# Evaluating Conversational Recommender Systems via Large Language Models

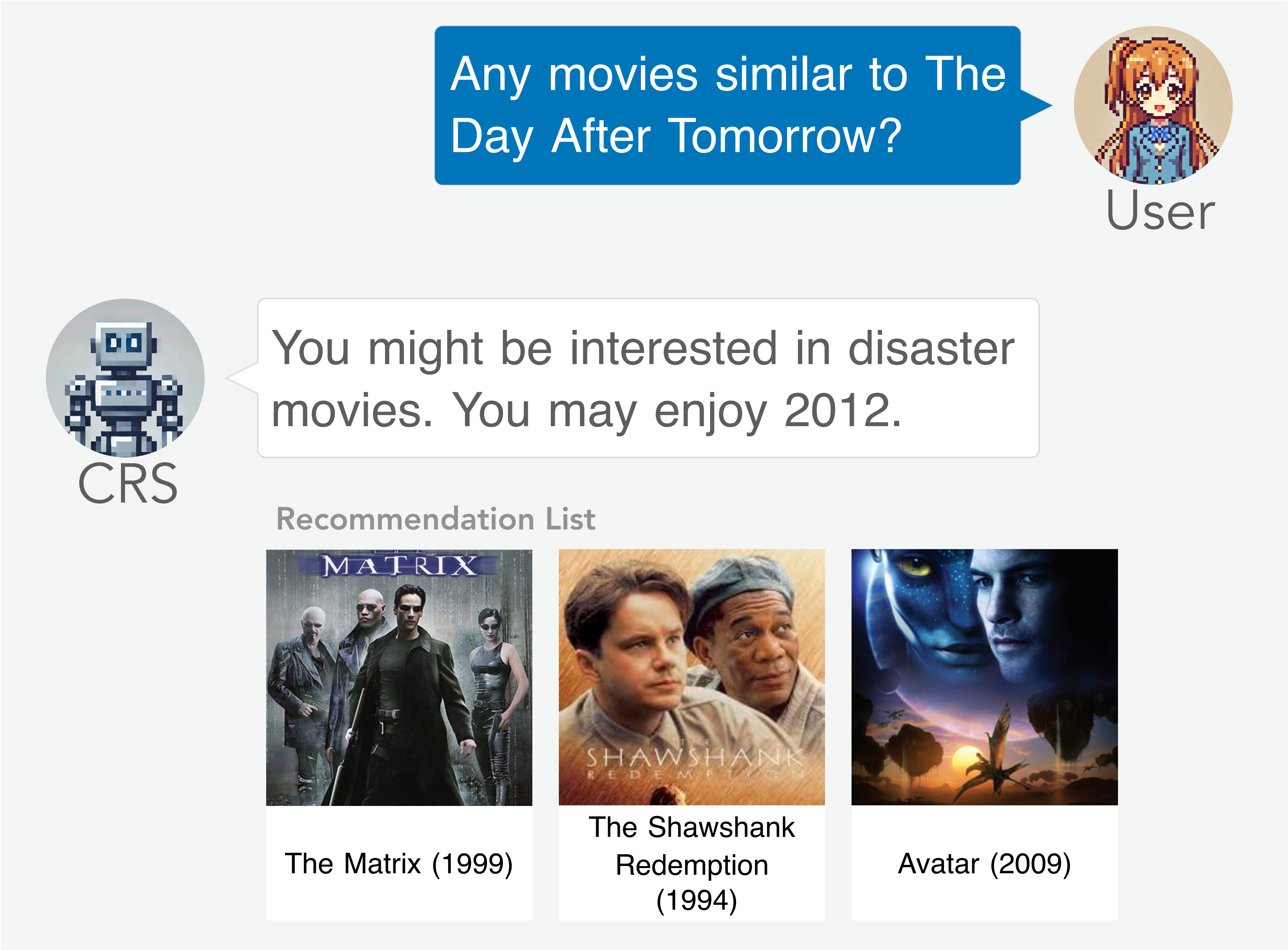
## A User-Centric Framework

Nuo Chen, Quanyu Dai, Xiaoyu Dong, Xiao-Ming Wu, Zhenhua Dong

# Introduction

# Conversational Recommender Systems

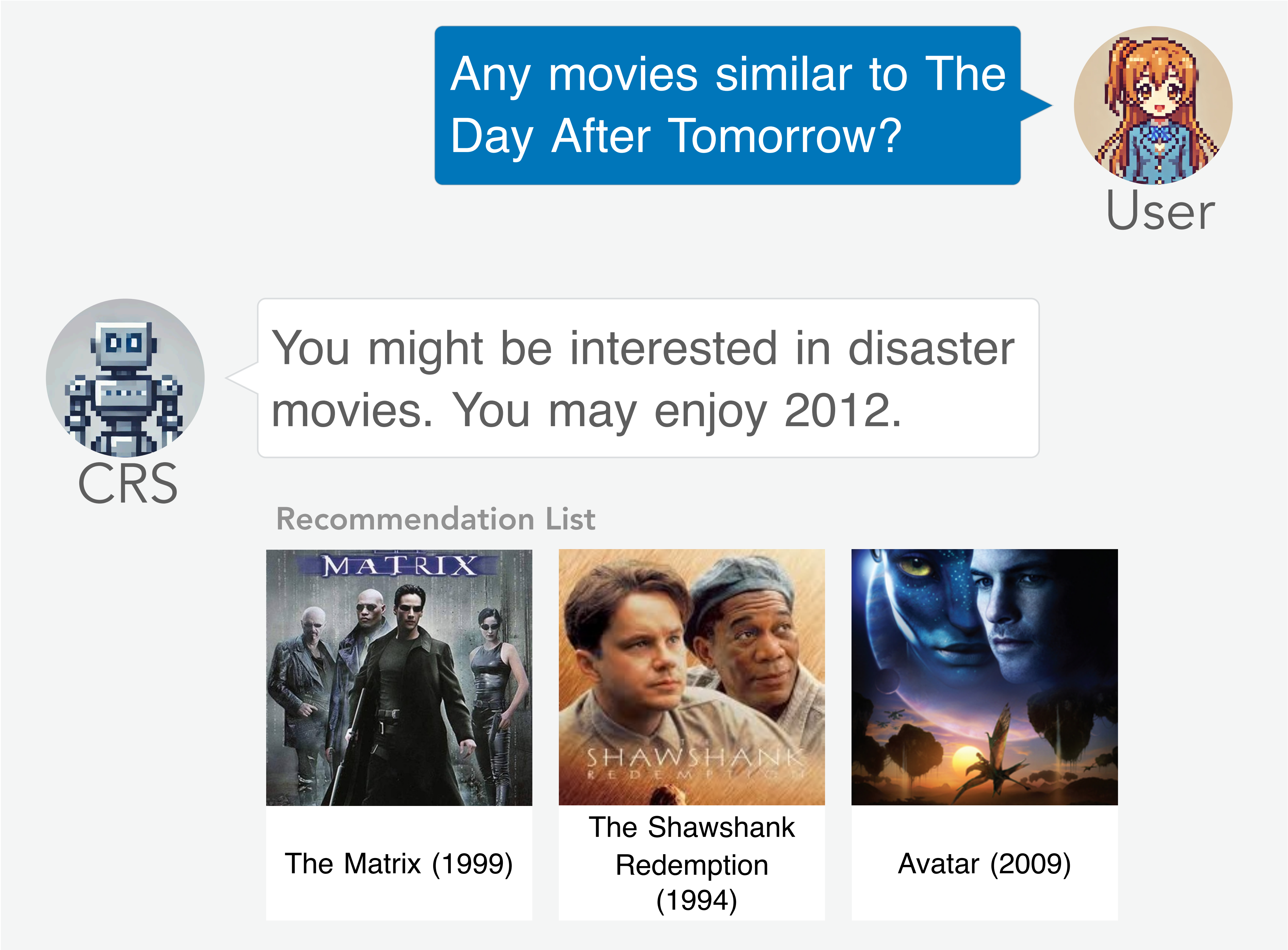
- A ***Conversational Recommender System (CRS)*** identifies user interests through **conversation.**
- A CRS not only provides item recommendations but also **manages dialogues** with users. (Jannach et al.,2021; Gao et al., 2021)
- Good CRS: **good recommendation** and **good dialogue management**




- Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and 733Li Chen. 2021. A survey on conversational recommender systems. ACM Comput. Surv., 54(5).
- Chongming Gao, Wenqiang Lei, Xiangnan He, Maarten de Rijke, and Tat-Seng Chua. 2021. Advances and challenges in conversational recommender systems: A survey. AI Open, 2:100–126.

# Challenges in Evaluating CRSs

- **Complexity** of the task
  - Traditional recommender systems
    - Item recommendations only.
  - CRSs
    - Not only **recommendation** but also **conversation.**
    - **Recommendation in conversation**
- How to evaluate?



# Research Gap

- Existing evaluation practice (e.g., Chen et al., 2019)
  - Treating item recommendation and dialogue management as **isolated** task
  - Using ***rule-based metrics***
  - **Drawbacks**
    - Fails to fully capture the **essence** of conversational recommendation
    - Rule-based evaluation metrics fail to align with actual user experience (Reiter, 2018; Chen et al., 2017)


- Qibin Chen, Junyang Lin, Yichang Zhang, Ming Ding,Yukuo Cen, Hongxia Yang, and Jie Tang. 2019. Towards knowledge-based recommender dialog system. In Proceedings of EMNLP-IJCNLP'19, pages 1803–1813.676.
- Ehud Reiter. 2018. A structured review of the validity of BLEU. Computational Linguistics, 44(3):393–401
- Ye Chen, Ke Zhou, Yiqun Liu, Min Zhang, and Shaoping Ma. 2017. Meta-evaluation of online and offline web search evaluation metrics. In Proceedings of SIGIR'17, page 15–24. 682.

# Research Gap

- Recent Advances in ***Large Language Models (LLMs)***
  - Enhanced nuanced natural language understanding
  - Significant potential in **aligning with human** text quality preferences (e.g., Liu et al., 2023)
  - Implications for CRSs
    - LLMs as a promising tool for intelligent **evaluation of CRSs**
- Only a few studies have investigated **LLM-based evaluation for conversational recommendation**


- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang,Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In Proceedings of EMNLP 2023, pp 2511–2522

# Our Contribution

- In this work, we propose a user-centric evaluation framework based on LLMs for CRS, namely **Conversational Recommendation Evaluator (CoRE)**
- CoRE consists of two main components:
  - (1) **LLM-As-Evaluator**.
    - Leverage LLM as evaluator to assign scores to 12 key factors influencing user experience in CRSs.
  - (2) **Multi-Agent Debater**.
    - A ***multi-agent*** debate framework with **four distinct roles** to discuss and synthesize the 12 evaluation factors into a unified overall performance score.

CRS
(Simulated) User
0
Interaction
Recommended Items
U I really liked Iron Man 3, can you recommend a movie similar to it?
S Iron Man 2 was written by Stan Lee and Stan Lee. Stan Lee also wrote The Avengers.
U I appreciate that you mentioned "Iron Man 2" and "The Avengers," but I'm actually looking for something similar to "Iron Man 3." It had a unique storyline, and I'm particularly interested in the journey of Tony Stark, including his challenges and character development. Could you suggest something more aligned with that?
S Sure, Tony Stark had a role in the Avengers, but his role in The Avengers was also in the movie Iron Man2.
U Thank you for the information, but I'm really looking for something that focuses on Tony Stark's character development and challenges, similar to the experience I had watching "Iron Man 3."
Iron Man 2
2012
Kick-Ass
Toy Story 3
Deadpool
The Incredible Hulk
Zombieland
Zero Dark Thirty
X-Men 2
Cowboys & Aliens
War
Shark Night
The Matrix
The Dark Knight
Multi-turn Conversation Log
Evaluation Guideline
Input
Input
LLMs
Output
1 LLM as Evaluator
Coherence 2
Recoverability 4
Proactiveness 2
Appropriate-ness 4
Naturalness 3
Grammar 4
Effectiveness 3
Novelty 1
Diversity 2
Semantic Relevance 1
Explainability 2
Groundness 4
12-Factor Evaluation
Scores & Reasons
Input
Else: Input Chat Buffer, Continue Discussion
Common User
Effectiveness 3
Coherence 2
Recoverability 4
Domain Expert
Novelty 1
Diversity 2
Groundness 4
Linguist
Appropriate-ness 4
Naturalness 3
Grammar 4
HCI Researcher
Semantic Relevance 1
Explainability 2
Proactiveness 2
LLM-based Agents Discussion
Statements & Scores
Chat Buffer
If (reach consensus or reach turn limit)
[Turn] 1 [Score] 70 [Reason] …
[Turn] 1 [Score] 20 [Reason] …
[Turn] 1 [Score] 90 [Reason] …
[Turn] 1 [Score] 20 [Reason] …
…
2 Multi-Agent Debater
67
Final Score

# Literature Review

# Evaluating Conversational Recommender Systems

- Traditional Methods (Chen et al., 2019; Wang et al., 2022a,b; Zhang et al., 2024; Feng et al., 2023):
  - Use rule-based measures (e.g., Recall, BLEU) for separate tasks
  - Often fail to capture the holistic user experience (Reiter, 2018; Chen et al., 2017)
- Limitations:
  - Isolated evaluation makes it hard to assess overall system performance
  - Difficulty in balancing recommendation accuracy with dialogue quality


- Qibin Chen, Junyang Lin, Yichang Zhang, Ming Ding, Yukuo Cen, Hongxia Yang, and Jie Tang. 2019. To- wards knowledge-based recommender dialog system. In Proceedings of EMNLP'19
- Ting-Chun Wang, Shang-Yu Su, and Yun-Nung Chen. 2022a. Barcor: Towards a unified framework for conversational recommendation systems. arXiv preprint
- Xiaolei Wang, Kun Zhou, Ji-Rong Wen, and Wayne Xin Zhao. 2022b. Towards unified conversational recommender systems via knowledge-enhanced prompt learning. In KDD'22
- Lu Zhang, Chen Li, Yu Lei, Zhu Sun, and Guanfeng Liu. 2024. An empirical analysis on multi-turn conversational recommender systems. In SIGIR'24
- Yue Feng, Shuchang Liu, Zhenghai Xue, Qingpeng Cai, Lantao Hu, Peng Jiang, Kun Gai, and Fei Sun. 2023. A large language model enhanced conversational recommender system. arXiv preprint

- Ehud Reiter. 2018. A structured review of the validity of BLEU. Computational Linguistics, 44(3):393–401
- Ye Chen, Ke Zhou, Yiqun Liu, Min Zhang, and Shaoping Ma. 2017. Meta-evaluation of online and offline web search evaluation metrics. In Proceedings of SIGIR'17, page 15–24. 682.

# Large Language Models as Evaluators

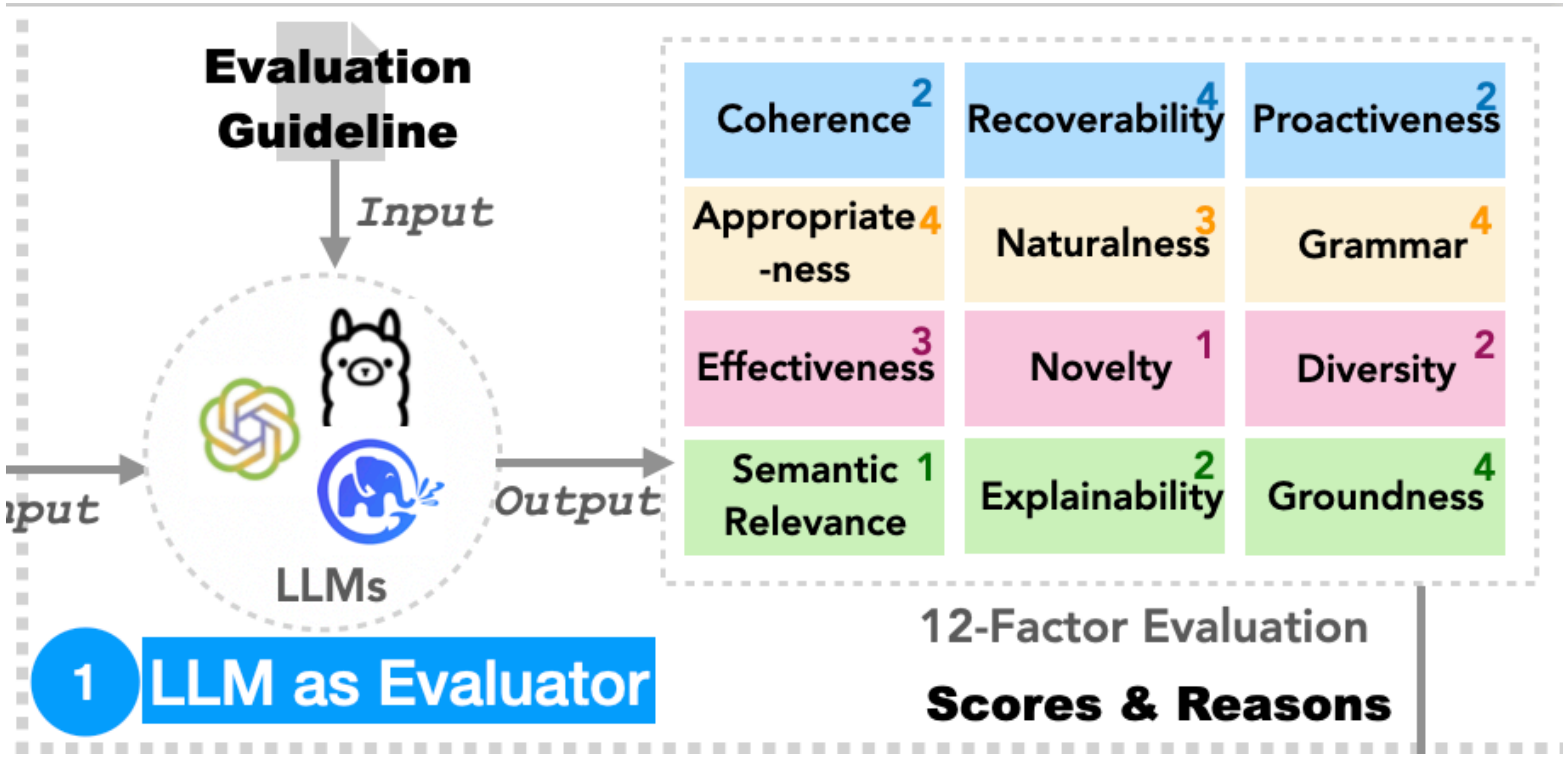
- **Motivation for LLMs:**
  - Strong natural language understanding capabilities (Liu et al., 2023; Fu et al., 2024; Chen et al., 2023b)
  - Proven potential to align with human evaluation of text quality (Chiang and Lee, 2023; Wang et al., 2024; Gao et al., 2023)
  - Early work shows LLMs can assess both recommendation relevance and dialogue fluency
- Limitations:
  - Few studies have used LLMs to integrate evaluation across both dimensions


- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *EMNLP'23*
- Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu. 2023b. Exploring the use of large language models for reference-free text quality evaluation: An empirical study. In *JCNLP-AACL 2023 (Findings)*
- *Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and 823 Yue Zhang. 2024. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization. ArXiv preprint*
- *Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. 2023. Human-like summarization evaluation with chatgpt. ArXiv preprint*
- Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In ACL'23
- Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei 708 Liu. 2024. GPTScore: Evaluate as you desire. In NAACL'24

# Proposed Framework

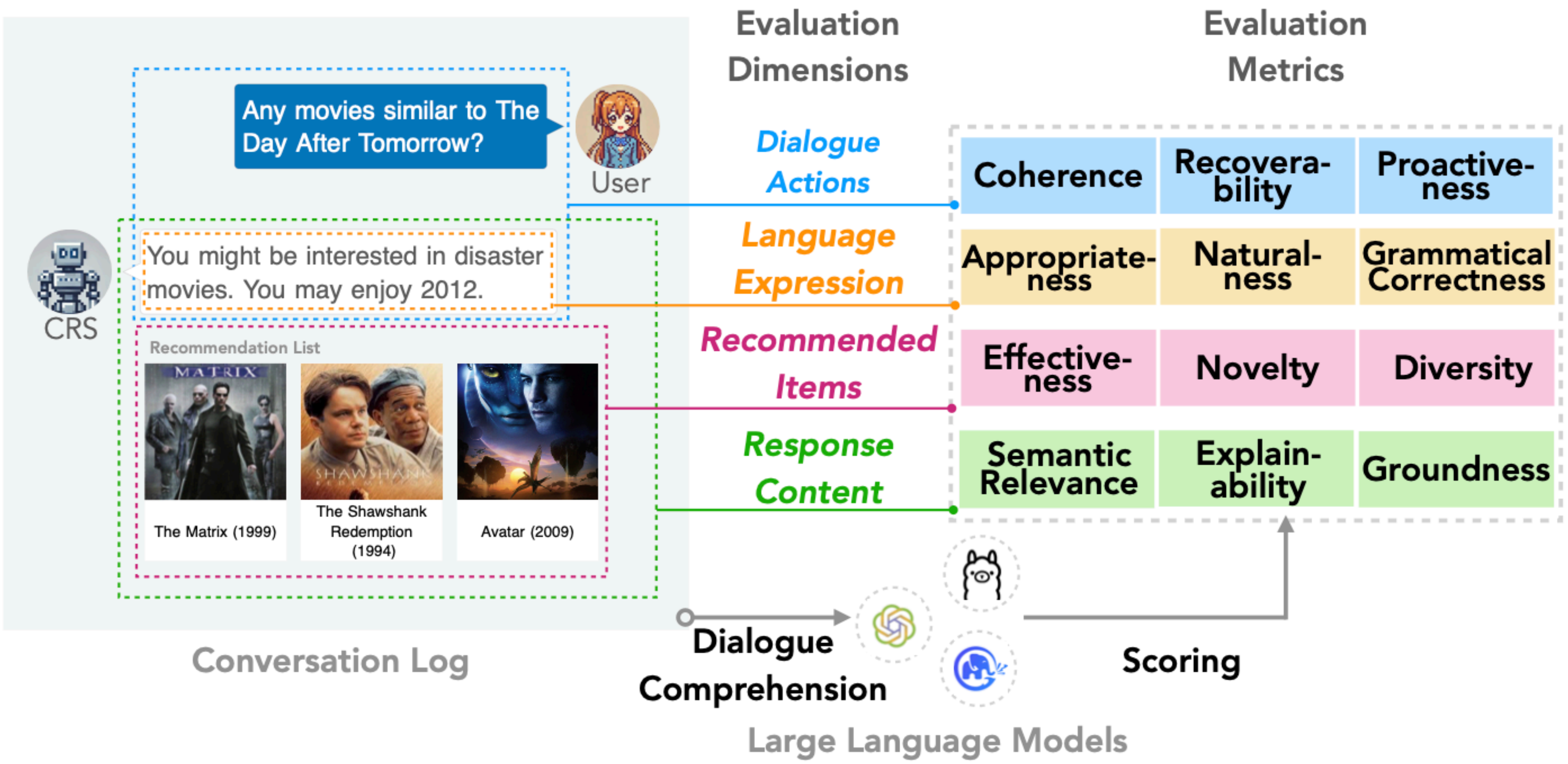
# LLM-As-Evaluator

- Evaluate CRS dialogues on **12 key factors** influencing user experience.
- Factors divided into 4 dimensions:
  - Dialogue Actions: Coherence, Recoverability, Proactiveness
  - Language Expression: Grammar, Naturalness, Appropriateness
  - Recommended Items: Effectiveness, Novelty, Diversity
  - Response Content: Semantic Relevance, Explainability, Groundness
  - Each factor scored (0–4) with rationale provided by the LLM.



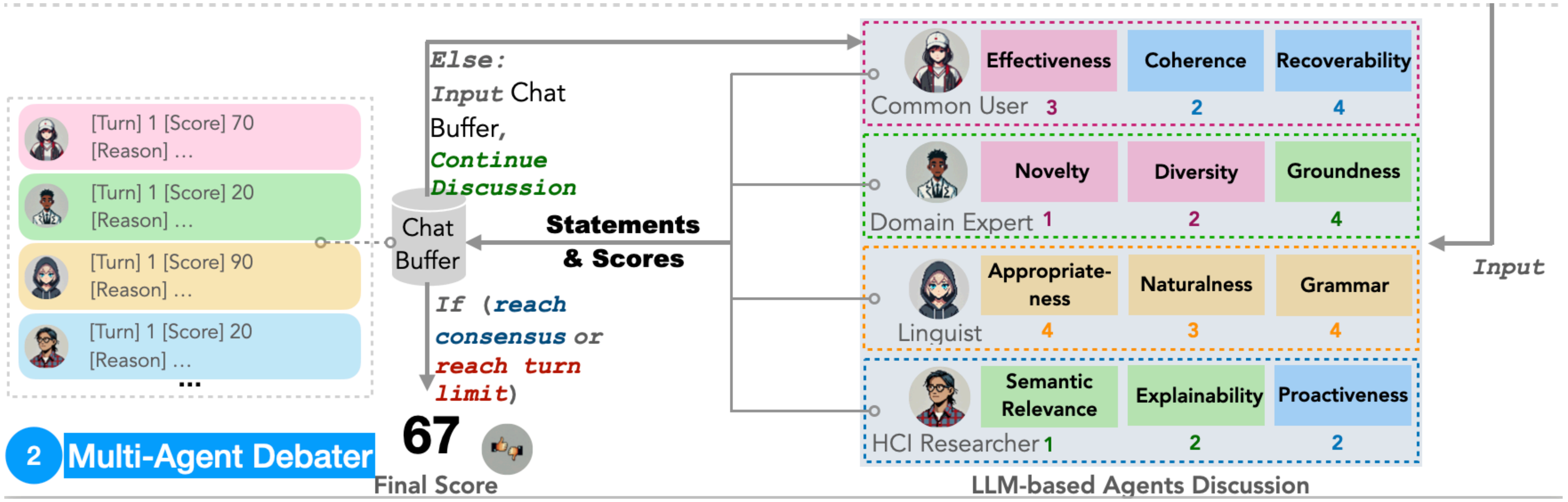
# LLM-As-Evaluator

- Evaluate CRS dialogues on **12 key factors** influencing user experience.
- Factors divided into 4 dimensions:
  - Dialogue Actions: Coherence, Recoverability, Proactiveness
  - Language Expression: Grammar, Naturalness, Appropriateness
  - Recommended Items: Effectiveness, Novelty, Diversity
  - Response Content: Semantic Relevance, Explainability, Groundness
  - Each factor scored (0–4) with rationale provided by the LLM.



# Multi-Agent Debater

- Four LLM agents (Common User, Domain Expert, Linguist, HCI Expert) simulate different perspectives.
- Debate and negotiate to synthesize a single overall score (0–100).
- Process:
  - Round-based scoring & justification.
  - Continue until consensus or max turns reached.
  - Final score = average of agents’ final scores.



# Experiments

# Experimental Settings

- **Datasets:**
  - ReDial (movie recommendations), OpenDialKG (multi-domain)
- **CRSs for Comparison**
  - BARCOR (Wang et al., 2022a), CHATCRS (Wang et al., 2023a ), KBRD (Chen et al., 2019), UniCRS (Wang et al., 2022b)
- **User Simulator** (Wang et al., 2023a)
  - Generates 3–5 turn conversations (based on GPT-4o-mini)


- Ting-Chun Wang, Shang-Yu Su, and Yun-Nung Chen. 2022a. Barcor: Towards a unified framework for conversational recommendation systems.
- Xiaolei Wang, Xinyu Tang, Xin Zhao, Jingyuan Wang, and Ji-Rong Wen. 2023a. Rethinking the evaluation for conversational recommendation in the era of large language models. In EMNLP'23
- Qibin Chen, Junyang Lin, Yichang Zhang, Ming Ding, Yukuo Cen, Hongxia Yang, and Jie Tang. 2019. Towards knowledge-based recommender dialog system. In EMNLP-IJCNLP'19
- Xiaolei Wang, Kun Zhou, Ji-Rong Wen, and Wayne Xin 815 Zhao. 2022b. Towards unified conversational recommender systems via knowledge-enhanced prompt learning. In KDD'22

CRS
(Simulated) User
0
Interaction
Recommended Items
U I really liked Iron Man 3, can you recommend a movie similar to it?
S Iron Man 2 was written by Stan Lee and Stan Lee. Stan Lee also wrote The Avengers.
U I appreciate that you mentioned "Iron Man 2" and "The Avengers," but I'm actually looking for something similar to "Iron Man 3." It had a unique storyline, and I'm particularly interested in the journey of Tony Stark, including his challenges and character development. Could you suggest something more aligned with that?
S Sure, Tony Stark had a role in the Avengers, but his role in The Avengers was also in the movie Iron Man2.
U Thank you for the information, but I'm really looking for something that focuses on Tony Stark's character development and challenges, similar to the experience I had watching "Iron Man 3."
Iron Man 2
2012
Kick-Ass
Toy Story 3
Deadpool
The Incredible Hulk
Zombieland
Zero Dark Thirty
X-Men 2
Cowboys & Aliens
War
Shark Night
The Matrix
The Dark Knight
Multi-turn Conversation Log
Evaluation Guideline
Input
Input
LLMs
Output
1 LLM as Evaluator
Coherence 2
Recoverability 4
Proactiveness 2
Appropriate-ness 4
Naturalness 3
Grammar 4
Effectiveness 3
Novelty 1
Diversity 2
Semantic Relevance 1
Explainability 2
Groundness 4
12-Factor Evaluation
Scores & Reasons
Input
Else: Input Chat Buffer, Continue Discussion
Common User
Effectiveness 3
Coherence 2
Recoverability 4
Domain Expert
Novelty 1
Diversity 2
Groundness 4
Linguist
Appropriate-ness 4
Naturalness 3
Grammar 4
HCI Researcher
Semantic Relevance 1
Explainability 2
Proactiveness 2
LLM-based Agents Discussion
Statements & Scores
Chat Buffer
If (reach consensus or reach turn limit)
[Turn] 1 [Score] 70 [Reason] ...
[Turn] 1 [Score] 20 [Reason] ...
[Turn] 1 [Score] 90 [Reason] ...
[Turn] 1 [Score] 20 [Reason] ...
...
2 Multi-Agent Debater
67
Final Score

# Results & Findings

- Factor-Level Evaluation:
  - LLM scores on 12 key factors show high correlation with human ratings.
- Overall Evaluation:
  - Multi-Agent Debate produces stable, reliable overall scores (0–100) that closely match human judgments.
  - The CoRE framework significantly outperforms traditional metrics (Recall, Persuasiveness) in reflecting true user experience.

| Factor | GPT-4o-mini | | GLM-4-Air | | LLaMa-3-8B | |
|---|---|---|---|---|---|---|
| | $r$ | $\tau_b$ | $r$ | $\tau_b$ | $r$ | $\tau_b$ |
| Coherence | $0.642^{\heartsuit}$ | $0.564^{\spadesuit}$ | $\underline{0.671}^{\heartsuit}$ | $\underline{0.585}^{\spadesuit}$ | $\mathbf{0.698}^{\heartsuit}$ | $\mathbf{0.617}^{\spadesuit}$ |
| Rec. | $\mathbf{0.624}^{\heartsuit}$ | $\mathbf{0.573}^{\spadesuit}$ | $\underline{0.609}^{\heartsuit}$ | $\underline{0.558}^{\spadesuit}$ | $0.609^{\heartsuit}$ | $0.558^{\spadesuit}$ |
| Proactiveness | $\mathbf{0.717}^{\heartsuit}$ | $\mathbf{0.655}^{\spadesuit}$ | $\underline{0.673}^{\heartsuit}$ | $\underline{0.603}^{\spadesuit}$ | $0.669^{\heartsuit}$ | $0.603^{\spadesuit}$ |
| Gra. | $\mathbf{0.723}^{\heartsuit}$ | $\mathbf{0.645}^{\spadesuit}$ | $\underline{0.626}^{\heartsuit}$ | $\underline{0.576}^{\spadesuit}$ | $0.626^{\heartsuit}$ | $0.575^{\spadesuit}$ |
| Naturalness | $\mathbf{0.743}^{\heartsuit}$ | $\mathbf{0.689}^{\spadesuit}$ | $\underline{0.679}^{\heartsuit}$ | $\underline{0.610}^{\spadesuit}$ | 0.543 | 0.492 |
| App. | $\mathbf{0.622}^{\heartsuit}$ | $\mathbf{0.612}^{\spadesuit}$ | 0.378 | 0.370 | $\underline{0.420}$ | $\underline{0.407}$ |
| Effectiveness | $\underline{0.736}^{\heartsuit}$ | $\underline{0.653}^{\spadesuit}$ | $\mathbf{0.742}^{\heartsuit}$ | $\mathbf{0.654}^{\spadesuit}$ | 0.470 | 0.411 |
| Novelty | $\underline{0.266}$ | $\underline{0.228}$ | **0.355** | **0.291** | 0.242 | 0.207 |
| Diversity | **0.424** | **0.398** | $\underline{0.211}$ | $\underline{0.191}$ | 0.058 | 0.052 |
| Sem. | $\underline{0.604}^{\heartsuit}$ | $\mathbf{0.554}^{\spadesuit}$ | $\mathbf{0.616}^{\heartsuit}$ | $\underline{0.542}^{\spadesuit}$ | 0.578 | $0.524^{\spadesuit}$ |
| Exp. | $\underline{0.729}^{\heartsuit}$ | $\underline{0.662}^{\spadesuit}$ | $0.689^{\heartsuit}$ | $0.611^{\spadesuit}$ | $\mathbf{0.755}^{\heartsuit}$ | $\mathbf{0.692}^{\spadesuit}$ |
| Groundness | $\underline{0.648}^{\heartsuit}$ | $\underline{0.581}^{\spadesuit}$ | $\mathbf{0.750}^{\heartsuit}$ | $\mathbf{0.680}^{\spadesuit}$ | 0.563 | $0.516^{\spadesuit}$ |

# Results & Findings

- Factor-Level Evaluation:
  - LLM scores on 12 key factors show high correlation with human ratings.
- Overall Evaluation:
  - Multi-Agent Debate produces stable, reliable overall scores (0–100) that closely match human judgments.
  - The CoRE framework significantly outperforms traditional metrics (Recall, Persuasiveness) in reflecting true user experience.

| **Factor** | **GPT-4o-mini** | | **GLM-4-Air** | | **LLaMa-3-8B** | |
|---|---|---|---|---|---|---|
| | $r$ | $\tau_b$ | $r$ | $\tau_b$ | $r$ | $\tau_b$ |
| Coherence | 0.642$^{\heartsuit}$ | 0.564$^{\spadesuit}$ | 0.671$^{\heartsuit}$ | 0.585$^{\spadesuit}$ | **0.698**$^{\heartsuit}$ | **0.617**$^{\spadesuit}$ |
| Rec. | **0.624**$^{\heartsuit}$ | **0.573**$^{\spadesuit}$ | 0.609$^{\heartsuit}$ | 0.558$^{\spadesuit}$ | 0.609$^{\heartsuit}$ | 0.558$^{\spadesuit}$ |
| Proactiveness | **0.717**$^{\heartsuit}$ | **0.655**$^{\spadesuit}$ | 0.673$^{\heartsuit}$ | 0.603$^{\spadesuit}$ | 0.669$^{\heartsuit}$ | 0.603$^{\spadesuit}$ |
| Gra. | **0.723**$^{\heartsuit}$ | **0.645**$^{\spadesuit}$ | 0.626$^{\heartsuit}$ | 0.576$^{\spadesuit}$ | 0.626$^{\heartsuit}$ | 0.575$^{\spadesuit}$ |
| Naturalness | **0.743**$^{\heartsuit}$ | **0.689**$^{\spadesuit}$ | 0.679$^{\heartsuit}$ | 0.610$^{\spadesuit}$ | 0.543 | 0.492 |
| App. | **0.622**$^{\heartsuit}$ | **0.612**$^{\spadesuit}$ | 0.378 | 0.370 | 0.420 | 0.407 |
| Effectiveness | 0.736$^{\heartsuit}$ | 0.653$^{\spadesuit}$ | **0.742**$^{\heartsuit}$ | **0.654**$^{\spadesuit}$ | 0.470 | 0.411 |
| Novelty | 0.266 | 0.228 | **0.355** | **0.291** | 0.242 | 0.207 |
| Diversity | **0.424** | **0.398** | 0.211 | 0.191 | 0.058 | 0.052 |
| Sem. | 0.604$^{\heartsuit}$ | **0.554**$^{\spadesuit}$ | **0.616**$^{\heartsuit}$ | 0.542$^{\spadesuit}$ | 0.578 | 0.524$^{\spadesuit}$ |
| Exp. | 0.729$^{\heartsuit}$ | 0.662$^{\spadesuit}$ | 0.689$^{\heartsuit}$ | 0.611$^{\spadesuit}$ | **0.755**$^{\heartsuit}$ | **0.692**$^{\spadesuit}$ |
| Groundness | 0.648$^{\heartsuit}$ | 0.581$^{\spadesuit}$ | **0.750**$^{\heartsuit}$ | **0.680**$^{\spadesuit}$ | 0.563 | 0.516$^{\spadesuit}$ |

# Results & Findings

- System Performance:
  - CHATCRS stands out as the best performer. Other systems (BARCOR, KBRD, UniCRS) exhibit weaknesses in semantic relevance and explainability.

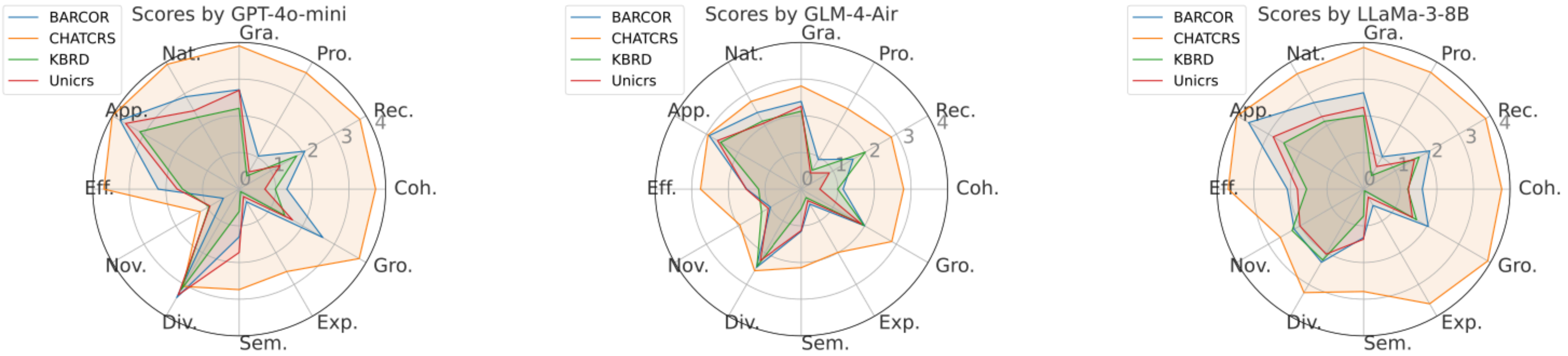



Figure 3: The scores of four systems on 12 factors are provided by GPT-4o-mini, LLaMa-3-8B, and GLM-4-Air. Coh. represents Coherence; Rec. represents Recoverability; Pro. represents Proactiveness; Gra. represents Grammatical Correctness; Nat. represents Naturalnsee; App. represents Appropriateness; Eff. represents Effectiveness; Nov. represents Novelty; Div. represents Diversity; Sem. represents Semantic Relevance; Exp. represents Explainability: Gro represents Groundness.

# Results & Findings

- System Performance:
  - CHATCRS stands out as the best performer. Other systems (BARCOR, KBRD, UniCRS) exhibit weaknesses in semantic relevance and explainability.

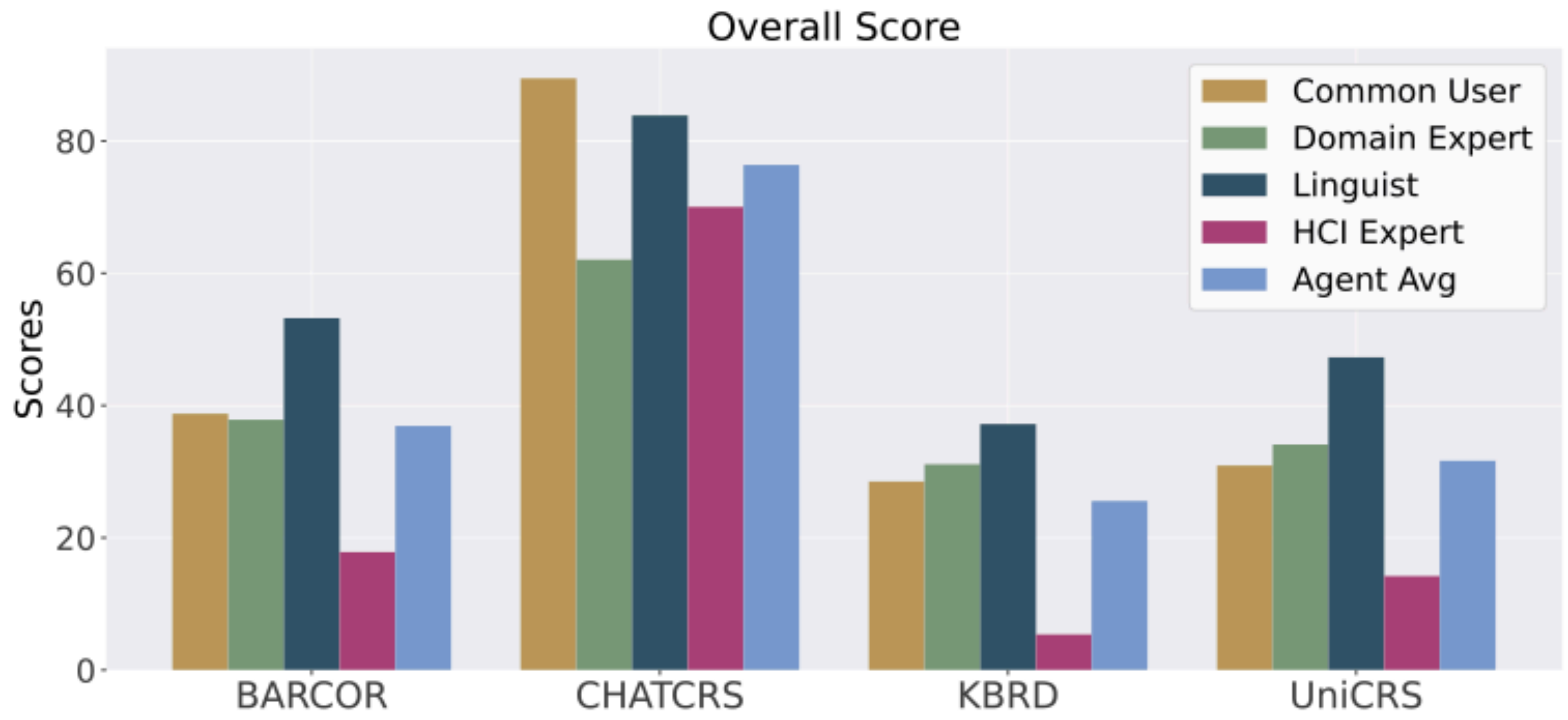



Figure 4: The scores of BARCOR, CHATCRS, KBRD and UniCRS after multi-agent discussion.

# Conclusion

# Conclusions

- Review:
  - Presented our user-centric CoRE framework for evaluating conversational recommender systems (CRS).
  - Demonstrated how LLMs can score 12 key factors and how a multi-agent debate synthesizes these into an overall score.
  - We benchmarked 4 CRSs on 2 datasets and collected real human data for validation.
- Main Findings:
  - High Agreement: CoRE’s evaluations closely match human ratings.
  - Effective LLM Evaluation: LLMs accurately capture user experience in CRS.
  - System Weaknesses: Neural network-based CRS show limitations in semantic relevance, explainability, and proactiveness.

*Thank you for your time.*