

# Exploring Multi-LLM Collaboration to Power Conversational Recommender System: A Case Study of Dietary Recommendation

Minhui Liang

Department of Computer Science  
City University of Hong Kong  
Hong Kong, China  
mhliang4-c@my.cityu.edu.hk

Yuhan Luo

Department of Computer Science  
City University of Hong Kong  
Hong Kong, China  
yuhanluo@cityu.edu.hk

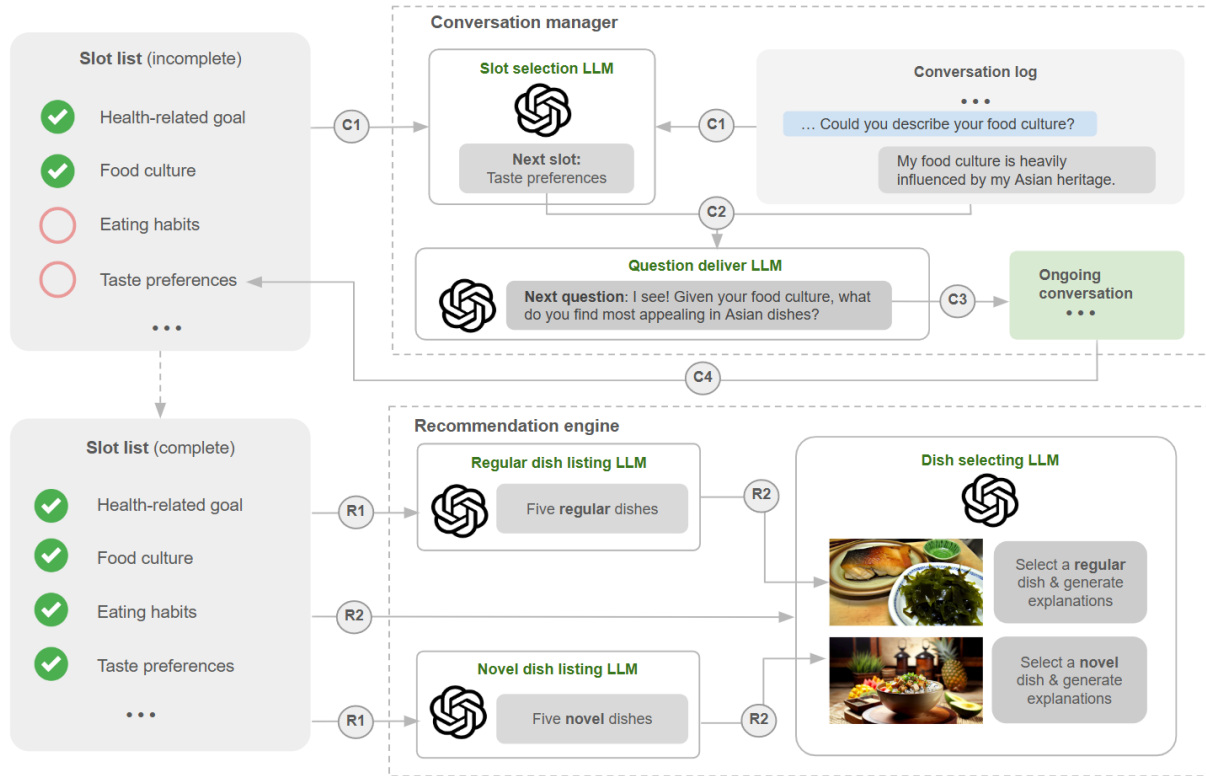


Figure 1: The multi-LLM collaboration pipeline in the conversation and recommendation phase.

## Abstract

Conversational recommender systems (CRS) are promising in delivering personalized recommendations by engaging users to share rich information about themselves, particularly in dietary recommendation, where various factors (e.g., food preferences, eating habits) needs to be considered. However, maintaining a coherent conversation for information collection and recommending healthy dishes tailored to different users remains challenging, even with

the emerging large language models (LLMs). In this study, we explore multi-LLM collaboration—where multiple LLMs specialize in subtasks of a complex problem—to enhance a dietary CRS. Through an online experiment ( $N = 161$ ), we compared multi-LLM collaboration with its single-LLM counterpart during the conversation and recommendation phases, evaluating system performance and participants’ experiences. We found multi-LLM collaboration equipped the conversation manager with greater adaptability to the conversation contexts, while powering the recommendation engine to deliver more nutritionally balanced and wide-range recommendations. Our discussion then focuses on the implications for designing user-centered CRS with LLMs.

## CCS Concepts

• Human-centered computing → HCI design and evaluation methods.

## Keywords

conversational recommender system, large language model, dietary recommendation

### ACM Reference Format:

Minhui Liang and Yuhao Luo. 2025. Exploring Multi-LLM Collaboration to Power Conversational Recommender System: A Case Study of Dietary Recommendation. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces (CUI '25)*, July 8–10, 2025, Waterloo, ON, Canada. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3719160.3737635>

## 1 Introduction and Backgrounds

Conversational recommender systems (CRS) are applications that make recommendations to individuals by gathering relevant information from users in natural language conversations [12]. Compared to traditional recommender systems that infer user preferences from online activities or rely on “one-shot” feedback (e.g., likes and dislikes) [7, 19], CRS actively refines its understanding of users through multi-turn dialogues, thereby providing more adaptive and tailored recommendations [12]. Particularly in the context of dietary recommendations, CRS excels in gathering rich information related to individuals’ diet, ranging from their health conditions and dietary restrictions to taste preferences and budgets, which are often unavailable in traditional recommender systems [1, 4, 20].

The recent rise of large language models (LLMs) holds the promise to scale up the development of CRS for broader audience: designers and developers can use natural language prompts to instruct the model’s data collection procedure [15, 26], while leveraging its extensive knowledge base to generate recommendations without external databases [5, 28]. However, building an effective LLM-powered CRS remains difficult. First, the challenge lies in structuring the conversation to gather sufficient information and respond to users in a coherent and engaging flow, while encouraging them to share more relevant information [13, 26]—balancing these objectives may exceed the capacity of a single prompt. Prior work shows that a single prompt often struggles to follow lengthy and complex instructions [15, 26]. Second, when generating open-ended items, such as recommendations, LLMs tend to focus on a narrow segment of the search space and make similar recommendations to different people, due to the generation mechanism that prioritizes items with the highest probabilities [2, 23].

In navigating the above challenges, we draw inspiration from recent research to explore the opportunities for multi-LLM collaboration, where multiple LLMs (or the same LLM with different prompts) are organized together in sequence or parallel to tackle complex tasks through a divide-and-conquer approach [6, 24, 25, 27]. One key advantage of this structure is allowing each LLM to focus on one specific sub-task without being burdened by overly lengthy prompts, enabling them to specialize in different aspects of the problem and complement each other’s generation space [6, 24, 25]. However, multi-LLM collaboration may increase the risk of error propagation (inaccuracies in one LLM’s output can cascade through the pipeline) and time cost [6, 24]. Not much work has explored how to design such a collaboration pipeline for CRS, and whether it can consistently improve the performance of conversation management and recommendation generation. Moreover, prior work primarily focused on assessing system metrics (e.g., Precision, Recall, and F1 score) [24], lacking empirical understanding of how end-users—who

engage with the CRS and receive the recommendations—perceive the performance of different system structures, which is crucial for developing more user-centered CRS.

In this light, we set out to explore LLM-powered CRS design for dietary recommendation, a context where diverse individualized needs, such as age, gender, taste preferences, and social lives, need to be carefully considered [3, 16, 17]. We developed SmartEats, using a  $2 \times 2$  design to examine single- versus multi-LLM structures in the conversation and recommendation phases, respectively. Specifically, we configured four versions of SmartEats: (1) single-LLM structures for both phases as baseline (PB), (2) multi-LLM conversation manager and single-LLM recommendation engine (PC), (3) single-LLM conversation manager and multi-LLM recommendation engine (PR), and (4) multi-LLM structures for both phases as a full version (PF). Here, multi-LLM structures consist of LLMs created using GPT-4 with distinct prompts, refined through collaboration with a professional nutritionist.

By comparing these four versions through an online between-subjects study ( $N = 161$ ), we found all versions successfully gathered the pre-defined information in over 97% cases (slot-filling rate), and the versions featuring multi-LLM conversation manager were able to adapt the question order based on the conversation context. While all versions achieved a recommendation acceptance rate of over 95%, the versions with the multi-LLM recommendation engine suggested a wider variety of dishes to different participants, which were also more nutritionally balanced. Additionally, participants in the groups with a multi-LLM recommendation engine reported it easier to communicate their preferences than in other groups.

Our findings contribute a systematic and empirical understanding of how multi-LLM collaboration compares to single-prompt structures in designing LLM-powered CRS, particularly regarding information collection, conversation management, recommendation quality, and user experience in the context of dietary recommendation. We also provide insights into the multi-LLM pipeline structure design, which can benefit CUI researchers and designers aiming to enhance CRS performance within and beyond the domain of dietary recommendation.

## 2 SmartEats Design and Development

In a CRS, interactions typically consist of two phases: the conversation phase, where the conversational agent (CA) engages with users and gathers information from them, and the recommendation phase, where the CA delivers recommendations [12]. Given the different objectives of these two phases, our design centered on developing a multi-LLM structured conversation manager and recommendation engine, and comparing them with their single-LLM counterparts, which were created using the same instruction content but consolidated into a single prompt. These LLM components were powered by GPT-4 with distinct prompts. To ensure the safety and appropriateness of the generated recommendations, we involved a professional nutritionist specialized in weight and chronic disease management to provide feedback during our design and internal test process.

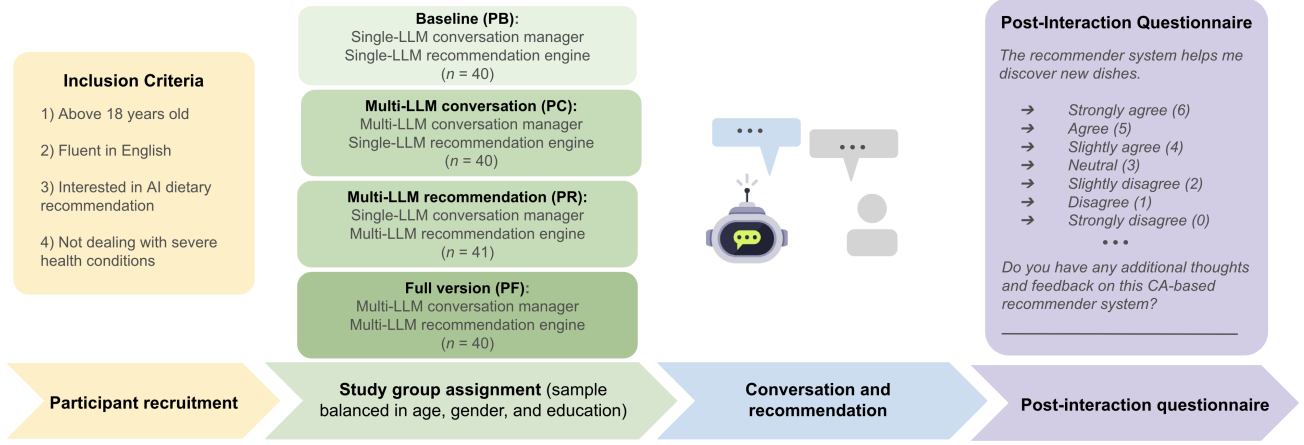


Figure 2: The procedure of the user study.

## 2.1 Conversation Manager

From prior research that identified key factors influencing people’s dietary practice [3, 16, 17], we defined 17 pieces of key information, referred to as “slots”, to collect from users during conversation. These slots can be grouped into seven categories: (1) demographics (age, height, weight, biological sex), (2) health-related goals, (3) recent emotional status, (4) eating habits (eating at regular times or only when one feels hungry, typical time spent on meals, commonly consumed food), (5) preferences (dietary restrictions, dining out or homemade meals, food culture, taste or cuisine), (6) food access (geographic location, budget), and (7) other lifestyle factors (eating alone or with others, exercise frequency).

To create a coherent, natural, and engaging conversation flow to gather these pieces of information, we created two LLMs working in sequence to form the multi-LLM conversation manager. As shown in Figure 1, first, a *slot selection* LLM is prompted to dynamically select a slot to be filled next, depending on the incomplete slot list and the conversation log ( $C1 \rightarrow C2$ ), and proceed to the recommendation phase when all questions corresponding to slots are delivered. In case the user expresses confusion about the question, the slot selection LLM will select the same slot again. Second, a *question deliver* LLM with a different prompt generates a question based on the selected slot and composes a message that contains responses to the user input in the previous round of conversation (e.g., acknowledging their situations) and naturally transits to the question ( $C2 \rightarrow C3$ ). Upon receiving a response from the user, the collected information will be saved in the slot list for later use ( $C4$ ).

## 2.2 Recommendation Engine

According to the USDA Healthy Eating Index [10], a healthy dietary recommendation should ensure nutritional balance in carbohydrates, proteins, and vegetables or fruits and cover various









food groups [8, 17], which our nutritionist collaborator also highlighted. Additionally, the recommendation engine should offer a wide range of foods to tailor to the unique needs of different individuals, while introducing novel recommendations that can bring surprises and joy [8].

Regarding this, we create multiple LLMs to work in parallel and sequence (see Figure 1). First, to create a wide range of recommendations, two different *dish listing* LLMs are prompted to generate dish candidates from the slots: one generates five dishes similar to the user’s regular diet (regular dish listing LLM), and the other generates five novel dishes that differs from the users’ typical meals (novel dish listing LLM). In both LLMs, we incorporated explicit instructions that the listed dishes should be nutritionally balanced and include different food groups in their main and side components ( $R1 \rightarrow R2$ ). Second, a *dish selecting* LLM selects one dish that best aligns with the person’s situation from each group and then recommends these two dishes to users. We also instructed this LLM to double-check the nutritional balance of the dishes. Along with the recommendations, the system generates images of the dishes with explanations to help users understand why the dishes are recommended to them. We implemented two rounds of recommendations, where users can provide feedback to refine the second round if unsatisfied with the first.

## 3 Method: A Between-Subjects Study

To examine the effectiveness of multi-LLM collaboration on CRS performance, we conducted a between-subjects study by comparing the single-LLM counterpart to the multi-LLM conversation manager and recommendation engine, as described in Figure 2. While messages and generated recommendations differed among the four groups, the interaction flows were the same. The study was approved by the university’s Ethics Review Committee.

**Table 1: Two conversation examples from participants in the PC and PB groups, both of whom expressed negative emotions in their responses. The PC version first acknowledged the participant’s feelings and then naturally transitioned to a question asking about their health goals. The PB version responded to the participant’s feelings in a more general way and then followed the question order listed in the prompt without transition. The bold texts corresponded to the predefined slots.**

| PC version (PC-7)   | PB version (PB-12)  |
|---|---|
|  ... How have you been <b>feeling emotionally</b> these days?<br> Bad, very stressed and anxious.<br> I am sorry to hear that you’re feeling stressed and anxious. If you’re open to it, managing stress and anxiety can also be supported through dietary choices. Do you have any specific <b>health goals</b> or aspects of your well-being that you would like to improve, which might also help with your stress and anxiety?<br> I would like to improve my diet to lose weight and increase fitness overall. |  ... How have you been <b>feeling emotionally</b> recently?<br> Slightly stressed with work, but other than that I feel good, just normal.<br> It’s quite common to feel stressed from work. Remember to take breaks and find time to relax, as stress can also affect your digestive health. Speaking of preferences, what <b>types of flavors or cuisines</b> do you enjoy the most?<br> I like Italian food, but I like most cuisines and try to have a range. |

**Participants:** we recruited participants on Prolific with the four inclusion criteria listed in Figure 2. The 161 participants who completed the study with validated responses aged from 18 to 69 ( $M = 34.80$ ,  $SD = 12.44$ ), including 84 females and 77 males, and were from the US (30) and the UK (131). At the end of the interaction with SmartEats, participants were asked to complete a post-interaction questionnaire on their perceived recommendation quality following prior literature in recommender systems (e.g., recommendation novelty, diversity across the dishes recommended each round, interaction adequacy, etc) [21], as well as their conversation experience (e.g., interaction satisfaction) [11, 21, 22]. Additionally, they shared open-ended thoughts and feedback with us. Participants spent 17 minutes on average to complete the entire study, each receiving EUR 2.25 as compensation.

**Data Analysis:** we compared the questionnaire responses of participants in the four groups using two-way ANOVA, treating the multi-LLM conversation manager and recommendation engine as the main effects. To further examine how these versions perform in managing the conversations, we analyzed participants’ conversation logs, deriving the slot-filling rate (i.e., the percentage of pre-defined slots that were filled by relevant user responses compared to the total number of slots) and question orders of different conversations (which reflects the conversation manager’s ability to dynamically adjust slot selections).

For the recommendation phase, we calculated the recommendation acceptance rate and examined whether the system generated nutritionally balanced meals—including vegetables or fruits, proteins, and carbohydrates [9, 10]. In the latter analysis, two researchers first independently analyzed a sample of recommended dishes (24.8%) and dummy-coded ‘1’ or ‘0’, indicating whether the dish was nutritionally balanced or not (i.e., ‘1’ for balanced dishes). Upon achieving a Cohen’s kappa of 0.83 and resolving discrepancies, the first author coded the remaining data. We also derived the range of dishes recommended to different participants in each group by calculating the text-similarity of dish names (e.g., the text-similarity between *Grilled lamb kebabs with quinoa salad* and *assorted bell peppers and spicy grilled chicken with mixed salad and quinoa*) [18]. A high level of name similarity among dishes recommended to different participants means a narrower range of recommended dishes. For both meal balance and recommendation range, we conducted two-way ANOVA tests to examine the effects of the multi-LLM collaboration. Additionally, we monitored the

time cost for generating responses during the conversation and recommendation phases, and reviewed participants’ responses to the last open-ended question to complements our quantitative data.

## 4 Results

In this section, we report the results of the performance and user experience with the four versions of SmartEats.

### 4.1 Conversation Phase

All groups achieved a slot-filling rate exceeding 97% (PB: 98.1%, PC: 97.2%, PR: 99.4%, PF: 98.2%), demonstrating the effectiveness of both single- and multi-LLM structures in information collection. The average time on processing user input and delivering questions did not significantly differ between the four groups (PB: 2.1 seconds, PC: 2.5 seconds, PR: 2.6 seconds, PF: 2.4 seconds), suggesting that multi-LLM collaboration did not increase time cost.

However, the single- and multi-LLM conversation manager showed difference in their ability to manage the conversation flow: all participants in the PB and PR groups received questions strictly follow the slot sequence listed in our prompt, despite the prompt specifying that the question order should be flexible and adapt to the conversation context; in contrast, 80 participants in the PC and PF groups received questions in 65 different orders, where the multi-LLM conversation manager demonstrated better adaptability by dynamically selecting questions (see Table 1). Although we did not observe significant difference in participants’ self-report conversation experiences among the four groups, some of those in the PC and PF groups highlighted that their conversation was natural and smooth (e.g., PF-29: “*the conversation was really smooth*”), while those in other groups felt the other way around: “*the questions were a bit random*” (PR-32), “*almost of a script*” (PB-32).

### 4.2 Recommendation Phase

All groups achieved an acceptance rate of over 95%, and the acceptance rate of the PR and PF groups reached 100% (PB: 97.5%, PC: 95.0%). Relatedly, participants in the PR and PF groups shared positive and excited comments: “*This is the first time I am able to do a survey that fits my dietary preferences!*” (PR-10). Similar to the conversation phase, the time cost on recommendation generation did not significantly differ among the four groups (PB: 31.4 seconds, PC: 29.4 seconds, PR: 33.0 seconds, PF: 28.5 seconds).

**Table 2: Participants’ perceptions of the recommendations. Statistical significance is marked with \* ( $p < .05$ ) and a marginally significant difference regarding recommendation diversity was observed, which is marked with † ( $p < .1$ ).**

| SmartEats versions                      | Group           | Recommendation diversity | Recommendation novelty | Interaction adequacy* | Eating intention      |
|---|-----------------|--------------------------|------------------------|-----------------------|-----------------------|
| With multi-LLM recommendation engine    | PF ( $n = 40$ ) | $M = 4.18, SD = 1.51$    | $M = 3.93, SD = 1.08$  | $M = 4.90, SD = 0.87$ | $M = 4.40, SD = 1.30$ |
|   | PR ( $n = 41$ ) | $M = 4.24, SD = 1.55$    | $M = 4.10, SD = 1.19$  | $M = 5.12, SD = 0.87$ | $M = 4.54, SD = 1.12$ |
| Without multi-LLM recommendation engine | PC ( $n = 40$ ) | $M = 3.53, SD = 1.77$    | $M = 3.65, SD = 1.32$  | $M = 4.50, SD = 1.38$ | $M = 4.20, SD = 1.38$ |
|   | PB ( $n = 40$ ) | $M = 4.03, SD = 1.52$    | $M = 3.95, SD = 1.24$  | $M = 4.78, SD = 1.05$ | $M = 4.35, SD = 1.19$ |

**Table 3: Nutritional balance and similarity scores of the recommended dishes. A higher similarity score indicates greater similarities among the dishes to different participants. Statistical significance is marked with \* ( $p < .05$ ), or \*\*\* ( $p < .001$ ).**

| SmartEats versions                      | Group           | Dish balance***       | Regular dish similarity*** | Novel dish similarity* |
|---|-----------------|-----------------------|----------------------------|------------------------|
| With multi-LLM recommendation engine    | PF ( $n = 40$ ) | $M = 0.89, SD = 0.31$ | $M = 0.46, SD = 0.26$      | $M = 0.21, SD = 0.11$  |
|   | PR ( $n = 41$ ) | $M = 0.90, SD = 0.30$ | $M = 0.44, SD = 0.22$      | $M = 0.19, SD = 0.06$  |
| Without multi-LLM recommendation engine | PC ( $n = 40$ ) | $M = 0.49, SD = 0.50$ | $M = 0.84, SD = 0.28$      | $M = 0.24, SD = 0.11$  |
|   | PB ( $n = 40$ ) | $M = 0.50, SD = 0.50$ | $M = 0.71, SD = 0.32$      | $M = 0.23, SD = 0.13$  |

Table 2 shows participants’ perceptions of the recommendations, in which PR and PF groups perceived better interaction adequacy ( $F = 4.976, p = .027$ ), meaning that the multi-LLM recommendation engine made it easier for participants to communicate their preferences to the system. With the multi-LLM recommendation engine, participants also perceived the dishes recommended to them as more diverse than the other two groups, with a marginal significance ( $F = 2.991, p = .086$ ). Furthermore, dishes recommended by the PR and PF versions exhibited smaller similarity scores, indicating they delivered a wider range of recommendations to different participants, including both regular and novel dishes (regular dishes:  $F = 63.758, p < .001$ ; novel dishes:  $F = 4.741, p = .031$ ).

As mentioned in section 2.2, we explicitly instructed all the versions of SmartEats to ensure the nutritional balance of the recommended dishes, which is an important aspect of a healthy diet. However, this instruction showed stronger effects on the PR and PF versions by including more nutritionally balanced dishes compared to PB and PC versions ( $F = 81.183, p < .001$ ). In other words, the multi-LLM recommendation engine demonstrated better effectiveness in adhering to the instruction details on nutritional balance, which was often overlooked by its single-LLM counterpart.

## 5 Discussions and Future Work

As the first step to exploring multi-LLM collaboration for enhancing the performance of CRS, our study uncovered nuanced yet important findings in the context of dietary recommendations. First, all versions achieved slot-filling rates exceeding 97%, indicating the latest GPT-4 with a single prompt could deliver all questions included in the prompt without omissions in most cases. This is an advancement compared to GPT-3, which was shown in prior study that achieved 79% slot-filling rate with a similar single-prompt design [26]. Nevertheless, the single-LLM conversation manager struggled to diversify the question orders, while the multi-LLM conversation manager exhibited greater adaptability to achieve this goal. Thus, this structure can benefit CAs for collecting long lists of

predefined information (e.g., pre-consultation [14]) through a more natural and engaging flow.

Moreover, although the multi-LLM recommendation engine did not significantly differ from its single-LLM counterpart in users’ recommendation acceptance rate, it demonstrated better recommendation quality regarding nutritional balance and recommendation range of dishes, as well as enhanced interaction adequacy. This approach expanded the space where the LLMs search for dishes by generating intermediate reasoning results to prevent a fixation on high-probability items [2, 23]. Lastly, multi-LLM collaboration did not necessarily increase time costs, indicating that resolving multiple sub-tasks may not sacrifice efficiency.

To conclude, the lessons learned from this work can be extended to broader contexts, such as recommending fitness plans and stress coping strategies. In the conversation phase, the multi-LLM conversation manager can dynamically select appropriate questions to maintain a coherent conversation flow, creating natural and engaging conversation experiences and eliciting richer information from users about their daily lives, health goals, and real-life challenges. In the recommendation phase, the multi-LLM recommendation engine can offer recommendations more appropriately, as the structure ensures better consistency in following prompt instruction details. These insights can inform CUI researchers and designers to develop CRS that deliver more engaging and tailored user experiences. Going forward, we aim to deploy SmartEats in a longitudinal study and examine its effects on individuals’ dietary behaviors in practice. This will include assessing how well individuals retain information from the recommendations and the extent to which they implement the suggestions in their daily lives.

## Acknowledgments

This project was supported by City University of Hong Kong (#9610597).

## References

- [1] Gediminas Adomavicius, Jesse Bockstedt, Shawn Curley, and Jingjing Zhang. 2011. Recommender systems, consumer preferences, and anchoring effects. In *RecSys 2011 workshop on human decision making in recommender systems*.

- CiteSeer, 35–42. <https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=4552ba5ac5cbb6c556ebd228c8f9116b46793bc2>
- [2] Keqin Bao, Jizhi Zhang, Yang Zhang, Xinyue Huo, Chong Chen, and Fuli Feng. 2024. Decoding Matters: Addressing Amplification Bias and Homogeneity Issue in Recommendations for Large Language Models. *ArXiv abs/2406.14900* (2024). <https://api.semanticscholar.org/CorpusID:270688514>
  - [3] Johnna Blair, Yuhan Luo, Ning F Ma, Sooyeon Lee, and Eun Kyoung Choe. 2018. OneNote Meal: A photo-based diary study for reflective meal tracking. In *AMIA Annual Symposium Proceedings*, Vol. 2018. American Medical Informatics Association, 252. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6371351/>
  - [4] Deepanjali Chowdhury, Ahana Roy, Sreenivasan Ramasamy Ramamurthy, and Nirmalya Roy. 2023. CHARLIE: A Chatbot That Recommends Daily Fitness and Diet Plans. In *2023 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*. 116–121. <https://doi.org/10.1109/PerComWorkshops56833.2023.10150359>
  - [5] Shishir Dwivedi, Nivedita Srivastava, Varun Rawal, and Deepali Dev. 2024. Healpal Chatmate: AI Driven Disease Diagnosis and Recommendation System. In *2024 2nd International Conference on Disruptive Technologies (ICDT)*. 1404–1408. <https://doi.org/10.1109/ICDT61202.2024.10489509>
  - [6] Shangbin Feng, Wenxuan Ding, Alisa Liu, Zifeng Wang, Weijia Shi, Yike Wang, Zejiang Shen, Xiaochuang Han, Hunter Lang, Chen-Yu Lee, et al. 2025. When One LLM Drools, Multi-LLM Collaboration Rules. *arXiv preprint arXiv:2502.04506* (2025). <https://arxiv.org/abs/2502.04506>
  - [7] Jill Freyne, Shlomo Berkovsky, and Gregory Smith. 2011. Recipe recommendation: accuracy and reasoning. In *International conference on user modeling, adaptation, and personalization*. Springer, 99–110. [https://doi.org/10.1007/978-3-642-22362-4\\_9](https://doi.org/10.1007/978-3-642-22362-4_9)
  - [8] Kazjon Grace, Elanor Finch, Natalia Gulbransen-Diaz, and Hamish Henderson. 2022. Q-Chef: The impact of surprise-eliciting systems on food-related decision-making. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>New Orleans</city>, <state>LA</state>, <country>USA</country>, </conf-loc>) (CHI '22). Association for Computing Machinery, New York, NY, USA, Article 11, 14 pages. <https://doi.org/10.1145/3491102.3501862>
  - [9] Health Eating Plate. 2024. <https://nutritionsource.hsph.harvard.edu/>.
  - [10] HEI. 2024. <https://www.fns.usda.gov/cnpp/healthy-eating-index-hei>.
  - [11] Victor Hung, Miguel Elvir, Avelino Gonzalez, and Ronald DeMara. 2009. Towards a method for evaluating naturalness in conversational dialog systems. In *2009 IEEE International Conference on Systems, Man and Cybernetics*. 1236–1241. <https://doi.org/10.1109/ICSMC.2009.5345904>
  - [12] Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. 2021. A Survey on Conversational Recommender Systems. *ACM Comput. Surv.* 54, 5, Article 105 (may 2021), 36 pages. <https://doi.org/10.1145/3453154>
  - [13] Eunkyung Jo, Yuin Jeong, Sohyun Park, Daniel A. Epstein, and Young-Ho Kim. 2024. Understanding the Impact of Long-Term Memory on Self-Disclosure with Large Language Model-Driven Chatbots for Public Health Intervention. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 440, 21 pages. <https://doi.org/10.1145/3613904.3642420>
  - [14] Brenna Li, Ofek Gross, Noah Crampton, Mamta Kapoor, Saba Tauseef Tetyana Skoropad, Mohit Jain, Khai Truong, and Alex Mariakakis. 2024. Beyond the Waiting Room: Patient's Perspectives on the Conversational Nuances of Pre-Consultation Chatbots. In *CHI 2024*. ACM. <https://www.microsoft.com/en-us/research/publication/beyond-the-waiting-room-patients-perspectives-on-the-conversational-nuances-of-pre-consultation-chatbots/>
  - [15] Zhuoyang Li, Minhui Liang, Ray Lc, and Yuhan Luo. 2024. StayFocused: Examining the Effects of Reflective Prompts and Chatbot Support on Compulsive Smartphone Use. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 247, 19 pages. <https://doi.org/10.1145/3613904.3642479>
  - [16] Yuhan Luo, Young-Ho Kim, Bongshin Lee, Naeemul Hassan, and Eun Kyoung Choe. 2021. FoodScrap: Promoting Rich Data Capture and Reflective Food Journaling Through Speech Input. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference* (Virtual Event, USA) (DIS '21). Association for Computing Machinery, New York, NY, USA, 606–618. <https://doi.org/10.1145/3461778.3462074>
  - [17] Yuhan Luo, Peiyi Liu, and Eun Kyoung Choe. 2019. Co-Designing Food Trackers with Dietitians: Identifying Design Opportunities for Food Tracker Customization. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) (CHI '19). Association for Computing Machinery, New York, NY, USA, 1–13. <https://doi.org/10.1145/3290605.3300822>
  - [18] Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing Order into Text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, Dekang Lin and Dekai Wu (Eds.). Association for Computational Linguistics, Barcelona, Spain, 404–411. <https://aclanthology.org/W04-3252/>
  - [19] Milica Milosavljevic, Vidhya Navalpakkam, Christof Koch, and Antonio Rangel. 2012. Relative visual saliency differences induce sizable bias in consumer choice. *Journal of consumer psychology* 22, 1 (2012), 67–74. <https://doi.org/10.1016/j.jcps.2011.10.002>
  - [20] Florian Pecune, Lucile Callebert, and Stacy Marsella. 2020. A Socially-Aware Conversational Recommender System for Personalized Recipe Recommendations. In *Proceedings of the 8th International Conference on Human-Agent Interaction* (Virtual Event, USA) (HAI '20). Association for Computing Machinery, New York, NY, USA, 78–86. <https://doi.org/10.1145/3406499.3415079>
  - [21] Pearl Pu, Li Chen, and Rong Hu. 2011. A user-centric evaluation framework for recommender systems. In *Proceedings of the Fifth ACM Conference on Recommender Systems* (Chicago, Illinois, USA) (RecSys '11). Association for Computing Machinery, New York, NY, USA, 157–164. <https://doi.org/10.1145/2043932.2043962>
  - [22] Nicole M. Radziwill and Morgan C. Benton. 2017. Evaluating Quality of Chatbots and Intelligent Conversational Agents. *ArXiv abs/1704.04579* (2017). <https://api.semanticscholar.org/CorpusID:6656629>
  - [23] Sangho Suh, Meng Chen, Bryan Min, Toby Jia-Jun Li, and Haijun Xia. 2024. Luminate: Structured Generation and Exploration of Design Space with Large Language Models for Human-AI Co-Creation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 644, 26 pages. <https://doi.org/10.1145/3613904.3642400>
  - [24] Shichao Sun, Ruifeng Yuan, Ziqiang Cao, Wenjie Li, and Pengfei Liu. 2024. Prompt Chaining or Stepwise Prompt? Refinement in Text Summarization. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 7551–7558. <https://doi.org/10.18653/v1/2024.findings-acl.449>
  - [25] Saranya Venkatraman, Nafis Irtiza Tripto, and Dongwon Lee. 2024. CollabStory: Multi-LLM Collaborative Story Generation and Authorship Analysis. *ArXiv abs/2406.12665* (2024). <https://api.semanticscholar.org/CorpusID:270562564>
  - [26] Jing Wei, Sungdong Kim, Hyunhoon Jung, and Young-Ho Kim. 2024. Leveraging Large Language Models to Power Chatbots for Collecting User Self-Reported Data. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 87 (apr 2024), 35 pages. <https://doi.org/10.1145/3637364>
  - [27] Tongshuang Wu, Ellen Jiang, Aaron Donsbach, Jeff Gray, Alejandra Molina, Michael Terry, and Carrie J Cai. 2022. PromptChainer: Chaining Large Language Model Prompts through Visual Programming. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) (CHI EA '22). Association for Computing Machinery, New York, NY, USA, Article 359, 10 pages. <https://doi.org/10.1145/3491101.3519729>
  - [28] Zhongqi Yang, Elahe Khatibi, Nitish Nagesh, Mahyar Abbasian, Iman Azimi, Ramesh Jain, and Amir M Rahmani. 2024. ChatDiet: Empowering Personalized Nutrition-Oriented Food Recommender Chatbots through an LLM-Augmented Framework. *arXiv preprint arXiv:2403.00781* (2024). <https://arxiv.org/abs/2403.00781>