

Towards Empathetic Conversational Recommender Systems

Zhaochun Ren | LIACS

December 13, 2024



**Universiteit
Leiden**
The Netherlands

Discover the world at Leiden University

Joint work with



Suzan Verberne
Leiden University



Shen Gao
Shandong University



Xiaoyu Zhang
Shandong University



Zhi Tian
Shandong University



Pengjie Ren
Shandong University



Maarten de Rijke
University of Amsterdam



Jiabao Fang
Shandong University



Xin Xin
Shandong University



Jie Zou
University of
Electronic Science and
Technology of China



Qintong Li
University of Hong Kong



Zhou Yang
Fuzhou University

Outline

- Part1: Knowledge-enhanced conversational recommender systems
- Part2: Empathetic dialogue systems
- Part3: Towards empathetic conversational recommender systems

Part1:

Knowledge-enhanced conversational recommender systems



Universiteit
Leiden
The Netherlands

Discover the world at Leiden University

Preliminaries

What is conversational recommender system (CRS)?

Traditional recommender systems



Conversational recommender systems
Multi-turn natural language interactions

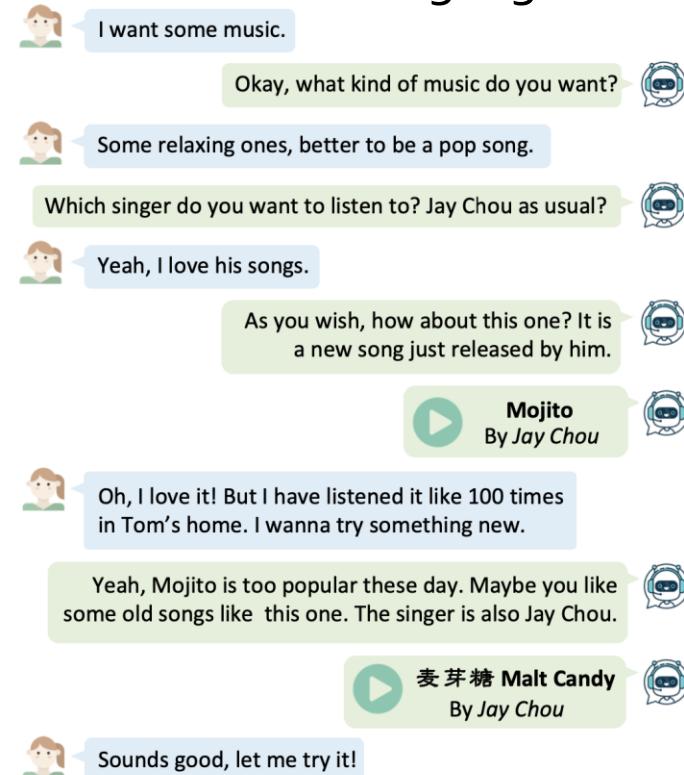


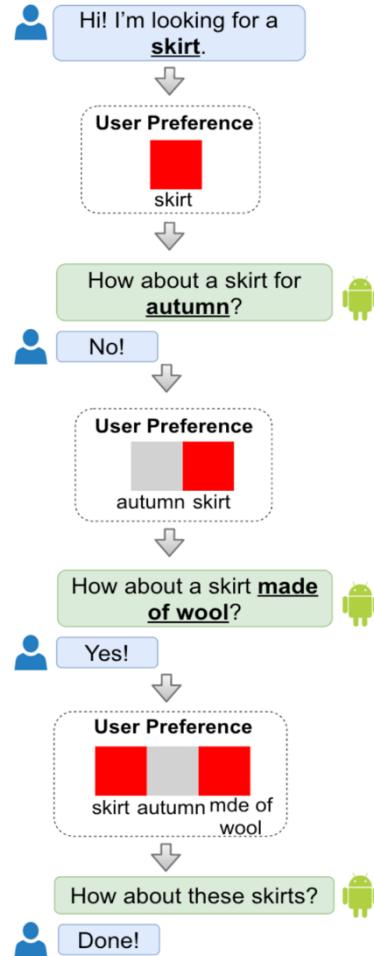
Fig. Conversational recommender system

(Source: Advances and Challenges in Conversational Recommender Systems: A Survey. In AI Open 2021)

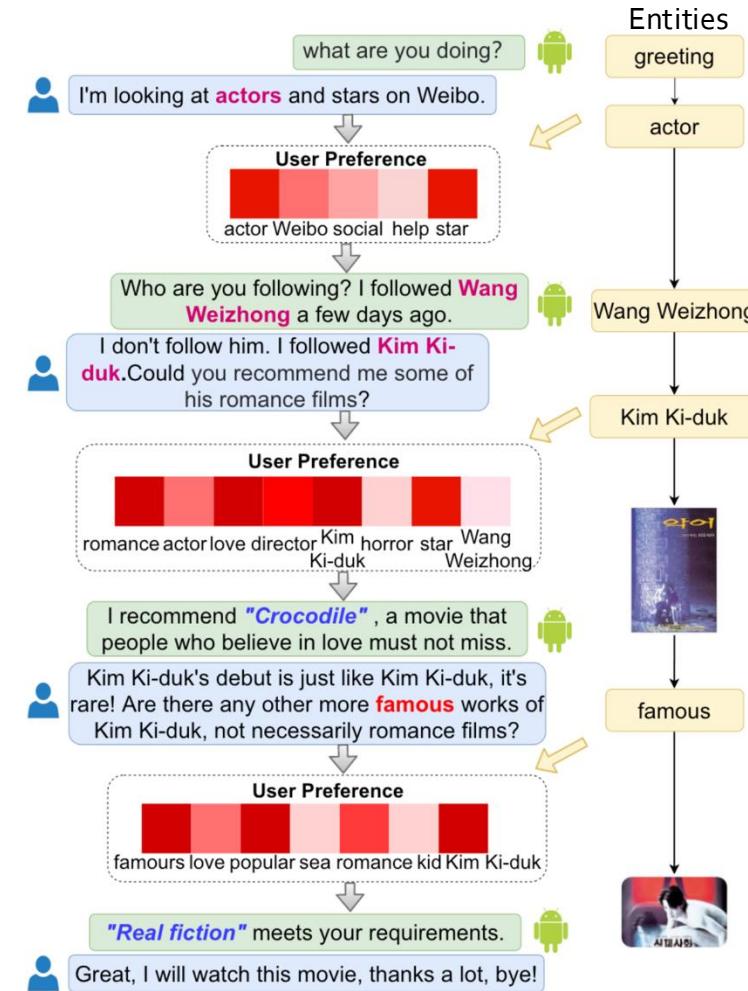
Preliminaries

Two kinds of CRS

Attribute-based CRSs



Generation-based CRSs ✓



Preliminaries

Two subtasks in CRS

Example from the ReDial dataset:

Recommender: How can I help you today?

Seeker: I would like to watch a fantasy movie.

Recommender: I recommend [redacted] ? or [redacted] ? Have you seen those?

Seeker: I have seen these, they are very good. Do you give another recommendation?

Recommender: I'm looking forward to seeing [redacted] ? but I'm not sure if you have seen it.

Item recommendation subtask

Evaluation Metrics:

Hit, Recall, Precision,
F1, NDCG, MRR

...

Preliminaries

Two subtasks in CRS

Example from the ReDial dataset:

Recommender: How can I help you today?

Seeker: I would like to watch a fantasy movie.

Recommender: ?

Seeker: I have seen these, they are very good. Do you give another recommendation?

Recommender: ?

Response generation subtask

Evaluation Metrics:

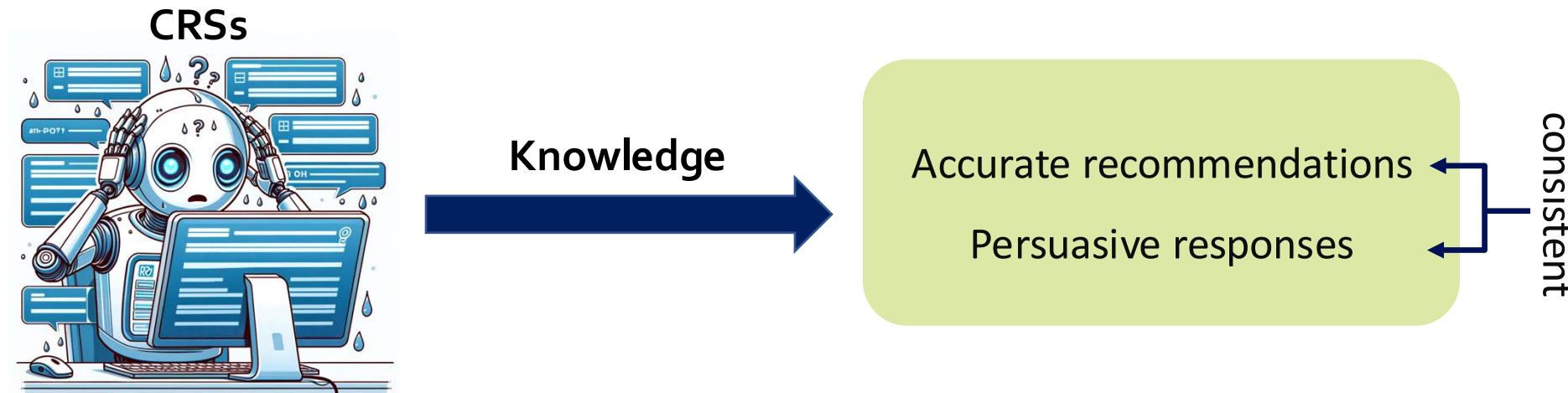
BLEU, ROUGE, PPL,
Fluency, Informativeness

...

Knowledge-enhanced CRSs

Challenges:

- Lack of sufficient contextual information.
- Semantic inconsistency between modules for the two subtasks.



From what data sources can we access knowledge?

Outline

Knowledge sources:

- Mining in-text knowledge
- Incorporating external knowledge graphs
- Using language models

Outline

Knowledge sources:

- Mining in-text knowledge
 - UPCR (SIGIR 2022): Across the sessions
- Incorporating external knowledge graphs
- Using language models

Motivation: capture user preference

How can we capture user preferences more accurately so as to provide accurate recommendation results?

Challenge 1

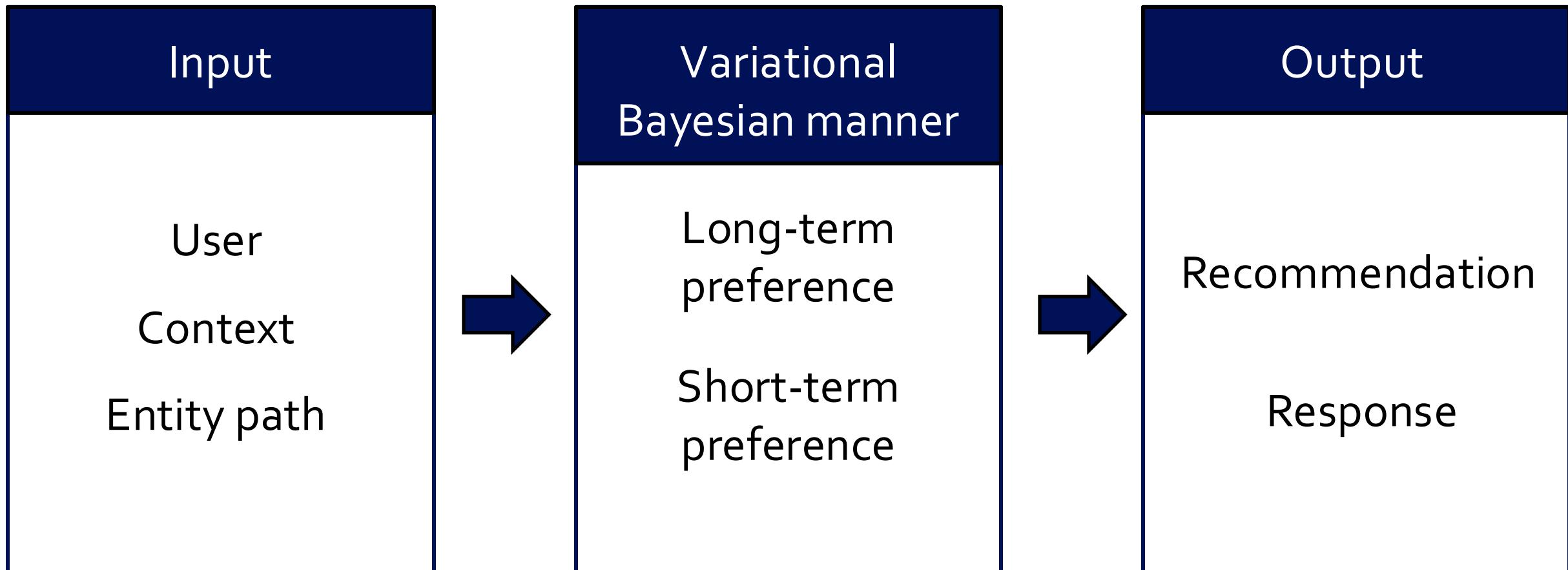
- Items or entities not enough to fully capture user preferences in conversational recommenders – only reflect part of information shared in conversation

Challenge 2

- Conversation only reflects user's short-term preferences, not their long-term preferences

UPCR: Proposed method

Problem formulation



UPCR: Main findings

Recommendation

Table 4: Automatic evaluation of recommendation on TG-Redial and REDIAL datasets. Bold face indicates best result. Significant improvements over best baseline results marked with * (t-test, $p < 0.05$).

Model	TG-Redial				REDIAL		
	NDCG		MRR		Recall		
	@10	@50	@10	@50	@1	@10	@50
Popularity	0.0015	0.0036	0.0011	0.0015	0.012	0.061	0.179
TextCNN	0.0144	0.0215	0.0119	0.0133	0.017	0.096	0.159
BERT	0.0246	0.0439	0.0182	0.0221	0.018	0.117	0.191
PostKS	0.0031	0.0048	0.0029	0.0038	0.019	0.122	0.236
KBRD	0.0064	0.0111	0.0040	0.0049	0.030	0.163	0.338
DCR	0.0261	0.0498	0.0129	0.0179	0.027	0.148	0.306
REDIAL	0.0006	0.0025	0.0003	0.0007	0.023	0.129	0.287
MGCG	0.0184	0.0412	0.0130	0.0210	0.027	0.121	0.264
TG-Redial	0.0348	0.0527	0.0240	0.0277	0.041	0.164	0.310
KGSF	0.0154	0.0259	0.0114	0.0135	0.039	0.183	0.378
CR-Walker	0.0565	0.0771	0.0489	0.0565	0.040	0.187	0.376
UPCR	0.0835*	0.1105*	0.0665*	0.0721*	0.046*	0.238*	0.484*

The proposed method outperforms all baselines on both datasets by tracking user preferences, more accurate recommendations can be produced (NDCG, MRR, Recall)

UPCR: Main findings

Generation

The proposed method performs fluency, diverse and informative reply since it considers user id, context, entity path and recommendation. (BLEU, Distinct, human evaluation).

Model	TG-Redial						REDIAL			
	Automatic				Human		Automatic			
	BLEU-1	BLEU-2	BLEU-3	Distinct-1	Distinct-2	Fluency	Informativeness	Distinct-2	Distinct-3	Distinct-4
PostKS	0.142	0.018	0.006	0.005	0.021	1.23	0.83	0.074	0.126	0.224
KBRD	0.221	0.028	0.009	0.004	0.008	1.17	1.12	0.086	0.153	0.265
DCR	0.128	0.021	0.007	0.008	0.021	0.92	0.91	0.081	0.138	0.233
REDIAL	0.069	0.008	0.002	0.015	0.062	1.22	1.02	0.082	0.143	0.245
MGCG	0.242	0.057	0.023	0.011	0.041	1.34	1.34	0.101	0.189	0.261
TG-Redial	0.280	0.065	0.031	0.021	0.094	1.45	1.36	0.086	0.153	0.216
KGSF	0.239	0.042	0.013	0.015	0.062	1.38	1.33	0.114	0.204	0.282
CR-Walker	0.271	0.059	0.028	0.019	0.081	1.43	1.34	0.163	0.289	0.365
Transformer	0.261	0.061	0.027	0.014	0.083	1.28	0.72	0.067	0.139	0.227
UPCR	0.316*	0.083*	0.043*	0.022	0.132*	1.44	1.51*	0.172*	0.363*	0.529*

Outline

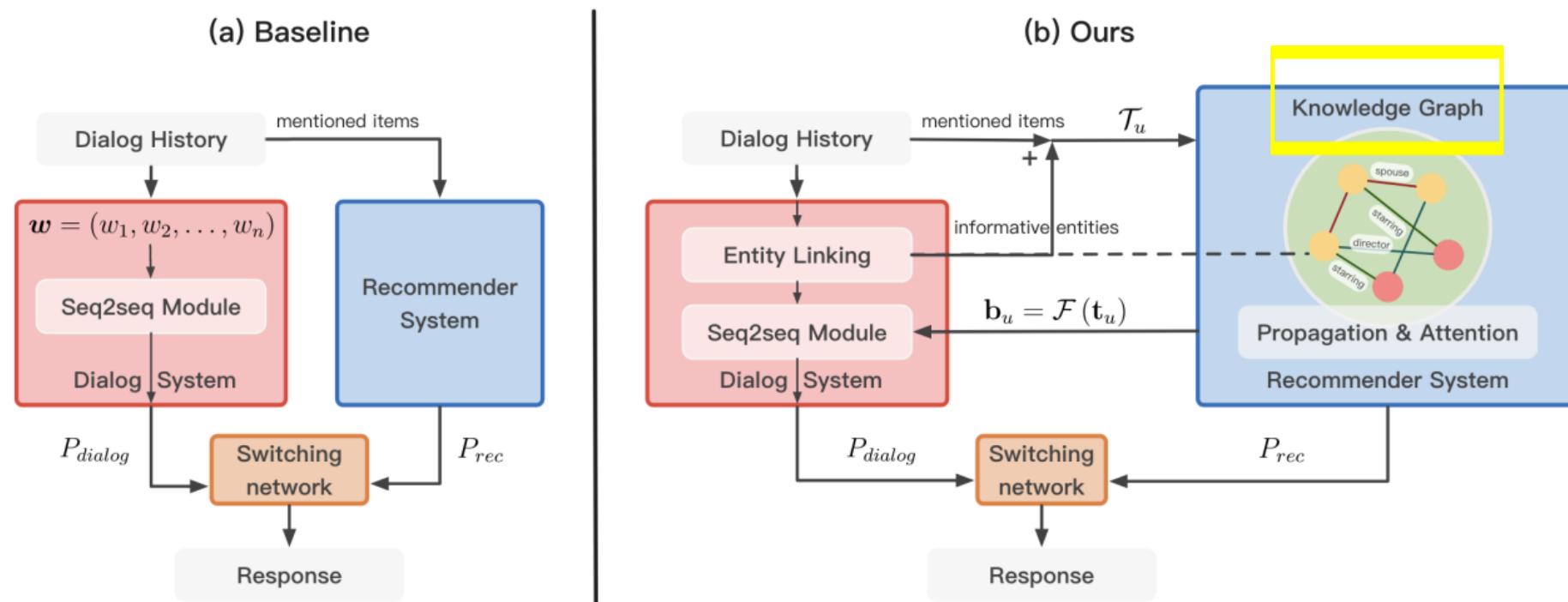
Knowledge sources:

- Mining in-text knowledge
- Incorporating external knowledge graphs
 - VRICR (WSDM 2023): External KGs with in-text knowledge
- Using language models

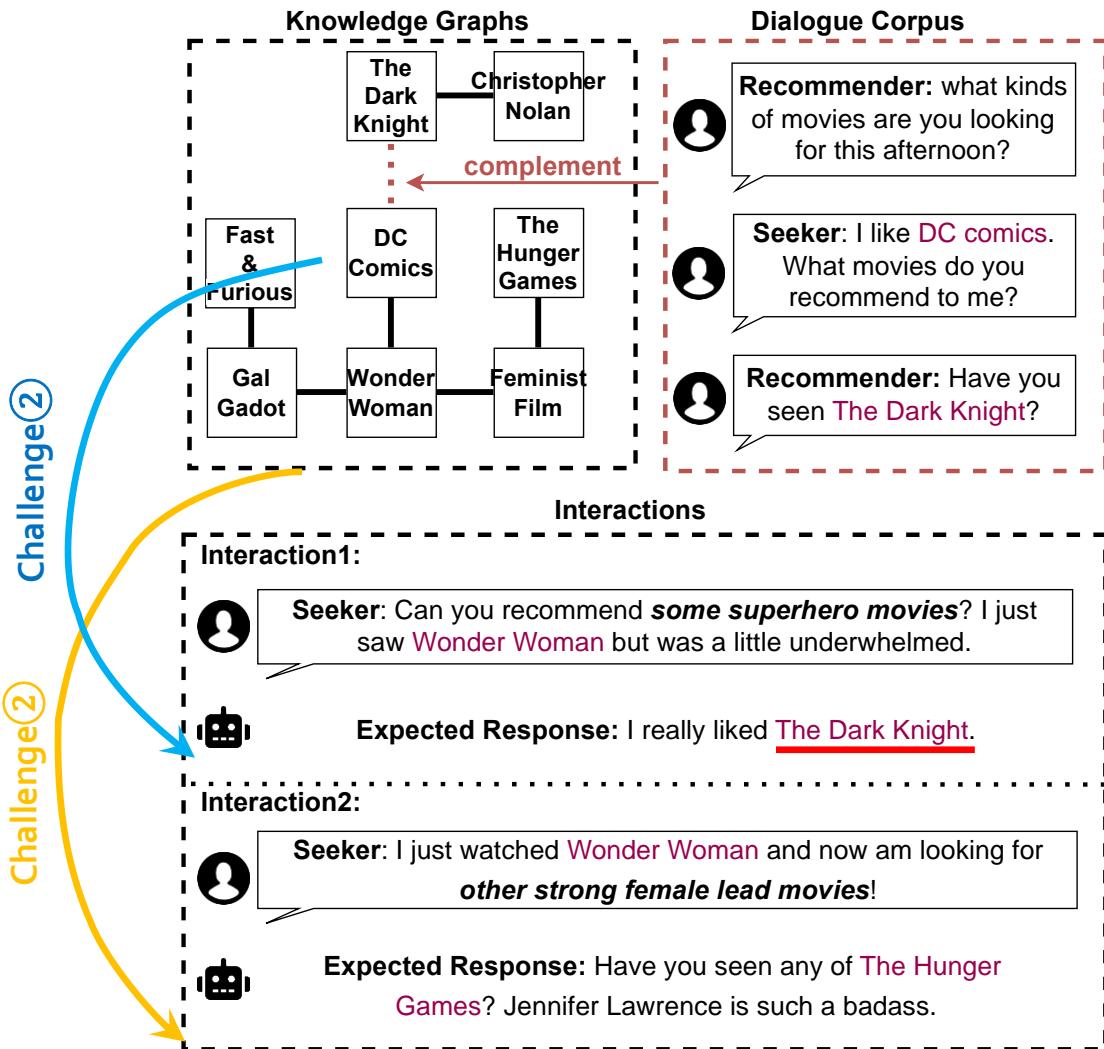
How to incorporate external KGs?

- KBRD (Chen et al., 2019)

Incorporate domain knowledge graph (DBpedia) to assist the CRS.



Motivation

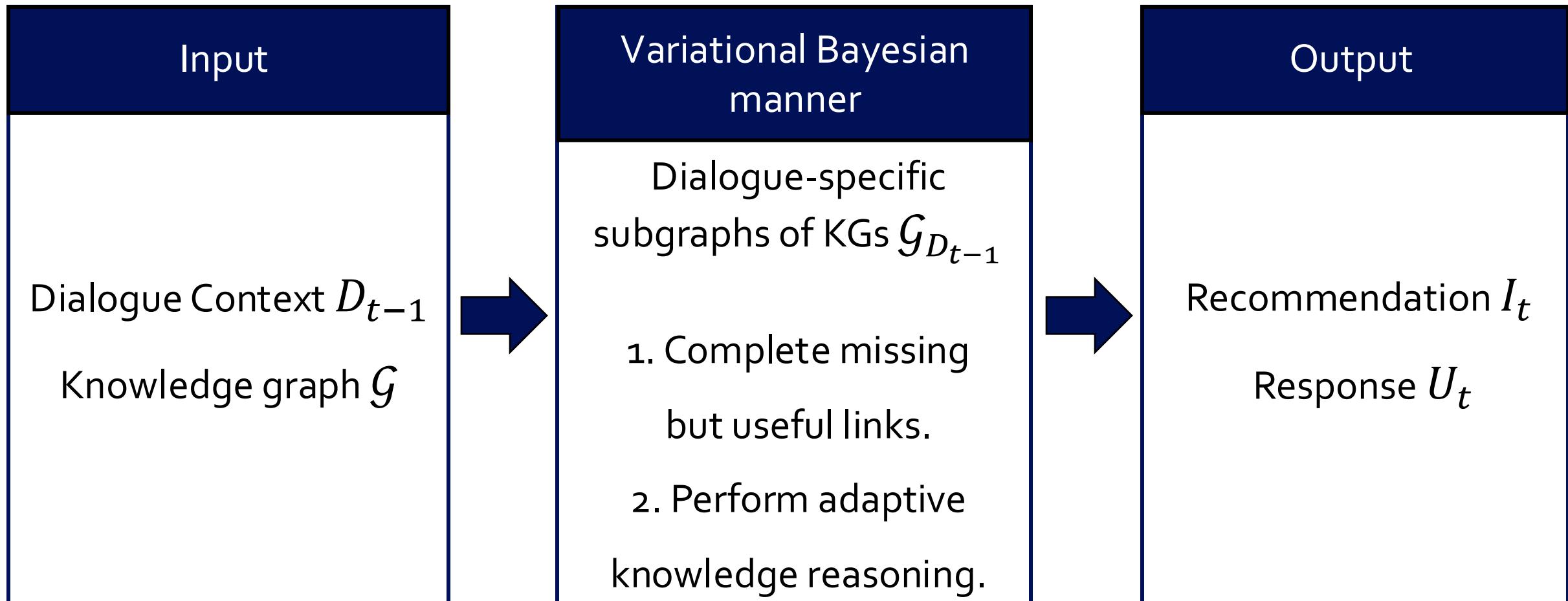


Challenges:

- ① Incomplete and sparse KGs.
- ② Utilize entire KGs for various interactions.

VRICR: Proposed method

Problem formulation



VRICR: Proposed method

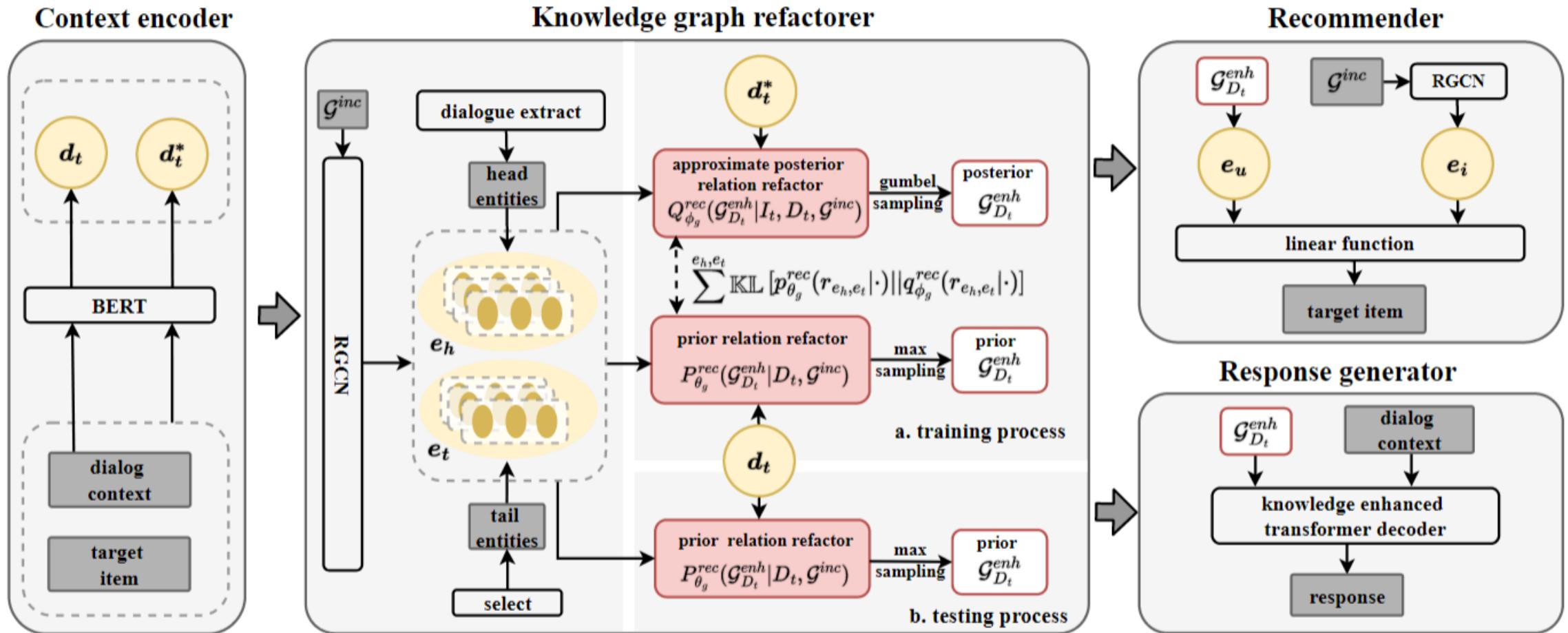


Figure 2: An overview of VRICR. VRICR has four components: a context encoder, a knowledge graph refactorer, a recommender and a response generator.

Outline

Knowledge sources:

- Mining in-text knowledge
- Incorporating external knowledge graphs
- Using language models
 - MACRS (2024): Multi-agent CRS framework.

Pretrained Language Models for CRSSs

- The parametrized knowledge in PLMs.
- **RID (Wang et al., 2021)**
Leverage the PLM to generate fluent and diverse dialogue responses.
- **BARCOR (Wang et al., 2022)**
Leverage the PLM as the recommender model and the response generator.

LLMs

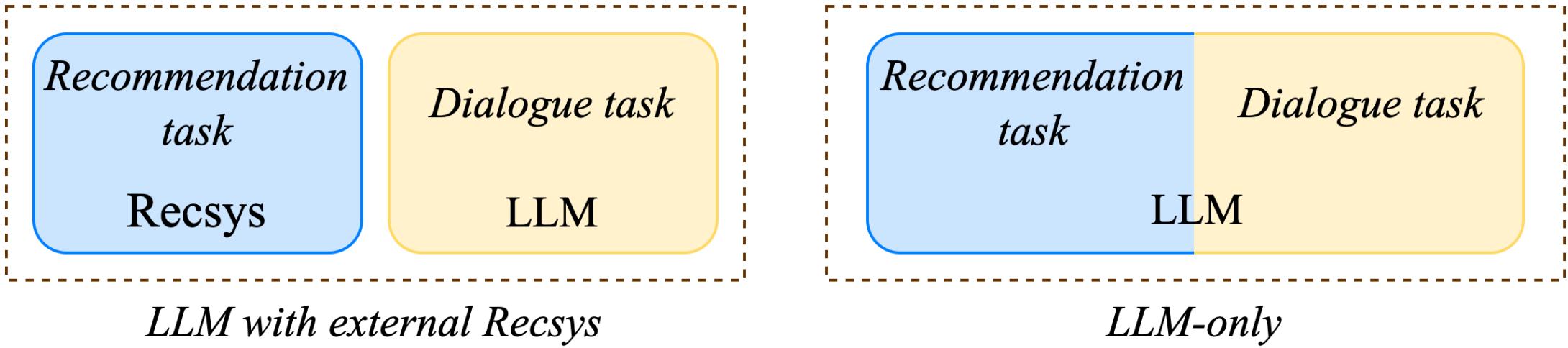


CRSs

Extensive parameterized world knowledge
Common-sense reasoning ability

How to use large language models for CRSs?

Large Language Models for CRSs



How to use large language models?

CRSs with LLMs: LLMs with external Recsys

- **InteRecAgent (Huang et al., 2023)**

Employ LLMs as the brain and the recommender models (SASREC) as tools.

- **LLMCRS (Feng et al., 2023)**

Collaborate LLMs with expert models (CRSs models such as KBRD, KGSF) to achieve the enhanced performance.

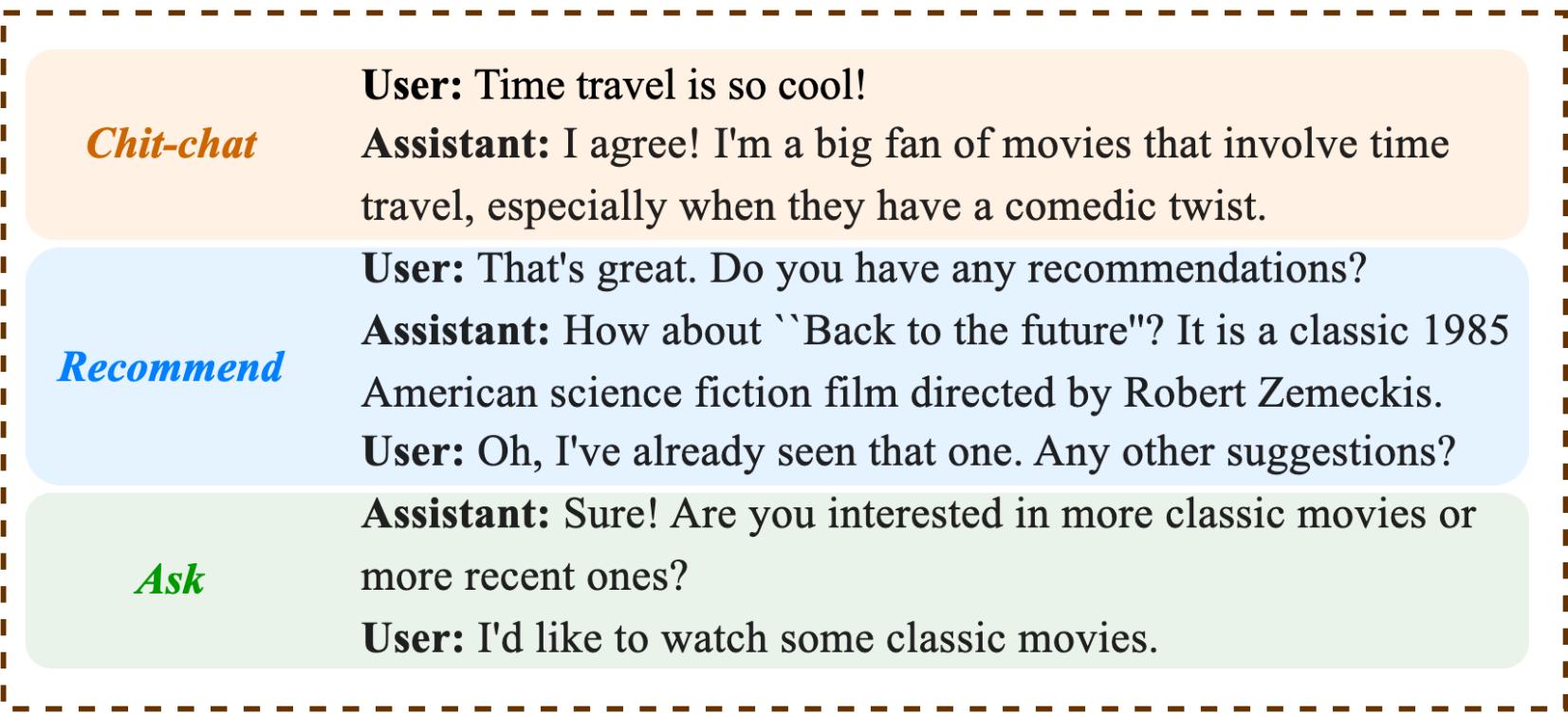
How to use large language models?

CRSs with LLMs: LLM-only

- Share **useful information** between the recommendation task and the dialogue task.
- **LLMs as Zero-shot CRSs (Gao et al., 2023)**
Prompt LLMs to provide recommendations based on the dialogue history.

MACRS: Challenges

How to plan the dialogue act ?



Efficient and engaging CRS should choose **the proper dialogue act** at each turn.

MACRS: Challenges

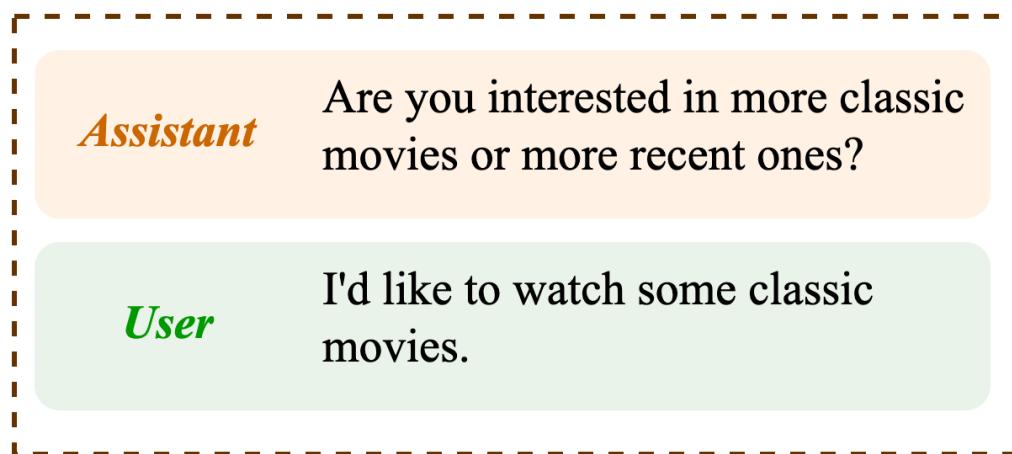
How to plan the dialogue act ?

- Single agent is hard to control the dialogue flow, leading to generating low-efficiency dialogue.
- Designing the collaborative methods of multi-agent to enhance the performance of the CRS.

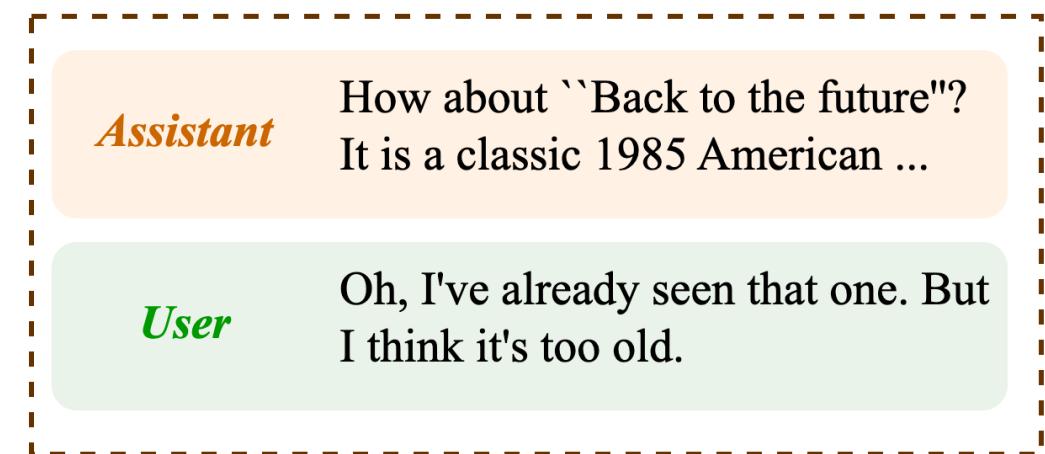
MACRS: Challenges

How to leverage user feedback?

- During dialogue, users can give feedback based on system responses.



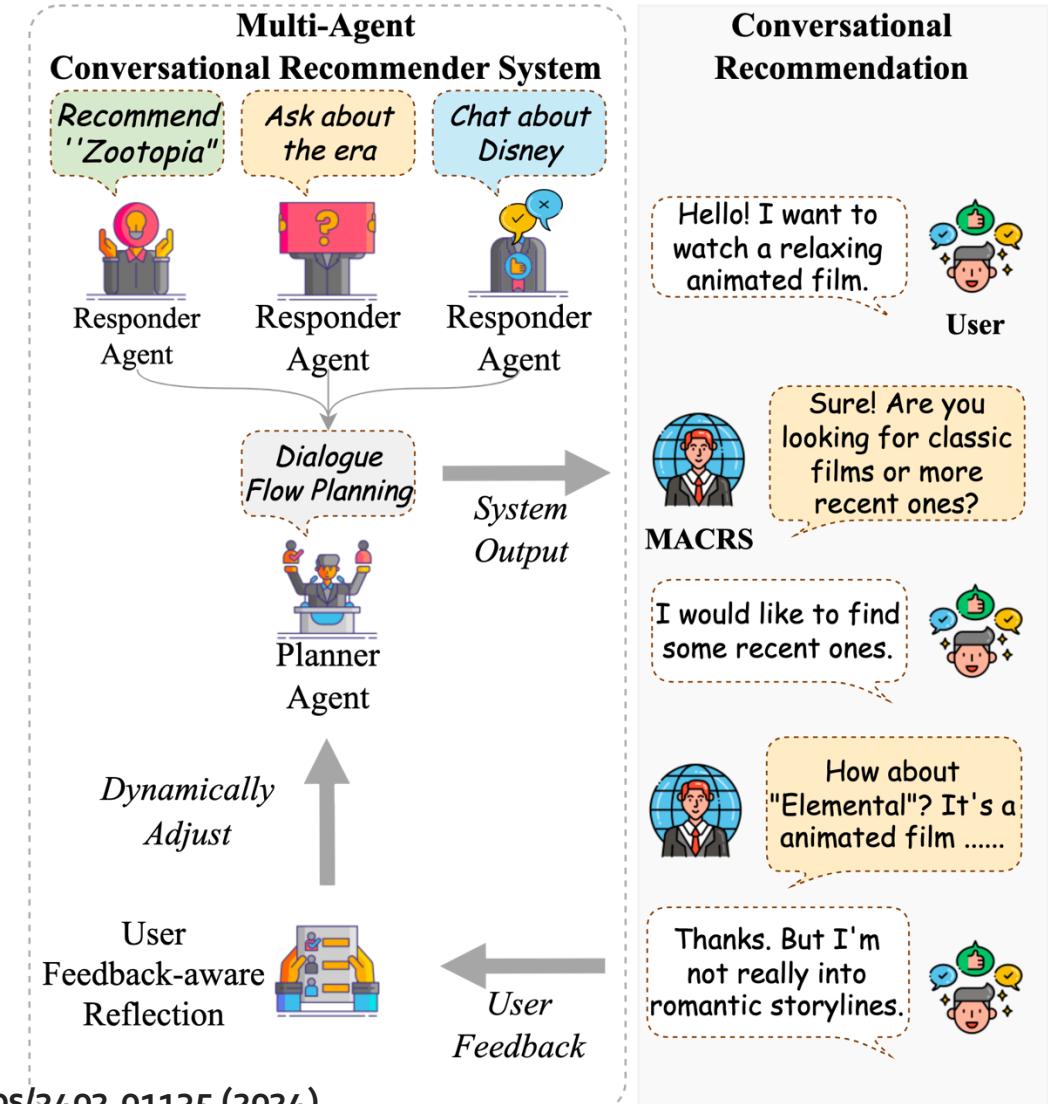
Clarification of user preferences



Satisfaction with recommended items

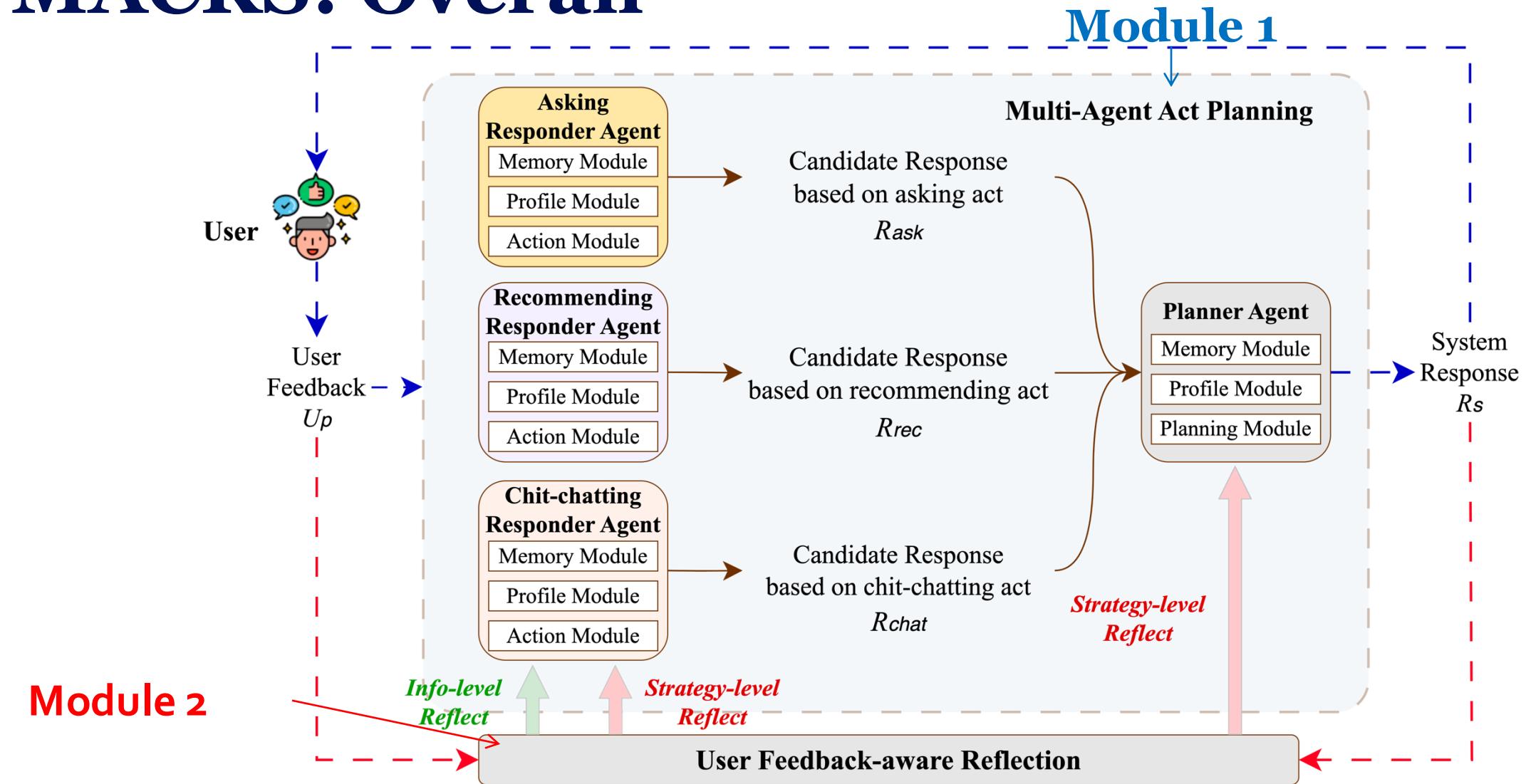
MACRS: Proposed method

- An LLM-only CRS can **efficiently plan** and **dynamically refine** its dialogue and recommendations.



Jiabao Fang, et al. 2024. A Multi-Agent Conversational Recommender System. CoRR abs/2402.01135 (2024).

MACRS: Overall



Four agents: Three responder agents and one planner agent

MACRS: Main findings

Recommendation

- Evaluation based on the user simulator.
- Select 100 samples from MovieLens dataset.

Method	Success Rate (↑)	Average Turns (↓)	Hit Ratio@5 (↑)	Hit Ratio@10 (↑)
<i>Traditional CRS Baselines</i>				
KBRD [2]	0.00	6.0	0.00	0.01
BARCOR [28]	0.03	5.9	0.03	0.07
<i>LLM-based CRS Baselines</i>				
Llama2	0.42	4.34	0.47	0.51
ChatGPT	0.39	4.24	0.44	0.51
<i>Our Proposed Method</i>				
MACRS-L	0.48	4.49	0.55	0.60
MACRS-C	0.61[‡]	4.19[†]	0.77[‡]	0.80[‡]
MACRS-C w/o IR	0.53	4.42	0.72	0.77
MACRS-C w/o SR	0.52	4.42	0.67	0.73
MACRS-C w/o SR+IR	0.51	4.40	0.64	0.66

MACRS: Main findings

Generation

Case study

Item Information

Target item: "Eraser (1996)"

Target item information: {genre: action, actor: Arnold Schwarzenegger, director: Chuck Russell, feature: witness protection}

Dialogue between ChatGPT and the user

User: Hello, I'm in the mood for an action movie. I really enjoy watching Arnold Schwarzenegger films. Can you recommend something for me?

ChatGPT: How about watching "Terminator 2: Judgment Day"? It's an iconic action film starring Arnold Schwarzenegger.

User: I haven't seen "Terminator 2: Judgment Day". But the director of it did not meet my expectations.

ChatGPT: If you haven't seen "Terminator 2: Judgment Day", how about watching "Predator"? It's another action film

User: I've seen "Predator", but the director of it did not meet my expectations.

FAIL

Dialogue between MACRS and the user

User: Hello, I'm in the mood for an action movie. I really enjoy watching Arnold Schwarzenegger films. Can you recommend something for me?

MACRS: Sure. Do you prefer more **recent action movies or older ones**?

User: I prefer older action movies.

MACRS: How about watching "Terminator 2: Judgment Day"? It's an **older** action film starring Arnold Schwarzenegger.

User: I haven't seen "Terminator 2: Judgment Day". But the director of it did not meet my expectations.

MACRS: What type of **director do you prefer** for action movies?

User: I prefer Chuck Russell for action movies.

MACRS: How about watching "Eraser"? It's an **older** action film starring Arnold Schwarzenegger and directed by **Chuck Russell**.

SUCCESS

Part2:

Empathetic dialogue systems



Universiteit
Leiden
The Netherlands

Discover the world at Leiden University

Empathetic Response Generation

EMPATHETICDIALOGUES dataset example



Fig. An Example of Empathetic Dialogue

(Source: Towards Empathetic Open-domain Conversation Models: a New Benchmark and Dataset)

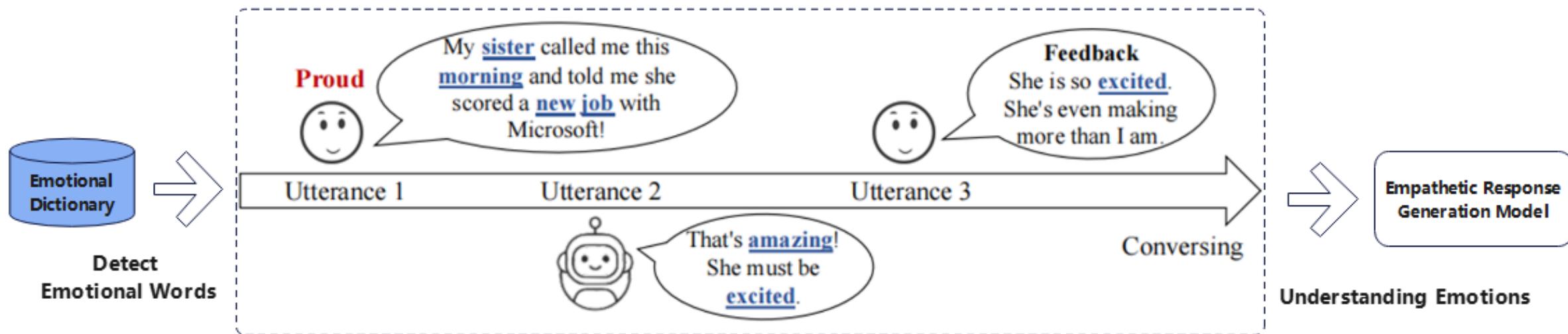
How to use external knowledge to enhance
empathetic response generation

Empathetic Response Generation

✓ External knowledge enhances dialogue understanding and expression.

- EmpDG (Li et al., 2020)

Utilized an emotion lexicon as an external knowledge base.



[1] Li et al. 2020. EmpDG: Multi-resolution Interactive Empathetic Dialogue Generation.

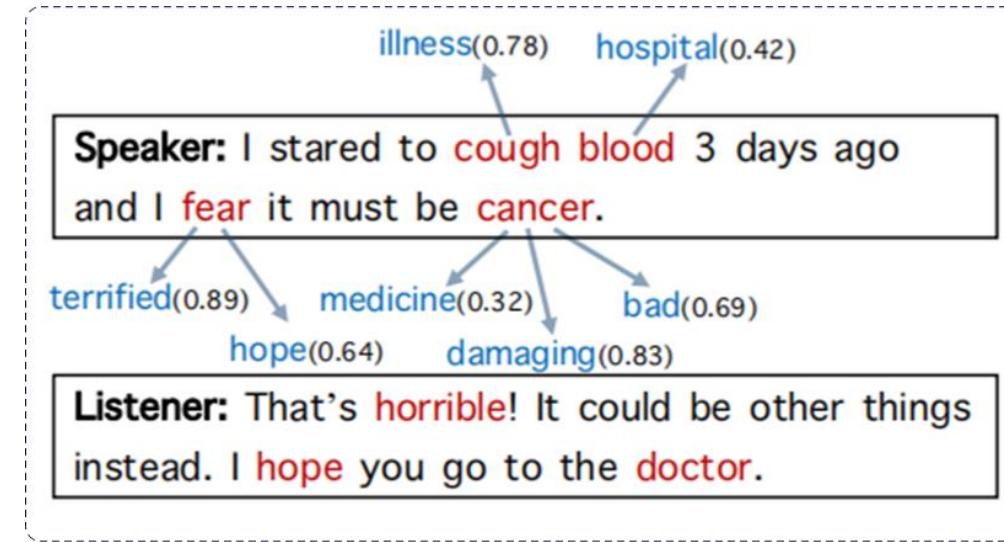
EmpDG: Main findings

Models	Accuracy	Perplexity ↓	Distinct-1	Distinct-2	Empathy	Relevance	Fluency
Transformer	-	33.91	1.17	4.70	3.15	3.42	3.64
EmoPrepend-1	0.3328	33.35	1.06	4.29	3.29	3.59	3.71
MoEL	0.3200	33.58	1.38	4.66	3.47	3.88	3.68
EmpDG	0.3431	34.18	1.81	6.94	3.58	3.91	3.67

Automatic and human evaluation results between baselines and EmpDG.

Challenges in leveraging external knowledge

Unable to perceive implicit emotions

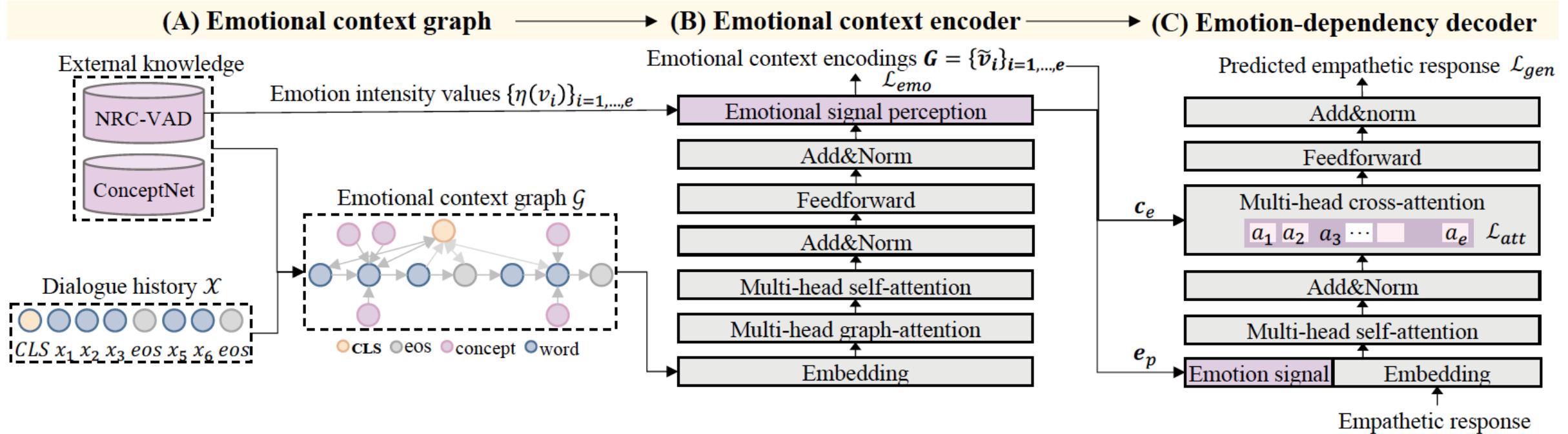


An example of empathetic dialogues with external knowledge

[1] Li et al. 2022. KEMP: Knowledge Bridging for Empathetic Dialogue Generation.

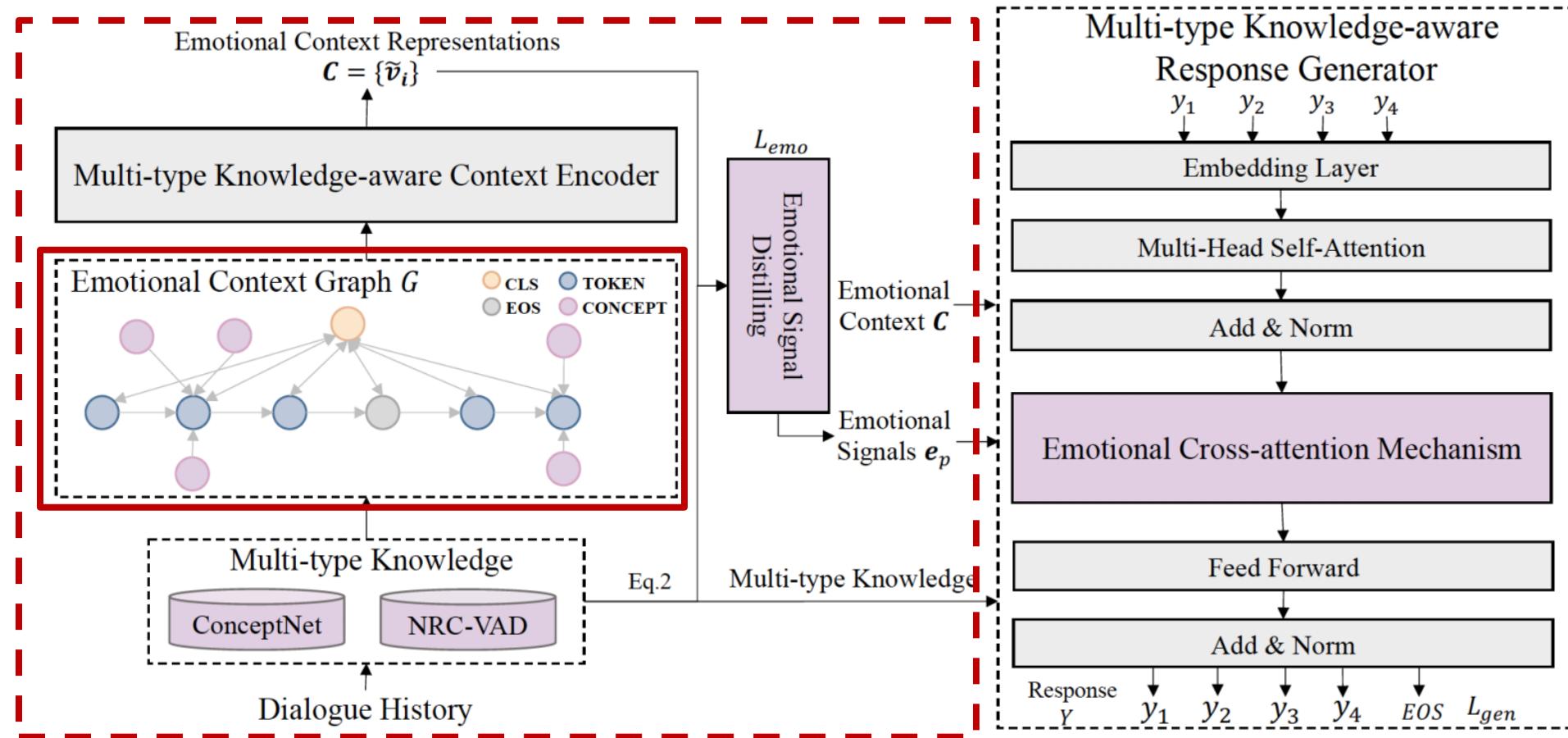
KEMP: Proposed Method

✓ Leveraging external knowledge to perceive implicit emotions.



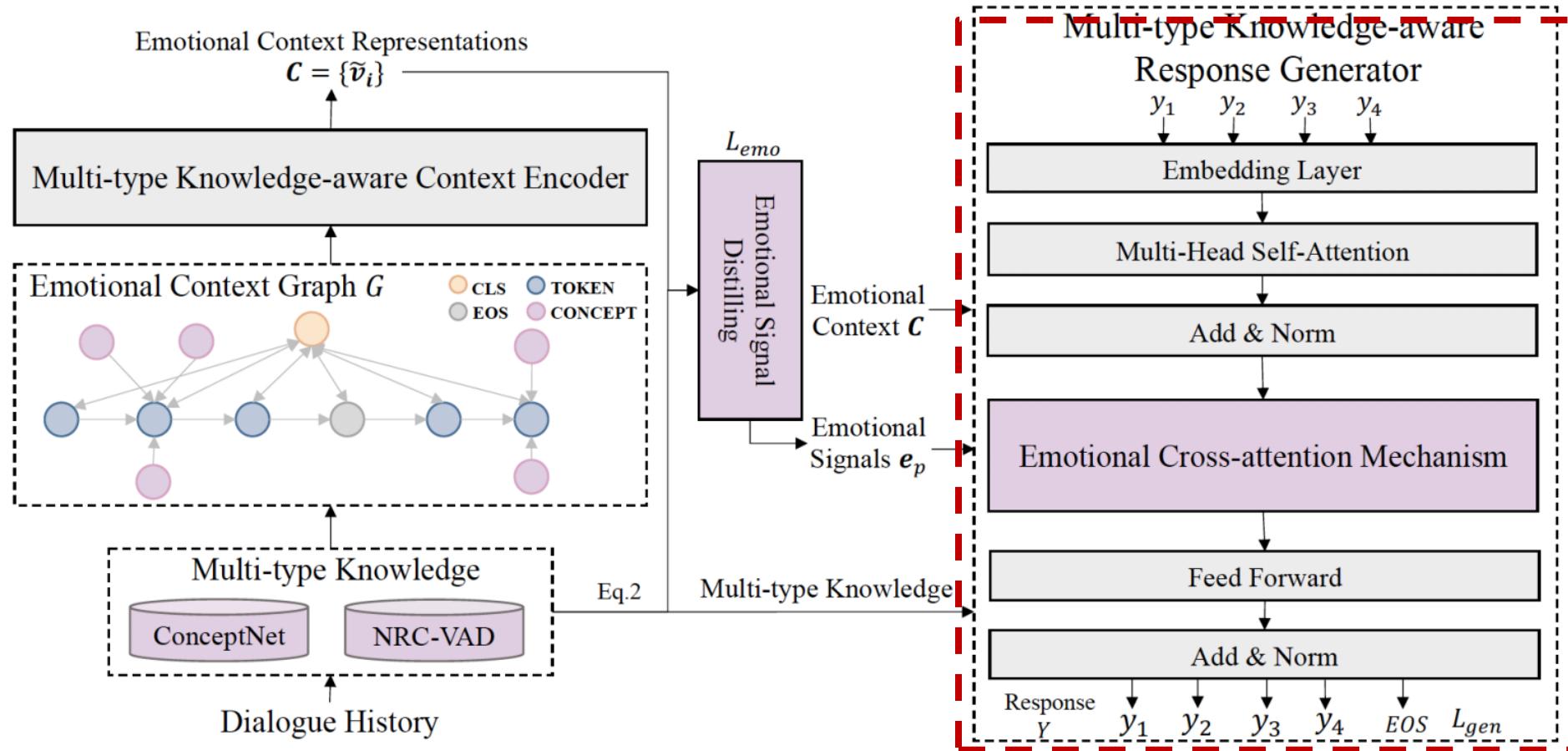
KEMP: Proposed Method

✓ Leveraging external knowledge to perceive implicit emotions.



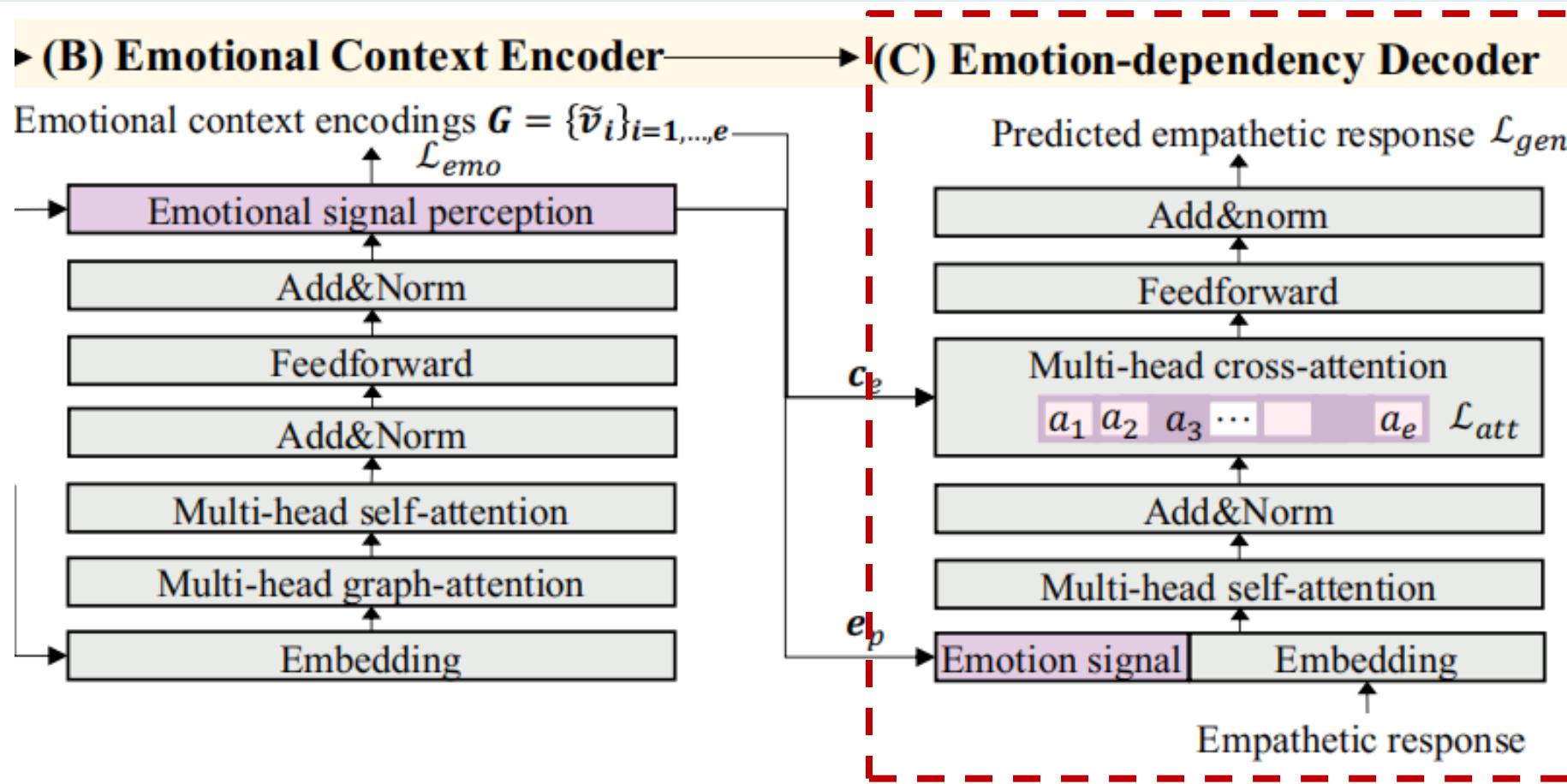
KEMP: Proposed Method

✓ Leveraging external knowledge to perceive implicit emotions.



KEMP: Proposed Method

✓ Leveraging external knowledge to perceive implicit emotions.



KEMP: Main findings

Models	Accuracy	Perplexity	Distinct-1	Distinct-2	Empathy	Relevance	Fluency
Transformer (Vaswani et al. 2017)	-	37.73	0.47	2.04	3.11	3.47	3.66
EmoPrepend-1 (Rashkin et al. 2019)	33.28	38.30	0.46	2.08	3.23	3.51	3.67
MoEL (Lin et al. 2019)	32.00	38.04	0.44	2.10	3.37	3.78	3.64
MIME (Majumder et al. 2020)	34.24	37.09	0.47	1.91	3.38	3.66	3.63
EmpDG (Li et al. 2020)	34.31	37.29	0.46	2.02	3.45	3.88	3.67
KEMP	39.31	36.89	0.55	2.29	3.49	3.92	3.65

KEMP: Main findings

Models	Accuracy	Perplexity	Distinct-1	Distinct-2	Empathy	Relevance	Fluency
Transformer (Vaswani et al. 2017)	-	37.73	0.47	2.04	3.11	3.47	3.66
EmoPrepend-1 (Rashkin et al. 2019)	33.28	38.30	0.46	2.08	3.23	3.51	3.67
MoEL (Lin et al. 2019)	32.00	38.04	0.44	2.10	3.37	3.78	3.64
MIME (Majumder et al. 2020)	34.24	37.09	0.47	1.91	3.38	3.66	3.63
EmpDG (Li et al. 2020)	34.31	37.29	0.46	2.02	3.45	3.88	3.67
KEMP	39.31	36.89	0.55	2.29	3.49	3.92	3.65

How to associate situations and content?

- ① Situations contain rich knowledge for emotion understanding
- ② How to associate situations and content for more accurate emotion understanding

Emotion: Furious

Situation: I was **driving** home and this **guy cut me off**. I had to **swerve** in order to **not hit him**.

Dialogue

Well I had to **swerve onto the shoulder** and almost **caused an accident**. **Some people** are just **jerks** on the road.

I know **the feeling**. I hate **driving** now. **Everyone** is looking in their phone.

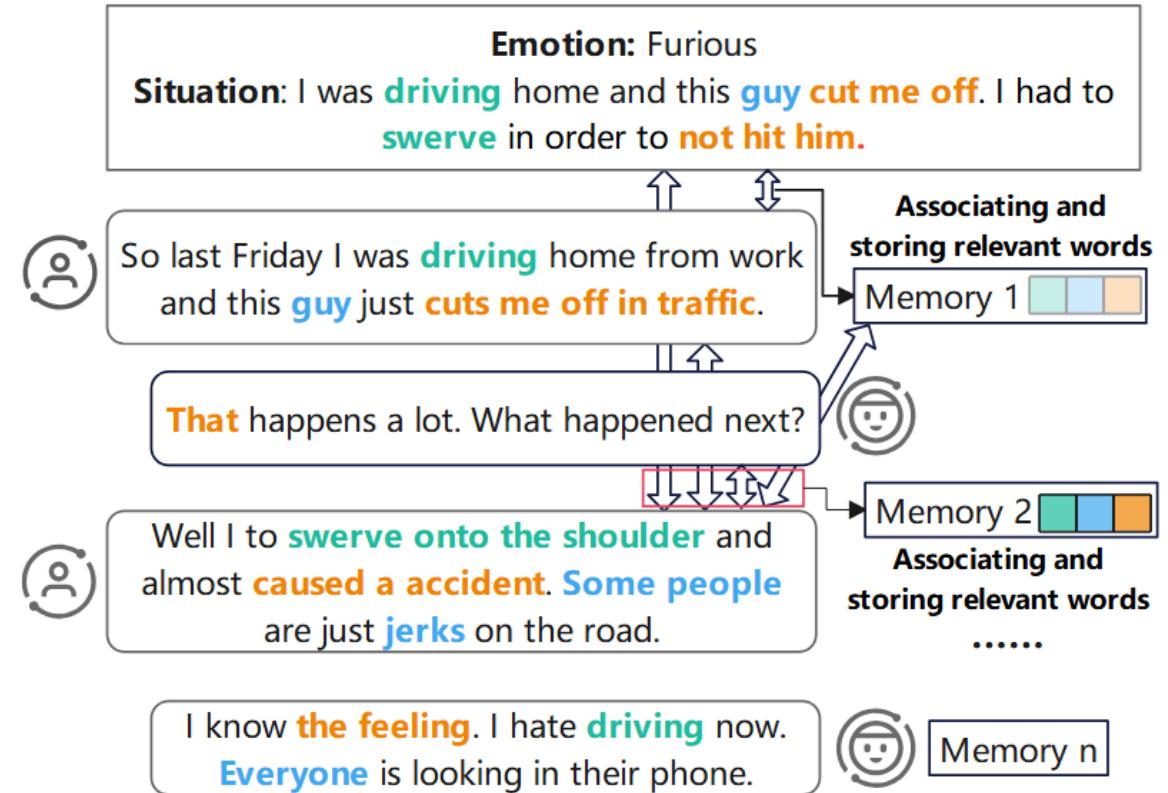
An example of situation enhancing dialogue Emotion understanding

[1] Yang et al. 2024. An Iterative Associative Memory Model for Empathetic Response Generation.

IAMM: Proposed Method

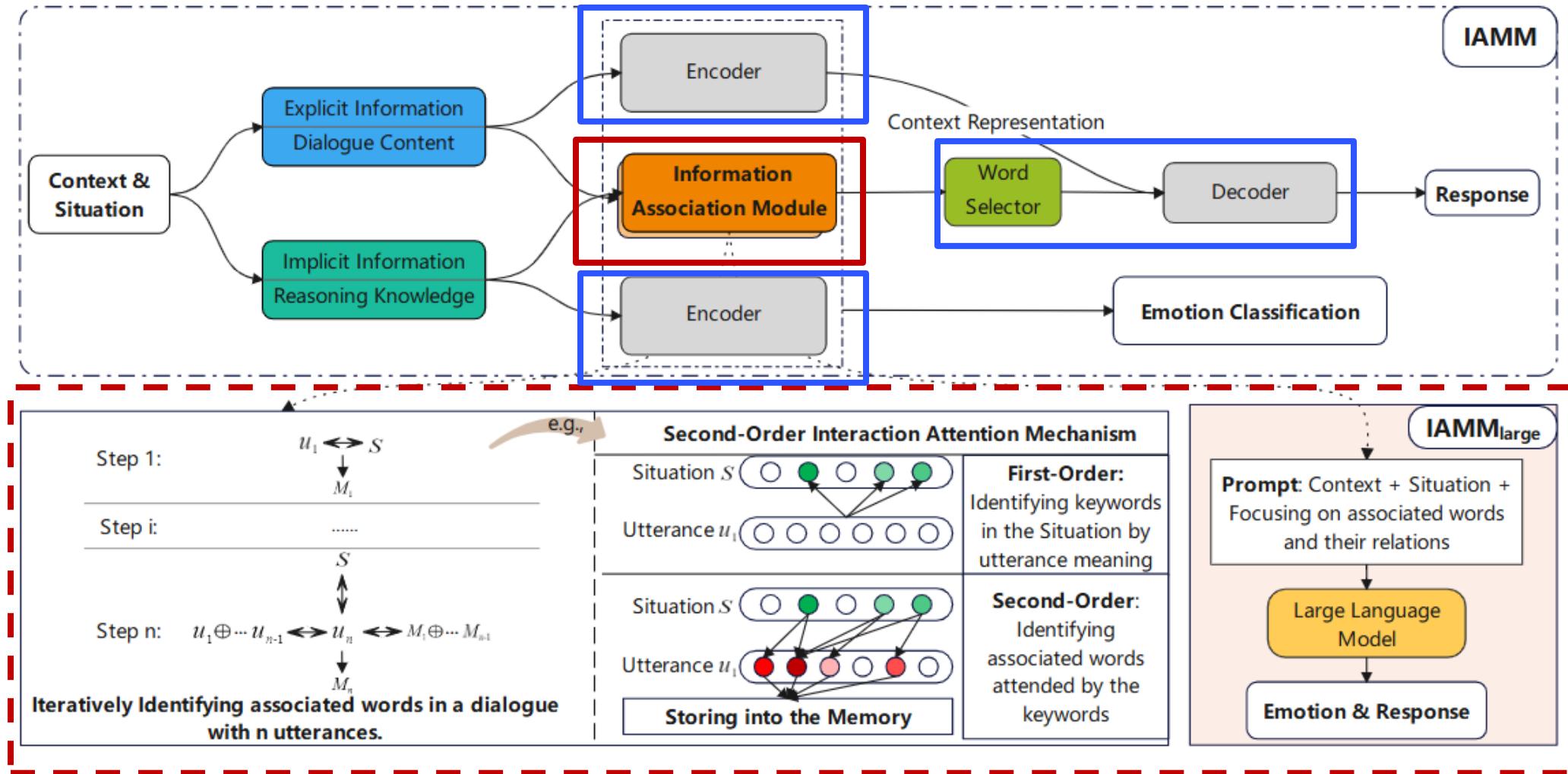
Motivation:

- ① Humans can more accurately perceive conversational emotions by iteratively integrating scene and dialogue knowledge.
- ② We mimic this iterative integration approach to merge scene knowledge.



An example of iterative knowledge integration

IAMM: Proposed Method



Part3:

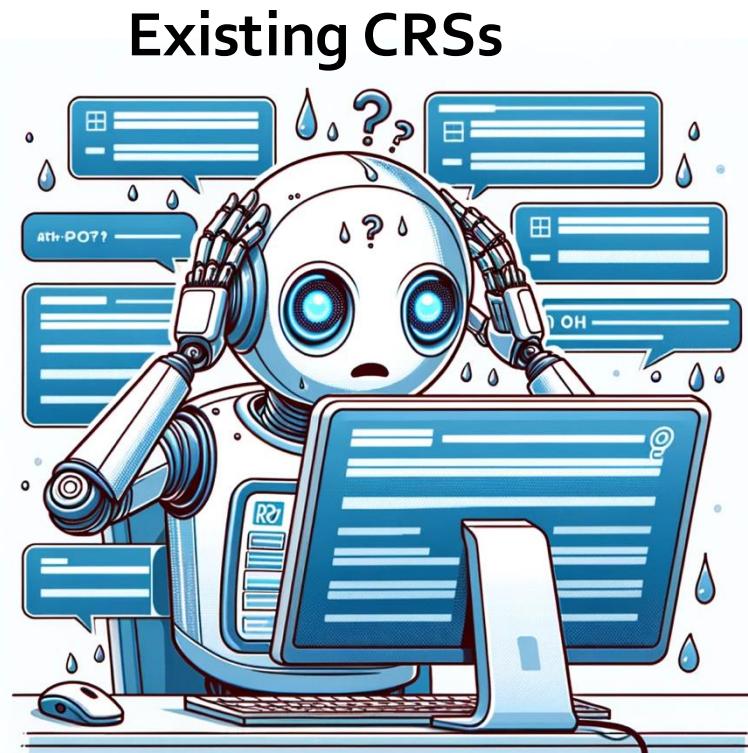
Towards empathetic conversational recommender systems



Universiteit
Leiden
The Netherlands

Motivation of this work

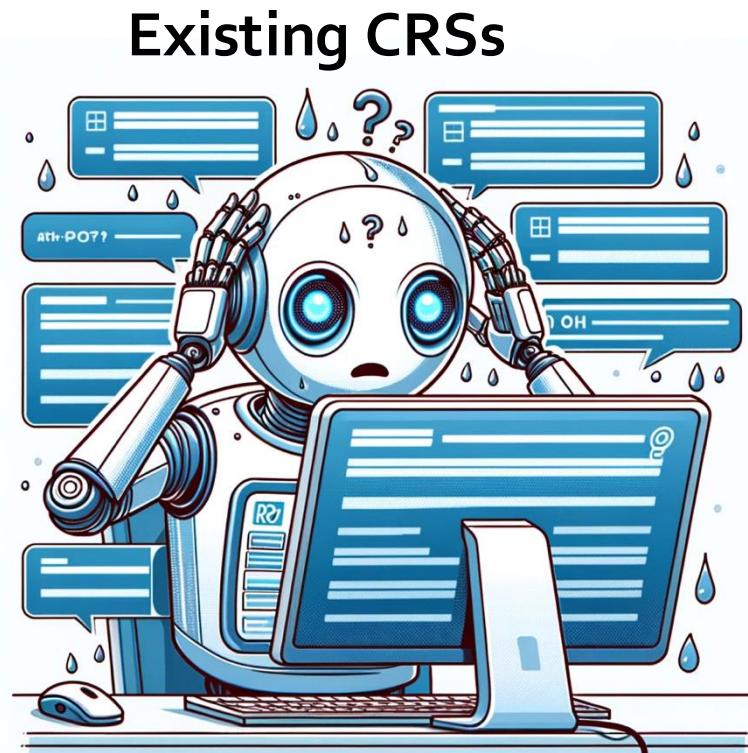
What is the ideal CRS like?



Rely on the benchmark dataset

Motivation of this work

What is the ideal CRS like?



Rely on the benchmark dataset

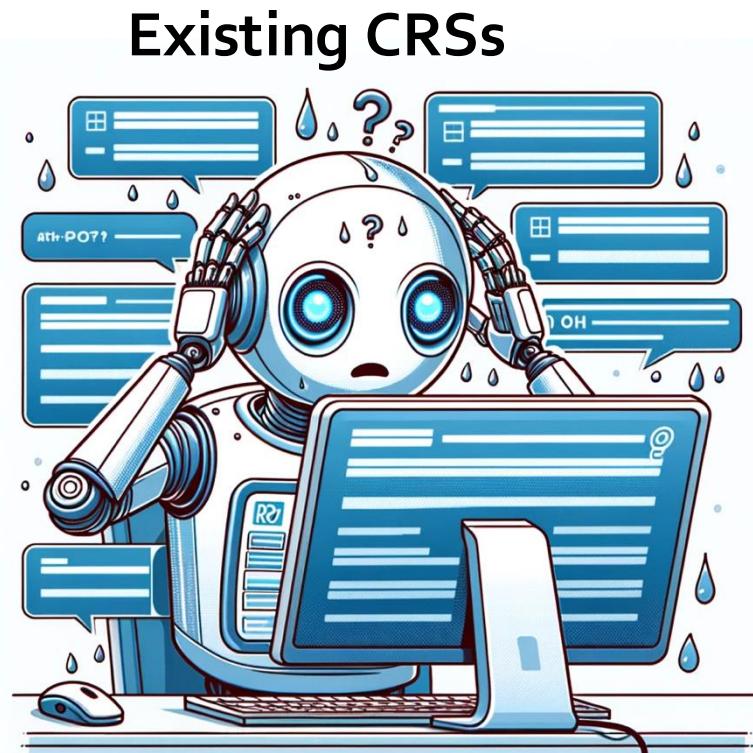
High recommender accuracy
User subjective satisfaction

User needs_[1]

[1] Pu et al. 2011. A User-Centric Evaluation Framework for Recommender Systems. RecSys'11.

Motivation of this work

What is the ideal CRS like?



Rely on the benchmark dataset



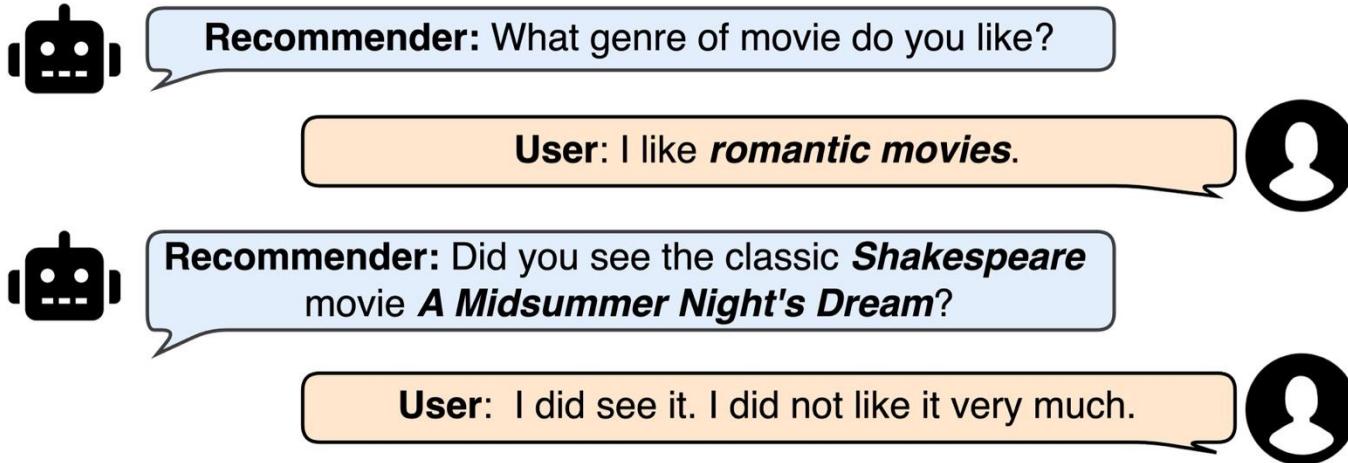
High recommender accuracy
User subjective satisfaction

User needs_[1]

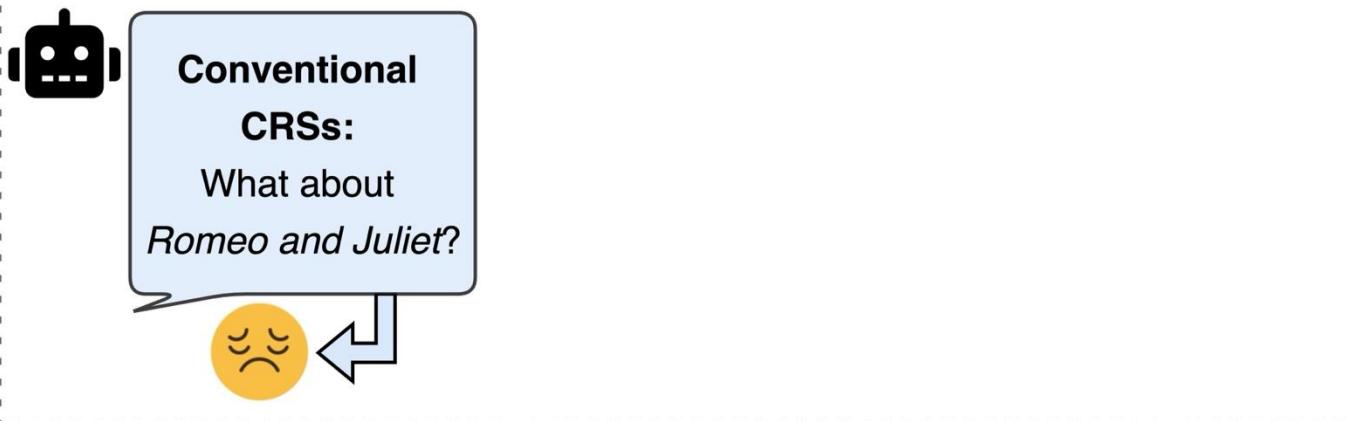
[1] Pu et al. 2011. A User-Centric Evaluation Framework for Recommender Systems. RecSys'11.

Motivation of this work

Dialogue History



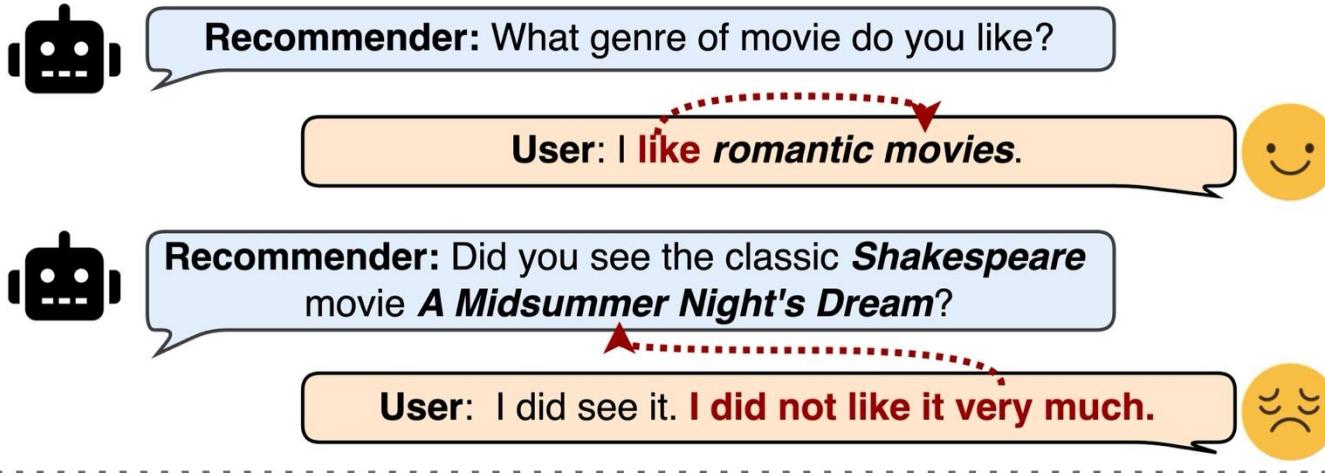
Systems Interaction



Misalignments with user needs

Motivation of this work

Dialogue History



Systems Interaction

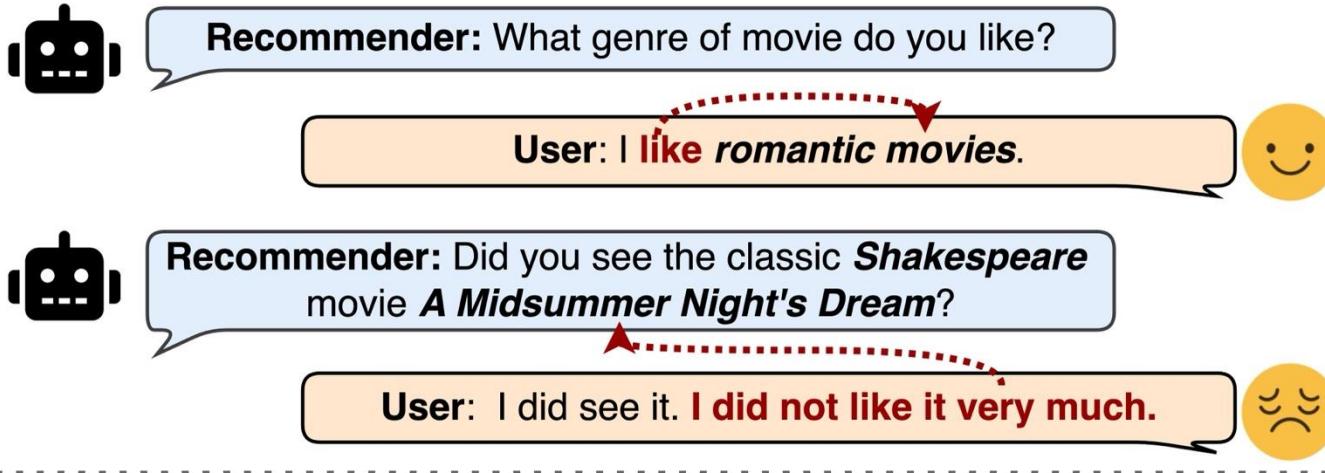


Misalignments with user needs

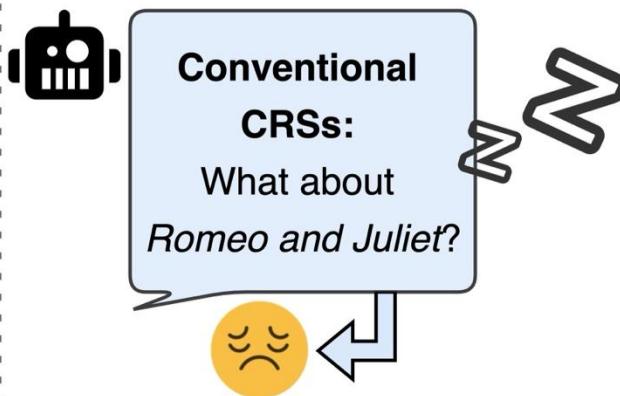
- ① Item recommendation:
Model preferences incorrectly.

Motivation of this work

Dialogue History



Systems Interaction



Misalignments with user needs

- ① Item recommendation:
Model preferences incorrectly.
- ② Response generation:
Diminish user satisfaction.

Motivation of this work

Dataset analysis

Redial

A large-scale CRS dataset, carefully curated by human workers.

Reasons

Recommended items

Items disliked by users: 4.9%_[3],
Items not explicitly stated by users : 14%_[3],
Entities mentioned but not reflect user preferences.

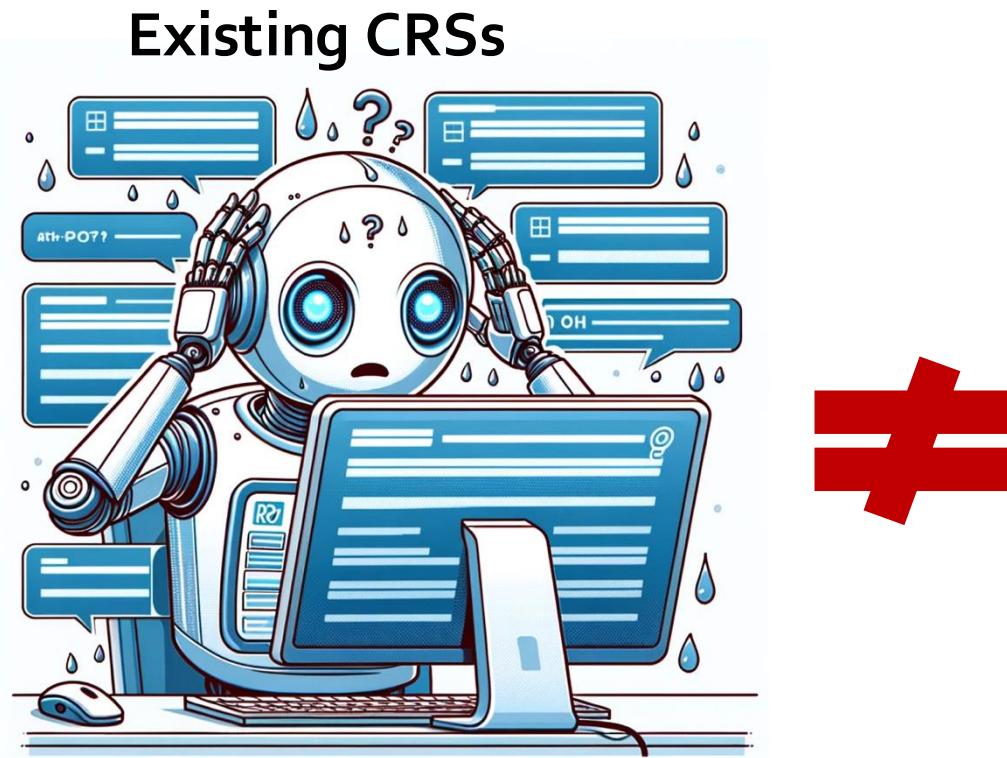
Standard responses

Short and lack narratives, resulting in little engagement.

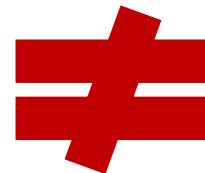
[3] Li et al. 2018. Towards Deep Conversational Recommendations. NeurIPS.

Motivation of this work

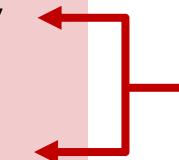
What is the ideal CRS like?



Rely on the benchmark dataset



High recommender accuracy
User subjective satisfaction



User needs [1]

[1] Pu et al. 2011. A User-Centric Evaluation Framework for Recommender Systems. RecSys'11.

Motivation of this work

How to deal with Misalignments?

Empathetic Response Generation

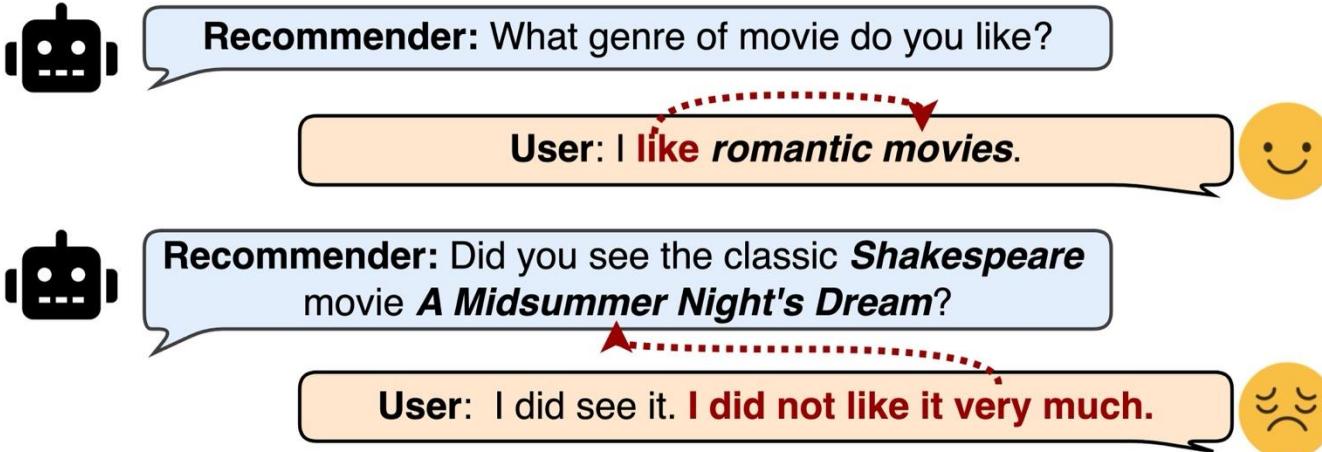


Fig. An Example of Empathetic Dialogue_[1]

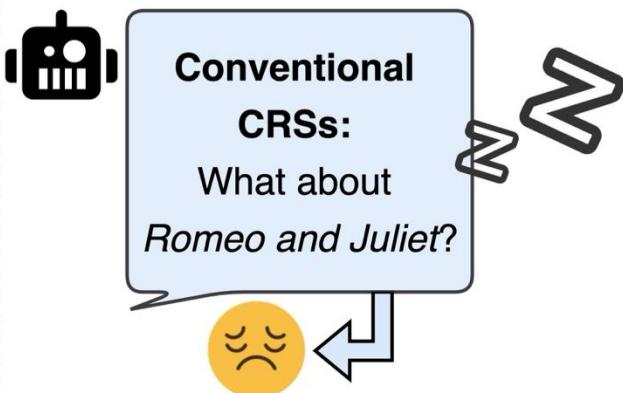
[1] Towards Empathetic Open-domain Conversation Models: a New Benchmark and Dataset : A Survey. In ACL 2019.

Motivation of this work

Dialogue History



Systems Interaction

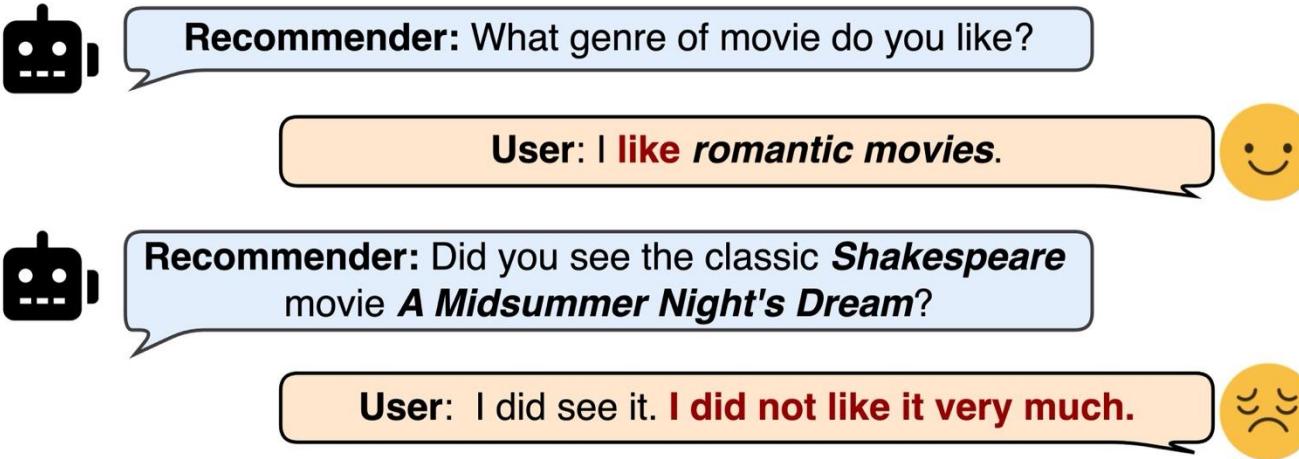


How to deal with Misalignments?

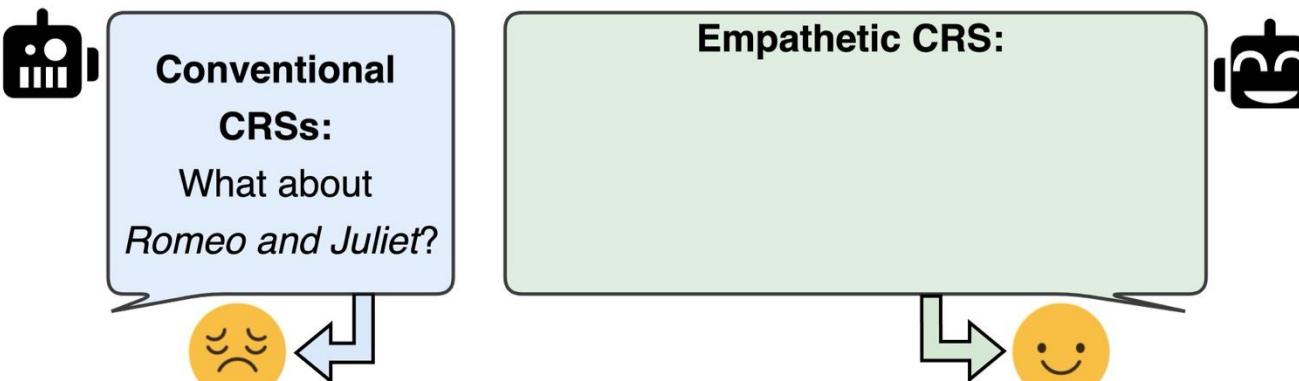
- ① Emotions are crucial in human decision-making processes.
- ② People favor agents simulating human beings by exhibiting emotions.

Motivation of this work

Dialogue History



Systems Interaction



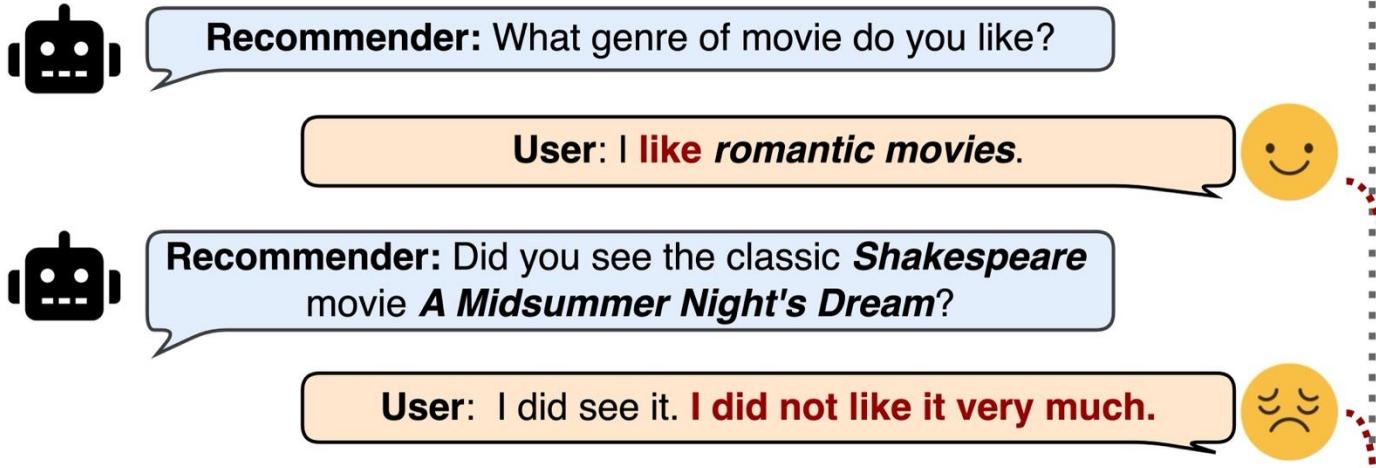
**Introduce Empathy
to bridge the gap**

① Capture emotions

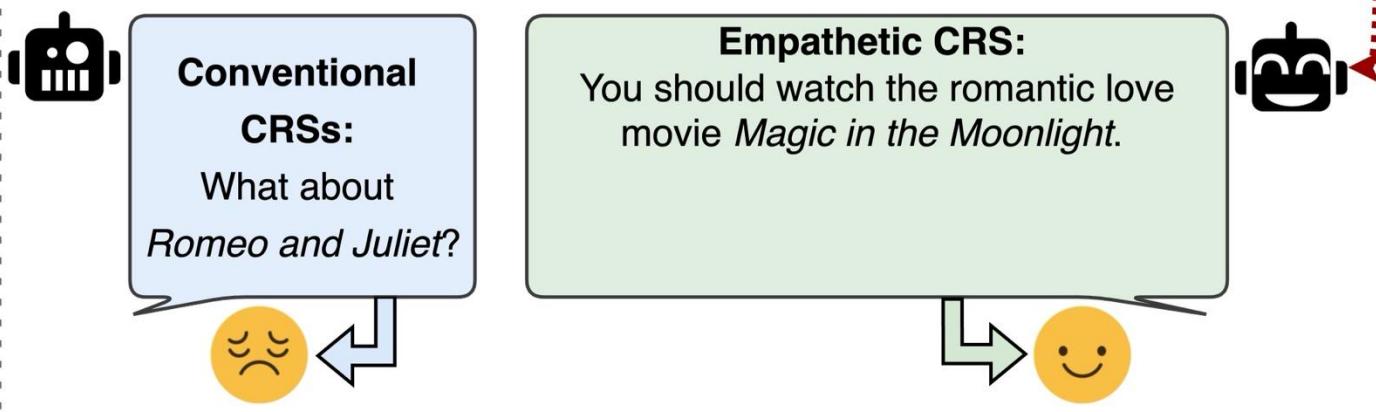
② Express emotions

Motivation of this work

Dialogue History



Systems Interaction

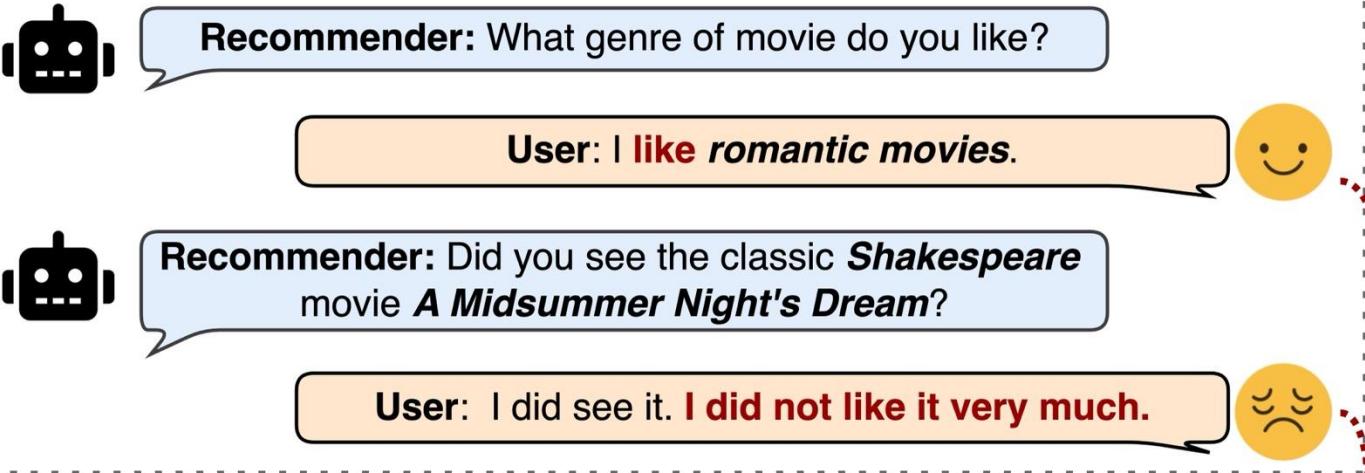


Introduce **Empathy** to bridge the gap

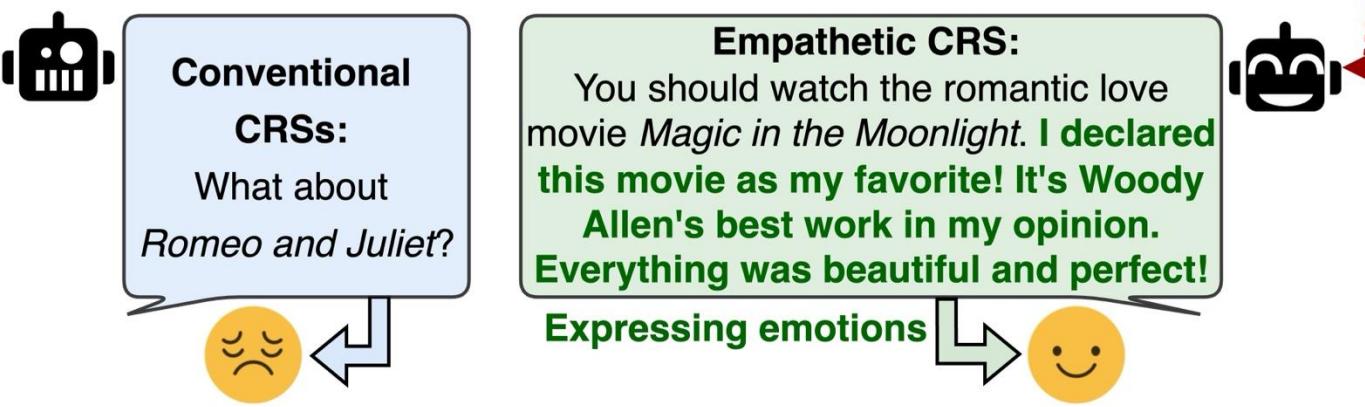
- ① Capture emotions:
Achieve accurate user preference modeling.
- ② Express emotions

Motivation of this work

Dialogue History



Systems Interaction



Introduce Empathy to bridge the gap

- ① Capture emotions:
Achieve accurate user preference modeling.
- ② Express emotions:
Enliven user experience and improve user satisfaction.

Empathetic Data Enlargement

Capture emotions: User emotion extraction

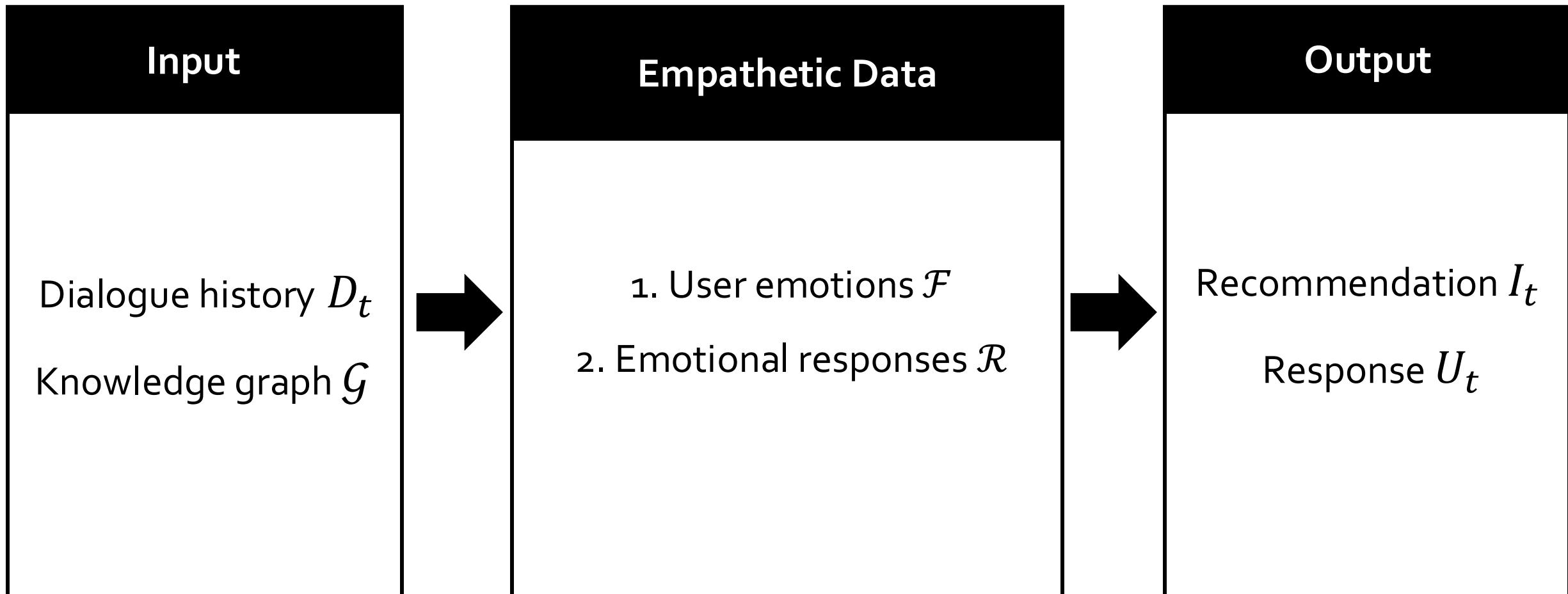
- ① Employ GPT-3.5-turbo to initially annotate some user utterances for emotions.
- ② Fine-tune a GPT-2 model based on the annotations.
- ③ Apply the model to obtain all utterance-level user emotions.

Express emotions: Emotional response construction

- ① Collect movie reviews from IMDb according to the items in the training dataset.
- ② Use top-rate(10/10) to filter the collected reviews and get the emotional review database.

Proposed framework: ECR

Problem Formulation



Proposed framework: ECR

Challenges

How to accurately model user preferences using emotions?

**How to generate emotional responses
contributing to user satisfaction?**

Proposed framework: ECR

Empathetic conversational recommender (ECR)

Emotion-aware item recommendation

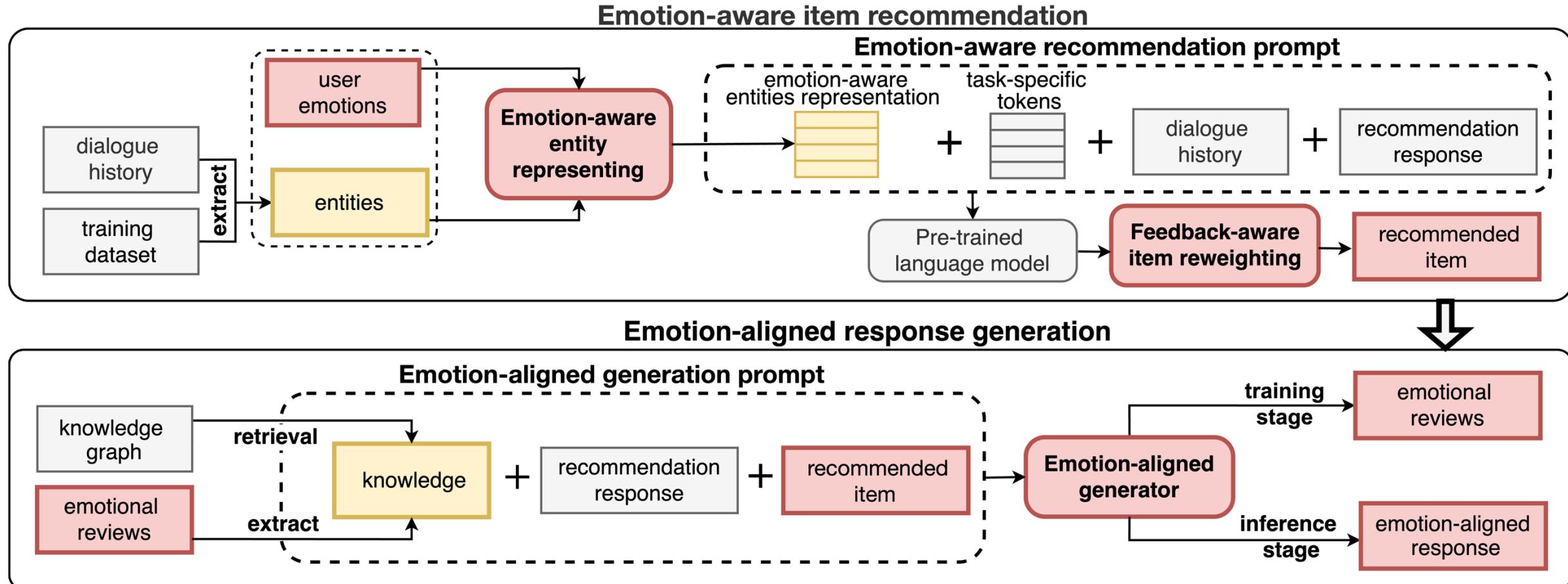
Integrate *user emotions* with entities in the utterance
to capture emotions.

Emotion-aligned response generation

Fine-tune a pre-trained language model with *emotional responses*
to express emotions.

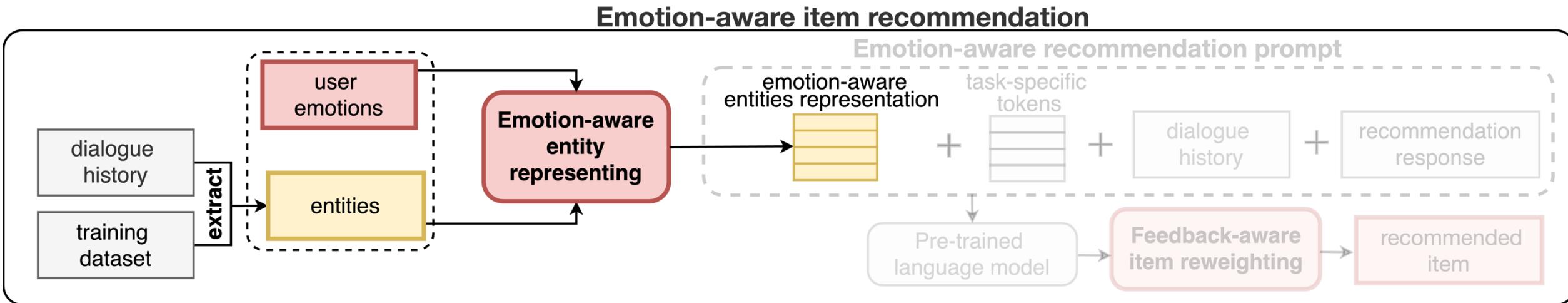
Proposed framework: ECR

Empathetic conversational recommender (ECR)



Proposed framework: ECR

Emotion-aware item recommendation module

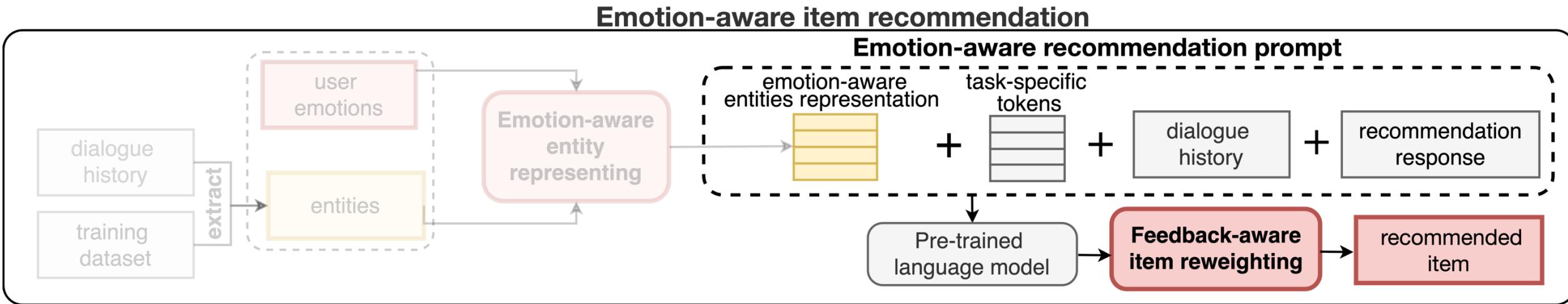


Emotion-aware entity representation:

- To model the effect of user emotions on the entities.
- Extracted entities fused with the user emotion.

Proposed framework: ECR

Emotion-aware item recommendation module

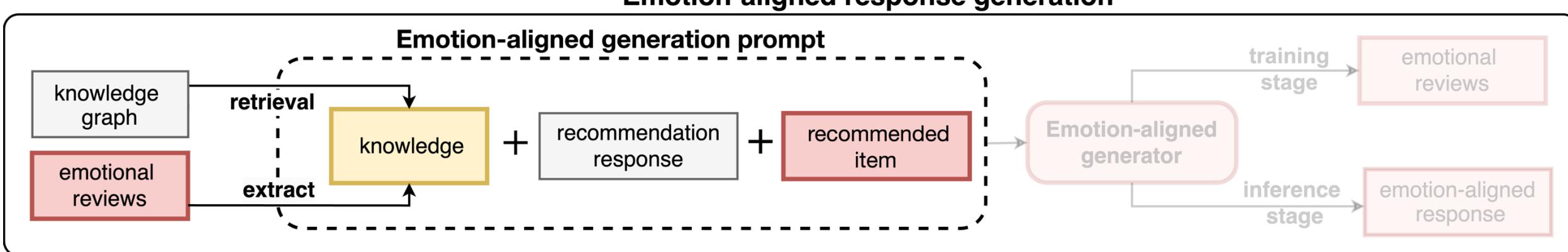


Emotion-aware recommendation prompt:

- To comprehensively model user preferences based on emotions.
- Emotion-aware entity representation.

Proposed framework: ECR

Emotion-aligned response generation module

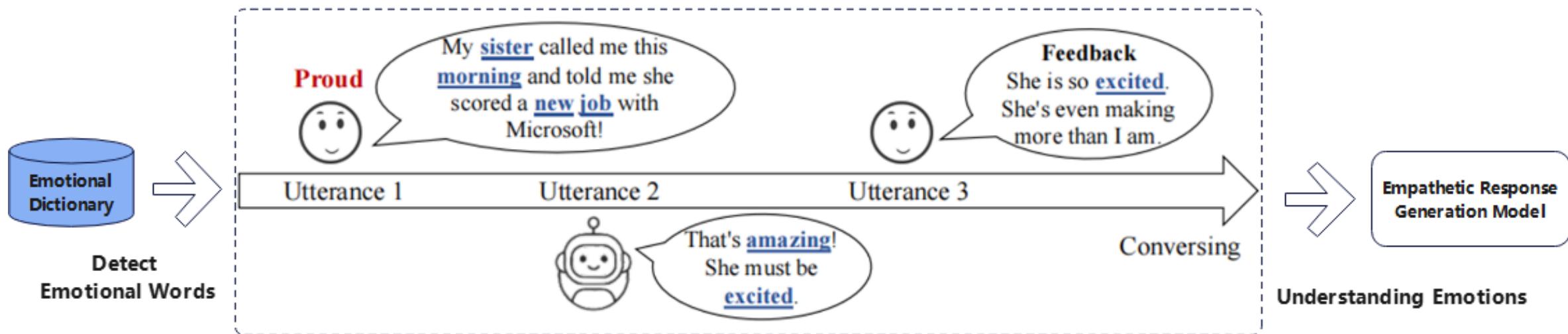


Emotion-aligned generation prompt:

Proposed framework: ECR

Emotion-aligned response generation module

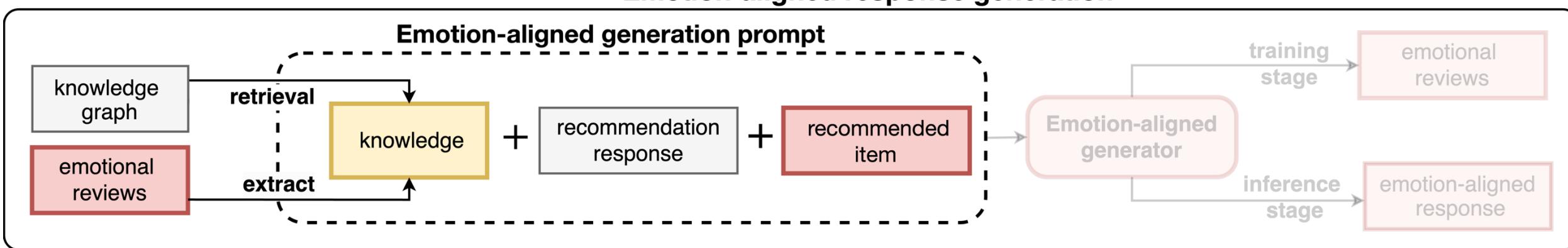
External knowledge enhances expression in Empathetic Response Generation_[1].



[1] EmpDG: Multi-resolution Interactive Empathetic Dialogue Generation. In COLING 2020.

Proposed framework: ECR

Emotion-aligned response generation module

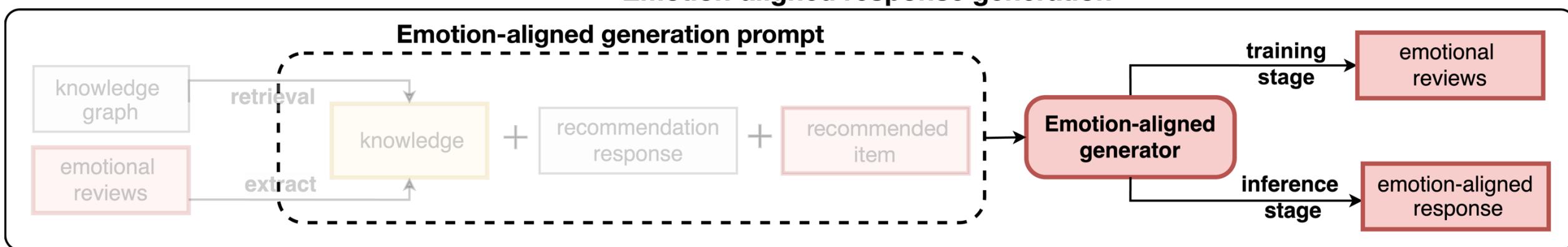


Emotion-aligned generation prompt:

- To satisfy users' information needs while mitigating hallucination.
- Knowledge retrieved and extracted from the knowledge graph and reviews.

Proposed framework: ECR

Emotion-aligned response generation module

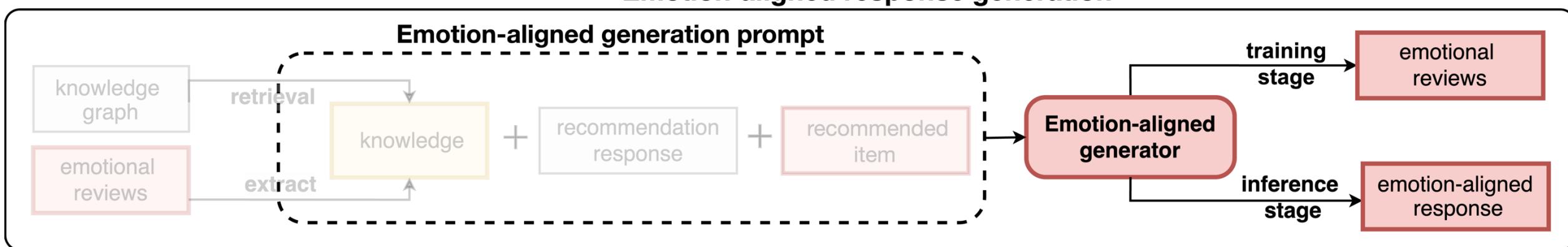


Emotion-aligned generator:

- To align the generator with persuasive dialogue style and emotions.
- Fine-tune a PLM using the constructed emotional reviews.

Proposed framework: ECR

Emotion-aligned response generation module



Emotion-aligned generator (DialoGPT and Llama 2-Chat):

- To align the generator with persuasive dialogue style and emotions.
- Fine-tune a PLM using the constructed emotional reviews.

Emotion-enhanced Evaluation

Objective metrics

- Recall@n (R@n)
- Recall_True@n (RT@n)
- AUC

Subjective metrics

- Emotional Intensity
- Emotional Persuasiveness
- Logic Persuasiveness
- Informativeness
- Lifelikeness

Emotion-enhanced Evaluation

Objective metrics

- Recall@n (R@n)
- Recall_True@n (RT@n)
- AUC

Subjective metrics

LLM-based scorer (GPT4-turbo) & three human annotators

- Emotional Intensity
- Emotional Persuasiveness
- Logic Persuasiveness
- Informativeness
- Lifelikeness

Experimental setup

Baselines

Recommendation

KBRD, KGSF, RevCore, UCCR and UniCRS

Generation

UniCRS, Llama 2-7B-Chat, GPT-3.5-turbo-instruct and
GPT-3.5-turbo

Main findings

Recommendation

Model	AUC	RT@1	RT@10	RT@50	R@1	R@10	R@50
KBRD	0.503	0.040	0.182	0.381	0.037	0.175	0.360
KGSF	0.513	0.043	0.195	0.383	0.040	0.182	0.361
RevCore	0.514	0.054	0.230	0.410	0.046	0.209	0.209
UCCR	0.499	0.038	0.208	0.423	0.039	0.198	0.407
UniCRS	0.506	0.052	0.229	0.439	0.047	0.212	0.414
ECR	0.541*	0.055	0.238*	0.452*	0.049	0.220*	0.428*

- ECR significantly outperforms all the baselines.

Main findings

Recommendation

Model	AUC	RT@1	RT@10	RT@50	R@1	R@10	R@50
KBRD	0.503	0.040	0.182	0.381	0.037	0.175	0.360
KGSF	0.513	0.043	0.195	0.383	0.040	0.182	0.361
RevCore	0.514	0.054	0.230	0.410	0.046	0.209	0.209
UCCR	0.499	0.038	0.208	0.423	0.039	0.198	0.407
UniCRS	0.506	0.052	0.229	0.439	0.047	0.212	0.414
ECR	0.541*	0.055	0.238*	0.452*	0.049	0.220*	0.428*

- ECR significantly outperforms all the baselines.
- These findings confirm the importance of capturing user emotions in enhancing CRSs ability to accurately estimate user preferences.

Main findings

Generation (LLM-based scorer)

Model	Emo Int	Emo Pers	Log Pers	Info	Life
UniCRS	0.400	0.942	0.793	0.673	2.241
GPT-3.5-turbo-instruct	1.706	3.043	3.474	2.975	4.182
GPT-3.5-turbo	2.215	3.754	4.782	4.147	5.338
Llama 2-7B-Chat	3.934	6.030	5.886	5.904	7.129
ECR[DialoGPT]	4.011	4.878	4.736	5.094	5.906
ECR[Llama 2-Chat]	6.826*	7.724*	6.702*	7.653*	8.063*

- LLMs in the zero-shot setting significantly outperform UniCRS fine-tuned on the entire dataset.
- This indicates the subpar quality of the dataset's standard responses.

Main findings

Generation (LLM-based scorer)

Model	Emo Int	Emo Pers	Log Pers	Info	Life
UniCRS	0.400	0.942	0.793	0.673	2.241
GPT-3.5-turbo-instruct	1.706	3.043	3.474	2.975	4.182
GPT-3.5-turbo	2.215	3.754	4.782	4.147	5.338
Llama 2-7B-Chat	3.934	6.030	5.886	5.904	7.129
ECR[DialoGPT]	4.011	4.878	4.736	5.094	5.906
ECR[Llama 2-Chat]	6.826*	7.724*	6.702*	7.653*	8.063*

- ECR[Llama 2-Chat] surpasses all the baselines.

Main findings

Generation (LLM-based scorer)

Model	Emo Int	Emo Pers	Log Pers	Info	Life
UniCRS	0.400	0.942	0.793	0.673	2.241
GPT-3.5-turbo-instruct	1.706	3.043	3.474	2.975	4.182
GPT-3.5-turbo	2.215	3.754	4.782	4.147	5.338
Llama 2-7B-Chat	3.934	6.030	5.886	5.904	7.129
ECR[DialoGPT]	4.011	4.878	4.736	5.094	5.906
ECR[Llama 2-Chat]	6.826	7.724	6.702	7.653	8.063

- ECR[Llama 2-Chat] surpasses all the baselines.
- ECR[DialoGPT] achieves a comparable performance with GPT-3.5-turbo.

Main findings

Generation (LLM-based scorer)

Model	Emo Int	Emo Pers	Log Pers	Info	Life
UniCRS	0.400	0.942	0.793	0.673	2.241
GPT-3.5-turbo-instruct	1.706	3.043	3.474	2.975	4.182
GPT-3.5-turbo	2.215	3.754	4.782	4.147	5.338
Llama 2-7B-Chat	3.934	6.030	5.886	5.904	7.129
ECR[DialoGPT]	4.011	4.878	4.736	5.094	5.906
ECR[Llama 2-Chat]	6.826*	7.724*	6.702*	7.653*	8.063*

- ECR[Llama 2-Chat] surpasses all the baselines.
- ECR[DialoGPT] achieves a comparable performance with GPT-3.5-turbo.
- ECR is human-like by expressing emotions, thereby improving user satisfaction.

Main findings

Generation (Human annotators)

Model	Emo Int	Emo Pers	Log Pers	Info	Life
UniCRS	0.947	0.775	1.158	0.38	1.805
GPT-3.5-turbo-instruct	2.048	2.555	3.265	1.822	3.648
GPT-3.5-turbo	2.89	3.678	5.323	3.233	5.125
Llama 2-7B-Chat	4.432	6.152	6.393	5.713	7.463
ECR[DialoGPT]	5.097	4.817	5.398	4.628	6.385
ECR[Llama 2-Chat]	7.130*	7.575*	7.403*	7.172*	8.468*

- Evaluations in the emotional dimension are highly consistent.

Main findings

Generation (Human annotators)

Model	Emo Int	Emo Pers	Log Pers	Info	Life
UniCRS	0.947	0.775	1.158	0.38	1.805
GPT-3.5-turbo-instruct	2.048	2.555	3.265	1.822	3.648
GPT-3.5-turbo	2.89	3.678	5.323	3.233	5.125
Llama 2-7B-Chat	4.432	6.152	6.393	5.713	7.463
ECR[DialoGPT]	5.097	4.817	5.398	4.628	6.385
ECR[Llama 2-Chat]	7.130*	7.575*	7.403*	7.172*	8.468*

- Evaluations in the emotional dimension are highly consistent.
- Discrepancies are primarily observed in logic persuasiveness.
- LLM still has a slight gap with humans when performing subjective tasks.

Main findings

Case study

- Standard response rarely expresses emotions and is uninformative.
- LLMs, i.e., GPT-3.5-turbo and Llama 2-7B-Chat, prioritize factual and logical interpretability.

Standard response: Do you like the Marvel movies, like Avengers: Infinity War?

GPT-3.5-turbo: If you enjoy *comedy*, I highly recommend watching "The Hangover." It's a hilarious movie that follows a group of friends on a wild and unforgettable adventure in Las Vegas. **I think you'll really enjoy it!**

Emo Int: 2; **Emo Pers:** 4; **Log Pers:** 4; **Info:** 5; **Life:** 5;

Llama 2-7B-Chat: Great! Based on your interests, I highly recommend "The Hangover"! **It's a hilarious comedy movie that will have you laughing out loud.** The movie follows a group of friends who travel to Las Vegas for a bachelor party and wake up the next morning with no memory of what happened the night before. They then have to retrace their steps and figure out what happened. **It's a really funny and entertaining movie that I think you'll enjoy.** Let me know if you have any other questions!

Emo Int: 2; **Emo Pers:** 5; **Log Pers:** 5; **Info:** 7; **Life:** 6;

Main findings

Case study

- Standard response rarely expresses emotions and is uninformative.
- LLMs, i.e., GPT-3.5-turbo and Llama 2-7B-Chat, prioritize factual and logical interpretability.
- Simulates the vivid conversational style between human beings, incorporating experiences, subjective viewpoints, and emotions.

ECR[DialoGPT]: Have you seen The Hangover? I have to say that I was very surprised at how good this movie was. I thought it was going to be a typical "chick flick" but it was much more than that. The story was great and the acting was great. I was laughing so hard I almost cried. I would recommend this movie to anyone who likes to laugh.

Emo Int: 5; Emo Pers: 6; Log Pers: 6; Info: 6; Life: 7;

ECR[Llama 2-Chat]: Have you seen The Hangover? I don't think I have ever laughed so hard at a movie. The acting by Ed Helms was excellent. I have seen every episode of The Office and I have always enjoyed him in it. The other actors did a great job too. I have been looking forward to this movie for months and I wasn't disappointed. The movie was very entertaining and I was laughing throughout the whole movie. The jokes were not crude and the story was good. I would definitely recommend this movie to anyone who enjoys a good comedy.

Emo Int: 4; Emo Pers: 7; Log Pers: 7; Info: 8; Life: 8 ;

Summary



Universiteit
Leiden
The Netherlands

Discover the world at Leiden University

Summary

Why do CRSs need knowledge?

- The job of CRSs is to provide recommendations to users through the real-time multi-turn interactions in natural language.
- We find challenges of meeting user expectations in CRSs .
 - Lack of sufficient contextual information.
 - Semantic inconsistency between the recommendation and generation subtasks.
- Incorporating knowledge addresses the above challenges.

Summary

Where does CRS' knowledge come from?

- **Mining in-text knowledge**
 - Extracted from user's natural language interactions within the ongoing session or across sessions.
 - Critical to deeply understanding individual user preference.
- **Incorporating external knowledge graphs.**
 - Provide sufficient domain background information and enhance understanding of user semantics.
 - External KGs fused with in-text knowledge provide more complete information.
- **Synthesizing emotional insights.**
 - Understanding and expressing emotions lead to recommendations that are better aligned with user expectations and responses that more enliven user experience.

Summary

Emotional insights play an important role in CRS

- Relook at the CRS and propose a new paradigm, introduce empathy within a CRS to bridge the gap between system outputs and user needs.
- Construct empathetic CRS dataset and propose new evaluation metrics, significantly facilitate the future research.
- Experimental results verify the effectiveness of our instantiated ECR in improving recommendation accuracy and user satisfaction.

Outlook

- **Human-centered CRS**
 - More challenging settings for recommending multiple items.
 - User study on empathetic conversational recommendation
- **CRSs with LLMs:**
 - 1. LLMs with external Recsys CRSs.**
 - LLMs are used for generation while an external recommender system is employed.
 - 2. LLMs-only CRSs.**
 - LLMs not only facilitate conversations but also directly generate recommendations.
 - Dividing the main CR task into simpler sub-tasks and having multiple LLM-based agents optimize systems performance.

Thanks for your attention !

Zhaochun Ren
z.ren@liacs.leidenuniv.nl

Code and Data :

UPCR: <https://github.com/tianz2020/UPCR>

VRICR: <https://github.com/zxd-octopus/VRICR>

EmpDG : <https://github.com/qtli/EmpDG>

KEMP: <https://github.com/qtli/KEMP>

IAMM: <https://github.com/zhouzhouyang520/IAMM>

ECR: <https://github.com/zxd-octopus/ECR>



**Universiteit
Leiden**
The Netherlands

Discover the world at Leiden University