Explanation of Revisions

Our paper was previously submitted to ARR in October 2024, December 2024, and February 2025. This revised manuscript addresses reviewer feedback.

This document outlines how we addressed the February 2025 feedback (Section 1) and summarizes manuscript improvements over previous cycles (Section 2).

1. Revisions from the Previous Submission (2025 Feb)

1.1 Meta-review

Below, we provide the mata review from the previous cycle.

Summary Of Reasons To Publish:

1. Domain Adaptability: The snippet-based framework is easily transferable across domains, reducing the need for domain-specific tuning.
2. User Simulation Innovation: The LLM-based simulator, evaluated for faithfulness and relevance (>90% alignment), provides a scalable alternative to human trials.

Summary Of Suggested Revisions:

1. Baseline Comparison: Clarify why classic CRS methods (e.g., attribute-based systems) were excluded as baselines. Discuss their incompatibility with the proposed data setting in detail.
2. Query Expansion Analysis: Include a dedicated section analyzing the impact of expansion types (paraphrase/support/opposite) on performance. Use metrics like BERTScore or human evaluation to quantify relevance.
3. User Simulator Robustness: Discuss potential risks if real users deviate from the simulator's behavior (e.g., providing off-topic responses). Suggest mitigation strategies.

1.2 Revisions

The current version addresses most of the feedback from the previous cycle. However, some points could not be addressed due to space constraints or lack of specification, as we did not hear back from reviewers despite requesting more details during the rebuttal period.

Suggestions	Our Revision
Baseline Comparison: Clarify why classic CRS methods (e.g., attribute-based systems)	**Clarified the comparisons** of our approach with existing studies in **Section 4.3.**

were excluded as baselines (Meta, Ka2S, ciEQ)	
Query expansion analysis: Evaluate the quality of expansion and analyze the impact of expansion types on performance. (Meta, Zd1B)	Performed a manual evaluation and included a discussion of the impact on performance in **Section 4.4**. Please see **Appendix A.1.3** and **B.2** for details.
User simulator robustness: Discuss the alignment of the user simulator and real users, higher user demands (Meta, Ka2S, ciEQ)	Clarified the implementation and evaluation of the user simulator in **Appendix A.2.** The user simulator used in our study **builds on existing frameworks** that have been shown to be effective for CRS evaluation. While we adapted it to our problem and manually evaluated its reliability **(Section 4.5)**, we do not consider this to be a primary research contribution of our work.
Snippet extraction analysis: Analyze potential hallucination and completeness (Ka2S, Zd1B)	Evaluated 30 reviews for each dataset (90 reviews in total) and found that decomposition is complete in many cases (21 for Yelp and Amazon Clothing, and 27 for Amazon Books). Clarified our manual evaluation results for hallucination, which was already included in the previous version (**Section 4.4**).
Improve the presentation of the paper (Ka2S, Zd1B)	Updated Figure 2 and tables. Clarified descriptions, especially the experimental setting and the interpretation of results, based on reviewer feedback.

2. Revisions Across ARR Cycles

In response to the reviewers' feedback, we have improved **the comprehensiveness of our evaluation,** including conducting a manual assessment of individual components, and have **clarified the scope and contributions of our research.**

Cycle	**Meta Review**	**Our Revision**
Oct-2024	**Summary Of Reasons To Publish:** ● This paper presents a well-reasoned approach to conversational recommendation systems. Using snippets extracted from customer reviews, SNIPREC eliminates the need for predefined attributes, allowing it to be scalable and adaptable across various domains. Another contribution of this study is incorporating a user simulator based on large	(1) Added experiments on the **Amazon Books** dataset. (2) Clarified that **the user simulator does not favor the proposed method.** (3) Added a discussion on **the reliability of LLM-based snippet extraction.** (early

	language models. This enables practical testing of CRS systems with minimal human involvement, enhancing scalability and reliability. Although the experimental results are limited to the Yelp dataset, they demonstrate that SNIPREC effectively captures user preferences and outperforms traditional document or sentence-based baselines. **Summary Of Suggested Revisions:** ● The paper introduces an intriguing framework. However, its evaluation is somewhat limited. Using a single Yelp dataset restricts the generalizability of the findings to other domains. (1)Incorporating additional datasets would significantly enhance the research. While the LLM-based user simulator is a valuable tool, (2)it operates under assumptions that may favor the snippet-based approach. This raises concerns about potential biases in the evaluation process. Including statistical significance tests and error bars would increase the credibility of the findings. (3)Also, relying on LLMs for snippet extraction carries the risk of noise or hallucination. Exploring alternative methods could help mitigate these risks. Lastly, please (4)improve the figures and tables and correct any typos the reviewers noted.	version of Section 4.4) (4) Improved the clarity of the paper and the presentation of tables and figures. Updated experiments to use the latest **LLaMA (v3.3).**
Dec-2024	**Summary Of Reasons To Publish:** ● SNIPREC leverages user-generated content to enhance the recommendation, which is intuitive and well-motivated. ● This approach eliminates domain-specific fine-tunning and can be easily deployed to various domains. **Summary Of Suggested Revisions:** ● There might be (1)hallucination in the LLM-generated snippets. It would be helpful to see some quantification or study of hallucination in snippets and how the hallucinated snippets affect the recommendation. ● There are some (2)concerns on the user simulator, e.g., whether it has hallucination in the evaluation and how it aligns with the real-world users. ● SNIPREC is limited by (3)its computational	(1) **Conducted manual and automatic analyses of item snippets,** finding hallucination rare (~3%) and most extracted snippets (~97%) conveying atomic meaning. (previous version of Section 4.4). (2) **Clarified our manual evaluation results for hallucination,** which was already included in the manuscript. (3) **Included inference time** for reader's reference, and clarified implementation details including computational resources. (Appendix B.3) (4) Improved the readability of the methodology section based

	cost. A computational analysis is suggested to demonstrate the effectiveness of this approach. • [4]The presentation of the methodology should be improved.	on feedback. Added experiments on the Amazon Clothing dataset.
Feb-2025	(See Section 1 of this document)	(See Section 1 of this document)