iui 2021
Where HCI Meets AI
26th Annual Conference on Intelligent User Interfaces

Critiquing for Music Exploration in Conversational Recommender Systems

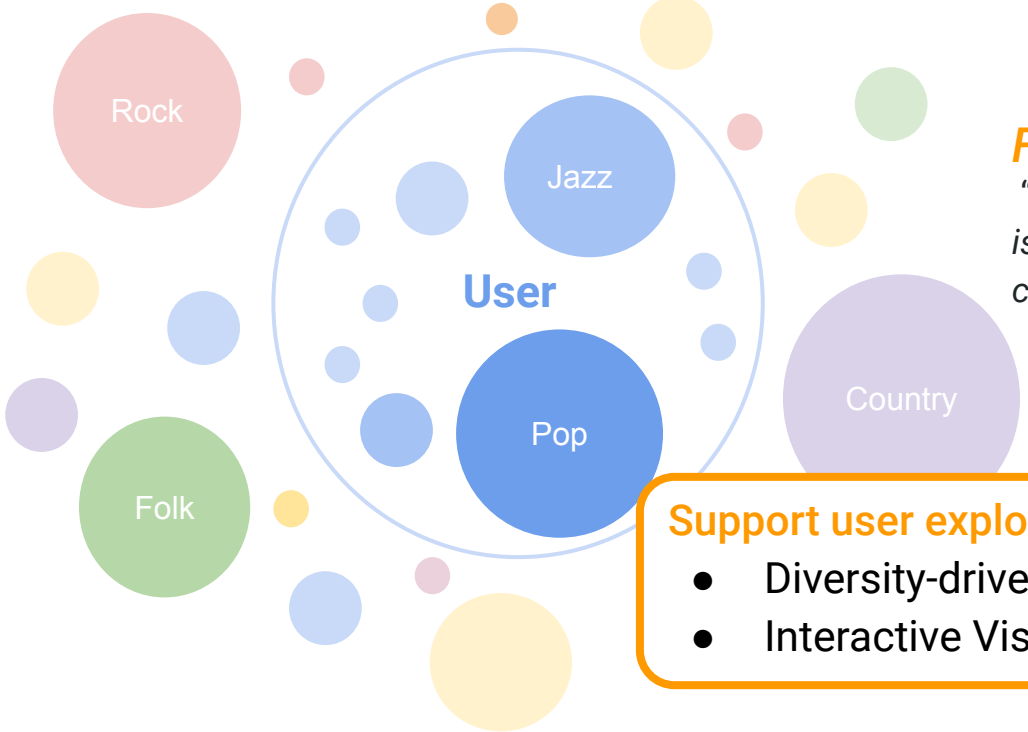
Wanling Cai
Hong Kong Baptist University
Yucheng Jin
Lenovo Research
Li Chen
Hong Kong Baptist University

香港浸會大學
HONG KONG BAPTIST UNIVERSITY

DEPARTMENT OF
COMPUTER SCIENCE
計算機科學系

Research联想研究院

Exploration in recommender systems



Filter Bubble

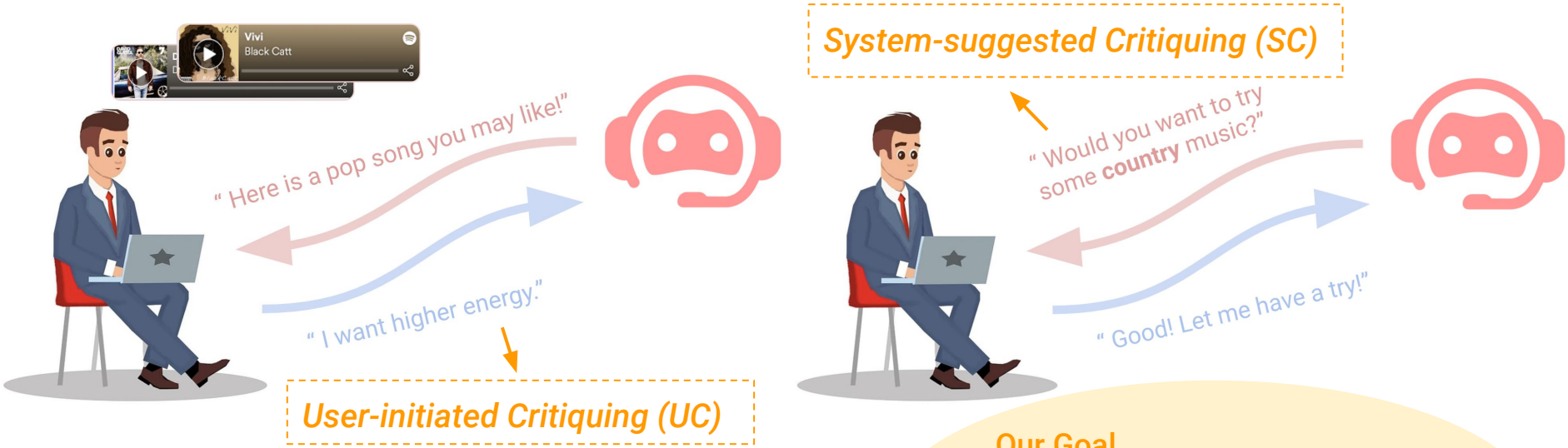
"the potential for online personalization to effectively isolate people from a diversity of viewpoints or content." -- Eli Pariser (Nguyen *et al.*, 2014)

Support user exploration in RS

- Diversity-driven Algorithms (e.g., Liang *et.al.*, 2019)
- Interactive Visualization (e.g., Verbert *et.al.*, 2013)

1. Tien T. Nguyen, Pik-Mai Hui, F. Maxwell Harper, Loren Terveen, and Joseph A. Konstan. Exploring the filter bubble: the effect of using recommender systems on content diversity. In WWW '14, pages 677–686, 2014.
2. Yu Liang and Martijn C. Willemsen. Personalized recommendations for music genre exploration. In UMAP '19, pages 276–284, 2019.
3. Katrien Verbert, Denis Parra, Peter Brusilovsky, and Erik Duval. Visualizing recommendations to support exploration, transparency and controllability. In IUI' 13, pages 351–362. 2013.

Critiquing-based Conversational Recommender Systems



Previous Studies (Jin *et al.*, 2019):

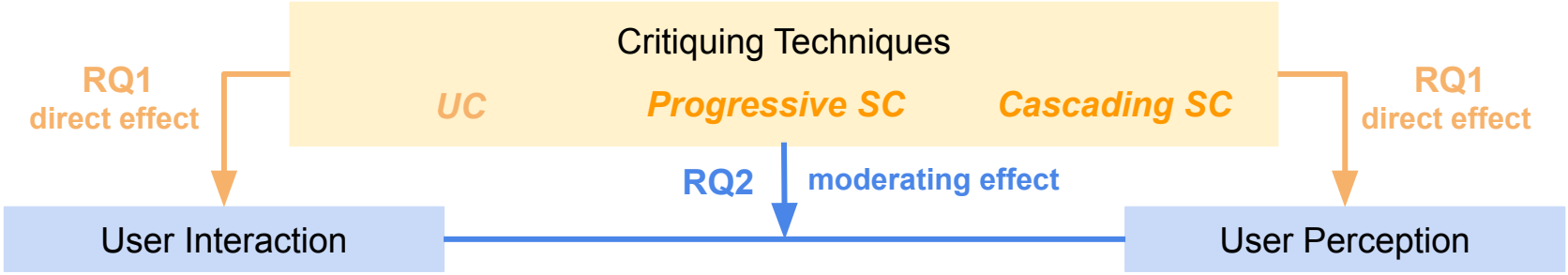
- User perceived *higher diversity* when using the music chatbot that supports both UC and SC

Our Goal

Stimulate **users' exploration of recommendations** through ***system-suggested critiques*** in conversational interaction

Yucheng Jin, Wanling Cai, Li Chen, Nyi Nyi Htun, and Katrien Verbert. MusicBot: Evaluating Critiquing-Based Music Recommenders with Conversational Interaction. In CIKM '19, pages 951–960, 2019.

Research Questions

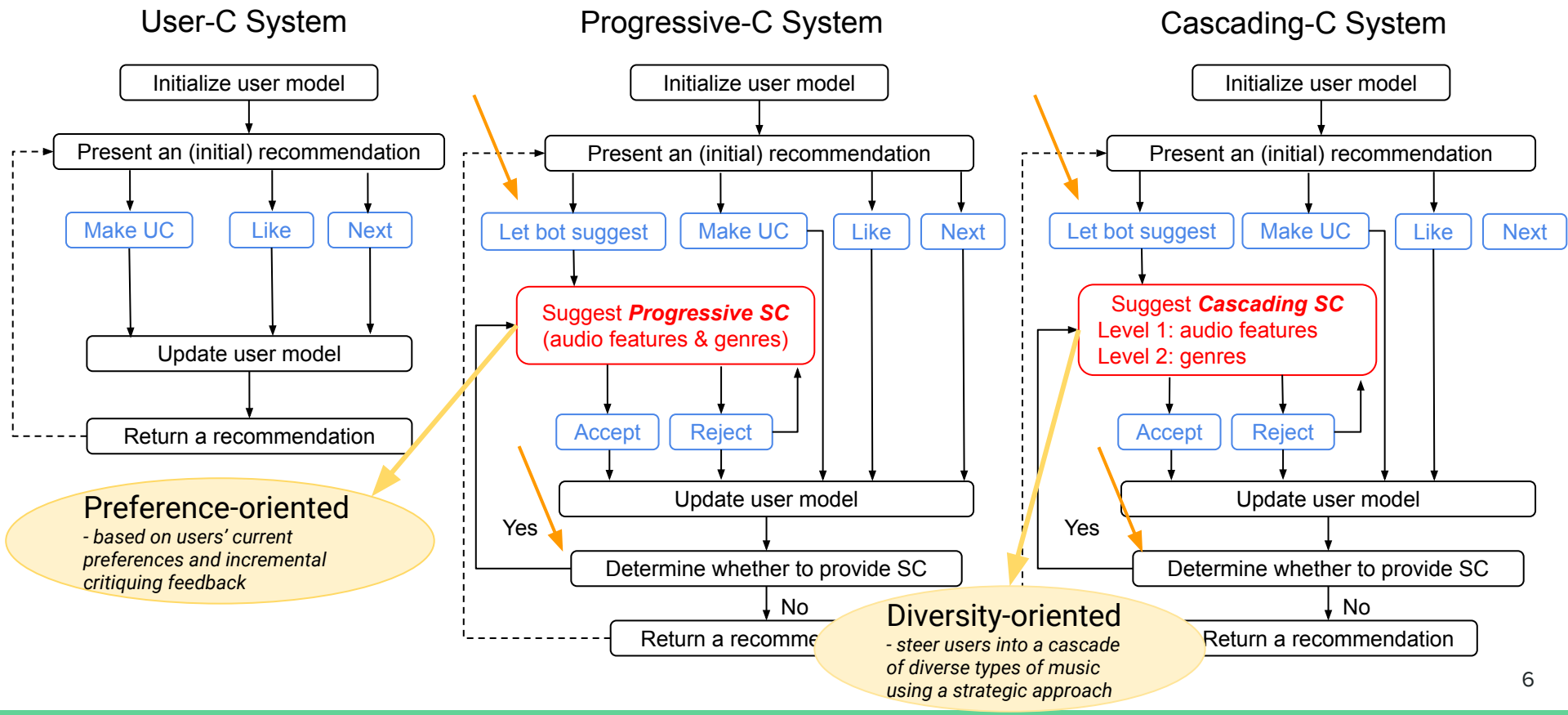


RQ1: *How do critiquing techniques influence users' exploration of music in a conversational recommender?*

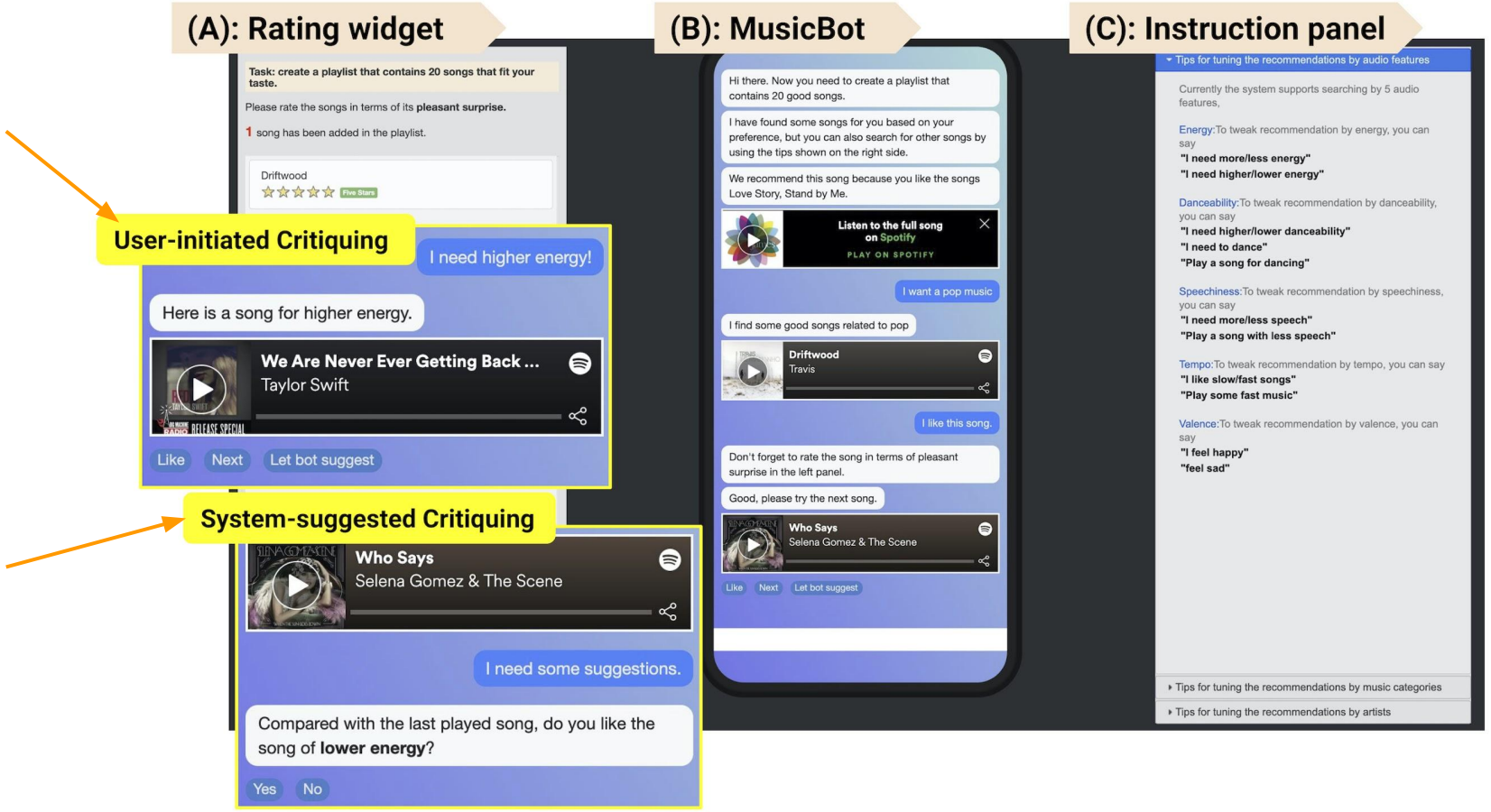
RQ2: *How do critiquing techniques moderate the relationship between user interaction behavior and user perception of music recommendations?*

System Design

System Design



MusicBot: Conversational User Interface



Experimental Design

User Experiment - Between-subjects design

Participants: 147 (107)

Recruitment: Prolific Platform

Task duration: about 25 mins

Reward: £2.4 per participant

User-C (35)

Progressive-C (36)

Cascading-C (36)

Age

- 19-25 (40)
- 26-35 (35)
- 36-50 (22)
- > 50 (10)

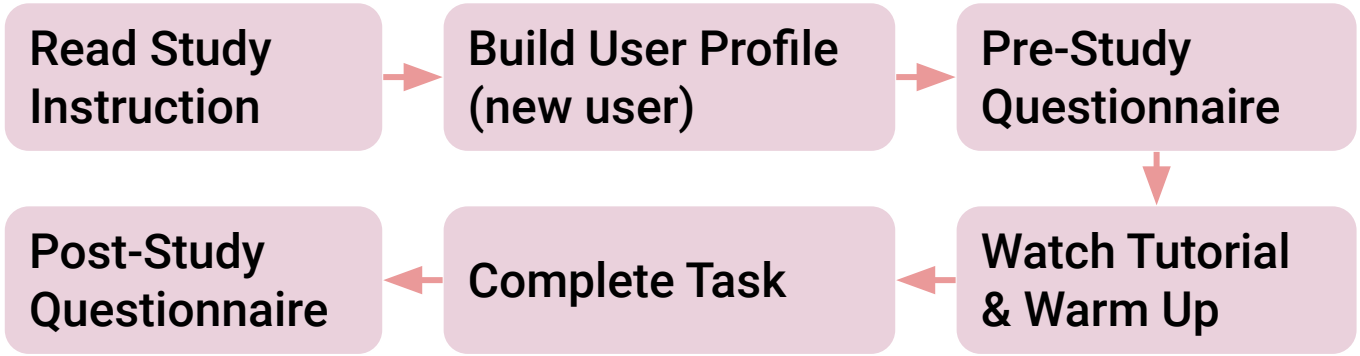
Gender

- Female = 52
- Male = 53
- Other = 2

User Experiment

Experimental task (two steps)

1) Use our Music chatbot to ***discover songs in different music types*** as much as possible, and create a playlist that contains **20** pieces of music that fit the user's taste, and then rate each song in terms of its ***pleasant surprise***.
2) Select **top-5 most preferred songs** from the created playlist.



Online Evaluation

User Perception (Post-study Questionnaire *7-point Likert Scale*)

Question items
Interest: Q1. The songs recommended to me matched my interests.
Novelty: Q2. The songs recommended to me are novel.
Music discovery: Q3. The music chatbot helped me discover new songs.
Diversity: Q4. The songs recommended to me are diverse.
Control: Q5. I feel in control of modifying my taste using this music chatbot.
Helpfulness: Q6. The music chatbot gave me good suggestions for helping me discover songs.
Engagement: Q7.I feel it is entertaining and interesting to engage in a dialogue with this music chatbot to discover songs.
Serendipity: Q8. The music chatbot provided me with recommendations that I had not considered in the first place but turned out to be a positive and surprising discovery.
Confidence: Q9. I am confident that I will like the songs in the created playlist (20 songs).
Pleasant surprise: Q10. The songs in the created playlist (20 songs) are overall pleasantly surprising to me.

***ResQue:* User-centric** evaluation framework for recommender systems (Pu *et al.*, 2011)

Pearl Pu, Li Chen, and Rong Hu. A user-centric evaluation framework for recommender systems. In RecSys '11, pages 157–164, 2011

Online Evaluation

User Interaction Behavior (Interaction Logs)

- Rating (stars) for the selected songs
- Duration

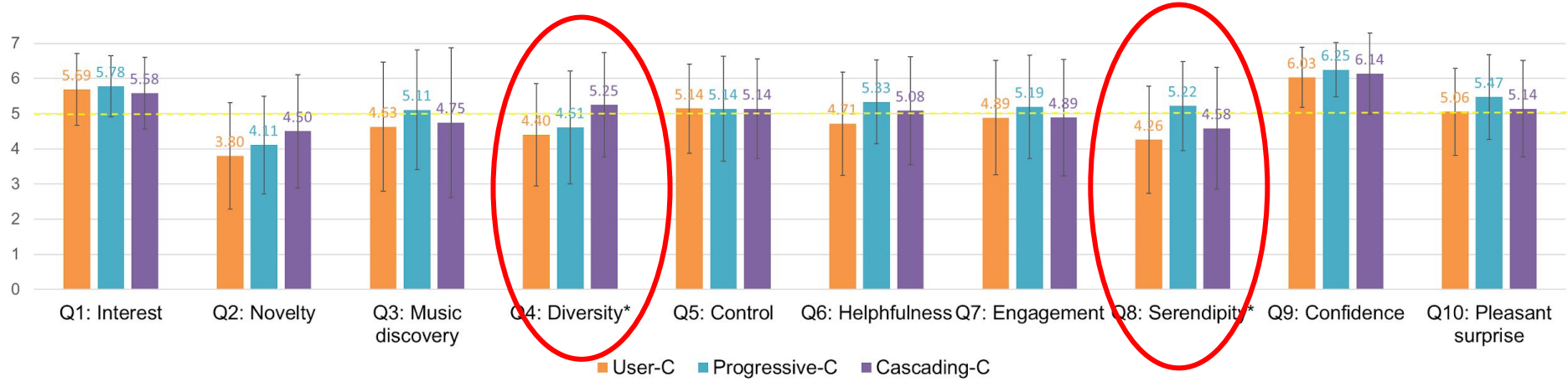
The number of

- Listened songs
- Dialogue turns
- Button clicks
- Messages by typing
- Words per utterance

Results & Discussion

Results - User Perception (RQ1)

Assessment results of statements related to user exploration



Significant results
(Kruskal-Wallis test; post-hoc Mann-Whitney test)

- Diversity (*Cascading-C* > *User-C*)
- Serendipity (*Progressive-C* > *User-C*)

	User-C		Progressive-C		Cascading-C	
	Mean	Std	Mean	Std	Mean	Std
Diversity	4.40	1.46	4.61	1.61	5.25	1.48
serendipity	4.26	1.52	5.22	1.27	4.58	1.73

Results - User Interaction Behavior (RQ1)

Descriptive Statistics for User Interaction Data

	User-C		Progressive-C		Cascading-C	
	Mean	Std	Mean	Std	Mean	Std
Interaction metrics						
#Listened songs	42.06	12.92	39.78	12.97	41.47	15.62
Duration (minutes)	10.95	4.43	12.04	4.59	12.47	5.28
#Dialogue turns (times)*	43.03	13.86	52.64	16.44	54.22	21.30
#Button (times)***	33.40	9.65	46.39	12.69	47.61	19.08
#Button-Next (times)	13.97	9.40	12.81	8.52	13.89	12.22
#Typing (times)*	9.94	8.17	6.42	7.62	6.78	5.40
#Words per utterance	3.32	1.12	2.72	1.72	3.66	1.54

Progressive-C and Cascading-C systems lead to more dialogue turns and button clicks.

Critiquing Behavior - identify the common **interaction patterns (IPs)** that lead to the use of UC and SC

IP1: *Recommend* → *Like* → *Recommend* → *Like* → *Recommend* → *Make UC* (56.84%, 54/95)

IP2: *Recommend* → *Next* → *Recommend* → *Next* → *Recommend* → *Make UC* (46.32%, 44/95)

IP3: *System Suggest Critiques* → *Accept SC* → *Recommend* → *Make UC* (36.07%, 22/61)

IP4: *System Suggest Critiques* → *Accept SC* →...→ *Let Bot Suggest* (48.89%, 22/45)

Proactive SC (proactively suggested by the system)

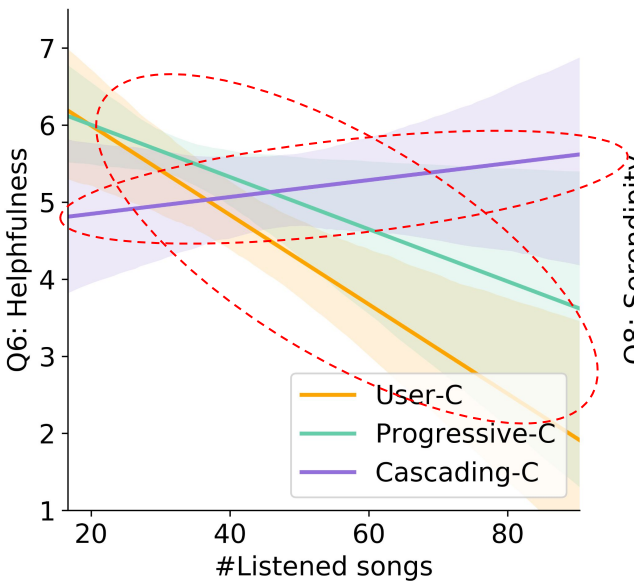
RQ1: *How do critiquing techniques influence users' exploration of music in a conversational recommender?*

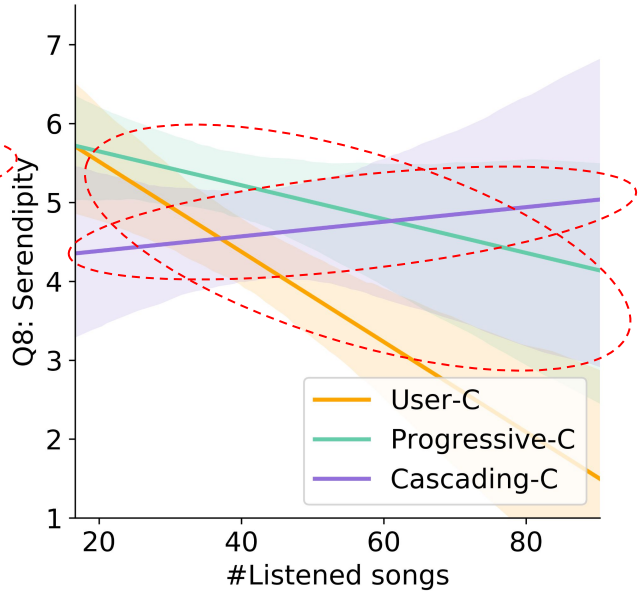
Our Findings

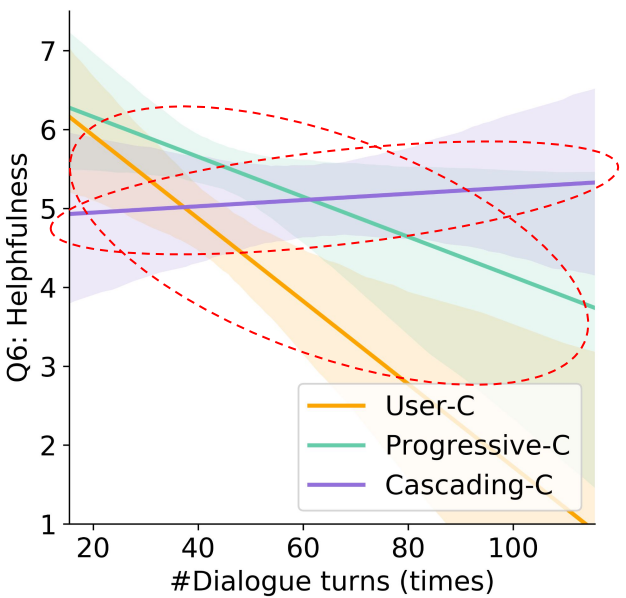
Both UC and SC facilitate the exploration of music recommendations with conversational interaction, while SC leads to ***higher perceived diversity and serendipity*** *of recommended songs and* ***more user interactions***.

Results - Moderation Effects (RQ2)

Moderation effects of critiquing technique on the relationship between user interaction and user perception







RQ2: *How do critiquing techniques **moderate** the relationship between user interaction behavior and user perception of music recommendations?*

Our Findings

*The critiquing techniques significantly moderate the relationships between some **user interaction metrics (e.g., number of listened songs, number of dialogue turns)** and **users' perceived helpfulness, serendipity, and pleasant surprise**.*

Implications of Our Work

Music Exploration

User-initiated Critiquing

allows users to explicitly initialize exploration when they have **a clear exploration goal**

System-suggested Critiquing

guides users to explore recommendations especially when they **have no specific goal**

Progressive SC (serendipity-oriented) & Cascading SC (diversity-oriented)

Progressive SC

→ *the initial period of exploration (short-term exploration)*

Users are more likely to accept the songs that are close to their current preferences.

Cascading SC

→ *the later period of exploring music (long-term exploration)*

After a period of exploration, users may expect to see more diverse types of song.

Conclusions

Conclusions

1. *Two kinds of **system-suggested critiquing technique** for encouraging users' exploration of music recommendations*
2. *Users perceive higher diversity of recommendations in **Cascading-C** and feel more serendipitous recommendations in **Progressive-C***
3. ***Significant moderation effects of critiquing techniques** on the relationships between some interaction metrics and user perception metrics*
4. ***Implications** for designing critiquing-based conversational recommender systems for music exploration*

Future Work

1. *To investigate how **personal characteristics such as personality** affect user exploration of music when interacting with different types of critiquing system*
2. *To verify the **generalizability of our findings** to other application domains (e.g., e-commerce, movie)*

Thanks! Q&A

Wanling Cai
cswlcai@comp.hkbu.edu.hk

Yucheng Jin
jinyc2@lenovo.com

Li Chen
lichen@comp.hkbu.edu.hk

香港浸會大學
HONG KONG BAPTIST UNIVERSITY

DEPARTMENT OF
COMPUTER SCIENCE
計算機科學系

Research联想研究院