Miao, Xiang Wang, Liang Chen, and Tat-Seng Chua. 2020b. Interactive Path Reasoning on Graph for Conversational Recommendation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2073–2083, New York, NY, USA. Association for Computing Machinery.

Belinda Z. Li, Alex Tamkin, Noah Goodman, and Jacob Andreas. 2023a. Eliciting Human Preferences with Language Models. *Preprint*, arXiv:2310.11589.

Bo Li, Gexiang Fang, Yang Yang, Quansen Wang, Wei Ye, Wen Zhao, and Shikun Zhang. 2023b. Evaluating ChatGPT's Information Extraction Capabilities: An Assessment of Performance, Explainability, Calibration, and Faithfulness. *Preprint*, arXiv:2304.11633.

Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards Deep Conversational Recommendations. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Tingting Liang, Chenxin Jin, Lingzhi Wang, Wenqi Fan, Congying Xia, Kai Chen, and Yuyu Yin. 2024. LLM-REDIAL: A Large-Scale Dataset for Conversational Recommender Systems Created from User Behaviors with LLMs. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 8926–8939, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Yu Lu, Junwei Bao, Yan Song, Zichen Ma, Shuguang Cui, Youzheng Wu, and Xiaodong He. 2021. RevCore: Review-Augmented Conversational Recommendation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1161–1173, Online. Association for Computational Linguistics.

Hanjia Lyu, Song Jiang, Hanqing Zeng, Yinglong Xia, Qifan Wang, Si Zhang, Ren Chen, Chris Leung, Jiajie Tang, and Jiebo Luo. 2024. LLM-Rec: Personalized Recommendation via Prompting Large Language Models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 583–612, Mexico City, Mexico. Association for Computational Linguistics.

Shengnan Lyu, Arpit Rana, Scott Sanner, and Mohamed Reda Bouadjenek. 2021. A Workflow Analysis of Context-driven Conversational Recommendation. In *Proceedings of the Web Conference 2021*, page 866–877, New York, NY, USA. Association for Computing Machinery.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating Content Selection in Summarization: The Pyramid Method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 145–152, Boston, Massachusetts, USA. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A New Benchmark for Natural Language Understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Florian Pecune, Shruti Murali, Vivian Tsai, Yoichi Matsuyama, and Justine Cassell. 2019. A Model of Social Explanations for a Conversational Movie Recommendation System. In *Proceedings of the 7th International Conference on Human-Agent Interaction*, page 135–143, New York, NY, USA. Association for Computing Machinery.

Noveen Sachdeva and Julian McAuley. 2020. How Useful are Reviews for Recommendation? A Critical Review and Potential Improvements. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 1845–1848, New York, NY, USA. Association for Computing Machinery.

Yueming Sun and Yi Zhang. 2018. Conversational Recommender System. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, page 235–244, New York, NY, USA. Association for Computing Machinery.

Zhiqing Sun, Shikhar Vashishth, Soumya Sanyal, Partha Talukdar, and Yiming Yang. 2020. A Re-evaluation of Knowledge Graph Completion Methods. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5516–5522, Online. Association for Computational Linguistics.

Lei Wang and Ee-Peng Lim. 2023. Zero-Shot Next-Item Recommendation using Large Pretrained Language Models. *Preprint*, arXiv:2304.03153.

Xiaolei Wang, Xinyu Tang, Xin Zhao, Jingyuan Wang, and Ji-Rong Wen. 2023. Rethinking the Evaluation for Conversational Recommendation in the Era of Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10052–10065, Singapore. Association for Computational Linguistics.

Zezhong Wang, Fangkai Yang, Lu Wang, Pu Zhao, Hongru Wang, Liang Chen, Qingwei Lin, and Kam-Fai Wong. 2024. SELF-GUARD: Empower the LLM to Safeguard Itself. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1648–1668, Mexico City, Mexico. Association for Computational Linguistics.

Miriam Wanner, Seth Ebner, Zhengping Jiang, Mark Dredze, and Benjamin Van Durme. 2024. A Closer Look at Claim Decomposition. In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics*, pages 153–175, Mexico City, Mexico. Association for Computational Linguistics.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, Yong Jiang, and Wenjuan Han. 2024. ChatIE: Zero-Shot Information Extraction via Chatting with ChatGPT. *Preprint*, arXiv:2302.10205.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-Pack: Packed Resources For General Chinese Embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 641–649, New York, NY, USA. Association for Computing Machinery.

Bei Xu and Yifan Xu. 2024. Personalized Review Recommendation based on Implicit dimension mining. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 86–91, Mexico City, Mexico. Association for Computational Linguistics.

Kerui Xu, Jingxuan Yang, Jun Xu, Sheng Gao, Jun Guo, and Ji-Rong Wen. 2021. Adapting User Preference to Online Feedback in Multi-round Conversational Recommendation. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 364–372, New York, NY, USA. Association for Computing Machinery.

Se-eun Yoon, Zhankui He, Jessica Echterhoff, and Julian McAuley. 2024. Evaluating Large Language Models as Generative User Simulators for Conversational Recommendation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1490–1504, Mexico City, Mexico. Association for Computational Linguistics.

Yizhe Zhang, Jiarui Lu, and Navdeep Jaitly. 2024. Probing the Multi-turn Planning Capabilities of LLMs via 20 Question Games. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 1495–1516, Bangkok, Thailand. Association for Computational Linguistics.

Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W. Bruce Croft. 2018. Towards Conversational Search and Recommendation: System Ask, User Respond. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, page 177–186, New York, NY, USA. Association for Computing Machinery.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

A Additional Details

This section details the implementation and evaluation. The manual evaluation was conducted by the authors of this paper, all of whom regularly work with large language models (LLMs), have experience in NLP research, and are fluent in English.

A.1 Snippets

A.1.1 Item Snippets from Pre-defined attributes (§3.1)

In addition to mining item snippets from reviews, we also transform predefined item attributes (e.g., item categories) into snippets using `gpt-4o-mini-2024-07-18` with manual validation by one of the authors. The prompts are included in the code attached to this submission. The transformation into an snippet is usually templatic. For example, the Yelp category "bubble tea" is transformed into the sentence "This place serves bubble tea." These generated snippets are also used in the document- and sentence-level baselines. For the document-level baseline, all attribute snippets for each item are combined into one document.

A.1.2 Evaluation of Item Snippets (§4.4)

This section details the manual evaluation of item snippets (§4.4). We focus on three criteria: (1) **hallucination**—whether the extracted snippets contain factual errors; (2) **atomicity**—whether the snippets represent minimal meaning units; and (3) **completeness**—whether the decomposition covers all key information from the original text. For (1), we also used `gpt-4o-2024-11-20` to automatically evaluate more data. Table 4 show annotated examples. We found that hallucinations were rare (3–4%), and the snippets generally achieve fair atomicity and completeness. Nevertheless, our end-to-end results (§4.3) demonstrate the effectiveness of the extracted snippets.

Hallucination Two of the authors manually annotated 100 snippets, and 3/100 are deemed as hallucinations. Overall, we found that our extractor is faithful to the given context. Following Wanner et al. (2024), we also employed the LLM-as-judge approach (Zheng et al., 2023) to more samples. We used `gpt-4o-2024-11-20` to judge whether a snippet can be inferred from a paragraph. We sampled 500 snippets from the decomposition results of each dataset, in total 1,500. The prompt used for the LLM-as-judge is as follows:

Prompt for Judging Hallucination

```
Work as a judge and determine if the following proposition
can be inferred from the given customer review in the
keywords. No prose.
-----------------------------------------------------------
Can the following proposition be inferred from the given
customer review in the restaurant domain?

Proposition: hypo
Customer Review: premise

Answer with 'yes' or 'no'.
```

Among these, the proportion of snippets containing hallucinations was 2.49% in the restaurant dataset, 3.90% in the book dataset, and 2.90% in the clothing dataset.

Atomicity Two annotators reviewed 100 samples using the following definition: A sentence (s) is not atomic if it can be decomposed into a conjunction of shorter sentences (s_1), (s_2), etc., such that any combination of these sentences is entailed by the original sentence. In other words, (s) entails (s_1), (s) entails (s_2), and so on. If these conditions are not met, then (s) is considered atomic. We apply this criterion to two levels: the aspect level (e.g., food, location) and the proposition level. The proposition-level atomicity corresponds to the decompositions considered in existing studies (Wanner et al., 2024). The two annotators independently evaluated the same 100 samples and resolved disagreements through discussion. 97/100 item snippets are atomic regarding the aspects they describe, though 43/100 contain multiple propositions. As mentioned in Section 4.4, while snippets are often not atomic at the proposition level, they are generally atomic at the aspect level.

Completeness We annotated a sample of the extracted snippets from 30 reviews. 27/30 of the reviews with Amazon book dataset, 21/30 of the reviews with Amazon clothes dataset and 21/30 of the reviews with Yelp dataset are judged as completely decomposed. The decomposition do not work that well with storytelling reviews, but this is also hard for human readers to get every information conveyed in a fully narrative story.

A.1.3 Evaluation of Query Snippets and Expansion (§4.4)

Three of the authors manually inspected a sample of extracted and expanded queries across 30 interaction rounds. The extraction was evaluated for **factual correctness** against the input text, while the expansion was assessed based on **plausibility**—

Category	Extracted Snippets	Original Review and Annotation Note	A/Asp.	A/Prop.	Halluc.
Good Examples —Atomic Snippets without Hallucination	The book is highly recommended for GRE preparation.	I had to take the GRE in two weeks because I decided at the last second to apply for business school. **Highly recommended!** If you're good at math, [...] —The snippet is factually accurate based on the review and cannot be further decomposed..	Yes	Yes	No
	The book provides explanations for recipe components.	It's a great cookbook with good recipes. But we enjoyed the "why" in the descriptions. For example, why did it need another egg yolk. So **you'll actually learn more about why recipes are constructed like they are.** —The snippet is factually accurate and contains no personal opinions or emotions,. It cannot be further decomposed.	Yes	Yes	No
Problematic Examples —Hallucination	Discount is offered on bulk purchases.	This place has such an amazing beer selection! I cleaned them out of their Southern Tier Pumpking the other night. **I ended up buying their last four bottles, so they hooked me up with a 10% discount.** The staff here is [...] —The discount was situational, not a standard result of a bulk purchase.	Yes	No	**Yes**
	The corner store is the best in the area.	I love this place. [...] **Fu-Wah is the best corner store in the area, at least in my opinion.** —The statement reflects a personal opinion rather than an generalizable fact.	Yes	Yes	**Yes**
Problematic Examples —Non-atomicity	This place is for the lovers and the dreamers.	I walk into the tune of Regulate by Nate Dogg and Warren G. [...] While you're at it, bring a friend because **this place is for the lovers, the dreamers, and me.** —Although the snippet is faithful to the original text and is focused on a single topic, it can be further decomposed into two propositions ("for lovers" and "for dreamers.")	Yes	**No**	No
	The menu fills the wall behind you and is made with individual letter pins.	At first I balked at the idea of spending $11 for a sandwich, but this place was highly recommended by local friends. [...] **The menu fills the wall behind you and is made with individual letter pins.** [...] —Although the snippet accurately describes a single topic (the restaurant's interior), it can be further split into two atomic sentences.	Yes	**No**	No

Table 4: Hallucination and Atomicity Analysis of Snippets (§A.1.2). The A/Asp. and A/Prop. columns indicate whether the extracted snippet is atomic (A) at the aspect (asp.) level and proposition (prop.) level, respectively. The Hallc. column indicates whether the snippet contains hallucination (an inaccurate statement with respect to the original review text).

whether the result could plausibly be inferred from the extracted snippet and its context. Although the evaluation of expansion is fundamentally subjective, ambiguous cases were resolved through discussion among the annotators to reach a consensus that fairly represented the context.

For GPT-4o-mini, 100% of the query snippets are hallucination-free, and 100% (opposite), 81% (paraphrase), and 71% (support) of the expanded queries were aligned. For LLaMA3.2-70B, 100% of the user query snippets are faithful, while in expansion, 93% (opposite), 83% (paraphrase), and 69% (support) of the expansions were aligned with the original snippets.

Below is an example of one turn of interaction in which one-third of the paraphrase and support queries are annotated as not aligned (denoted by *). Nevertheless, they still make some sense: for instance, *100% cotton* pants are generally comfortable. Therefore, these snippets can still be helpful for the overall recommendation performance. This highlights the complexity of the role of snippet expansion (or more generally, review expansion) in recommendation systems.

Utterance: *I'm looking for casual pants that are soft and comfortable.*

Snippet 1: The item is casual pants

- **Para:** The garment is a pair of casual trousers.
- **Supp:** The pants are made of lightweight cotton.
- **Opp:** The item is formal trousers

Snippet 2: The casual pants are soft

- **Para:** *The relaxed trousers are gentle to the touch.
- **Supp:** The fabric is soft and breathable.
- **Opp:** The casual pants are rough.

Snippet 3: The casual pants are comfortable

- **Para:** The relaxed-fit pants are cozy.
- **Supp:** *The fabric is made of 100% cotton.
- **Opp:** The casual pants are uncomfortable.

A.2 User Simulator

A.2.1 Implementation (§4.2)

We build upon prior LLM-based user simulator methods for CRS evaluation (Yoon et al., 2024; Liang et al., 2024; Kim et al., 2024). Specifically, following Kim et al. (2024), we prompt an LLM simulator to emulate the author of a positive customer review. As input, we also provide information about the target item (its description, attributes, etc.) and LLM-generated summaries of popular reviews. See the prompt template below:

Prompt for user simulators

````
You are a Seeker who is looking for a restaurant/book recommendation. No prose, be concise and casual in the conversation.
---------------------------------------------------
# Role-Play Task: Seeker
You will play the role of a Seeker looking for a restaurant/book recommendation. You will interact with a Recommender to find a restaurant/book that suits your preferences.

**Instructions:**

- The Recommender will ask for your preferences to identify a restaurant/book that aligns with your tastes. Your role is to provide responses as hints based on the details below.
- Express your preferences by answering the Recommender's questions based on the information provided below.

**Details of Your Favorite Restaurant/Book:**

{item info}
{item review summary}

**Your Opinion About It:**

```
{review text}
```

**Dialogue Context:**

{dialogue context}
Seeker:

Now, generate a response in the role of the Seeker based on the information provided above.

**Response Guidelines:**

- Your response should be concise, typically one sentence. Avoid giving multiple preference details at once.
- If the data provided lacks specifics to answer the Recommender, communicate no particular preference. Try not to invent details.
- Tailor your responses around 'Your Opinion About It' rather than 'Details of Your Favorite Restaurant/Book' when there is conflicting information.
- Keep your answers limited to the question asked.
- Do not reveal the name of your favorite restaurant/book or any personal or street names.
- Focus on answering the Recommender's questions. Do not proactively ask questions such as, "What kinds of restaurants/books are there?" or "Can you tell me about different cuisines in the area?"
````

Although the overall structure follows previous work (See Table 14 of (Kim et al., 2024)), we made several revisions to better align with human behavior and improve evaluation reliability on our observations of existing user simulators. In particular, we included instructions to prevent the user simulators from being overly expressive or formal. Furthermore, to avoid revealing the target item's name during the conversation, we replaced explicit names in the input data (item information, review summaries, and seed reviews) with neutral nouns as preprocessing, using *gpt-4o-mini-2024-07-18*.

A.2.2 Evaluation of User Simulator (§4.2)

As described in Section 4.2, five of the authors evaluated the user simulator's response over 50 turns and found it reliable and natural. In this section, we describe the details of evaluation guidelines and inter-annotator agreements.

Guidelines: The annotators were provided with clear instructions, which included the definition of an effective user simulator: it should accurately answer the asked question without adding unnecessary information and remain consistent with the provided context, such as the review or item details. Additionally, the guidelines contains metric definitions (see below) and annotated examples for clarity. The annotation criteria were developed through annotation practice and discussions among the annotators.

- Relevant: Does user's answer address system's question fully and appropriately, without omitting information? For instance, if system asks about food but user neither mentions a relevant preference at all nor explicitly states there is no preference, the response is not relevant.
- On-Topic: Does user's answer stay limited to the topic asked by system, without providing extra information? For instance, if system asks about food and user provides information other than food, it's off-topic. If answered topics covered more than asked topics, it is likely off-topic.
- Faithful: Is user's answer supported by the provided information? Stating "no preference" even if the information contains preferences is considered inaccurate.

Inter-Annotator Agreement: The distribution of collected judgments was highly skewed, making the use of widely used metrics like Fleiss' Kappa inappropriate for quantifying inter-annotator agreements. Therefore, we used Gwet's AC1, a method known for its robustness to class imbalance. The scores were 98.37 (Relevant), 97.48 (On-Topic), and 84.55 (Faithful), showing strong agreements among the annotators.

A.3 Use of External Data and Tools

In this study, we used the following English datasets. We have reviewed the terms of use for each dataset and confirmed that our usage complies with their guidelines.

- **Yelp Open Dataset** (Yelp)[12]: This dataset is released under the Yelp license for research purposes.
- **Amazon Reviews** (Hou et al., 2024)[13]: Although specific licensing information is not clearly provided, this dataset has been made publicly available for research purposes.

Personally identifiable information, such as user IDs, has already been anonymized in these datasets, and no demographic information is included. Furthermore, we masked person names during our data preprocessing (see Appendix A.2.1 for details).

We used LLMs as supporting tools for coding and writing. Specifically, we used OpenAI's GPT model for proofreading and polishing this paper as part of the writing process. Additionally, we used GitHub Copilot to assist with coding tasks. The idea and execution of the research are entirely our own original work.

B Additional Experiment Results

In this section, we include further experimental results that are not included in the main manuscript due to space restrictions.

B.1 Full Performance Comparison

Tables 7, 8 and 9 provide a comprehensive overview of the system performance across all compared methods.

B.2 Ablation Study: Query Snippet Expansion

We perform an ablation study to understand the contribution of different snippet expansion types. We remove each type of expansion and evaluate the resulting performance using the Yelp dataset with GPT-4o-mini. As shown in Table 5, removing either *paraphrase* and *opposite* expansions leads to largest performance drops across Hits@10 and MRR. This result is consistent with our analysis in Appendix. A.1.3, where these two expansion types are shown to align more closely with the semantics of the original snippets. Nevertheless, retaining all expansion types yields the best overall performance, suggesting that even less contributive *support* provide complementary signals.

B.3 Inference Time

SNIPREC takes around 5–10 seconds per turn on a single GPU (see §4.1). Table 6 provides a breakdown of the inference time. Over 70% of the time

[12] https://www.yelp.com/dataset

[13] https://amazon-reviews-2023.github.io/

	Hits@1	Hits@5	Hits@10	MRR
Full Expansion	0.189	0.360	0.464	0.277
- Paraphrase	0.123	0.209	0.292	0.176
- Support	0.127	0.241	0.323	0.191
- Opposite	0.118	0.215	0.294	0.176
- All (No Exp.)	0.197	0.341	0.454	0.277

Table 5: Ablation study on query snippet expansion. Removing *paraphrase* and *opposite* expansions show more performance degradation, as aligned with Appendix. A.1.3.

Module	**Restaurants**	**Books**	**Clothing**
parser^{API}	0.87 ($\pm$0.55)	1.08 ($\pm$0.69)	0.93 ($\pm$0.52)
expander^{API}	2.28 ($\pm$4.71)	3.36 ($\pm$2.50)	3.29 ($\pm$6.67)
retriever	0.59 ($\pm$0.52)	0.70 ($\pm$0.70)	1.01 ($\pm$0.60)
validator^{GPU}	0.76 ($\pm$0.53)	0.99 ($\pm$0.59)	0.96 ($\pm$0.59)
scorer	$< 10_{\text{ms}}$	$< 10_{\text{ms}}$	$< 10_{\text{ms}}$
responder^{API}	0.68 ($\pm$0.49)	0.78 ($\pm$0.57)	0.65 ($\pm$0.40)

Table 6: Breakdown of inference time (in seconds). The standard deviation is reported in parentheses. "API" and "GPU" refer to the modules whose runtime is primarily influenced by the latency of API calls and the performance of GPUs, respectively.

is spent on the parser, expander, and responder components, mainly due to LLM API call latency. The validator accounts for roughly 13% of the time, with its runtime varying based on the number of retrieved snippets and GPU performance. The retriever may slow down with more snippets, although modern dense retrievers like faiss are highly scalable.

C List of Prompts

We provide the key prompts used in SnipRec, with system prompts shown in the top section and user prompts in the bottom section. The full list of prompts, along with the system implementation, is available in the supplementary material of this submission (see the included README for navigation). This section presents representative prompts used for the Yelp dataset. Note that prompts for other datasets follow the same structure with minor wording variations (e.g., "restaurant" for Yelp and "book" for Amazon).

C.1 Item Snippet Extraction (§ 3.1)

Prompt for decomposing reviews into item snippets (Yelp)

````
You are a helpful assistant. Follow the instructions. No prose.
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
As a language genius, you are tasked with reading restaurant reviews and extracting and summarizing atomic, simple, short and coherent sentences that contain factual descriptions or subjective opinions related to a specific aspect of the restaurant, and subsequently you also need to (1) identify the topic of each atomic sentence and (2) the attitude towards the restaurant in terms of this proposition. The topics should be the aspects relevant to restaurant domains. You should use multiple atomic propositions if the content is about several topics, but combine similar content into one proposition.

Ensure these sentences carry information that effectively help differentiate various restaurants. If there is already a proposition with a similar meaning, ignore the redundant information. Ignore the irrelevant chatter, narratives and descriptions unrelated to the properties of the restaurant in the reviews. Try to use original texts from the reviews and but do summarize them if they are verbose. Be sure to cover the whole review. Try to eliminate any references of "I" or "reviewer", but focus on "restaurant". Follow the exemplar format to extract.

{examples}

Analyze this review:
```{review}```
````

C.2 Query Snippet Extraction (§ 3.2)

Prompt for decomposing user responses into query snippets (Yelp)

````
You are a helpful assistant.
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
You are provided with a pair of a question and a response related to restaurant recommendations. The question asks about preferences for a restaurant, and the response may provide specific information about the user's preferences or indicate that there is no particular preference.

**Your task:**

1. **Extract Intents:** Identify any stated preferences or dislikes about restaurants in the response.
2. **Convert to Requirement Statements:** Write short, complete sentences that objectively describe a requirement for the restaurant search. Avoid using subjective phrases like "the user likes" or "the user wants." Instead, write factual statements.
3. **Annotate:**
- `prop`: A brief description of the restaurant feature the user prefers or dislikes.
- `sentiment`: `preference` or `dislike`.

{examples}

**Guidelines:**

- Disregard information in the response that doesn't express a preference or dislike.
- Annotate only clear and specific intents.
- Each intent should address a single topic; separate multiple topics into individual intents.
- Known intents (listed in the provided set) should not be repeated. Ensure that any new intent you extract is distinct and does not overlap with the known intents.
- If the response is vague or indicates no specific preference (e.g., "I'm open to...", "I'm not specifically looking for..."), return an empty list (‘[]‘).

**Analyze the following question-response pair:**
````
````

Turn	Hits@1	Hits@5	Hits@10	Avg Pos	Avg MRR	Hits@1 95ci	Hits@5 95ci	Hits@10 95ci	Avg Pos 95ci	Avg MRR 95ci
Baseline - Document GPT-4o-mini (k=100)										
1	0.020	0.091	0.149	757.709	0.065	0.01-0.03	0.07-0.11	0.13-0.17	694.98-820.44	0.06-0.08
2	0.054	0.127	0.206	511.415	0.105	0.04-0.07	0.11-0.15	0.18-0.23	457.44-565.39	0.09-0.12
3	0.077	0.169	0.270	364.596	0.135	0.06-0.09	0.15-0.19	0.24-0.30	319.54-409.65	0.12-0.15
4	0.084	0.170	0.290	337.750	0.144	0.07-0.10	0.15-0.19	0.26-0.32	293.83-381.67	0.13-0.16
5	0.089	0.167	0.283	307.381	0.148	0.07-0.11	0.14-0.19	0.26-0.31	266.76-348.01	0.13-0.17
Baseline - Document LLaMA-3.3 70B (k=100)										
1	0.018	0.090	0.153	704.959	0.065	0.01-0.03	0.07-0.11	0.13-0.18	646.76-763.16	0.06-0.08
2	0.041	0.093	0.167	575.275	0.081	0.03-0.05	0.07-0.11	0.14-0.19	520.48-630.07	0.07-0.09
3	0.055	0.146	0.230	418.054	0.110	0.04-0.07	0.12-0.17	0.20-0.26	371.06-465.05	0.10-0.12
4	0.067	0.153	0.246	362.794	0.123	0.05-0.08	0.13-0.18	0.22-0.27	317.75-407.84	0.11-0.14
5	0.074	0.167	0.257	337.193	0.131	0.06-0.09	0.14-0.19	0.23-0.28	294.50-379.89	0.11-0.15
Baseline - Sentence GPT-4o-mini (k=500)										
1	0.029	0.110	0.180	159.568	0.082	0.02-0.04	0.09-0.13	0.16-0.20	137.92-181.22	0.07-0.09
2	0.083	0.220	0.299	116.661	0.158	0.07-0.10	0.19-0.25	0.27-0.33	98.23-135.09	0.14-0.18
3	0.126	0.255	0.359	104.103	0.202	0.11-0.15	0.23-0.28	0.33-0.39	86.23-121.97	0.18-0.22
4	0.149	0.285	0.381	89.070	0.222	0.13-0.17	0.26-0.31	0.35-0.41	73.23-104.91	0.20-0.24
5	0.161	0.280	0.386	78.523	0.231	0.14-0.18	0.25-0.31	0.36-0.42	66.30-90.75	0.21-0.25
Baseline - Sentence LLaMA-3.3 70B (k=100)										
1	0.029	0.110	0.181	236.378	0.083	0.02-0.04	0.09-0.13	0.16-0.20	203.85-268.91	0.07-0.09
2	0.078	0.170	0.238	197.668	0.136	0.06-0.09	0.15-0.19	0.21-0.26	168.19-227.15	0.12-0.15
3	0.112	0.227	0.332	145.162	0.180	0.09-0.13	0.20-0.25	0.30-0.36	122.62-167.70	0.16-0.20
4	0.138	0.273	0.380	118.890	0.215	0.12-0.16	0.25-0.30	0.35-0.41	99.23-138.55	0.19-0.24
5	0.150	0.288	0.396	98.083	0.227	0.13-0.17	0.26-0.32	0.37-0.43	82.44-113.72	0.21-0.25
Baseline - Snippet GPT-4o-mini (k=1000)										
1	0.043	0.126	0.194	134.714	0.097	0.03-0.06	0.11-0.15	0.17-0.22	118.15-151.27	0.08-0.11
2	0.109	0.231	0.337	95.198	0.181	0.09-0.13	0.20-0.26	0.31-0.37	80.84-109.56	0.16-0.20
3	0.148	0.286	0.393	90.195	0.226	0.13-0.17	0.26-0.31	0.36-0.42	74.67-105.72	0.20-0.25
4	0.171	0.294	0.390	78.418	0.243	0.15-0.19	0.27-0.32	0.36-0.42	66.24-90.60	0.22-0.27
5	0.190	0.305	0.414	74.093	0.257	0.17-0.21	0.28-0.33	0.38-0.44	62.37-85.82	0.23-0.28
Baseline - Snippet LLaMA-3.3 70B (k=100)										
1	0.034	0.122	0.191	204.307	0.090	0.02-0.05	0.10-0.14	0.17-0.22	176.03-232.58	0.08-0.10
2	0.085	0.191	0.271	174.967	0.148	0.07-0.10	0.17-0.22	0.24-0.30	148.24-201.69	0.13-0.17
3	0.118	0.263	0.362	131.685	0.195	0.10-0.14	0.24-0.29	0.33-0.39	111.34-152.03	0.17-0.21
4	0.160	0.290	0.400	109.844	0.233	0.14-0.18	0.26-0.32	0.37-0.43	91.10-128.59	0.21-0.25
5	0.185	0.316	0.415	96.587	0.255	0.16-0.21	0.29-0.34	0.38-0.45	80.43-112.74	0.23-0.28
SnipRec GPT-4o-mini (k=500)										
1	0.034	0.119	0.191	152.083	0.088	0.02-0.05	0.10-0.14	0.17-0.22	133.05-171.12	0.08-0.10
2	0.097	0.218	0.305	111.660	0.166	0.08-0.12	0.19-0.24	0.28-0.33	94.83-128.49	0.15-0.18
3	0.152	0.288	0.387	93.688	0.228	0.13-0.17	0.26-0.32	0.36-0.42	79.16-108.22	0.21-0.25
4	0.180	0.315	0.429	81.767	0.257	0.16-0.20	0.29-0.34	0.40-0.46	68.26-95.27	0.23-0.28
5	0.197	0.341	0.454	77.189	0.277	0.17-0.22	0.31-0.37	0.42-0.48	63.21-91.17	0.25-0.30
SnipRec LLaMA-3.3 70B (k=500)										
1	0.032	0.127	0.209	143.613	0.091	0.02-0.04	0.11-0.15	0.18-0.23	126.61-160.61	0.08-0.10
2	0.089	0.201	0.282	154.001	0.156	0.07-0.11	0.18-0.23	0.25-0.31	134.28-173.72	0.14-0.17
3	0.131	0.270	0.366	135.404	0.205	0.11-0.15	0.24-0.30	0.34-0.40	114.79-156.02	0.18-0.23
4	0.160	0.326	0.424	104.493	0.244	0.14-0.18	0.30-0.36	0.39-0.45	87.90-121.08	0.22-0.27
5	0.185	0.333	0.438	93.701	0.263	0.16-0.21	0.30-0.36	0.41-0.47	78.24-109.16	0.24-0.29
SnipRec + Expansion GPT-4o-mini (k=500)										
1	0.037	0.132	0.212	129.337	0.096	0.03-0.05	0.11-0.15	0.19-0.24	113.89-144.79	0.08-0.11
2	0.090	0.229	0.337	98.499	0.169	0.07-0.11	0.20-0.26	0.31-0.37	84.04-112.95	0.15-0.19
3	0.143	0.306	0.406	92.005	0.227	0.12-0.16	0.28-0.33	0.38-0.44	76.01-108.00	0.21-0.25
4	0.172	0.346	0.440	75.537	0.260	0.15-0.20	0.32-0.38	0.41-0.47	63.02-88.05	0.24-0.28
5	0.189	0.360	0.464	65.323	0.277	0.16-0.21	0.33-0.39	0.43-0.49	54.61-76.03	0.25-0.30
SnipRec + Expansion LLaMA-3.3 70B (k=500)										
1	0.038	0.138	0.223	139.549	0.100	0.03-0.05	0.12-0.16	0.20-0.25	122.82-156.28	0.09-0.11
2	0.092	0.205	0.301	127.229	0.159	0.07-0.11	0.18-0.23	0.27-0.33	110.47-143.99	0.14-0.18
3	0.137	0.280	0.381	116.490	0.214	0.12-0.16	0.25-0.31	0.35-0.41	98.69-134.29	0.19-0.24
4	0.175	0.329	0.440	95.486	0.256	0.15-0.20	0.30-0.36	0.41-0.47	79.67-111.30	0.23-0.28
5	0.199	0.341	0.459	83.463	0.278	0.17-0.22	0.31-0.37	0.43-0.49	69.99-96.94	0.25-0.30

Table 7: Full system performance at each turn for the restaurant dataset.

Turn	Hits@1	Hits@5	Hits@10	Avg Pos	Avg MRR	Hits@1 95ci	Hits@5 95ci	Hits@10 95ci	Avg Pos 95ci	Avg MRR 95ci
Baseline - Document GPT-4o-mini (k=100)										
1	0.020	0.076	0.133	1979.563	0.063	0.01-0.03	0.06-0.09	0.11-0.15	1790.70-2168.43	0.05-0.07
2	0.062	0.205	0.305	1201.196	0.138	0.05-0.08	0.18-0.23	0.28-0.33	1045.12-1357.27	0.12-0.15
3	0.106	0.268	0.371	956.239	0.188	0.09-0.13	0.24-0.30	0.34-0.40	812.28-1100.20	0.17-0.21
4	0.130	0.284	0.403	846.761	0.210	0.11-0.15	0.26-0.31	0.37-0.43	712.86-980.67	0.19-0.23
5	0.148	0.292	0.422	792.695	0.227	0.13-0.17	0.26-0.32	0.39-0.45	658.59-926.80	0.21-0.25
Baseline - Document LLaMA-3.3 70B (k=100)										
1	0.016	0.080	0.147	1868.929	0.063	0.01-0.02	0.06-0.10	0.13-0.17	1686.99-2050.87	0.05-0.07
2	0.062	0.174	0.295	1264.248	0.127	0.05-0.08	0.15-0.20	0.27-0.32	1103.34-1425.16	0.11-0.14
3	0.100	0.230	0.321	1021.103	0.173	0.08-0.12	0.20-0.26	0.29-0.35	875.37-1166.84	0.15-0.19
4	0.124	0.257	0.339	882.698	0.197	0.10-0.14	0.23-0.28	0.31-0.37	746.00-1019.40	0.18-0.22
5	0.138	0.261	0.360	859.349	0.209	0.12-0.16	0.23-0.29	0.33-0.39	726.88-991.82	0.19-0.23
Baseline - Sentence GPT-4o-mini (k=100)										
1	0.103	0.225	0.299	1506.339	0.169	0.08-0.12	0.20-0.25	0.27-0.33	1333.48-1679.19	0.15-0.19
2	0.209	0.351	0.415	1136.255	0.282	0.18-0.23	0.32-0.38	0.38-0.45	982.63-1289.88	0.26-0.31
3	0.257	0.404	0.472	1021.802	0.327	0.23-0.28	0.37-0.43	0.44-0.50	869.68-1173.92	0.30-0.35
4	0.285	0.413	0.486	946.894	0.349	0.26-0.31	0.38-0.44	0.45-0.52	803.36-1090.43	0.32-0.38
5	0.306	0.421	0.479	825.236	0.364	0.28-0.33	0.39-0.45	0.45-0.51	692.67-957.80	0.34-0.39
Baseline - Sentence LLaMA-3.3 70B (k=1000)										
1	0.101	0.220	0.288	562.415	0.165	0.08-0.12	0.19-0.25	0.26-0.32	453.63-671.20	0.15-0.18
2	0.185	0.335	0.408	360.629	0.259	0.16-0.21	0.31-0.36	0.38-0.44	282.28-438.98	0.24-0.28
3	0.237	0.353	0.439	303.673	0.299	0.21-0.26	0.32-0.38	0.41-0.47	233.82-373.52	0.27-0.32
4	0.263	0.352	0.429	289.565	0.317	0.24-0.29	0.32-0.38	0.40-0.46	222.26-356.87	0.29-0.34
5	0.274	0.366	0.422	247.523	0.325	0.25-0.30	0.34-0.40	0.39-0.45	198.17-296.88	0.30-0.35
Baseline - Snippet GPT-4o-mini (k=1000)										
1	0.113	0.237	0.302	493.710	0.179	0.09-0.13	0.21-0.26	0.27-0.33	399.35-588.07	0.16-0.20
2	0.222	0.352	0.428	350.655	0.291	0.20-0.25	0.32-0.38	0.40-0.46	274.10-427.21	0.27-0.32
3	0.274	0.421	0.495	292.019	0.348	0.25-0.30	0.39-0.45	0.46-0.53	224.40-359.64	0.32-0.37
4	0.305	0.429	0.499	320.447	0.369	0.28-0.33	0.40-0.46	0.47-0.53	246.70-394.19	0.34-0.40
5	0.325	0.446	0.504	255.871	0.385	0.30-0.35	0.42-0.48	0.47-0.54	200.36-311.38	0.36-0.41
Baseline - Snippet LLaMA-3.3 70B (k=500)										
1	0.116	0.243	0.311	758.719	0.186	0.10-0.14	0.22-0.27	0.28-0.34	636.88-880.55	0.17-0.21
2	0.205	0.361	0.431	568.457	0.282	0.18-0.23	0.33-0.39	0.40-0.46	462.04-674.87	0.26-0.31
3	0.252	0.381	0.461	507.324	0.317	0.23-0.28	0.35-0.41	0.43-0.49	406.86-607.79	0.29-0.34
4	0.276	0.384	0.451	452.367	0.334	0.25-0.30	0.35-0.41	0.42-0.48	357.41-547.32	0.31-0.36
5	0.294	0.395	0.453	460.492	0.349	0.27-0.32	0.36-0.43	0.42-0.48	365.74-555.25	0.32-0.38
SNIPREC GPT-4o-mini (k=1000)										
1	0.155	0.322	0.391	455.690	0.235	0.13-0.18	0.29-0.35	0.36-0.42	368.47-542.91	0.21-0.26
2	0.259	0.430	0.492	328.777	0.339	0.23-0.29	0.40-0.46	0.46-0.52	255.68-401.87	0.31-0.36
3	0.325	0.483	0.559	306.239	0.400	0.30-0.35	0.45-0.51	0.53-0.59	230.10-382.38	0.37-0.43
4	0.357	0.485	0.550	251.113	0.420	0.33-0.39	0.45-0.52	0.52-0.58	190.80-311.42	0.39-0.45
5	0.368	0.489	0.552	258.158	0.429	0.34-0.40	0.46-0.52	0.52-0.58	193.37-322.94	0.40-0.46
SNIPREC LLaMA-3.3 70B (k=1000)										
1	0.149	0.302	0.370	605.365	0.225	0.13-0.17	0.27-0.33	0.34-0.40	496.77-713.96	0.20-0.25
2	0.240	0.389	0.466	401.158	0.314	0.21-0.27	0.36-0.42	0.44-0.50	316.25-486.06	0.29-0.34
3	0.289	0.413	0.491	346.228	0.353	0.26-0.32	0.38-0.44	0.46-0.52	269.90-422.55	0.33-0.38
4	0.314	0.420	0.491	326.659	0.369	0.29-0.34	0.39-0.45	0.46-0.52	254.03-399.29	0.34-0.40
5	0.323	0.430	0.495	329.054	0.376	0.29-0.35	0.40-0.46	0.46-0.53	256.44-401.67	0.35-0.40
SNIPREC + Expansion GPT-4o-mini (k=1000)										
1	0.150	0.318	0.384	333.929	0.230	0.13-0.17	0.29-0.35	0.35-0.41	268.01-399.85	0.21-0.25
2	0.246	0.414	0.489	276.166	0.329	0.22-0.27	0.38-0.44	0.46-0.52	214.54-337.79	0.30-0.35
3	0.311	0.464	0.535	219.170	0.387	0.28-0.34	0.43-0.49	0.50-0.57	168.26-270.08	0.36-0.41
4	0.335	0.471	0.540	225.405	0.401	0.31-0.36	0.44-0.50	0.51-0.57	172.79-278.02	0.37-0.43
5	0.353	0.482	0.545	209.771	0.414	0.32-0.38	0.45-0.51	0.51-0.58	161.52-258.02	0.39-0.44
SNIPREC + Expansion LLaMA-3.3 70B (k=1000)										
1	0.152	0.321	0.394	340.011	0.234	0.13-0.17	0.29-0.35	0.36-0.42	271.78-408.24	0.21-0.26
2	0.249	0.403	0.489	235.439	0.326	0.22-0.28	0.37-0.43	0.46-0.52	183.89-286.99	0.30-0.35
3	0.291	0.438	0.520	208.621	0.362	0.26-0.32	0.41-0.47	0.49-0.55	162.37-254.87	0.34-0.39
4	0.315	0.443	0.513	204.563	0.379	0.29-0.34	0.41-0.47	0.48-0.54	157.46-251.67	0.35-0.41
5	0.331	0.439	0.513	197.468	0.389	0.30-0.36	0.41-0.47	0.48-0.54	158.00-236.94	0.36-0.42

Table 8: Full system performance at each turn for the book dataset.

Turn	Hits@1	Hits@5	Hits@10	Avg Pos	Avg MRR	Hits@1 95ci	Hits@5 95ci	Hits@10 95ci	Avg Pos 95ci	Avg MRR 95ci
Baseline - Document GPT-4o-mini (k=100)										
1	0.010	0.050	0.107	1934.348	0.047	0.00-0.02	0.04-0.06	0.09-0.13	1750.77-2117.93	0.04-0.05
2	0.035	0.092	0.133	1515.333	0.072	0.02-0.05	0.07-0.11	0.11-0.15	1352.60-1678.07	0.06-0.08
3	0.042	0.100	0.150	1519.049	0.079	0.03-0.05	0.08-0.12	0.13-0.17	1354.95-1683.15	0.07-0.09
4	0.052	0.110	0.163	1341.672	0.090	0.04-0.07	0.09-0.13	0.14-0.19	1179.50-1503.84	0.08-0.10
5	0.059	0.118	0.167	1254.036	0.098	0.04-0.07	0.10-0.14	0.14-0.19	1099.20-1408.87	0.08-0.11
Baseline - Document LLaMA-3.3 70B (k=100)										
1	0.011	0.058	0.114	1875.500	0.050	0.00-0.02	0.04-0.07	0.09-0.13	1692.44-2058.56	0.04-0.06
2	0.034	0.110	0.150	1574.288	0.076	0.02-0.05	0.09-0.13	0.13-0.17	1403.32-1745.26	0.06-0.09
3	0.048	0.115	0.185	1351.646	0.091	0.03-0.06	0.10-0.13	0.16-0.21	1192.86-1510.43	0.08-0.10
4	0.057	0.125	0.178	1234.909	0.099	0.04-0.07	0.10-0.15	0.15-0.20	1082.26-1387.56	0.08-0.11
5	0.062	0.117	0.170	1242.217	0.099	0.05-0.08	0.10-0.14	0.15-0.19	1088.56-1395.87	0.08-0.11
Baseline - Sentence GPT-4o-mini (k=1000)										
1	0.031	0.113	0.170	780.693	0.082	0.02-0.04	0.09-0.13	0.15-0.19	653.85-907.53	0.07-0.09
2	0.064	0.115	0.160	706.770	0.099	0.05-0.08	0.10-0.13	0.14-0.18	597.16-816.38	0.08-0.11
3	0.078	0.132	0.186	653.231	0.114	0.06-0.09	0.11-0.15	0.16-0.21	553.00-753.46	0.10-0.13
4	0.085	0.141	0.187	594.313	0.120	0.07-0.10	0.12-0.16	0.16-0.21	503.94-684.69	0.10-0.14
5	0.090	0.138	0.192	564.795	0.125	0.07-0.11	0.12-0.16	0.17-0.22	481.93-647.66	0.11-0.14
Baseline - Sentence LLaMA-3.3 70B (k=1000)										
1	0.032	0.110	0.167	748.171	0.083	0.02-0.04	0.09-0.13	0.14-0.19	629.35-866.99	0.07-0.09
2	0.053	0.112	0.158	704.295	0.094	0.04-0.07	0.09-0.13	0.14-0.18	596.25-812.34	0.08-0.11
3	0.068	0.126	0.183	691.014	0.105	0.05-0.08	0.11-0.15	0.16-0.21	584.23-797.80	0.09-0.12
4	0.083	0.133	0.191	631.332	0.117	0.07-0.10	0.11-0.15	0.17-0.22	536.46-726.21	0.10-0.13
5	0.088	0.141	0.191	596.028	0.123	0.07-0.11	0.12-0.16	0.17-0.22	506.10-685.95	0.10-0.14
Baseline - Snippet GPT-4o-mini (k=500)										
1	0.039	0.128	0.200	1081.991	0.093	0.03-0.05	0.11-0.15	0.18-0.22	935.58-1228.40	0.08-0.11
2	0.071	0.142	0.194	951.478	0.115	0.06-0.09	0.12-0.16	0.17-0.22	820.43-1082.52	0.10-0.13
3	0.090	0.154	0.199	929.788	0.128	0.07-0.11	0.13-0.18	0.17-0.22	801.13-1058.45	0.11-0.15
4	0.097	0.162	0.203	858.729	0.136	0.08-0.12	0.14-0.18	0.18-0.23	733.62-983.84	0.12-0.15
5	0.107	0.162	0.212	771.016	0.142	0.09-0.13	0.14-0.18	0.19-0.24	656.00-886.03	0.12-0.16
Baseline - Snippet LLaMA-3.3 70B (k=100)										
1	0.045	0.133	0.207	2085.765	0.098	0.03-0.06	0.11-0.15	0.18-0.23	1895.93-2275.60	0.08-0.11
2	0.079	0.139	0.185	1785.214	0.117	0.06-0.10	0.12-0.16	0.16-0.21	1610.68-1959.75	0.10-0.13
3	0.093	0.152	0.193	1715.030	0.128	0.07-0.11	0.13-0.17	0.17-0.22	1540.30-1889.76	0.11-0.15
4	0.103	0.152	0.205	1639.961	0.136	0.08-0.12	0.13-0.17	0.18-0.23	1466.08-1813.84	0.12-0.15
5	0.108	0.156	0.208	1555.878	0.141	0.09-0.13	0.13-0.18	0.18-0.23	1382.13-1729.63	0.12-0.16
SnipRec GPT-4o-mini (k=1000)										
1	0.041	0.133	0.212	1048.920	0.099	0.03-0.05	0.11-0.15	0.19-0.24	906.72-1191.12	0.09-0.11
2	0.062	0.156	0.227	840.274	0.121	0.05-0.08	0.13-0.18	0.20-0.25	709.96-970.59	0.10-0.14
3	0.081	0.170	0.241	716.963	0.136	0.06-0.10	0.15-0.19	0.21-0.27	600.57-833.36	0.12-0.15
4	0.091	0.168	0.232	634.158	0.142	0.07-0.11	0.14-0.19	0.21-0.26	527.93-740.39	0.12-0.16
5	0.095	0.175	0.253	606.619	0.146	0.08-0.11	0.15-0.20	0.23-0.28	501.88-711.36	0.13-0.16
SnipRec LLaMA-3.3 70B (k=1000)										
1	0.037	0.141	0.215	825.146	0.096	0.03-0.05	0.12-0.16	0.19-0.24	696.72-953.57	0.08-0.11
2	0.069	0.165	0.233	708.483	0.124	0.05-0.08	0.14-0.19	0.21-0.26	589.36-827.61	0.11-0.14
3	0.080	0.158	0.241	535.007	0.131	0.06-0.10	0.14-0.18	0.21-0.27	438.02-631.99	0.11-0.15
4	0.083	0.164	0.250	507.373	0.135	0.07-0.10	0.14-0.19	0.22-0.28	415.12-599.63	0.12-0.15
5	0.091	0.170	0.249	514.354	0.143	0.07-0.11	0.15-0.19	0.22-0.28	419.79-608.91	0.12-0.16
SnipRec + Expansion GPT-4o-mini (k=500)										
1	0.040	0.128	0.208	979.550	0.097	0.03-0.05	0.11-0.15	0.18-0.23	841.81-1117.29	0.08-0.11
2	0.065	0.162	0.231	759.479	0.123	0.05-0.08	0.14-0.18	0.20-0.26	640.86-878.10	0.11-0.14
3	0.090	0.167	0.232	656.578	0.141	0.07-0.11	0.14-0.19	0.21-0.26	545.53-767.63	0.12-0.16
4	0.101	0.168	0.226	576.390	0.149	0.08-0.12	0.14-0.19	0.20-0.25	475.24-677.54	0.13-0.17
5	0.108	0.174	0.232	498.779	0.152	0.09-0.13	0.15-0.20	0.21-0.26	410.83-586.73	0.13-0.17
SnipRec + Expansion LLaMA-3.3 70B (k=1000)										
1	0.033	0.136	0.218	558.727	0.094	0.02-0.04	0.11-0.16	0.19-0.24	464.37-653.09	0.08-0.11
2	0.072	0.167	0.239	507.867	0.129	0.06-0.09	0.14-0.19	0.21-0.27	416.63-599.10	0.11-0.15
3	0.086	0.173	0.246	439.198	0.140	0.07-0.10	0.15-0.20	0.22-0.27	359.31-519.08	0.12-0.16
4	0.098	0.175	0.255	423.658	0.148	0.08-0.12	0.15-0.20	0.23-0.28	346.45-500.86	0.13-0.17
5	0.109	0.177	0.254	426.539	0.154	0.09-0.13	0.15-0.20	0.23-0.28	349.63-503.45	0.14-0.17

Table 9: Full system performance at each turn for the clothing dataset.

```
- Question: "{question}"
- Response: "{response}"
- Known intents: {intents}
```

C.3 Query Snippet Expansion (§ 3.2)

Prompt for paraphrasing snippets

```
You are a helpful assistant. Strictly follow the format of the examples; do not provide anything other than the direct answer.
```

```
Paraphrase a given sentence.
The sentence should be an atomic, simple, short and coherent sentence that contain factual descriptions or subjective opinions related to a specific aspect of the restaurant/book.

Paraphrase this sentence: "{sentence}"
```

Prompt for generating supporting snippets (Yelp)

```
You are a helpful assistant. Strictly follow the format of the examples; do not provide anything other than the direct answer.
```

```
Generate a sentence that could serve as evidence for a given sentence.
The sentence should be an atomic, simple, short and coherent sentence that contain factual descriptions or subjective opinions related to a specific aspect of the restaurant.

<example>
Given sentence: "the restaurant is located in a bad neighborhood"
the restaurant is near bad crime area.

Given sentence: "the place is vegetarian-friendly."
the menu contains some veggie options.

Given sentence: "the place is good for family dinner."
high chairs are available.
<\example>

Given sentence: "{sentence}"
```

Prompt for generating opposing snippets (Yelp)

```
You are a helpful assistant. Strictly follow the format of the examples; do not provide anything other than the direct answer.
```

```
Generate the sentence of opposite meaning in restaurant domain following the examples.

<example>
What's the opposite of this sentence: "this place has sweet options like Cannolis"
This place lacks sweet options like Cannolis.

What's the opposite of this sentence: "this place is too crowded."
This place is very spacious.
<\example>

What's the opposite of this sentence: "{sentence}"
```

C.4 Clarification Question Generation (§ 3.4)

Prompt for generating clarification questions

```
You are a Recommender chatting with a Seeker to provide restaurant/book recommendation. Your task is to ask questions for understanding the Seeker's preference.
```

```
# Role-Play Task: Recommender
You will play the role of a Recommender helping a Seeker find a restaurant/book that suits the Seeker's preferences.

Based on the conversation log provided below, identify the most relevant aspect of the Seeker's preferences that will help refine the search for a suitable restaurant/book. Your question should focus only on one topic. Do not ask about multiple topics at once.

**Dialogue Context:**
{context}
Recommender:

Now, generate a response in the role of the Recommender.

**Response Guidelines:**

- Your response should be concise, typically one sentence. Avoid asking multiple questions at once.
- Do not ask for a restaurant/book name or any personal or street names.
- Respond directly and concisely to the scenario without repeating the instructions or adding unrelated details. Use question types that give the human user more flexibility, allowing for creative and open-ended answers while staying relevant to the context.
```

D Illustrative Examples

D.1 User Simulation Example

We present an example of user simulation. Below are the selected item review, attribute information, and summary of its other favorable reviews, which are used to build the context provided to the simulator as in Appendix A.2.1.

Example context (Yelp) provided to user simulator prompt (§ A.2.1)

```
{review text}
```

```
Tried the cart at a street. I ordered the General Tso's chicken stir fry. I talked to our server about how spicy and he was like it's not that spicy but it definitely has a kick (I'm also a wimp haha) but the flavor is great. The guy working was super nice and personable while helping me order. The veggies taste fresh and are crisp and I am very happy with the size of the portion for $10. I'll definitely have a bit for leftovers!"
```

```
{item info}
```

```
Category: asian fusion
- Alcohol: none
- Ambience/touristy: False
- Ambience/hipster: False
- Ambience/romantic: False
- Ambience/divey: False
- Ambience/intimate: False
- Ambience/trendy: False
- Ambience/upscale: False
- Ambience/classy: False
- Ambience/casual: False
- BYOB: No
- BikeParking: No
- BusinessAcceptsBitcoin: Yes
- BusinessAcceptsCreditCards: Yes
- BusinessParking: None
```

```
- Caters: Yes
- DogsAllowed: Yes
- GoodForKids: No
- GoodForMeal/dessert: False
- GoodForMeal/latenight: False
- GoodForMeal/lunch: False
- GoodForMeal/dinner: False
- GoodForMeal/brunch: False
- GoodForMeal/breakfast: False
- HappyHour: No
- HasTV: No
- OutdoorSeating: Yes
- RestaurantsDelivery: Yes
- RestaurantsGoodForGroups: No
- Price Range: 11—30
- RestaurantsReservations: No
- RestaurantsTableService: No
- RestaurantsTakeOut: Yes
- WheelchairAccessible: No
- WiFi: no
```

```
{item review summary}
```

```
What people generally like: Customers appreciate the delicious food, particularly the drunken noodles and chili garlic sauce, which many found flavorful and well-prepared. The freshness of the vegetables and the generous portion sizes receive high praise, with some meals lasting enough for two servings. The friendly and helpful staff enhance the dining experience, creating a positive atmosphere at this restaurant. Convenient location adds to the appeal, making it a popular lunch spot for locals. Many reviewers recommend trying the different combinations and sauces available.
```

Using the above exemplar item, we also provide sample ratings for selected user responses on three dimensions: relevancy, on-topicness, faithfulness.

- **Question:** What type of atmosphere are you hoping for, like casual, upscale, or something in between?
- **Answer:** I'm looking for something casual.
- **Rating:** Relevant ✓, On-topic ✓, Faithful ×
- **Question:** What specific dishes or flavors are you most interested in trying at an Asian fusion restaurant?
- **Answer:** I'm interested in trying dishes like drunken noodles and anything with chili garlic sauce.
- **Rating:** Relevant ✓, On-topic ✓, Faithful ✓
- **Question:** Are you looking for a specific type of cuisine within Asian fusion, like Thai, Chinese, or Japanese influences?
- **Answer:** I'm open to any influences, but I really enjoyed the General Tso's chicken stir fry.
- **Rating:** Relevant ×, On-topic ×, Faithful ×

D.2 Conversational Recommendation Examples

Tables 10 and 11 present two example cases of conversational recommendation from the Yelp and Amazon book datasets, each spanning five turns.

GPT-4o-mini		
1	Recommender	Hello, what category of restaurant are you looking for?
	Seeker	I'm looking for a place that serves cheesesteaks.
	Snippets	**(+)** *cheesesteaks are served. the menu features a variety of cheesesteak sandwiches. Cheesesteaks are offered.* **(-)** *Cheesesteaks are not served.*
	Rank	11
2	Recommender	What type of atmosphere do you prefer for your cheesesteak experience, casual or more upscale?
	Seeker	Casual is definitely my vibe for cheesesteaks.
	Snippets	**(+)** *the atmosphere is casual. customers often wear jeans and t-shirts. The vibe is relaxed.* **(-)** *The atmosphere is formal.*
	Rank	10
3	Recommender	What is your preferred location or neighborhood for enjoying your casual cheesesteak?
	Seeker	I'm open to any location, just looking for a good cheesesteak spot.
	Snippets	**(+)** **(-)**
	Rank	10
4	Recommender	What is your preferred price range for a casual cheesesteak?
	Seeker	I'm looking for a price range between $11 and $30 for a cheesesteak.
	Snippets	**(+)** *the price range is between $11 and $30. most main dishes cost around $20. The prices vary from $11 to $30.* **(-)** *The price range is not between $11 and $30.*
	Rank	7
5	Recommender	What specific toppings or ingredients do you enjoy on your cheesesteak?
	Seeker	I enjoy classic toppings, but I'm not picky about specifics.
	Snippets	**(+)** *classic toppings are enjoyed. pepperoni and mushrooms are popular choices. Traditional toppings are appreciated.* **(-)** *Classic toppings are disliked.*
	Rank	0
LLaMa-3.3 70B		
1	Recommender	Hello, what category of restaurant are you looking for?
	Seeker	I'm looking for a place that serves cheesesteaks.
	Snippets	**(+)** *the place serves cheesesteaks. the menu includes a philly cheesesteak sandwich. They offer cheesesteaks here.* **(-)** *The place does not serve cheesesteaks.*
	Rank	13
2	Recommender	What kind of atmosphere are you envisioning for your cheesesteak dining experience?
	Seeker	I'm looking for a casual atmosphere.
	Snippets	**(+)** *the place has a casual atmosphere. people wear jeans and t-shirts when dining there. The restaurant has a relaxed ambiance.* **(-)** *The place has a formal atmosphere.*
	Rank	11
3	Recommender	How important is it for you to have a wide variety of cheesesteak options, such as different types of cheese or meats, at the restaurant?
	Seeker	I don't have a specific preference for the variety of cheesesteak options, just a solid cheesesteak is good for me!
	Snippets	**(+)** **(-)**
	Rank	11
4	Recommender	What is your preferred price range for this casual cheesesteak dining experience?
	Seeker	I'm looking for something in the $11-$30 range.
	Snippets	**(+)** *the price range is between $11 and $30. most meals cost around twenty dollars. Prices fall within the $11 to $30 range.* **(-)** *The price range is outside of $11 and $30, or more specifically: The price range is either under $11 or over $30.*
	Rank	2
5	Recommender	Are you open to trying a restaurant that is a bit of a walk from public transportation, or would you prefer one that is more conveniently located near a bus or train stop?
	Seeker	I'm okay with a bit of a walk from public transportation.
	Snippets	**(+)** *the place is within walking distance from public transportation. a bus stop is right in front of the restaurant. The restaurant is easily accessible on foot from the nearest public transport stop.* **(-)** *The place is far from public transportation.*
	Rank	7

Table 10: User and SnipRec + Expansion interaction excerpts from Yelp dataset. Rank represents the current position of the target item among all candidates.

		GPT-4o-mini
1	Recommender	Hello, which category of books are you looking for?
	Seeker	I'm looking for a biography or memoir, especially related to arts and literature.
	Snippets	(+) *It chronicles the life of a famous individual. It discusses various literary movements. It features discussions on famous artists. It recounts personal experiences from the author's life. The book is a life story. The book is a personal narrative. The book pertains to literature. The book pertains to the arts. the book is a biography. the book is a memoir. the book is related to arts. the book is related to literature.* (-) *The book is a work of fiction. The book is unrelated to literature. The book is unrelated to the arts.*
	Rank	7038
2	Recommender	What specific time period or artist's life are you most interested in exploring through a biography or memoir?
	Seeker	I'm interested in biographies that capture the spirit of historical events, even if they take some liberties with the facts.
	Snippets	(+) *It vividly describes key moments from the past. The author alters historical events for dramatic effect. The book alters some of the facts. The book reflects the essence of historical events. the book captures the spirit of historical events. the book takes some liberties with the facts.* (-) *The book adheres strictly to the facts. The book fails to convey the essence of historical events.*
	Rank	8743
3	Recommender	What specific historical events or figures in the arts and literature are you particularly drawn to?
	Seeker	I'm drawn to biographies that explore royal history and the complexities of personal relationships, even if they aren't entirely factually accurate.
	Snippets	(+) *It details the lives of several monarchs. The book examines the history of royalty. The book examines the intricacies of personal connections. The characters experience deep emotional conflicts. the book explores royal history. the book explores the complexities of personal relationships.* (-) *The book avoids the intricacies of personal relationships. The book ignores royal history.*
	Rank	242
4	Recommender	What specific royal figures or historical events related to the arts and literature are you most interested in?
	Seeker	I'm particularly interested in biographies that delve into royal figures and their personal relationships, even if they take some creative liberties with the facts.
	Snippets	(+) *It explores the lives of several kings and queens. The book examines individual connections. The book explores royal characters. The characters have deep emotional connections. the book delves into royal figures. the book explores personal relationships* (-) *The book avoids any mention of royal figures. The book ignores personal relationships.*
	Rank	70
5	Recommender	Which specific royal figures or dynasties are you most fascinated by in terms of their personal relationships and artistic contributions?
	Seeker	I'm really fascinated by Queen Elizabeth II and her relationships, especially how they played out against the backdrop of royal duties.
	Snippets	(+) *It explores her reign and significant events. It provides insights into her friendships and rivalries. The book examines Queen Elizabeth II's connections. The book focuses on Queen Elizabeth II. The characters navigate court politics. The story takes place amid royal responsibilities. the book explores Queen Elizabeth II's relationships. the book is about Queen Elizabeth II. the book is set against the backdrop of royal duties.* (-) *The book ignores Queen Elizabeth II's relationships. The book is not about Queen Elizabeth II. The book is set against the backdrop of ordinary life.*
	Rank	9
		LLaMa-3.3 70B
1	Recommender	Hello, which category of books are you looking for?
	Seeker	I'm looking for biographies or memoirs, especially related to arts and literature.
	Snippets	(+) *the book is a biography. the book is a memoir. the book is related to arts and literature. It describes the subject's life events in chronological order. The author recounts their personal experiences. It discusses various famous authors and their works. The book tells the story of someone's life. The book is an autobiography. The book focuses on creative fields such as art and literature.* (-) *The book is not a biography. The book is not a memoir. The book is unrelated to arts and literature.*
	Rank	203
2	Recommender	What specific aspects of arts and literature, such as the life of a particular artist, writer, or movement, interest you the most in a biography or memoir?
	Seeker	I'm really into royal history and the personal lives of historical figures.
	Snippets	(+) *the book is about royal history. the book is about the personal lives of historical figures. It describes the lives of several monarchs. It describes the daily routines of famous leaders. The book focuses on the history of royalty. The book explores the private lives of famous historical individuals.* (-) *The book is not about royal history. The book is not about the personal lives of historical figures.*
	Rank	60
3	Recommender	What time period or historical era would you prefer the royal history or personal lives of historical figures to be from in the biography or memoir you're looking for?
	Seeker	I'm particularly interested in the 20th century, especially the British royal family.
	Snippets	(+) *the book is set in the 20th century. the book is about the British royal family. The story takes place during World War II. The story revolves around the lives of kings and queens of England. The story takes place during the 20th century. The book focuses on the British monarchy.* (-) *The book is not set in the 20th century. The book is not about the British royal family.*
	Rank	20
4	Recommender	What kind of experiences or aspects of the British royal family's lives in the 20th century would you like the biography or memoir to focus on, such as their struggles, achievements, or personal relationships?
	Seeker	I'm interested in their struggles and personal relationships, especially how they navigated challenges during that time.
	Snippets	(+) *the book focuses on the struggles of the British royal family. the book focuses on the personal relationships of the British royal family. The story revolves around the personal lives of kings and queens. The story explores the intimate interactions between Queen Elizabeth and her siblings. The book explores the challenges faced by the British monarchy. The book explores the interpersonal dynamics within the British monarchy.* (-) *The book focuses on something other than the struggles of the British royal family. The book ignores the personal relationships of the British royal family.*
	Rank	27
5	Recommender	How much emphasis would you like the biography or memoir to place on historical context and the broader social and political climate of the time, versus a more intimate, personal focus on the individual's experiences?
	Seeker	I'm looking for a balance between personal experiences and historical context, especially how they dealt with challenges.
	Snippets	(+) *the book provides a balance between personal experiences and historical context. the book includes how the individual dealt with challenges. The author weaves together personal anecdotes with thorough historical research. The author provides personal anecdotes about overcoming obstacles. The book strikes a balance between the author's personal stories and the broader historical background. The book discusses the individual's methods for overcoming obstacles.* (-) *The book lacks a balance between personal experiences and historical context. The book excludes how the individual dealt with challenges.*
	Rank	39

Table 11: User and SnipRec + Expansion interaction excerpts from Amazon book dataset. Rank represents the current position of the target item among all candidates.