IJCAI ECAI BREMEN 2026

The ACM Conference Series on Recommender Systems

Technical University of Munich

TU WIEN

CD Lab RecSys

Politecnico di Bari

A Tutorial on Recent Advances in Generative Conversational Recommender Systems

Ahmadou Wagne[1]

Thomas E. Kolb[1]

Ashmi Banerjee[2]

Fatemeh Nazary[3]

Julia Neidhardt[1]

Yashar Deldjoo[3]

[1] TU Wien, Austria
[2] TU Munich, Germany
[3] Polytechnic University of Bari, Italy

Also Take a look at our Gen-RecSys Work

Conversational Recommender Systems Using Generative Models (Gen-CRS): A Literature Review

AHMADOU WAGNE, TU Wien, Vienna, Austria
THOMAS KOLB, TU Wien, Vienna, Austria
ASHMI BANERJEE, Technical University of Munich, Munich, Germany
FATEMEH NAZARY, Polytechnic University of Bari, Bari, Italy
JULIA NEIDHARDT, TU Wien, Vienna, Austria
YASHAR DELDJOO, Department of Electrical Engineering and Information Technology, Polytechnic University of Bari, Bari, Italy

Generative models are profoundly impacting how conversational recommender systems (CRS) are conceptualized, designed, and deployed in practice, enabling mixed-initiative, context-aware, and tool-augmented recommendation workflows that go beyond rigid traditional pipelines. This survey provides a comprehensive review of Generative Conversational Recommender Systems (Gen-CRS) from 2018–2025. We organize the field into three analytical layers: (i) architectural and methodological foundations, covering unified, modular, and agentic system designs, strategies for adapting foundation models, and methods for recommendation and response generation, (ii) knowledge and data foundations, detailing how item-level information (structured catalogs, knowledge graphs, unstructured and multi-modal content), user-level signals (short-term preferences, long-term profiles, and persona representations), and dialogue corpora and logs are integrated into generative workflows, and (iii) evaluation methodologies, structured around what is evaluated (output types), which quality dimensions are measured (e.g., task effectiveness, conversational quality, efficiency, user trust, and ethical concerns), how evaluation is conducted (offline metrics, simulation, user studies, online testing), who evaluates (humans, automated metrics, LLM-based judges), and which stakeholders are considered (consumers, item providers, platforms and society). Our analysis highlights both the capabilities and the risks introduced by generative components, including challenges in grounding, catalog fidelity, controllability, safety, and bias. We argue for responsible and transparent system design, emphasizing validated knowledge integration and evaluation protocols that reflect the full complexity of conversational interactions and system behaviors, and we outline open research challenges that must be addressed to develop reliable and trustworthy Gen-CRS. This survey is also informed by, and partly the result of, tutorial feedback we collected at ACM RecSys 2025 in Prague [42].[1]

CCS Concepts: • **Information systems** → **Recommender systems**; • **Human-centered computing** → *Interactive systems and tools*; • **Computing methodologies** → Artificial Intelligence.

Additional Key Words and Phrases: Generative AI, LLM, Conversational Recommender System, Dialogue System

[1] https://www.youtube.com/watch?v=oWLr23-wRRE

Authors' Contact Information: Ahmadou Wagne, TU Wien, Vienna, Austria; e-mail: ahmadou.wagne@tuwien.ac.at; Thomas Kolb, TU Wien, Vienna, Austria; e-mail: thomas.kolb@tuwien.ac.at; Ashmi Banerjee, Technical University of Munich, Munich, Germany; e-mail: ashmi.banerjee@tum.de; Fatemeh Nazary, Polytechnic University of Bari, Bari, BA, Italy; e-mail: fatemeh.nazary@poliba.it; Julia Neidhardt, TU Wien, Vienna, Vienna, Austria; e-mail: julia.neidhardt@tuwien.ac.at; Yashar Deldjoo, Department of Electrical Engineering and Information Technology, Polytechnic University of Bari, Bari, Italy; e-mail: deldjooy@acm.org.

ACM 2770-6699/2026/7-ART
https://doi.org/10.1145/3828551

SURVEY
July 2026
JUST ACCEPTED

Conversational Recommender Systems Using Generative Models (Gen-CRS): A Literature Review

Ahmadou Wagne, Thomas Kolb, Ashmi Banerjee, Fatemeh Nazary, Julia Neidhardt, Yashar Deldjoo

https://doi.org/10.1145/3828551

Generative models are profoundly impacting how conversational recommender systems (CRS) are conceptualized, designed, and deployed in practice, enabling mixed-initiative, context-aware, and tool-augmented recommendation workflows that go beyond rigid ...

Link: https://dl.acm.org/doi/10.1145/3828551

Agenda

- Introduction
- Core Systems & Components
- Foundation Model Integration & Generative Paradigms
- Knowledge and Data Foundation

Coffee Break

- Simulation & Evaluation
- Open Challenges & Future Directions

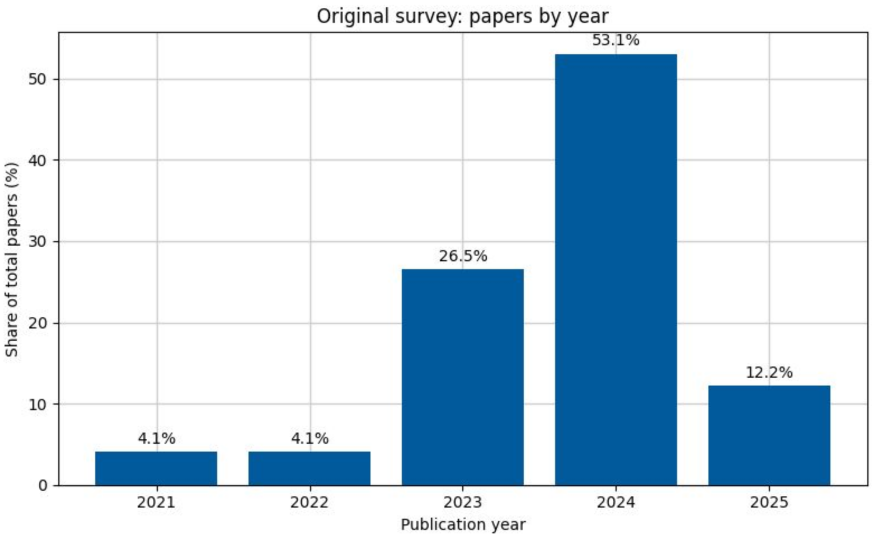
Agenda

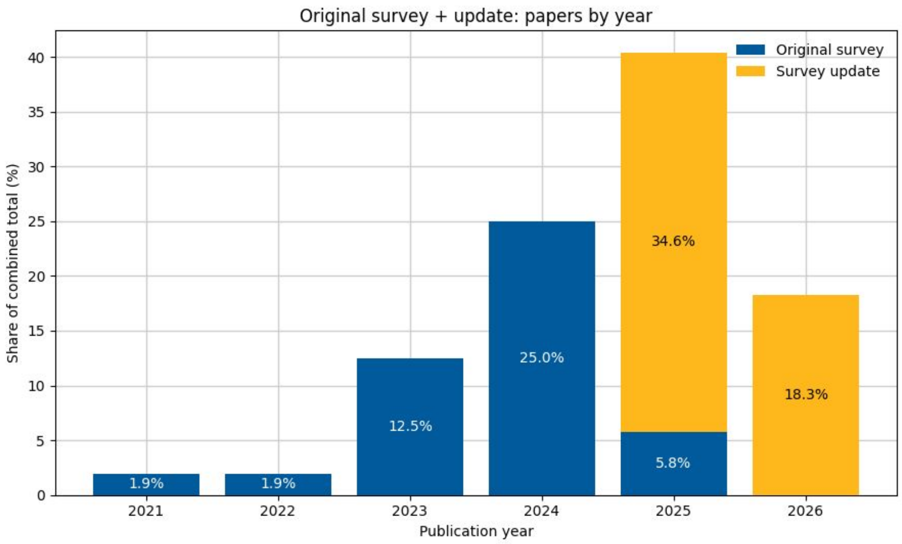
- **Introduction**
- Core Systems & Components
- Foundation Model Integration & Generative Paradigms
- Knowledge and Data Foundation
- Simulation & Evaluation
- Open Challenges & Future Directions

Julia Neidhardt

Introduction

Rapid Expansion of the Literature





Updated resource available: https://doi.org/10.48436/gc46a-qmw81 (TUW Research Data Repo.)

Conversational recommendation

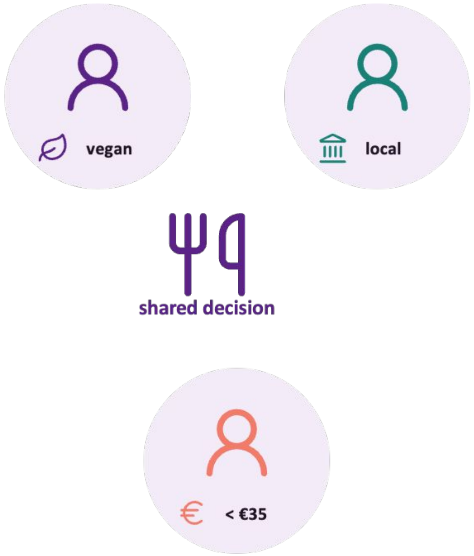
powered by Generative AI (Gen-CRS)

Scenario

After the conference: where should the three of us have dinner in Bremen?

One shared decision, with preferences and constraints distributed across three people.

- near the city centre
- table around 19:30 (7:30 pm)



A suitable recommendation must account for the group's preferences, maintain the conversational context, and use current venue data.

Problem

Relevant information is distributed across sources

Group preferences

dietary requirements
budget
atmosphere
acceptable travel time

Item information

menus and allergens
location
price level
noise and accessibility

Dynamic context

opening hours
table availability
travel conditions
changes within the group

The user must otherwise integrate these sources and resolve conflicting constraints manually.

Dialogue

Clarification supports preference and constraint elicitation

USER We are three people looking for dinner in Bremen after the conference.

SYSTEM Do you have any dietary requirements, budget limits, or preferences regarding the type of restaurant?

USER Yes. One person needs a vegan option, one would prefer something local, and one wants to spend less than €35.

SYSTEM Is the €35 limit per person? And are vegan and budget strict requirements, while local character is a preference?

Multi-turn dialogue transforms an underspecified request into an explicit decision model.

Before and Now

From fragmented coordination to a shared conversational state

Before: The user coordinates the decision

- Collects the group's preferences
- Searches across multiple sources
- Compares menus, prices and locations
- Checks availability and revises the shortlist

vegan local < €35

Gen-CRS: One conversation maintains context

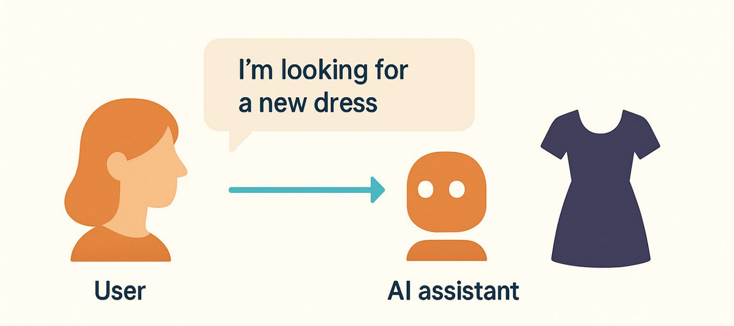
- Captures the group's preferences
- Clarifies requirements and preferences
- Retrieves current venue information
- Explains the trade-offs behind the shortlist

vegan · local · < €35 · city centre · 19:30

The system reduces manual coordination while keeping the recommendation grounded in current information.

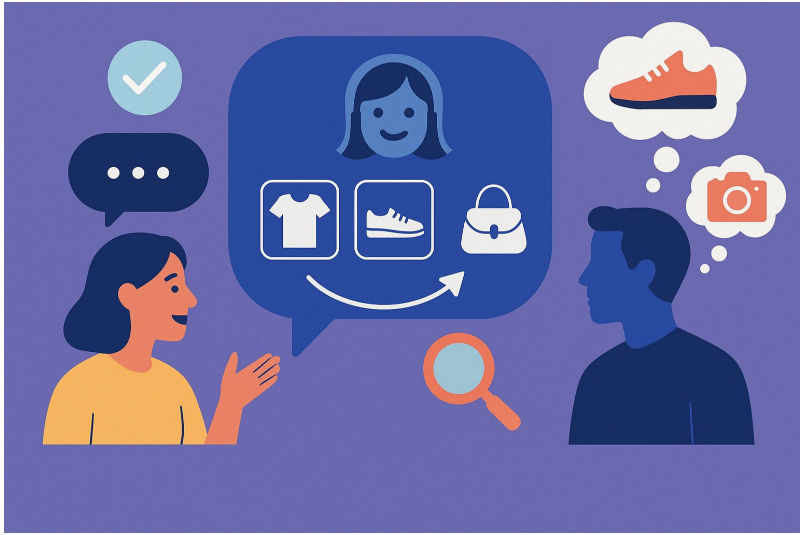
What is a Conversational Recommender System (CRS)?

A Conversational Recommender System (CRS) is an interactive system that supports users in achieving **recommendation-related goals** through **multi-turn dialogue**.



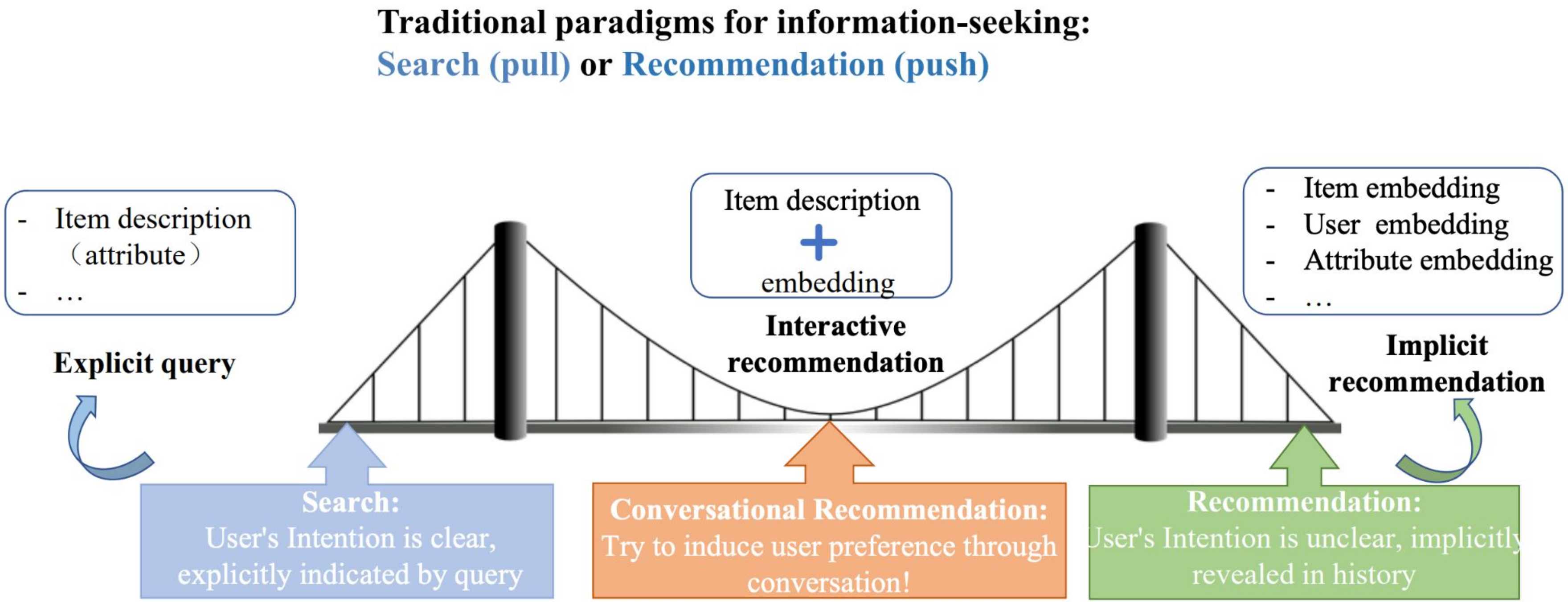
Advantages

- Assist decision-making and information-seeking
- Support product space exploration e.g., discover unexpected but relevant items
- Elicit users' nuanced or hidden preferences



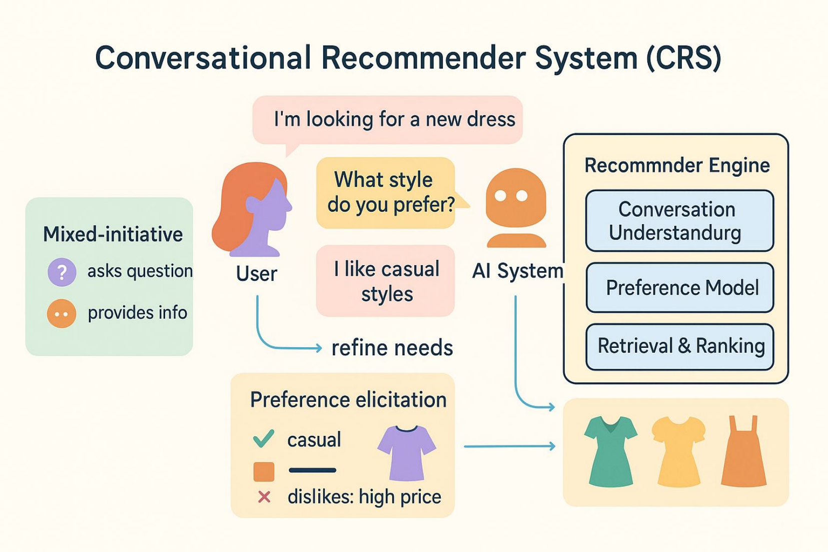
Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. A Survey on Conversational Recommender Systems. ACM Computing Surveys, vol. 54(5), Article 105, pages 1-36, 2021.

Search, Interactive Recommendation, RecSys



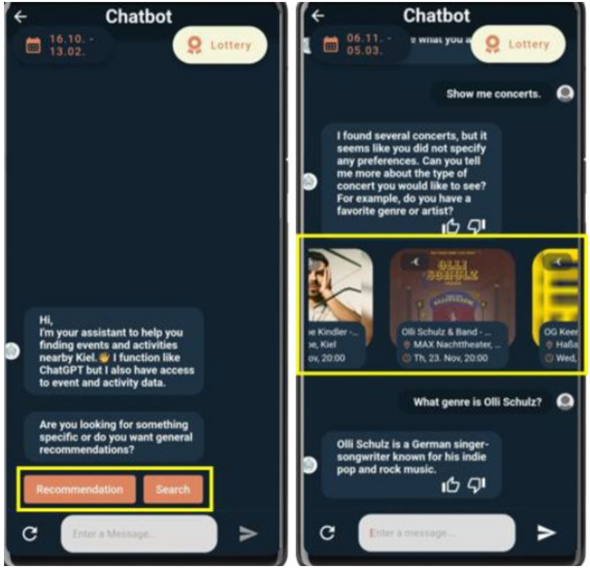
Key Characteristics

- **Multi turn interaction** – conversation persists across rounds.
- **Mixed initiative** – system and user alternate turns and roles.
- **Natural language** – voice or text as the primary interface.
- **Dynamic preference elicitation** – ask, refine and adapt
- **Grounding and tool use** – integration of external knowledge sources and live data



Conversational AI - High Level Categorization

1. Goal-driven (task-oriented)



Goal-driven (task-oriented): aiming to assist users to complete specific tasks

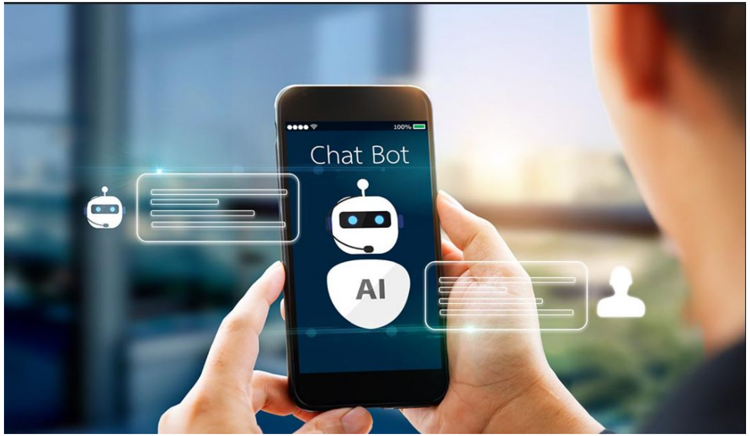
- Conversational information access: tasks with underlying information need, which can be satisfied through a conversation
- Includes task of **search**, **recommendation**, and QA (boundaries often blurred)

Courtesy of image: EventChat

Conversational AI - High Level Categorization

2. Non Goal-driven (chatbots)

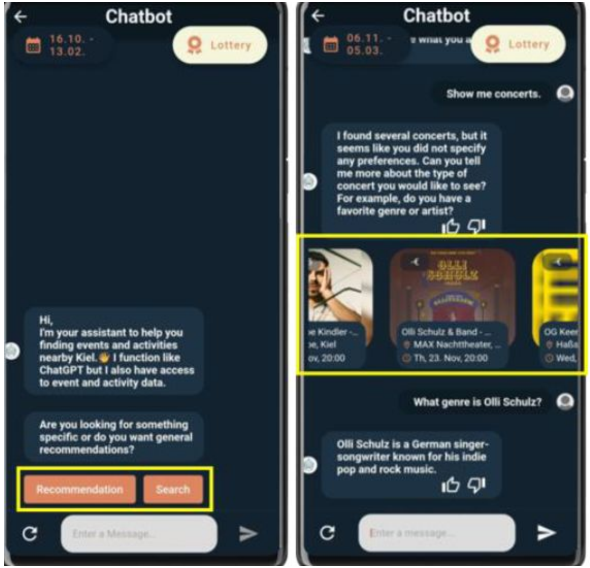
Non Goal-driven (chatbots): aiming to carry out an extended conversation **("chit-chat")** usually with the purpose of **entertainment**.



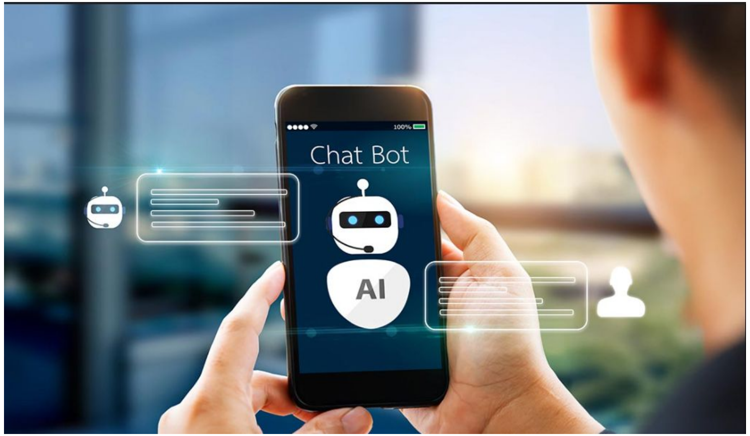
Courtesy of image: [What Is an AI Chatbot, and How Do They Work? - Caltech]

Conversational AI - High Level Categorization

Goal-driven (task-oriented)



Non Goal-driven (chatbots)



Our focus

Overview of Conversational AI

CRS (Conversational RecSys)

CQA (Conversational Q&A)

CS (Conversational Search)

Social Chatbot

Gen-CRS

CRS vs Conversational Search

- **Common**: rank or surface relevant options in multi-turn interaction
- **Key difference**: CRS builds a user model & personalised; search focuses on query understanding & retrieval

CRS vs Conversational QA

- **Common**: dialogue to resolve information needs
- **Key difference**: QA answers factual questions; CRS elicits preferences and recommends subjective items

CRS vs Social Chatbot

- **Common**: free-form conversational exchange
- **Key difference**: CRS is goal-oriented (task success); social chat aims for open-ended chit-chat & engagement

CRS vs Gen-CRS (LLM-powered)

- **Common**: goal-oriented recommendation via dialogue
- **Key difference**: Traditional CRS uses rules/templates & separate rankers; Gen-CRS uses LLMs for free-form NLG, mixed-initiative, and tool use (with grounding to reduce hallucination)

Evolution of Conversational RS

2017–2018

Neural CRS Emergence

First neural conversational recommenders appear; separate dialogue and rec modules.

2018–2021

Evolution

Research integrates knowledge graphs & critique-based recommendation.

2022–2023

LLM Revolution

GPT-3 & ChatGPT enable generative dialogue & zero-shot recommendation.

2023–2025

GenCRS Explosion

Unified, modular & agentic architectures flourish with LLMs & tools.

Evolution of Conversational RS

2017–2018
Neural CRS Emergence
First neural conversational recommenders appear; separate dialogue and rec modules.

2018–2021
Evolution
Research integrates knowledge graphs & critique-based recommendation.

2022–2023
LLM Revolution
GPT-3 & ChatGPT enable generative dialogue & zero-shot recommendation.

2023–2025
GenCRS Explosion
Unified, modular & agentic architectures flourish with LLMs & tools.

Where the field stands (2025–2026)

- Tool invocation and integration of external APIs
- Multi-agent architectures and task decomposition
- Persistent memory and long-term personalisation
- User simulators and LLM-based evaluation

Some traditional approaches…

Traditional approaches rarely involved "conversation" as we might normally think of it:

1. **Thompson et al., 2004** (query refinement): Elicits users' preferences and constraints with regard to **item attributes**;

Example of a user model

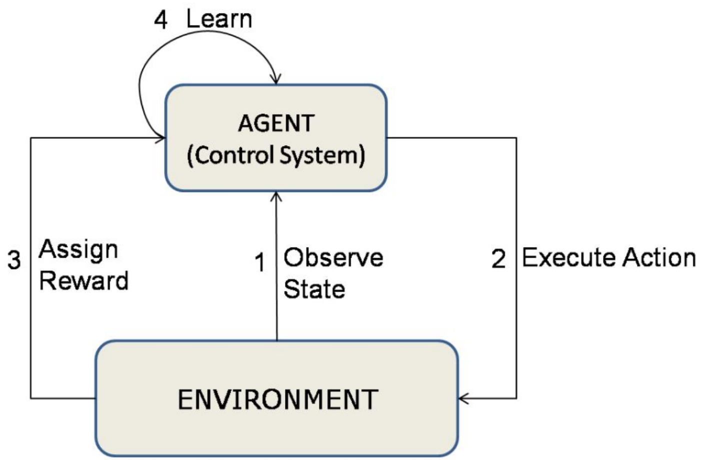
User Name		Homer					
Attributes	w_i	Values and probabilities					
Cuisine	0.4	Italian	French	Turkish	Chinese	German	English
		0.35	0.2	0.25	0.1	0.1	0.0
Price Range	0.2	one	two	three	four	five	
		0.2	0.3	0.3	0.1	0.1	
...	...	...					
Parking	0.1	Valet		Street		Lot	
		0.5		0.4		0.1	
Item Nbr.		0815	5372	7638	...		6399
Accept/Present		23 / 25	10 / 19	33 / 36	...		12 / 23

Thompson, C. A., Goker, M. H., & Langley, P. (2004). A personalized system for conversational recommendations. Journal of Artificial Intelligence Research, 21, 393-428.

Some traditional approaches…

Traditional approaches rarely involved "conversation" as we might normally think of it:

2. Mahmood and Ricci, 2009 (reinforcement learning): Queries users about **recommendation attributes** during each round; learns a **policy** to choose queries to efficiently yield a desirable recommendation



Mahmood, T., & Ricci, F. (2009). Learning Adapted Interaction Strategies in Conversational Recommender Systems. *International Doctorate School in Information and Communication Technologies, PhD Dissertation, 226*.

Some traditional approaches…

Traditional approaches rarely involved "conversation" as we might normally think of it:

3. Christakopoulou et al., 2016 (iterative recommendation): Collects feedback about recommended items in order to **iteratively learn user preferences**; explores various query strategies to elicit preferences quickly

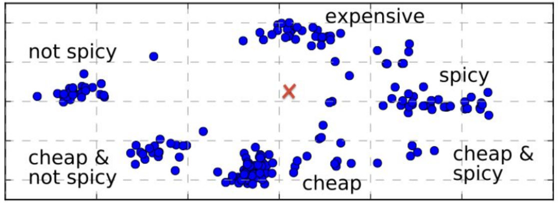


Table 3: Question selection strategies evaluated.

Greedy: $j^* = \arg\max_j y_{ij}$
A trivial *exploit*-only strategy: Select the item with highest estimated affinity mean.

Random: $j^* = \mathrm{random}(1, N)$
A trivial *explore*-only strategy.

Maximum Variance (MV): $j^* = \arg\max_j \epsilon_{ij}$
A *explore*-only strategy, variance reduction strategy: Select the item with the highest noisy affinity variance.

Maximum Item Trait (MaxT): $j^* = \arg\max_j \|\mathbf{v}_j\|_2$
Select the item whose trait vector $\mathbf{v}_j$ contains the most information, namely has highest L2 norm $\|\mathbf{v}_j\|_2 = \sqrt{v_{j1}^2 + v_{j2}^2 + \ldots + v_{jd}^2}$.

Minimum Item Trait (MinT): $j^* = \arg\min_j \|\mathbf{v}_j\|_2$
Select the item with trait vector with least information.

Upper Confidence (UCB): $j^* = \arg\max_j y_{ij} + \epsilon_{ij}$
Based on UCB1 [3]: Pick the item with the highest upper confidence bound, namely mean plus variance (95% CI)

Thompson Sampling (TS) [5]: $j^* = \arg\max_j \hat{y}_{ij}$
For each item, sample the noisy affinity from the posterior. Select item with the maximum sampled value.

Christakopoulou, K., Radlinski, F., & Hofmann, K. (2016, August). Towards conversational recommender systems. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 815-824).

Actual conversation…

Li et al. (2018) sought approaches more closely matching “free-form” conversation. Roughly:

- Dialogs (around 10k) are constructed by crowd workers, who assume roles of a *recommender* or *seeker*;
- Conversations between the recommender and the seeker are tagged in terms of the movies mentioned, as well as explicit feedback (has the seeker seen the movies mentioned and did they like them);
- Train a dialog generation model that can fulfil the role of the recommender;
- Preferences can then be estimated and the output controlled to reference specific movies

Table 5: Sample conversation from validation set

SEEKER:	hi !
HUMAN:	hi ! how is your morning going ?
HRED:	hi , what kind of movies do you like
OURS:	hello ! what kind of movie are you looking for ?
SEEKER:	great , thanks ! i was thinking of starting the day off with a great comedy , any suggestions ?
HUMAN:	i love the hangover it always makes me laugh . and i bet lots of people are hungover this morning ! airplane ! is another classic of mine .
HRED:	that 's a good one ! have you seen the lego movie ? it 's a great movie !
OURS:	i have n't seen that one ! what kind of movies do you like ? i can help you with that .
SEEKER:	i 've seen the hangover it 's really good ! i thought about watching knocked up have you seen it ?
HUMAN:	i enjoyed that movie . i enjoyed spy it 's a well done parody .
HRED:	it 's about a man who gets revenge on a cyborg and a hockey team . it 's really good . it 's really good . i also like the girl on the train
OURS:	i have n't seen that one . i will have to check it out . i hope i was able to help you find a good one ! have a great day !
SEEKER:	thanks ! i 'll check it out !
HUMAN:	sure ! have a great day !
HRED:	you 're welcome ! have a great day !
OURS:	have you seen revenge of the nerds ? it 's really good . it 's a classic .
SEEKER:	you too .

(from ReDial)

Actual conversation…

Li et al. (2018)'s approach has a number of virtues:

- Actually looks (more or less) like "real" conversation, especially compared to approaches that came before
- Contributes a (now widely used) benchmark dataset for training and evaluation
- Elegant / principled in terms of how the model is trained and the objective it's trained for (i.e., reach a goal movie in the fewest possible number of steps)

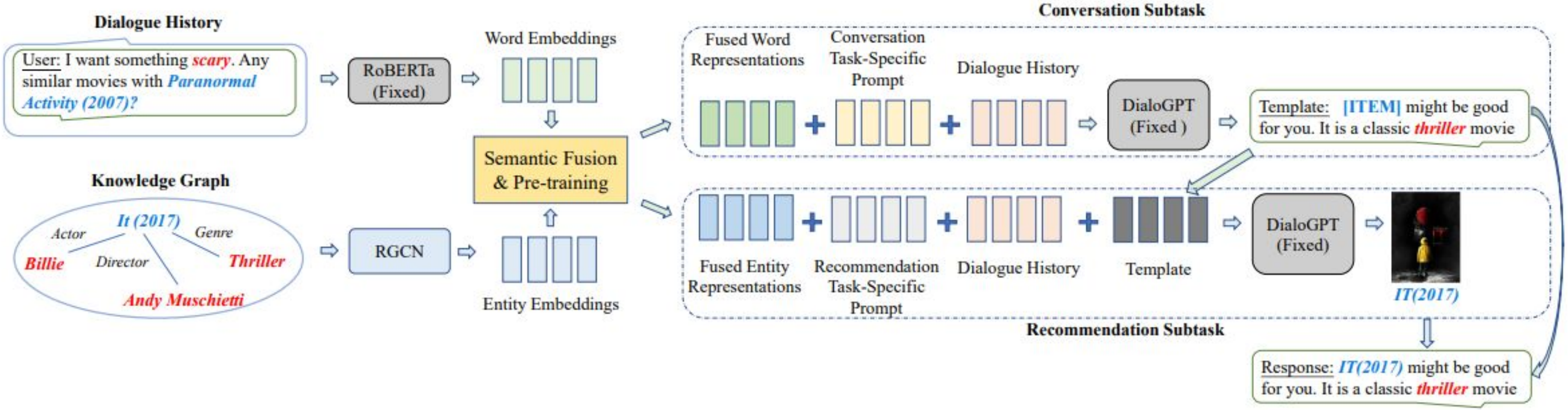
Actual conversation…

Though it also has some **limitations**

- Conversations aren’t particularly “real”: the users aren’t actually seeking some item, but play a synthetic game in which they are told which item to seek
- It’s unclear to what extent the data collection effort could be applied in other settings, in particular ones not based on “general knowledge” (i.e., for which crowd workers would struggle to engage in synthetic conversations)
- Even within movies, it’s hard to tell how closely conversations in ReDial (or similar efforts) represent “organic” conversations

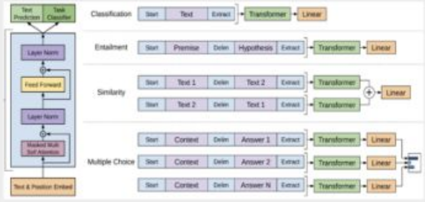
“LM+RecSys” approaches (UniCRS; Wang et al., 2022)

(Fairly) recent attempts incorporate knowledge grounding, and arguably (among a few others) represented the pre-LLM state-of-the-art

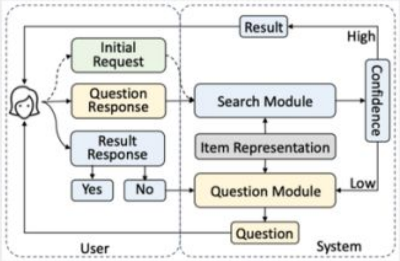


(UniCRS)

End-to-End Architecture
E.g., Sequence-to-Sequence models, Generative Language Models (GPT)

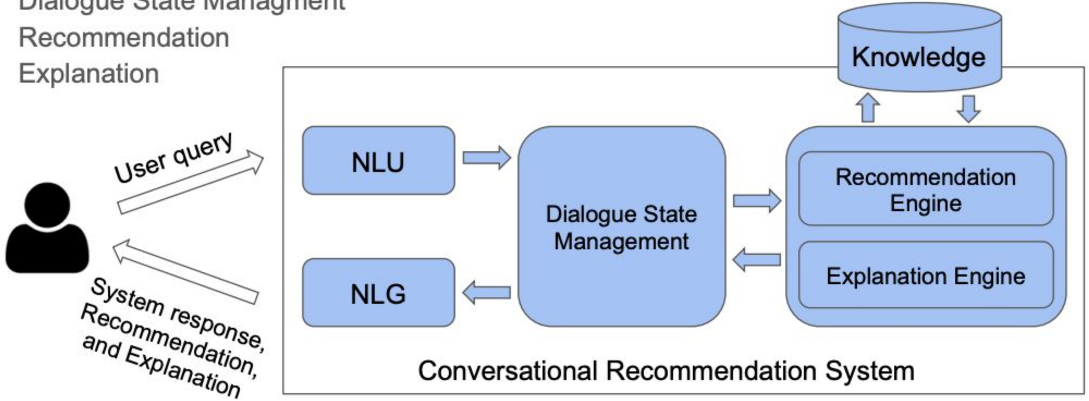


Modularized Architecture
e.g., Conversational Agent as Linked Functional Modules

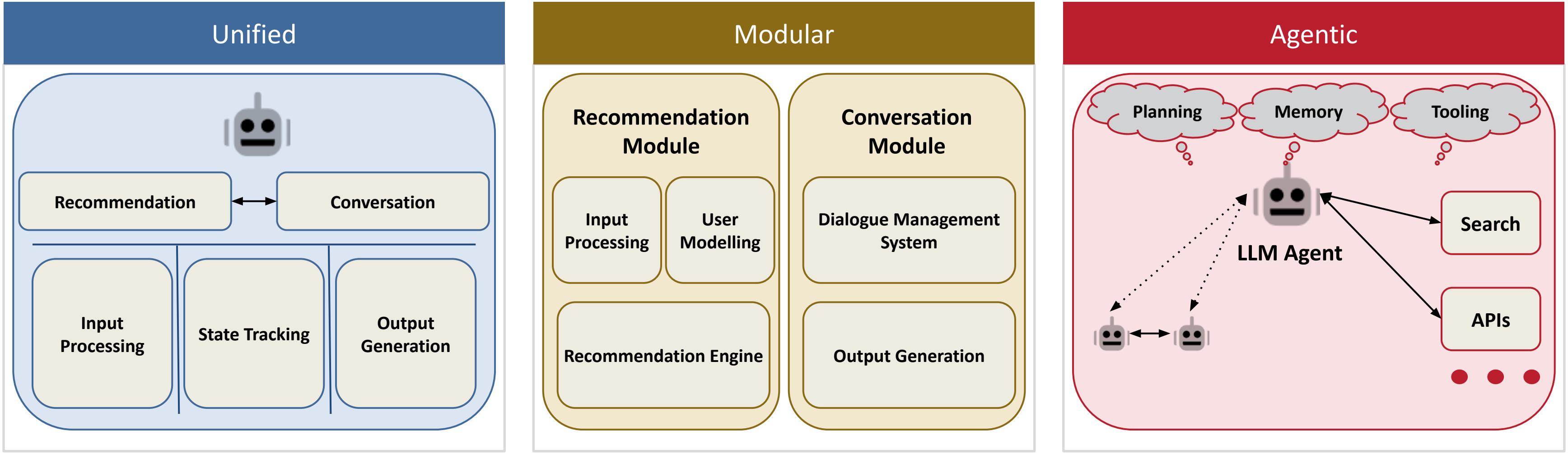


Data-Flow Architecture
E.g., Dialogue State as Dataflow Graphs (DataFlow)

- Four Major Modules
  - Natural Language Understanding/Generation
  - Dialogue State Managment
  - Recommendation
  - Explanation



Transition to Gen-CRS



Ahmadou Wagne

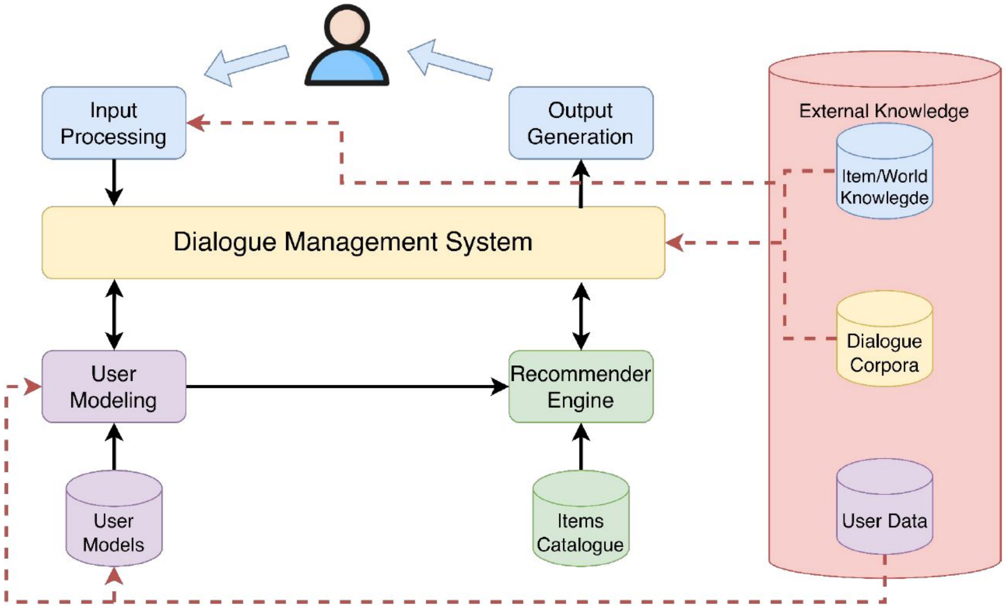
Core Systems & Components

Agenda

- Introduction
- **Core Systems & Components**
    - **System Architecture**
    - Dialogue Initiative
    - Recommendation Generation
    - Response Generation
- Foundation Model Integration & Generative Paradigms
- Knowledge and Data Foundation
- Simulation & Evaluation
- Open Challenges & Future Directions

System Architecture

General CRS Architecture

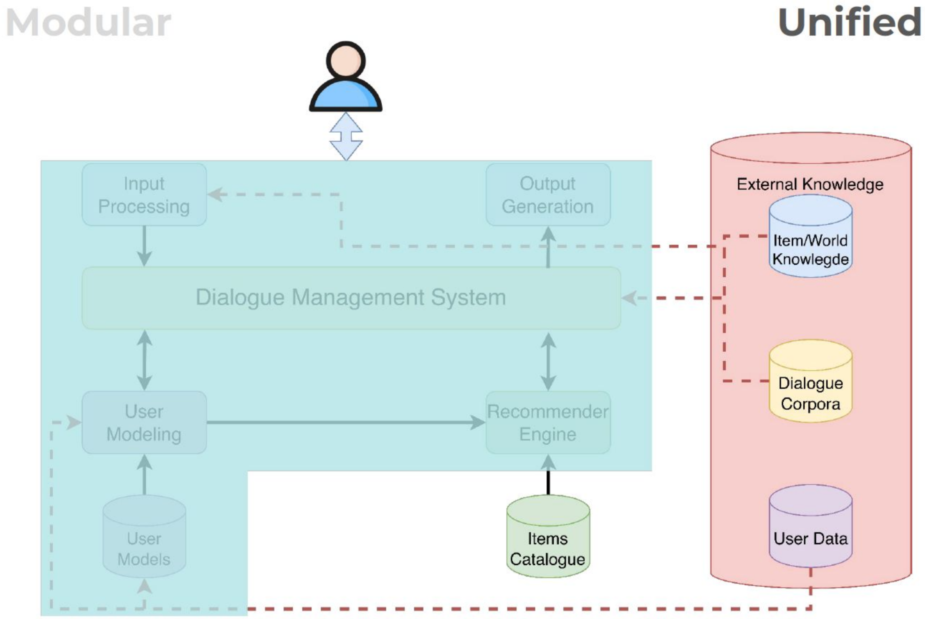


D. Jannach, A. Manzoor, W. Cai, and L. Chen, "A Survey on Conversational Recommender Systems," *ACM Comput. Surv.*, vol. 54, no. 5, p. 105:1-105:36, May 2021, doi: 10.1145/3453154

G. Faggioli, N. Ferro, and S. Merlo. 2025, "Conversational Information Retrieval and Recommender Systems," In Advances in Information Retrieval: 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6–10, 2025, 280–290. https://doi.org/10.1007/978-3-031-88720-8_44

System Architecture

General CRS Architecture

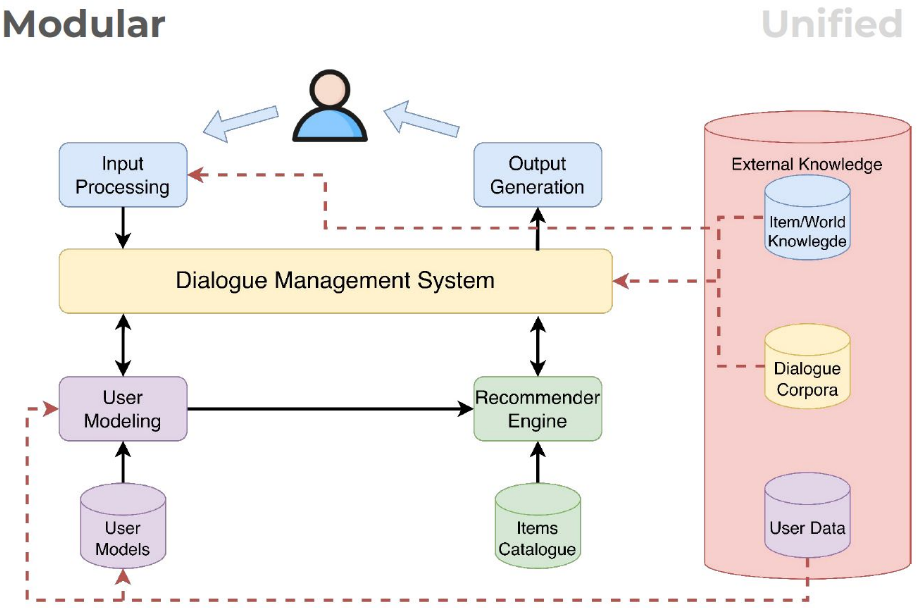


D. Jannach, A. Manzoor, W. Cai, and L. Chen, "A Survey on Conversational Recommender Systems," *ACM Comput. Surv.*, vol. 54, no. 5, p. 105:1-105:36, May 2021, doi: 10.1145/3453154

G. Faggioli, N. Ferro, and S. Merlo. 2025, "Conversational Information Retrieval and Recommender Systems," In Advances in Information Retrieval: 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6–10, 2025, 280–290. https://doi.org/10.1007/978-3-031-88720-8_44

System Architecture

General CRS Architecture



D. Jannach, A. Manzoor, W. Cai, and L. Chen, "A Survey on Conversational Recommender Systems," *ACM Comput. Surv.*, vol. 54, no. 5, p. 105:1-105:36, May 2021, doi: 10.1145/3453154

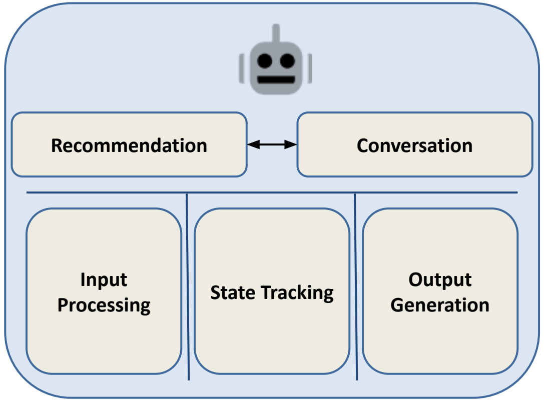
G. Faggioli, N. Ferro, and S. Merlo. 2025, "Conversational Information Retrieval and Recommender Systems," In Advances in Information Retrieval: 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6–10, 2025, 280–290. https://doi.org/10.1007/978-3-031-88720-8_44

Key Questions

- What are different architectural paradigms in Gen-CRS?
- Which role do generative models play?
- How are multi-turn dialogues managed?
- What methods are used to generate recommendations and responses?
- What are the trade-offs between control and fluency, and where does each paradigm sit?

System Architecture

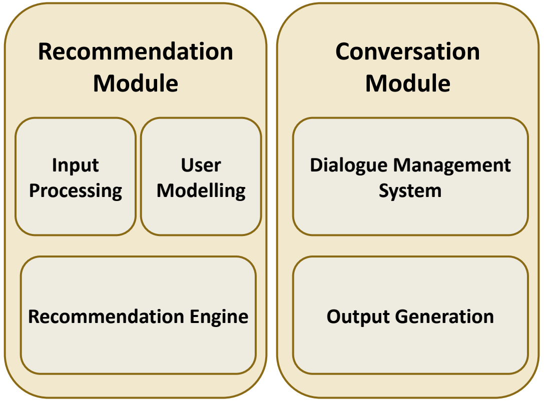
Unified



- Single model -> multiple subtasks
- Strong content/context knowledge → rich semantics & explanations
- Commonly weak alignment with collaborative signals

Highest simplicity

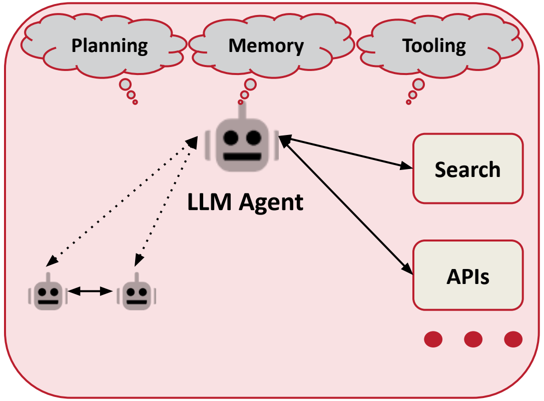
Modular



- LLM focuses on NLP and NLG subtasks
- Expert models for recommendation
- Best for control and grounding

Highest control

Agentic

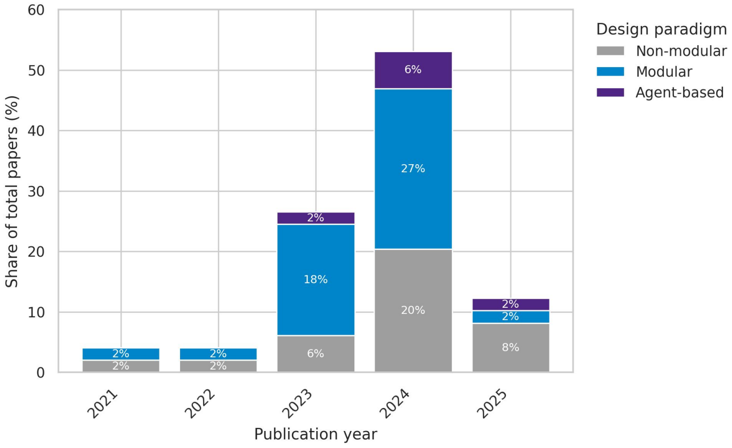


- Task decomposition orchestrated by central planning agent
- Dynamic integration of various (external) tools
- Multi-step planning & proactivity

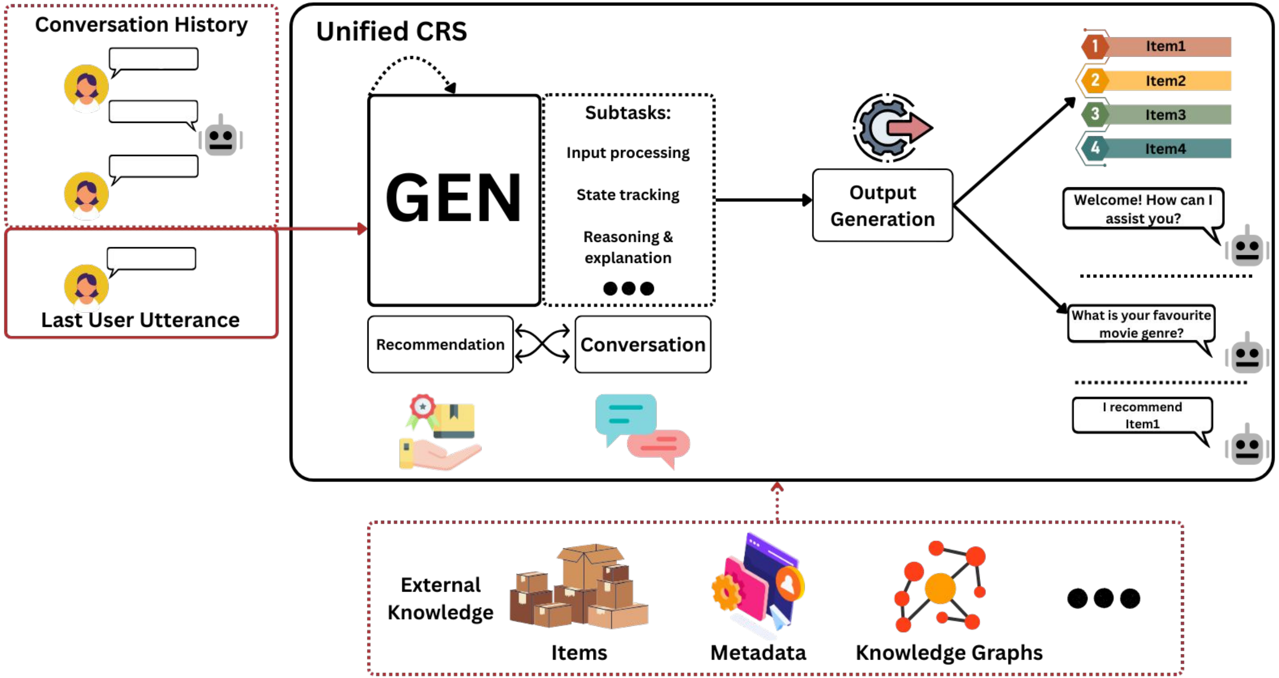
Highest flexibility

System Types

- Gen-CRS as **emerging** topic over the past years
- **Modular** systems **dominate** the research about Gen-CRS
- **Agentic** systems gained traction, while unified systems mostly feature **early investigations** of LLMs as CRS
- Most Gen-CRS are **standalone** applications, while agentic systems are often integrated in existing platforms

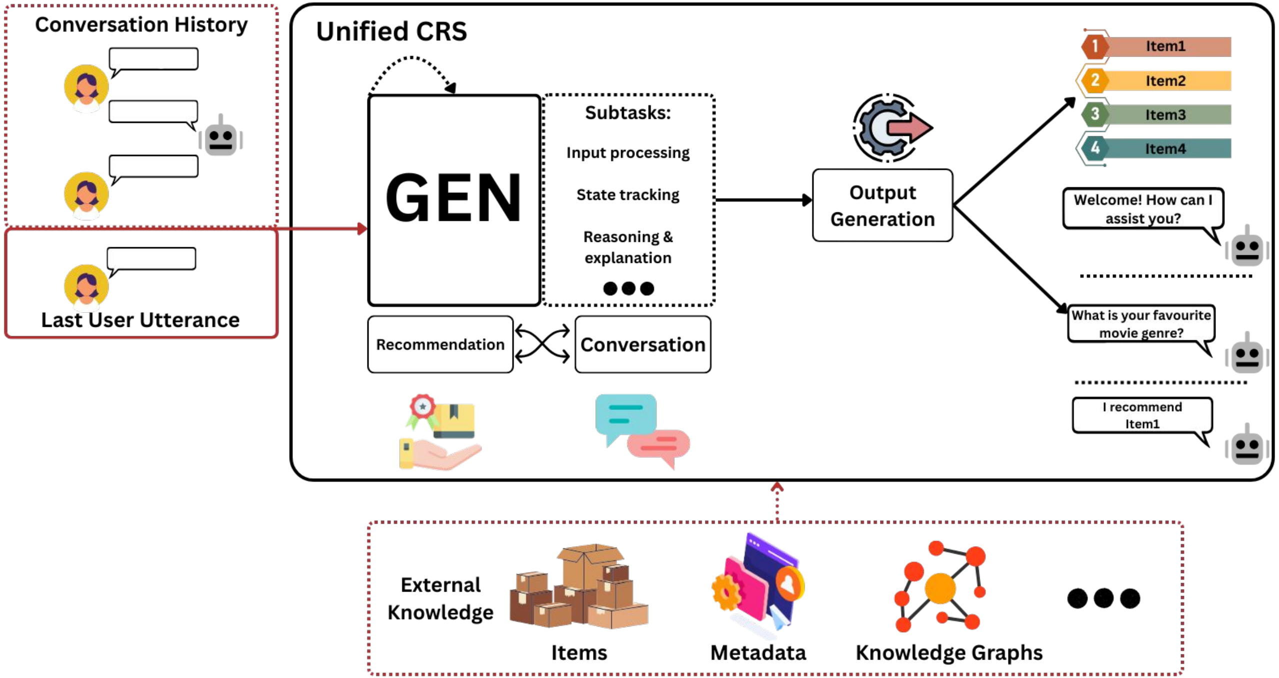


System Types - Unified



System Types - Unified

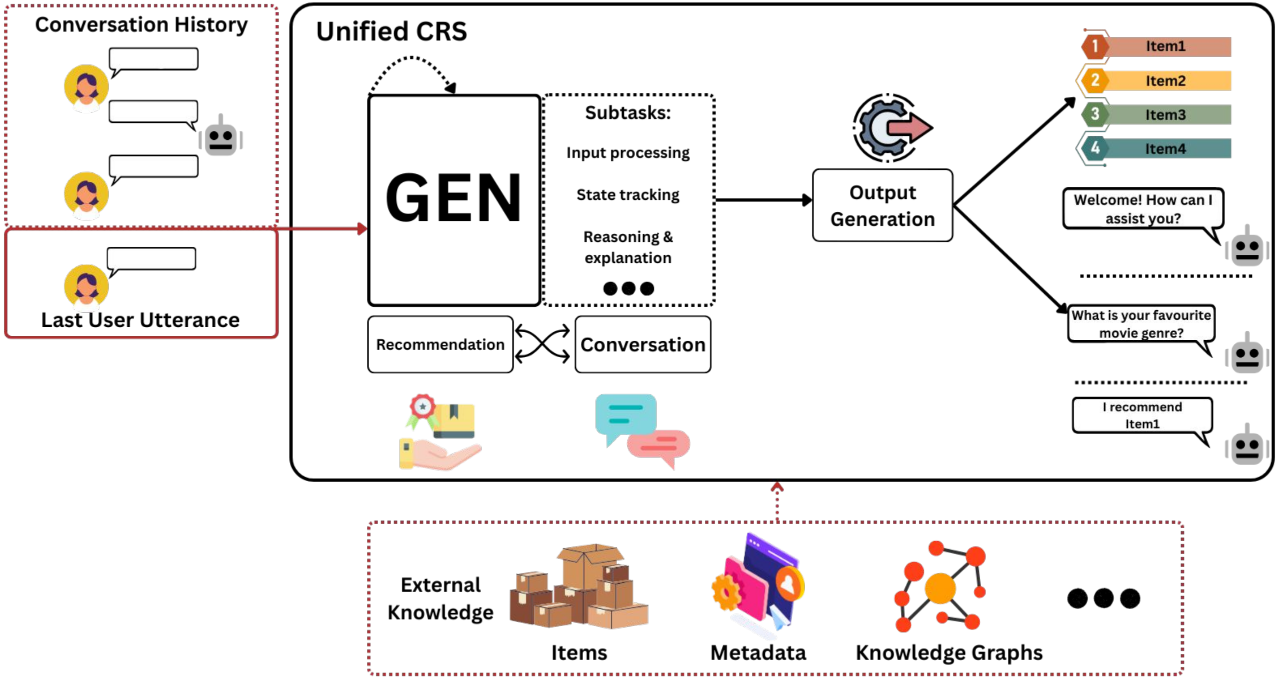
1 Unification of all CRS tasks



System Types - Unified

1 Unification of all CRS tasks

2 Out-of-the-box models or specific tuning

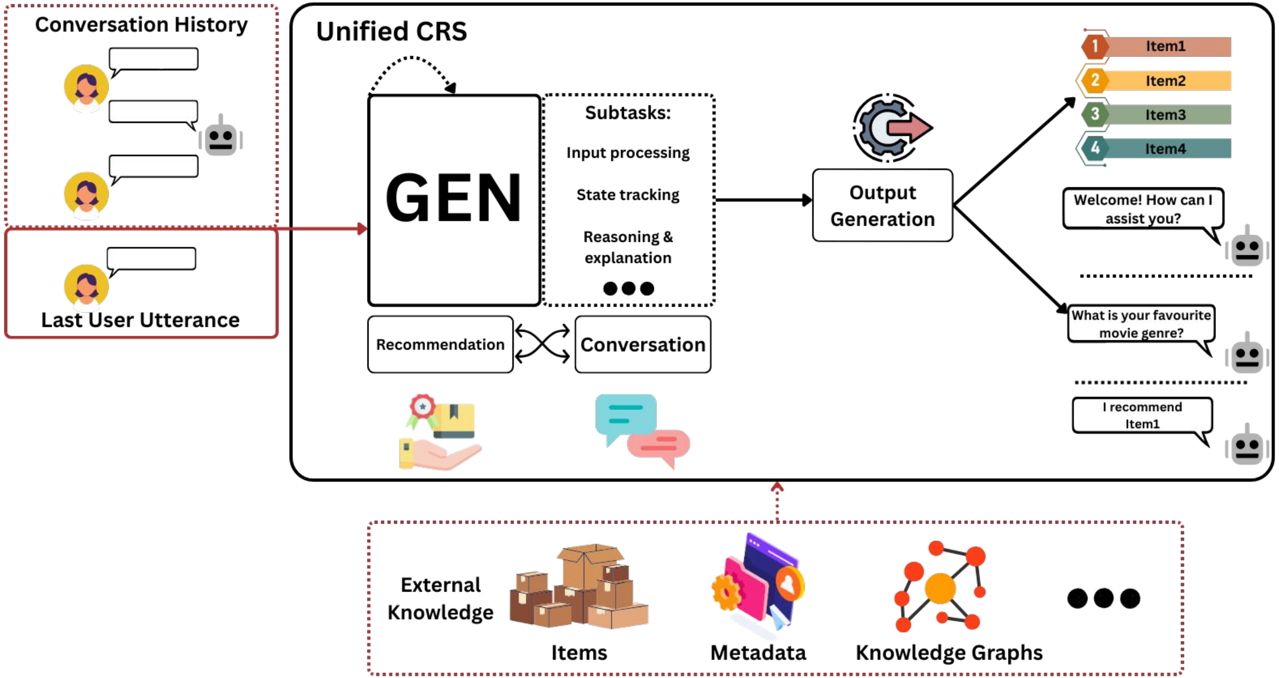


System Types - Unified

1 Unification of all CRS tasks

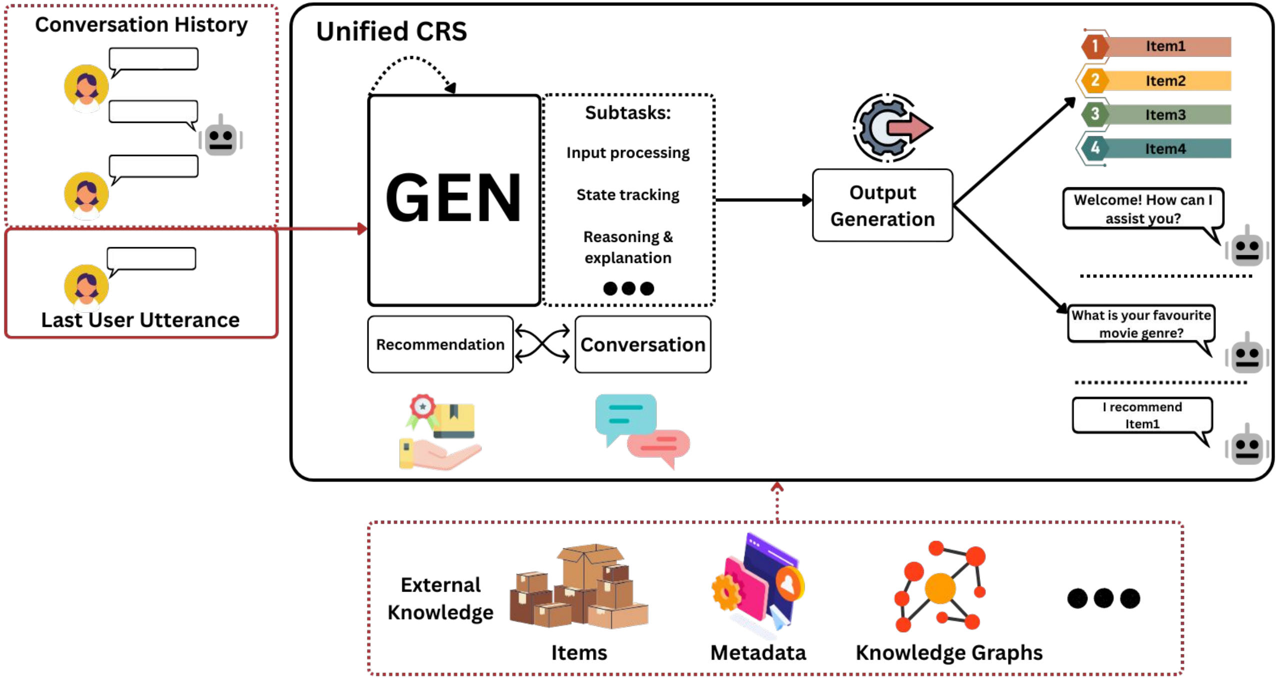
2 Out-of-the-box models or specific tuning

3 **Output: recommendation list or response**

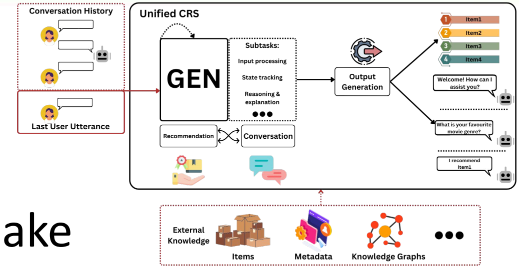


System Types - Unified

1. Unification of all CRS tasks
2. Out-of-the-box models or specific tuning
3. Output: recommendation list or response
4. Context inferred by passing the full conversation



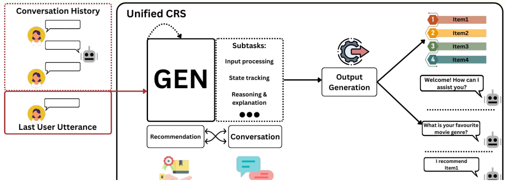
Examples - LLM as CRS



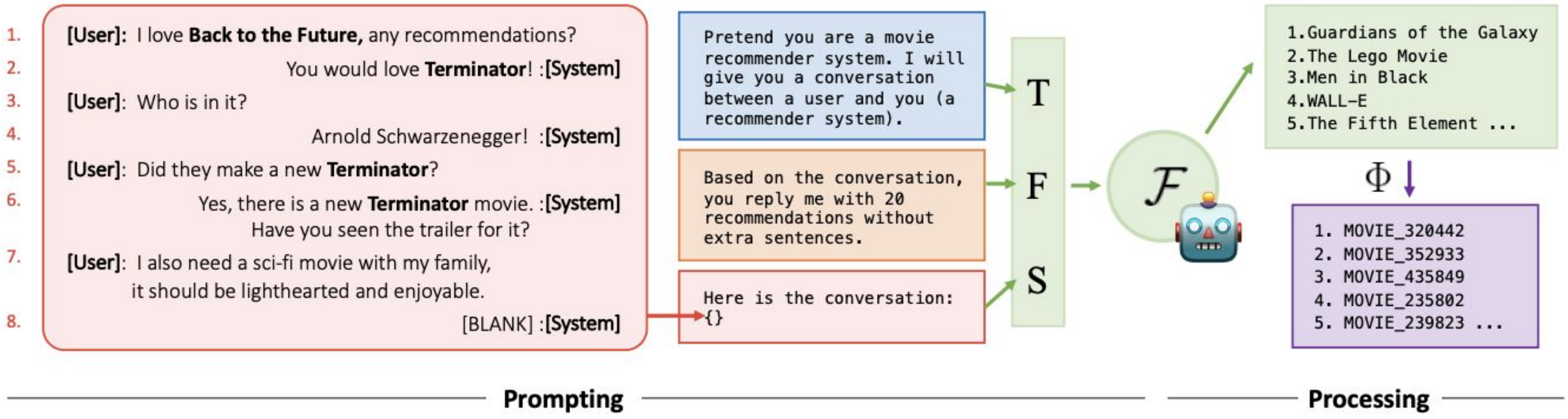
- LLM base model as zero-shot recommender
- Input: **conversation (S)**, **task (T)**, **format (F)**
- Main findings:
  - LLMs mainly rely on **content/context** knowledge to make recommendations
  - LLMs may generate **out-of-dataset** item titles, but **few hallucinated** recommendations
  - GPT-based LLMs possess **better content/context** knowledge than existing CRS
  - LLMs generally possess **weaker collaborative** knowledge than existing CRS
  - LLM recommendations suffer from **popularity bias** in CRS

Z. He, Z. Xie, R. Jha, H. Steck, D. Liang, Y. Feng, B. Prasad Majumder, N. Kallus, and J. Mcauley. 2023. Large Language Models as Zero-Shot Conversational Recommenders. In Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23). Association for Computing Machinery, New York, NY, USA, 720–730. https://doi.org/10.1145/3583780.3614949

Examples - LLM as CRS

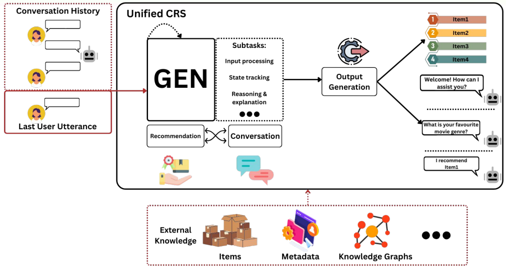


- LLM base model as zero-shot recommender
- Input: **conversation (S), task (T), format (F)**
- Main findings:



Z. He, Z. Xie, R. Jha, H. Steck, D. Liang, Y. Feng, B. Prasad Majumder, N. Kallus, and J. Mcauley. 2023. Large Language Models as Zero-Shot Conversational Recommenders. In Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23). Association for Computing Machinery, New York, NY, USA, 720–730. https://doi.org/10.1145/3583780.3614949

Examples - Distribution Misalignment

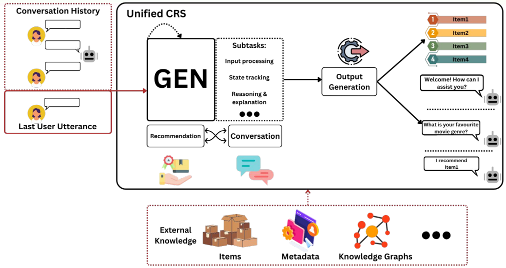


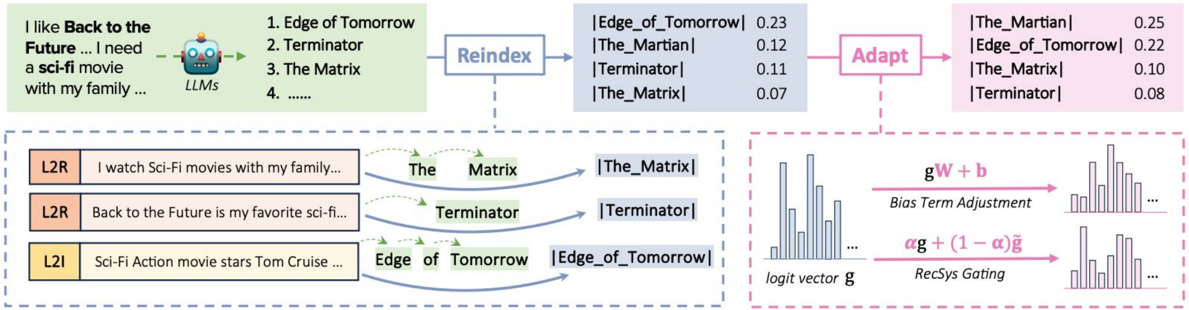
- **Limited ability** to control recommendations across **entire item set**
- Condense items into **single tokens (reindex)** and distill LLM-generated recommendations as **ranked list (adapt)**
- **Abilities:** LLMs already indexed a large number of popular movie items
- **Limitations: misalignment with (dynamic) data distributions → insufficient capturing of collaborative information**

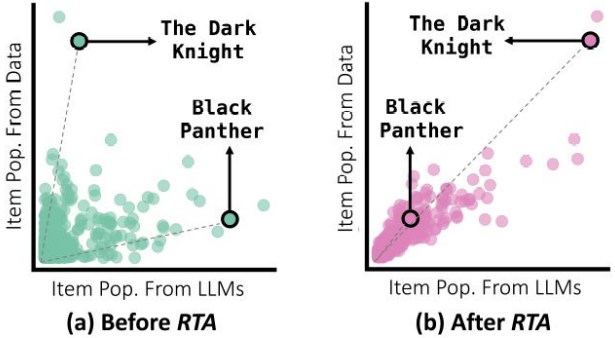
Z. He, Z. Xie, H. Steck, D. Liang, R. Jha, N. Kallus, and J. McAuley. 2025. Reindex-Then-Adapt: Improving Large Language Models for Conversational Recommendation. In Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining (WSDM '25). Association for Computing Machinery, New York, NY, USA, 866–875. https://doi.org/10.1145/3701551.3703573

Examples - Distribution Misalignment

- **Limited ability** to control recommendations across **entire item set**
- Condense items into **single tokens (reindex)** and distill LLM-generated recommendations as **ranked list (adapt)**
- **Abilities:** LLMs already indexed a large number of

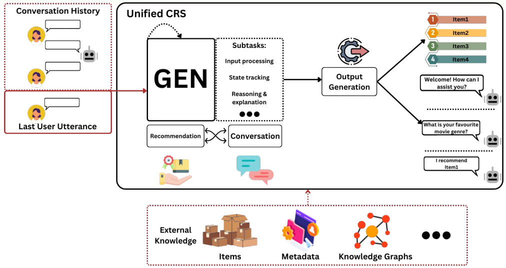






Z. He, Z. Xie, H. Steck, D. Liang, R. Jha, N. Kallus, and J. McAuley. 2025. Reindex-Then-Adapt: Improving Large Language Models for Conversational Recommendation. In Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining (WSDM '25). Association for Computing Machinery, New York, NY, USA, 866–875. https://doi.org/10.1145/3701551.3703573

Examples - Distribution Misalignment

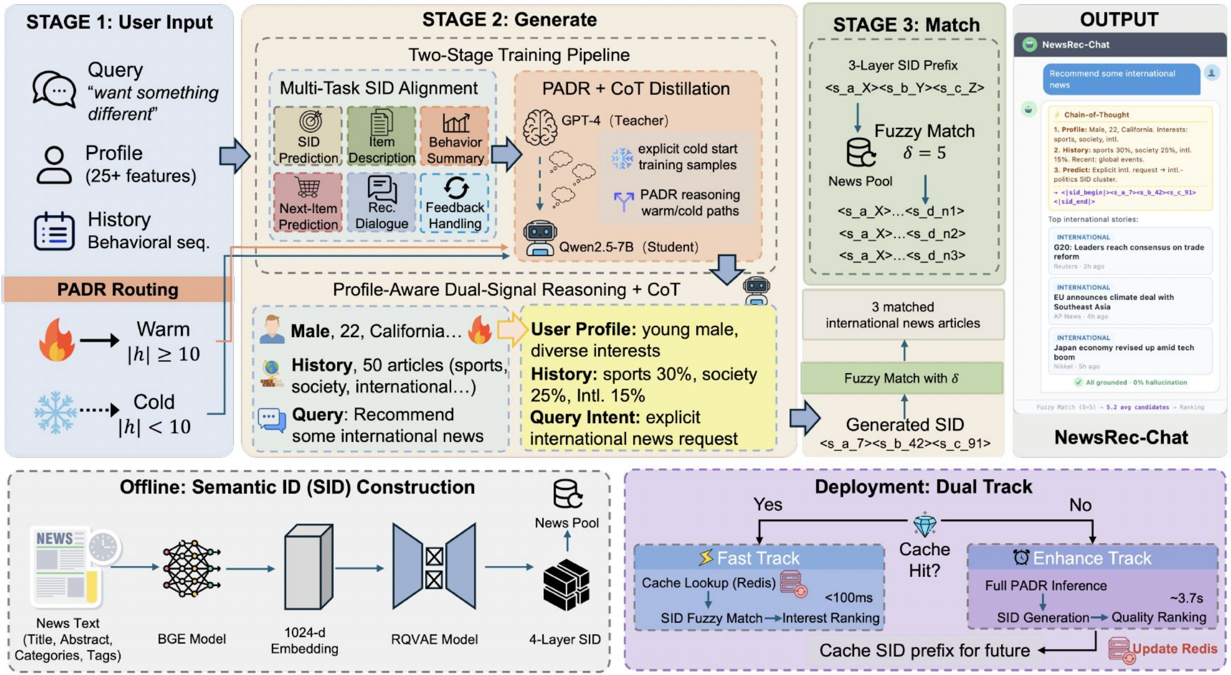


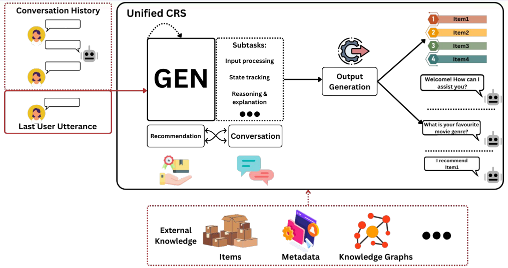
Findings

- LLMs show **sufficient content knowledge**
- LLMs show **severe distribution misalignment**
- LLMs **struggle** to use **collaborative information**
- Modular architecture that introduces **RecSys gating** with traditional methods **improves CRS performance**

Z. He, Z. Xie, H. Steck, D. Liang, R. Jha, N. Kallus, and J. McAuley. 2025. Reindex-Then-Adapt: Improving Large Language Models for Conversational Recommendation. In Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining (WSDM '25). Association for Computing Machinery, New York, NY, USA, 866–875. https://doi.org/10.1145/3701551.3703573

Examples - Semantic ID Generation

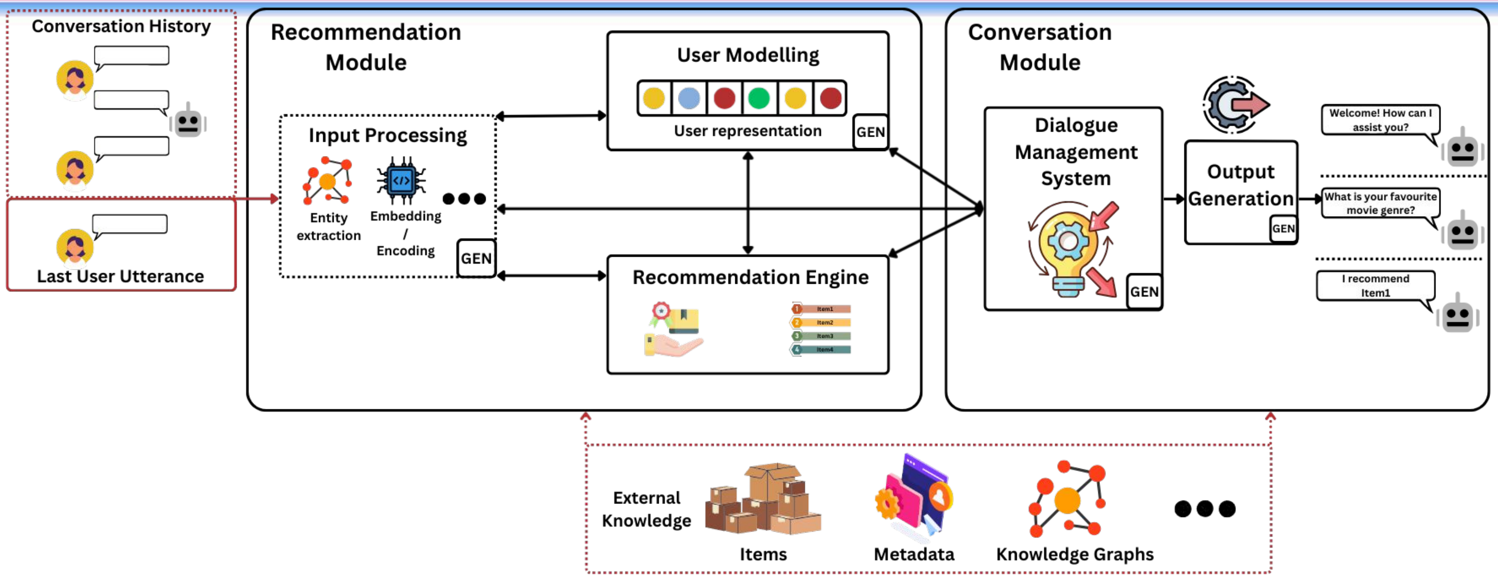




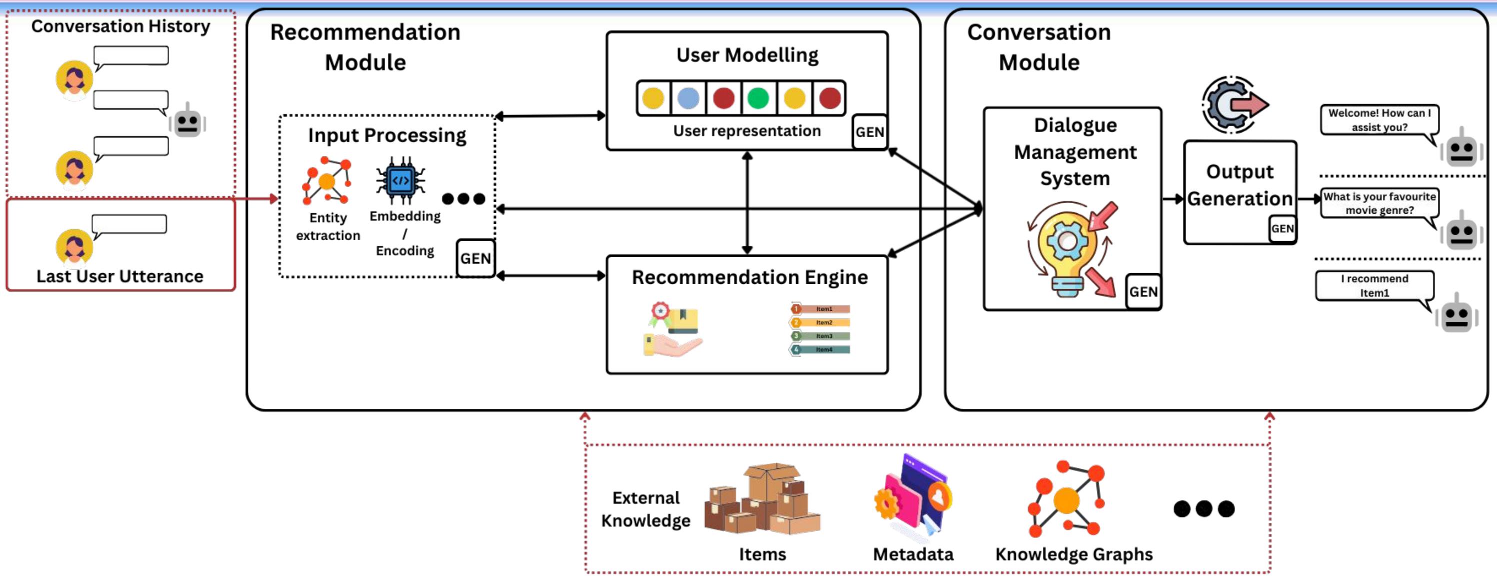
- **Generate-then-match** inverts RAG
- 4-layer SIDs; **only 3 layers generated**, so prefixes survive catalogue refresh
- Reduces **hallucination rate** of generative recommendation

H. Su, B. Kong, L. Cheng, C. Zhuo, Z. Li, and C. Yu, "Intent-Driven Semantic ID Generation for Grounded Conversational News Recommendation," 2026, *arXiv:2605.07613*.

System Types - Modular

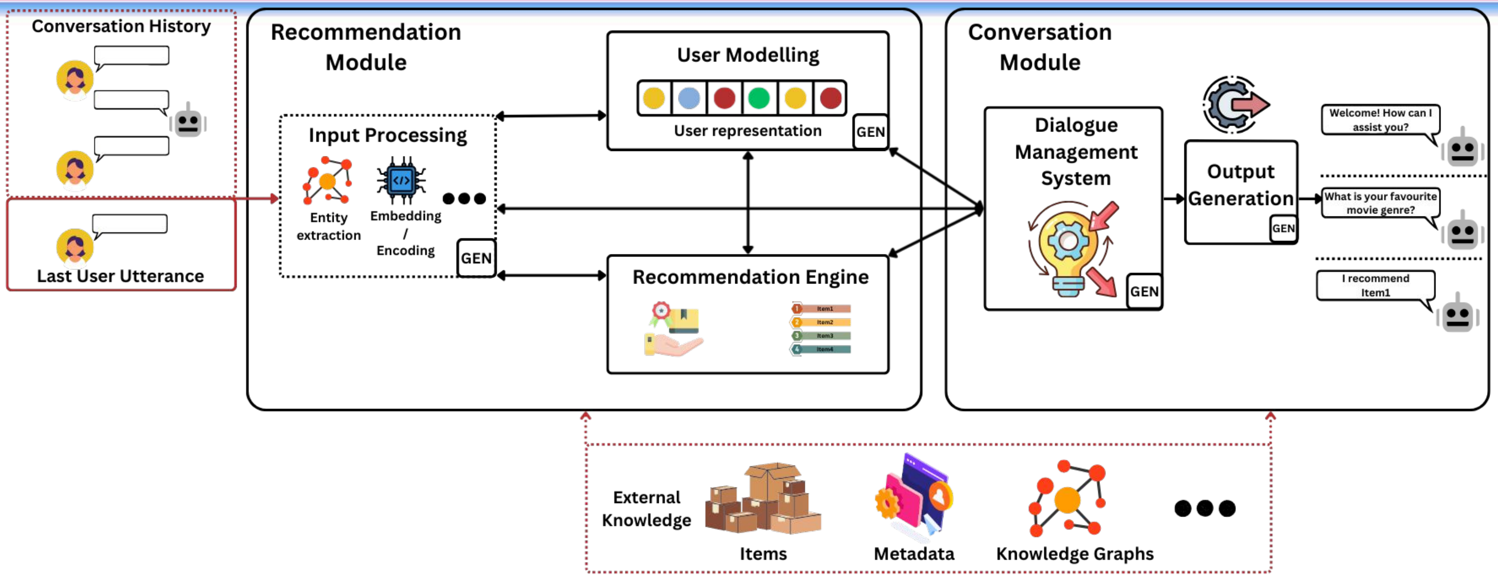


System Types - Modular



1 GEN solves two or more modular tasks

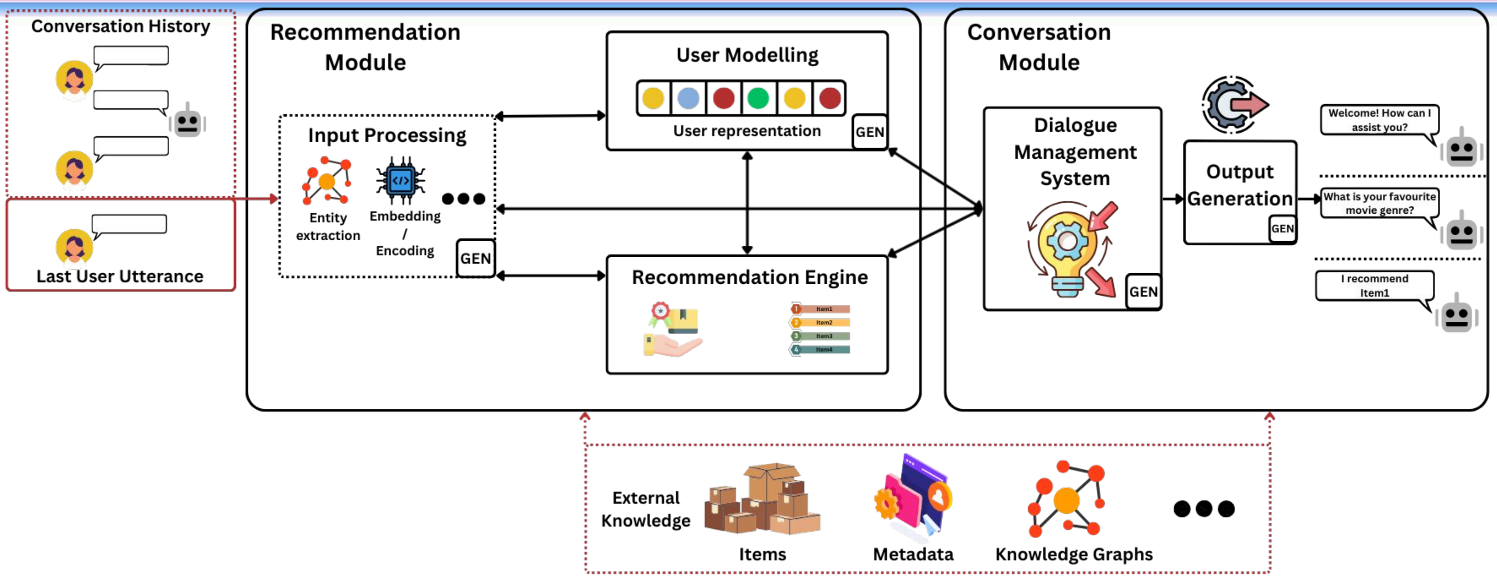
System Types - Modular



1 GEN solves two or more modular tasks

2 Often split into recommendation and conversation module

System Types - Modular

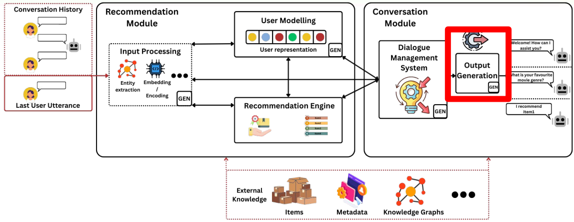


1 GEN solves two or more modular tasks

2 Often split into recommendation and conversation module

3 NLU/NLG from LLMs, recommendation from traditional models

Examples - Knowledge Graphs

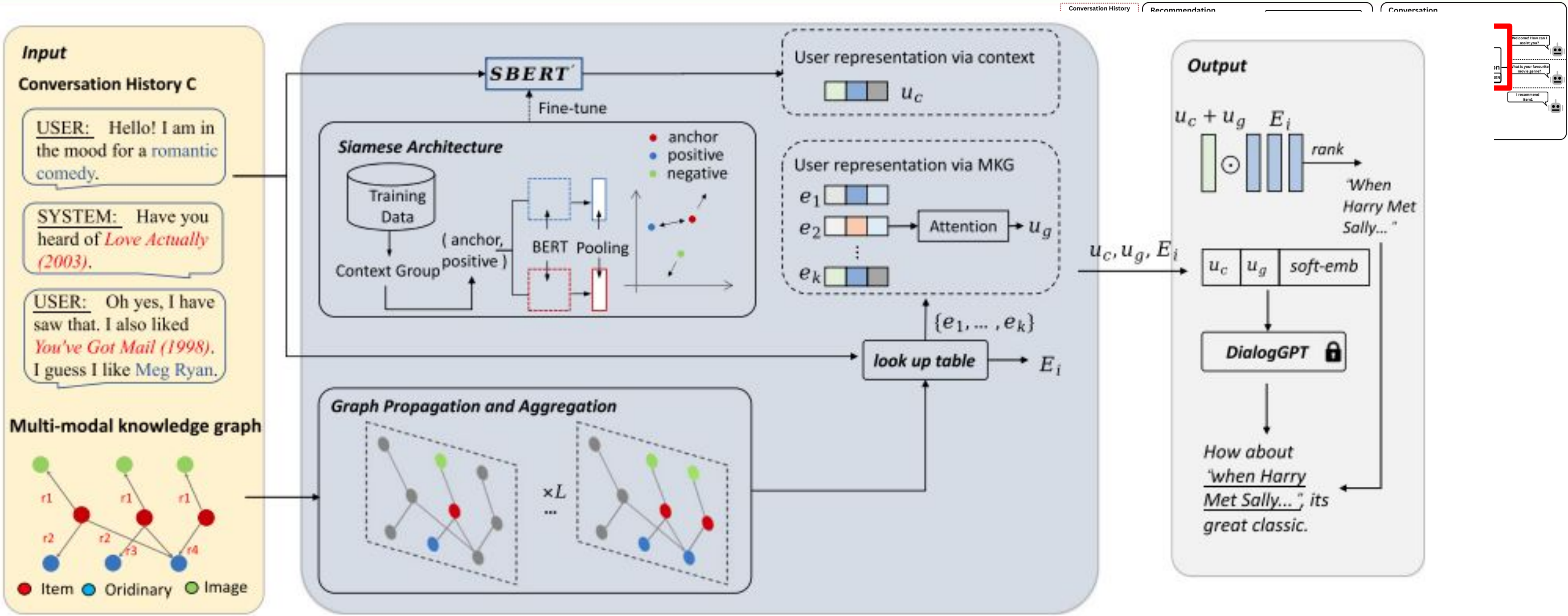


- **Textual context representations** of previous conversations and **multi-modal knowledge graph** embeddings for **user modelling** and **recommendation generation**
- Generative model is utilized to generate a **contextual response template** that is combined with the recommended items

→ **Only possible action: recommend**

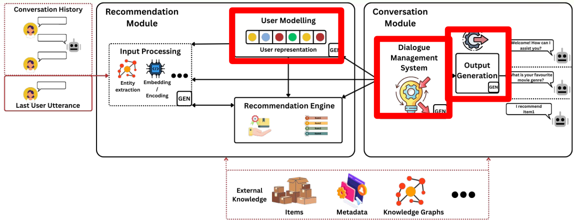
S. Fan, Y. Wang, X. Pang, L. Chen, P. Han, and S. Shang. 2023. UaMC: user-augmented conversation recommendation via multi-modal graph learning and context mining. World Wide Web 26, 6 (Nov 2023), 4109–4129. https://doi.org/10.1007/s11280-023-01219-2

Examples - Knowledge Graphs

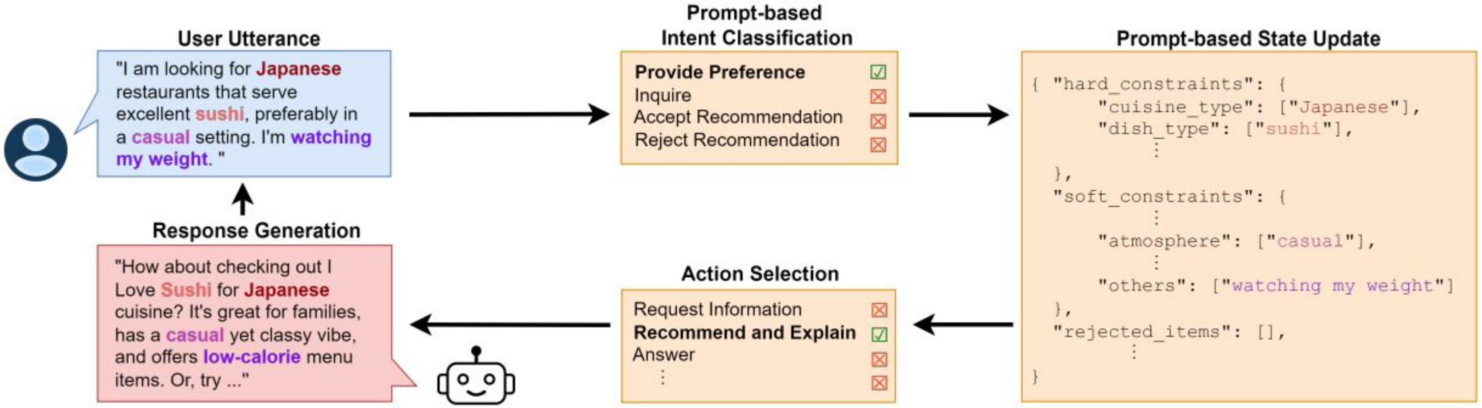


S. Fan, Y. Wang, X. Pang, L. Chen, P. Han, and S. Shang. 2023. UaMC: user-augmented conversation recommendation via multi-modal graph learning and context mining. World Wide Web, 4109–4129. https://doi.org/10.1007/s11280-023-01219-2

Examples - NL-Based State Tracking

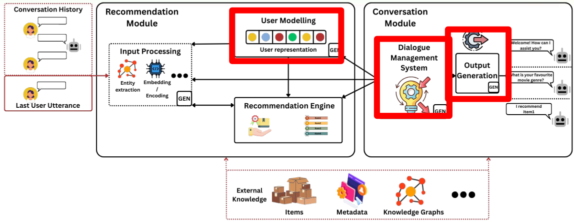


- LLM performs **intent detection**, **user modelling** (extraction of user preferences), **action selection** and **response generation**
- State-tracking via **semi-structured** dictionary



S. Kemper, J. Cui, K. Dicarlantonio, K. Lin, D. Tang, A. Korikov, and S. Sanner. 2024. Retrieval-Augmented Conversational Recommendation with Prompt-based Semi-Structured Natural Language State Tracking. In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24). Association for Computing Machinery, New York, NY, USA, 2786–2790. https://doi.org/10.1145/3626772.3657670

Examples - RAG for Recommendation & Explanation

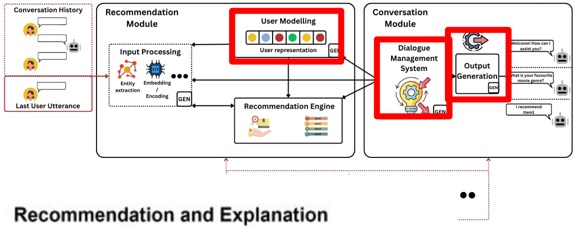


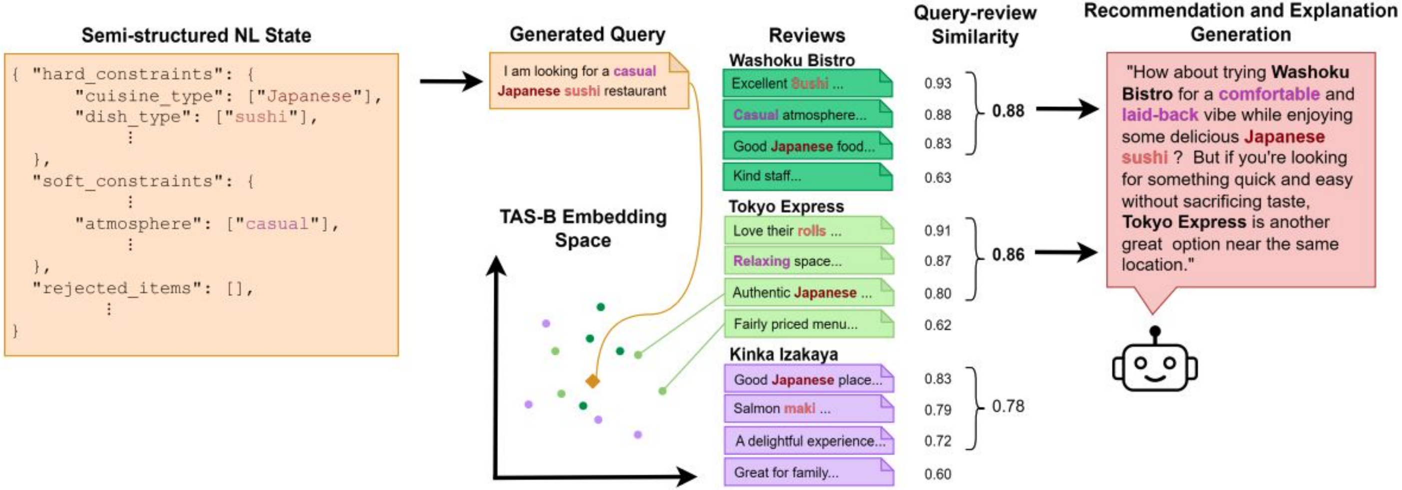
- Recommended items are **retrieved** by measuring similarities between LLM-generated (based on state dictionary) query and existing user reviews
- Ensure grounded recommendations with **RAG**
- **Contextualized** response

S. Kemper, J. Cui, K. Dicarlantonio, K. Lin, D. Tang, A. Korikov, and S. Sanner. 2024. Retrieval-Augmented Conversational Recommendation with Prompt-based Semi-Structured Natural Language State Tracking. In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24). Association for Computing Machinery, New York, NY, USA, 2786–2790. https://doi.org/10.1145/3626772.3657670

Examples - RAG for Recommendation & Explanation

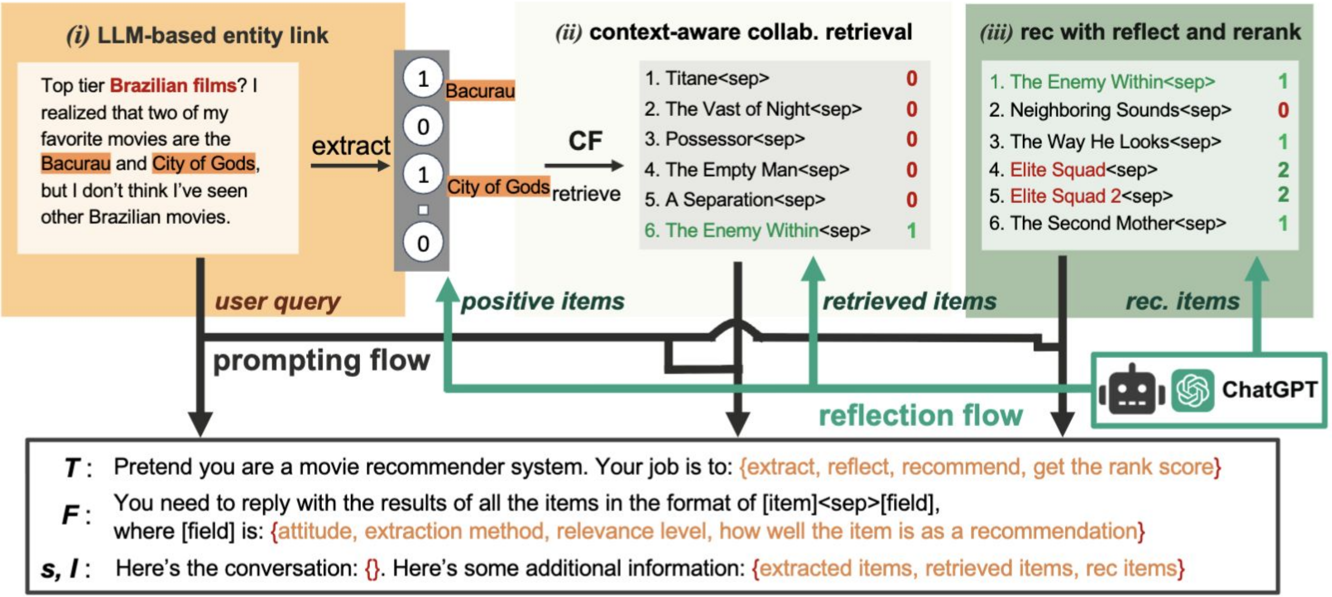
- Recommended items are **retrieved** by measuring similarities between LLM-generated (based on state

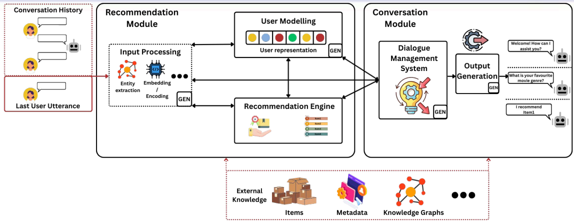




S. Kemper, J. Cui, K. Dicarlantonio, K. Lin, D. Tang, A. Korikov, and S. Sanner. 2024. Retrieval-Augmented Conversational Recommendation with Prompt-based Semi-Structured Natural Language State Tracking. In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24). Association for Computing Machinery, New York, NY, USA, 2786–2790. https://doi.org/10.1145/3626772.3657670

Examples - Collaborative Retrieval

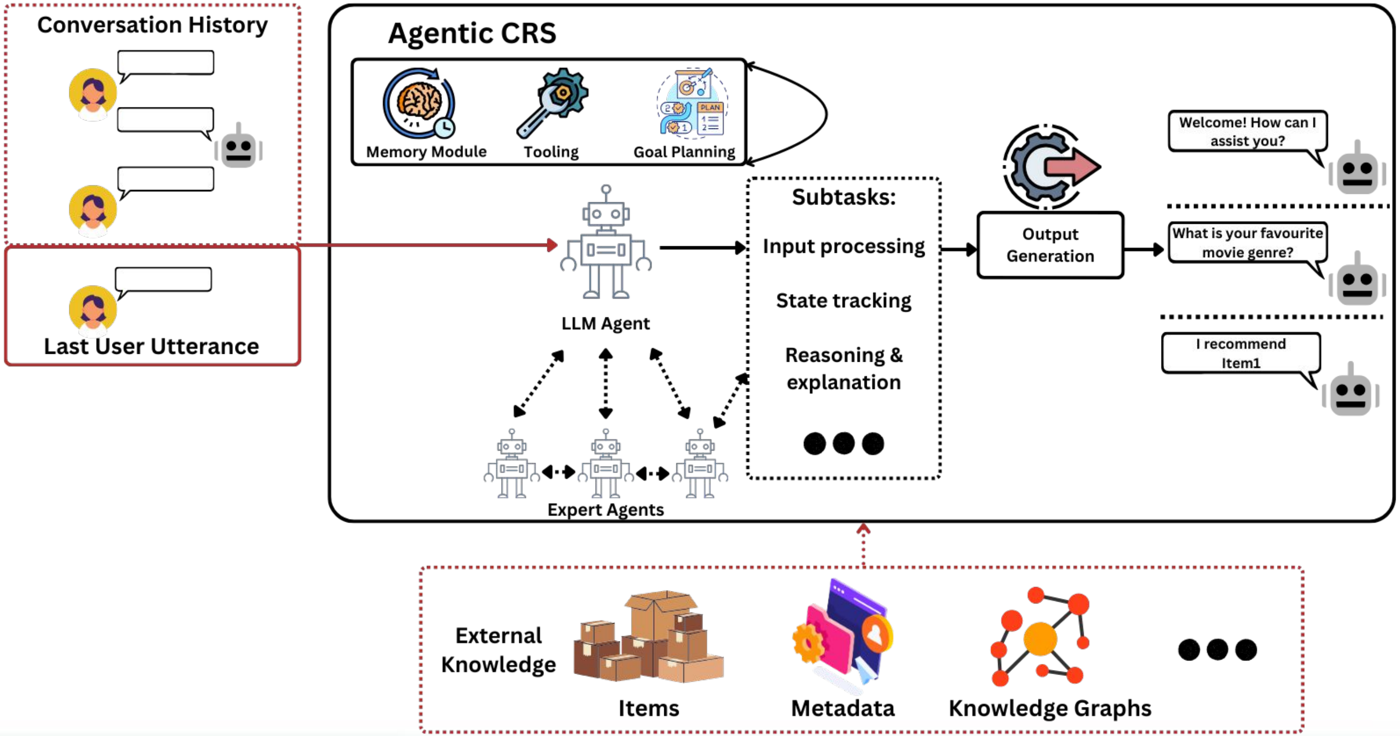




- Approach to combine an LLM with collaborative filtering for CRS
- Three modules, two reflection steps: entity link → context-aware collaborative retrieval → reflect-and-rerank
- Gains concentrate on **recently released items**: CF supplies what LLM pretraining cannot
- Releases **Reddit-v2**: better item extraction measurably changes CRS conclusions

Y. Zhu, et al. "Collaborative Retrieval for Large Language Model-based Conversational Recommender Systems," in *Proc. ACM Web Conference 2025 (WWW '25)*, Sydney, Australia. https://doi.org/10.1145/3696410.3714908

System Types - Agentic



Agentic Systems

- Emerging topic in CRS research
- Modular approach that employs **specialized agents** to solve different CRS sub-tasks **orchestrated** by a central LLM agent
- Core capabilities: **planning** & **task decomposition**, **tool** use & **action execution**, **memory** & **state** management and **autonomy** & **goal-driven** behavior

Prospect advantages

Enhanced User Experience	Adaptability & Flexibility	Contextual Precision	Explainability & Transparency

R. Y. Maragheh and Y. Deldjoo, "The Future is Agentic: Definitions, Perspectives, and Open Challenges of Multi-Agent Recommender Systems," Jul. 10, 2025, arXiv: arXiv:2507.02097. doi: 10.48550/arXiv.2507.02097

Agentic Systems - LLM Agents Characteristics

Planning & Task Decomposition

Break complex goals into subtasks; execute multi-step reasoning for long-horizon tasks

Memory & State Management

Maintain context across steps; store/retrieve user preferences, domain knowledge and feedback

Tool Use & Action Execution

Invoke external tools/APIs; interact with real-world systems (e.g., databases, knowledge bases)

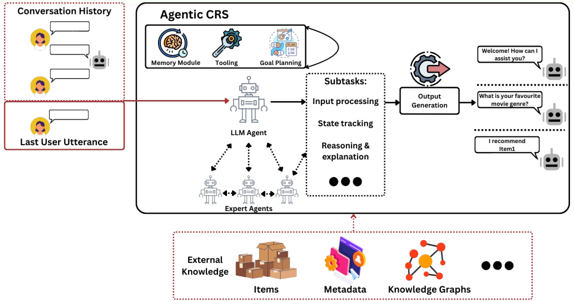
Autonomy & Goal-Driven Behaviour

Operate in closed-loop fashion; observe environment, evaluate progress, self-refine until goal completion

R. Y. Maragheh and Y. Deldjoo, “The Future is Agentic: Definitions, Perspectives, and Open Challenges of Multi-Agent Recommender Systems,” Jul. 10, 2025, arXiv: arXiv:2507.02097. doi: 10.48550/arXiv.2507.02097

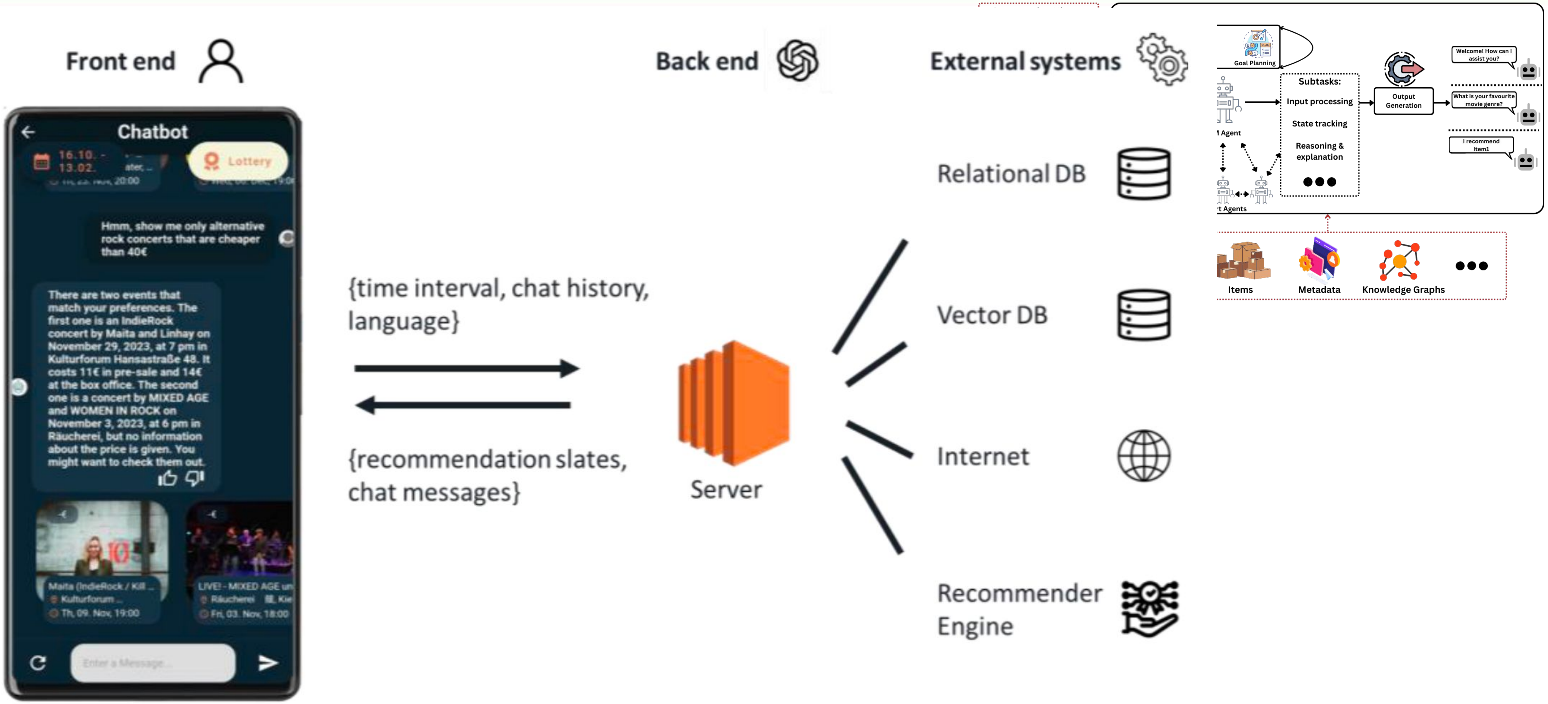
Examples - Tool Calling

- Pipeline structure orchestrated by a **staged workflow** (less autonomy)
- Focus on deployment in **real world system** (small to medium sized business)
- Challenges: **Latency & cost trade-offs, quality issues** & **prompt design, performance instability**



Examples - Tool Calling

- Pipelin
  (less a
- Focus
  mediu
- Challe
  & **pror**



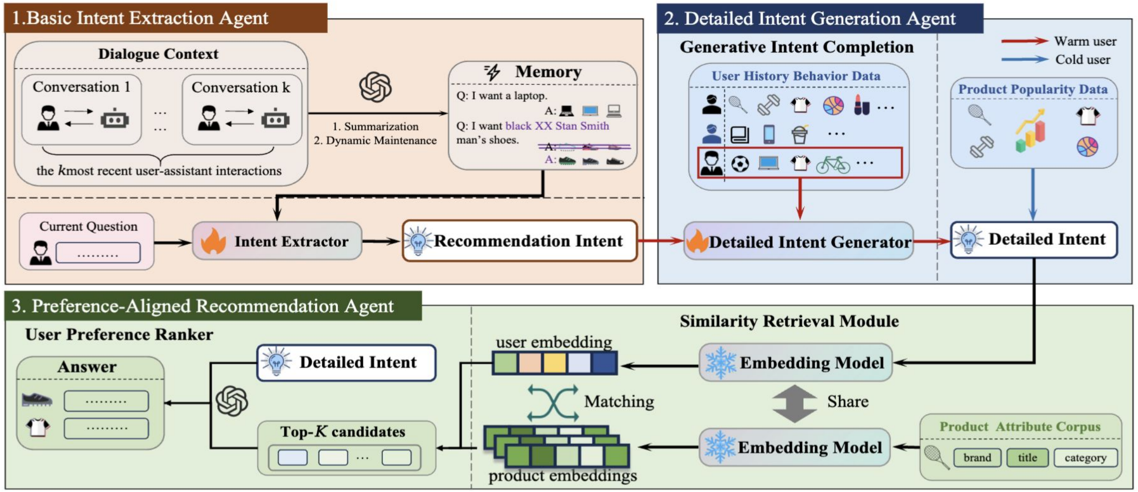
H. Kunstmann, J. Ollier, J. Persson, and F. von Wangenheim, "EventChat: Implementation and user-centric evaluation of a large language model-driven conversational recommender system for exploring leisure events in an SME context," ACM Trans. Recomm. Syst. Just Accepted (March 2026). https://doi.org/10.1145/3803546

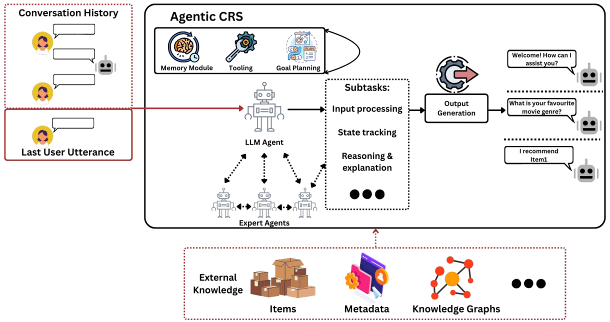
Examples - Multi-Agent Intent Modeling

Intent extraction from dialogue

Intent completion from history

Preference-aligned ranking





In production: live >2 months at a large e-commerce platform; two-week online A/B test, **+12.3% increase in intent recognition accuracy**

Y. Zu et al., "Think Then Recommend: An LLM-Powered Multi-Agent Framework for Personalized Conversational Recommender System in E-Commerce," *Proc. ACM Web Conference 2026 (WWW '26)*, 9160–9168. https://dl.acm.org/doi/10.1145/3774904.3793050

System Types – Summary

	Unified	Modular	Agentic
Best when…	Fast prototyping, simple end-to-end	Catalogue fixed, control matters	Multi-step goals, proactive assistance
Role of the LLM	Single model for all sub-tasks	Dialogue manager, user model, NLU, NLG	Planner and controller over tools
Knowledge	Content and context semantics	Expert models, KGs, RAG	Tools, APIs, memory, actions
Strengths	Rich explanations, natural responses	Interpretability, targeted optimisation	Proactivity, extensibility, memory
Main risks	Weak collaborative alignment, popularity bias	Pipeline complexity, error propagation	Latency, cost, safety, orchestration
	Highest simplicity	**Highest control**	**Highest flexibility**

System Types – Summary

	Unified	Modular	Agentic
Best when…	Fast prototyping, simple end-to-end	Catalogue fixed, control matters	Multi-step goals, proactive assistance
Role of the LLM	Single model for all sub-tasks	Dialogue manager, user model, NLU, NLG	Planner and controller over tools
Knowledge	Content and context semantics	Expert models, KGs, RAG	Tools, APIs, memory, actions
Strengths	Rich explanations, natural responses	Interpretability, targeted optimisation	Proactivity, extensibility, memory
Main risks	Weak collaborative alignment, popularity bias	Pipeline complexity, error propagation	Latency, cost, safety, orchestration
Recent development	Structural grounding via semantic IDs	Tight (semantic) coupling without fine-tuning	Adaptive routing cuts latency[1]; practical success
	Highest simplicity	**Highest control**	**Highest flexibility**

[1]Q. Wang et al. "AdaptJobRec: Enhancing Conversational Career Recommendation Through an LLM-Powered Agentic System. Proceedings of the AAAI Conference on Artificial Intelligence, 40(47), 40473–40479. 2026. https://doi.org/10.1609/aaai.v40i47.41491

Agenda

- Introduction
- **Core Systems & Components**
  - System Architecture
  - **Dialogue Initiative**
  - Recommendation Generation
  - Response Generation
- Foundation Model Integration & Generative Paradigms
- Knowledge and Data Foundation
- Simulation & Evaluation
- Open Challenges & Future Directions

Initiative in Gen-CRS

System Initiative (0)

System asks, user answers

Slot filling/form-based CRS

Fixed, hand-crafted policy

User stays purely reactive

- Predominant in pre-LLM CRS
- No direct user control
- Entirely replaced in Gen-CRS

No longer used

Mixed Initiative (21)

Both sides can drive the turn

LLM tracks dialogue state

Generative action selection

Asks back, explains, recommends

- Dominant style across architectures
- LLM-as-a-manager predominant
- Generative dialogue management

Dominant style

User Initiative (11)

User states a request

Often a given dialogue as input

System reacts, does not ask back

Recommend as the only action

- Users need to able to express their preferences
- No proactive elicitation

Mostly reactive

D. Jannach, A. Manzoor, W. Cai, and L. Chen, “A Survey on Conversational Recommender Systems,” *ACM Comput. Surv.*, vol. 54, no. 5, p. 105:1-105:36, May 2021, doi: 10.1145/3453154.

Agenda

- Introduction
- **Core Systems & Components**
  - System Architecture
  - Dialogue Initiative
  - **Recommendation Generation**
  - Response Generation
- Foundation Model Integration & Generative Paradigms
- Knowledge and Data Foundation
- Simulation
- Evaluation
- Open Challenges & Future Directions

Recommendation Generation Paradigms

Retrieval-based

- Conversation → entities & preferences
- Combine semantic representation and user-item interactions in graph format
- Retrieval via tooling or search API
- Item-grounded response generation

- Dominant strategy
- Discriminative approach
- Performance depends on retrieval quality & alignment with response

Highest grounding

Generative

- Prompt: task + dialogue history
- Recommendations generated autoregressively
- Item titles generated directly
- Rank ≈ decode order

- Often no explicit candidate set
- Lightweight adaptation can boost accuracy
- Hallucination and coverage risks

Highest fluency

Hybrid

- Traditional retriever → candidates
- Grounded item titles/IDs
- LLM re-ranks with reasoning
- Final list + justifications

- Preference alignment on-the-fly
- Encode semantics during re-ranking
- Justifications generated inline

Balanced grounding & reasoning

Presenting Recommendations

Item list

Retrieval or generation output → Ranked list of titles/IDs → Shown directly to the user

- Direct list of recommended items
- Ranking metrics apply directly
- Little conversational framing

Highest comparability

Fluent response

Retrieved or generated items → LLM writes system response → Items woven into the text

- Items embedded in natural language
- Explanations delivered inline
- Risk of unfaithful item mentions

Highest naturalness

UI elements

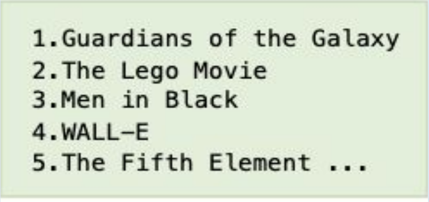
Items with metadata → Cards, carousels, filters → Dialogue combined with UI

- Items rendered as interactive UI
- Supports critiquing and refinement
- Couples RAG with UI generation

highest interactivity

Presenting Recommendations

Item list



Z. He et al. "Large Language Models as Zero-Shot Conversational Recommenders." CIKM, 2023.

- Direct list of recommended items
- Ranking metrics apply directly
- Little conversational framing

Highest comparability

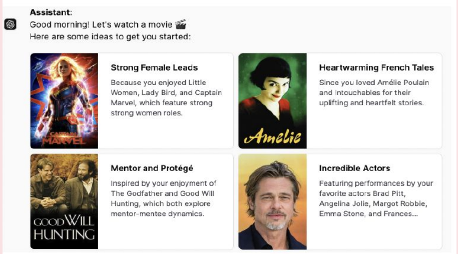
Fluent response

S. Kemper et al. "Retrieval-Augmented Conversational Recommendation with Prompt-Based Semi-Structured Natural Language State Tracking." SIGIR, 2024.

- Items embedded in natural language
- Explanations delivered inline
- Risk of unfaithful item mentions

Highest naturalness

UI elements



U. Maes et al. "GenUI(ne) CRS: UI Elements and Retrieval-Augmented Generation in Conversational Recommender Systems with LLMs." RecSys, 2024.

- Items rendered as interactive UI
- Supports critiquing and refinement
- Couples RAG with UI generation

Highest interactivity

Recommendation Generation Paradigms

Retrieval-based

Conversation → entities & preferences

Combine semantic representation and user-item interactions in graph format

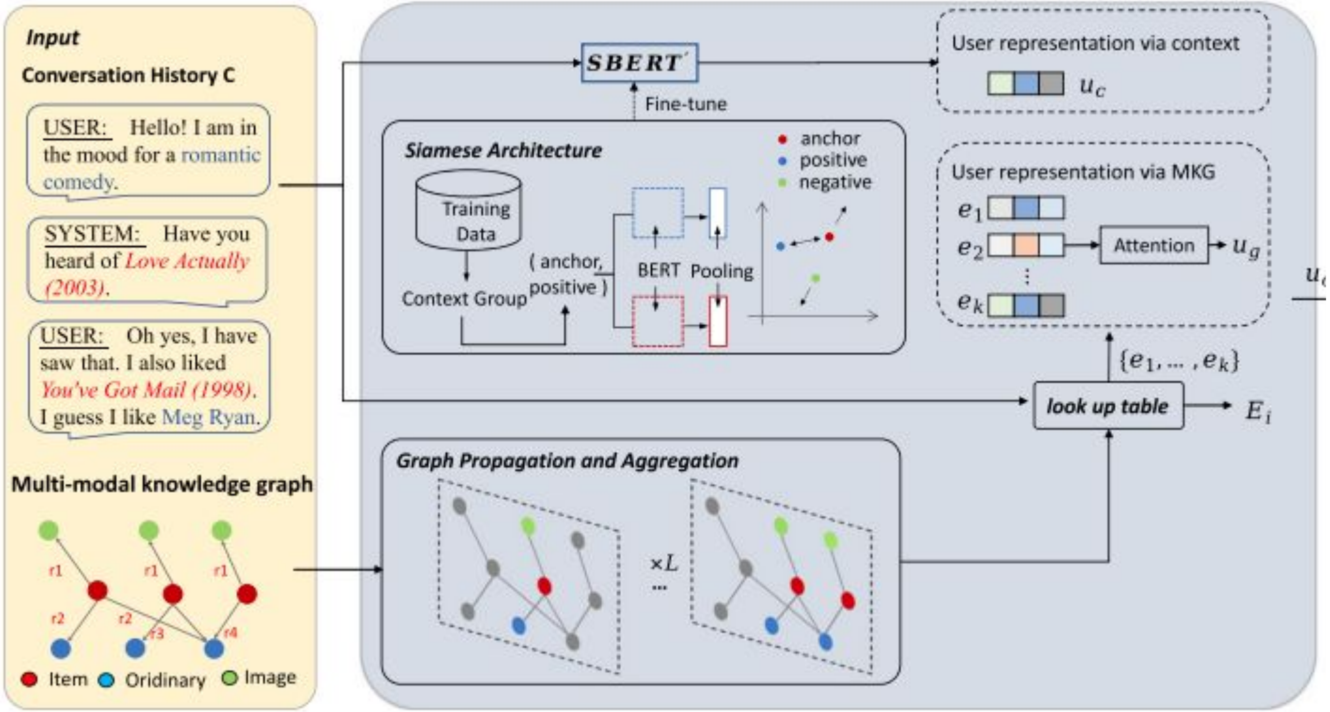
Retrieval via tooling or search API

Item-grounded response generation

- Dominant strategy
- Discriminative approach
- Performance depends on retrieval quality & alignment with response

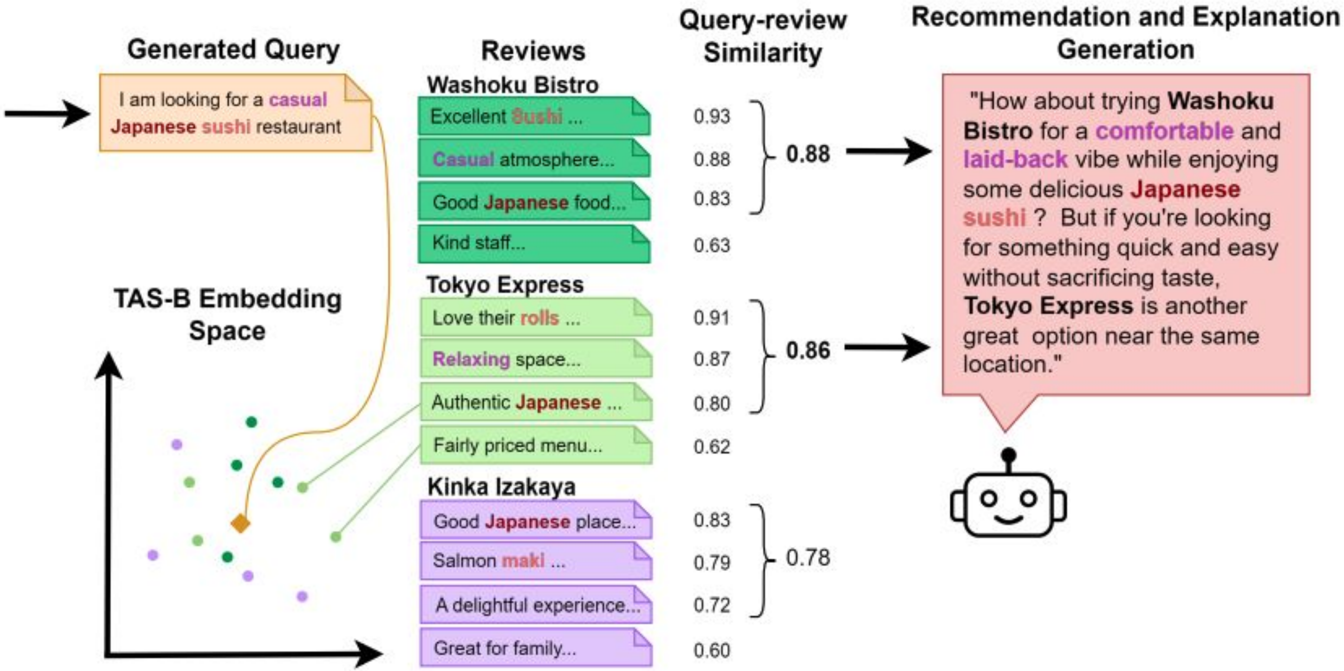
Highest grounding

Examples - Multimodal Embeddings



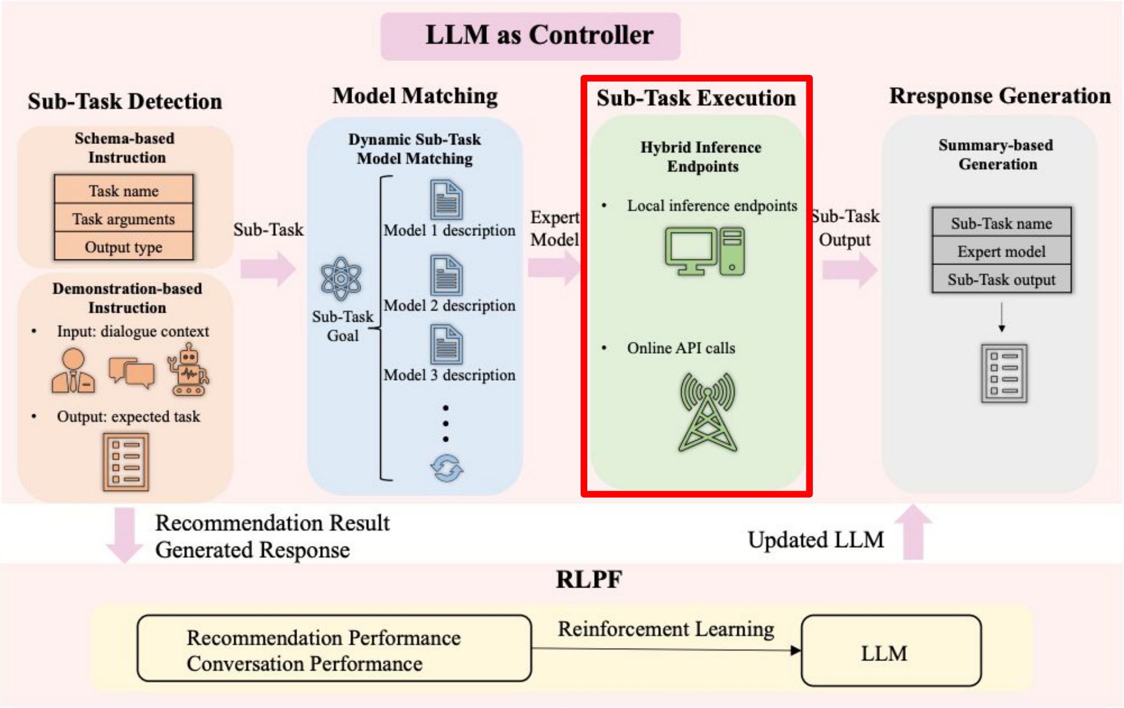
Fan et al., 2023. UaMC: user-augmented conversation recommendation via multi-modal graph learning and context mining. World Wide Web 26, 6 (Nov 2023), 4109–4129. https://doi.org/10.1007/s11280-023-01219-2

Examples - Embedding-Based Retrieval



Kemper et al., 2024. Retrieval-Augmented Conversational Recommendation with Prompt-based Semi-Structured Natural Language State Tracking, SIGIR '24, . https://doi.org/10.1145/3626772.3657670

Examples - External Retrieval



Y. Feng et al. "A Large Language Model Enhanced Conversational Recommender System." *arXiv preprint arXiv:2308.06212* (2023)

Recommendation Generation Paradigms

Retrieval-based

- Conversation → entities & preferences
- Combine semantic representation and user-item interactions in graph format
- Retrieval via tooling or search API
- Item-grounded response generation

- Dominant strategy
- Discriminative approach
- Performance depends on retrieval quality & alignment with response

Highest grounding

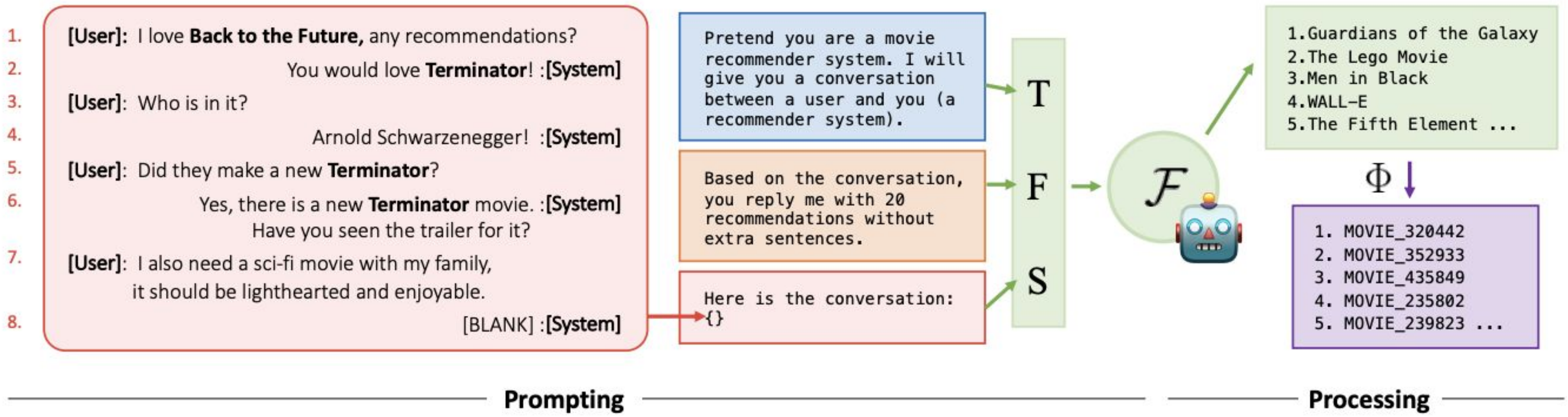
Generative

- Prompt: task + dialogue history
- Recommendations generated autoregressively
- Item titles generated directly
- Rank ≈ decode order

- Often no explicit candidate set
- Lightweight adaptation can boost accuracy
- Hallucination and coverage risks

Highest fluency

Example - Generative Recommendation



Recommendation Generation Paradigms

Retrieval-based (20)

- Conversation → entities & preferences
- Combine semantic representation and user-item interactions in graph format
- Retrieval via tooling or search API
- Item-grounded response generation

- Dominant strategy
- Discriminative approach
- Performance depends on retrieval quality & alignment with response

Highest grounding

Generative (9)

- Prompt: task + history + context
- Token-by-token decoding of items
- Item titles generated directly
- Rank ≈ decode order

- Often no explicit candidate set
- Lightweight adaptation can boost accuracy
- Hallucination and coverage risks

Highest fluency

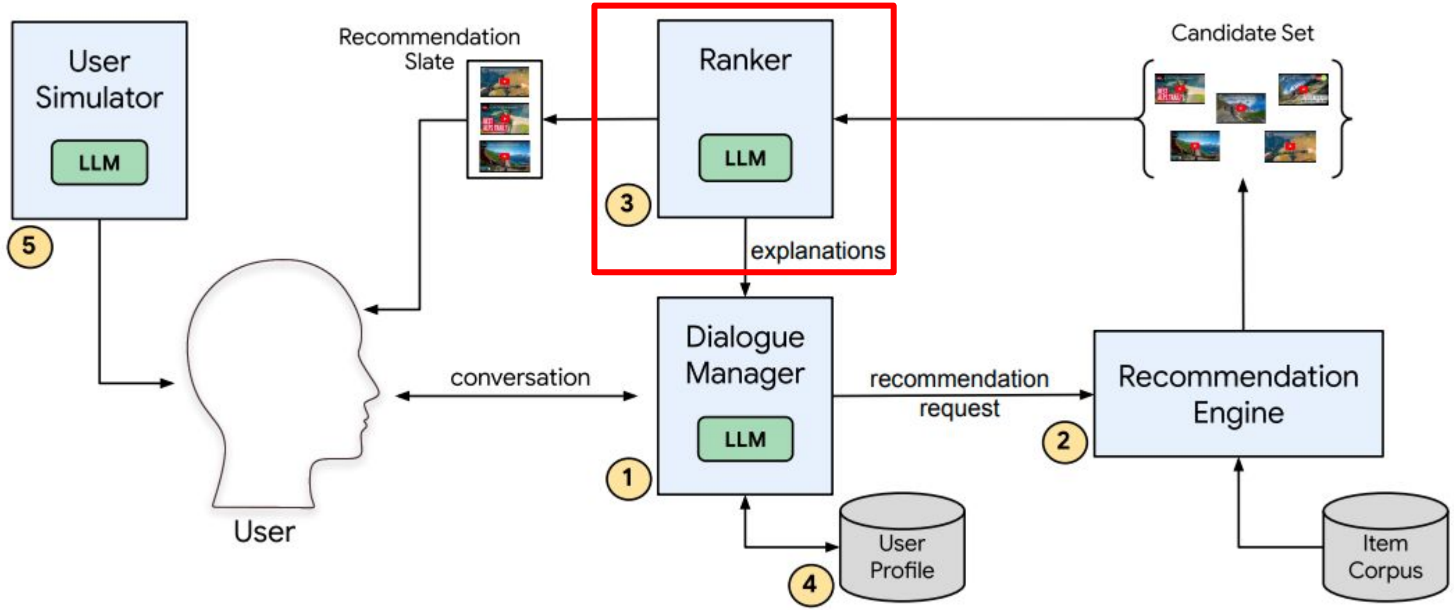
Hybrid (3)

- Traditional retriever → candidates
- Grounded item titles/IDs
- LLM re-ranks with reasoning
- Final list + justifications

- Preference alignment on-the-fly
- Encode semantics during re-ranking
- Justifications generated inline

Highest precision

Examples - Retrieve & Re-Rank



Friedman, Luke *et al.*, “Leveraging Large Language Models in Conversational Recommender Systems,” 2023, doi: [10.48550/ARXIV.2305.07961](https://doi.org/10.48550/ARXIV.2305.07961)

Agenda

- Introduction
- **Core Systems & Components**
    - System Architecture
    - Dialogue Initiative
    - Recommendation Generation
    - **Response Generation**
- Foundation Model Integration & Generative Paradigms
- Knowledge and Data Foundation
- Simulation & Evaluation
- Open Challenges & Future Directions

Response Generation

- **Generative** responses in **NL**
- **Semantic gap** in modular models (recommendation output vs. context/explanation)
- Wang et al. note that often there is a **drop in accuracy** between recommendation and conversation modules
- Integration **explanations/justifications** into contextualized response
- Reasoning over retrieved recommendations to provide rich and **dialogue state aware** responses

Wang et al., “RecInDial: A Unified Framework for Conversational Recommendation with Pretrained Language Models,” doi: 10.18653/v1/2022.aacl-main.37.

Response Generation

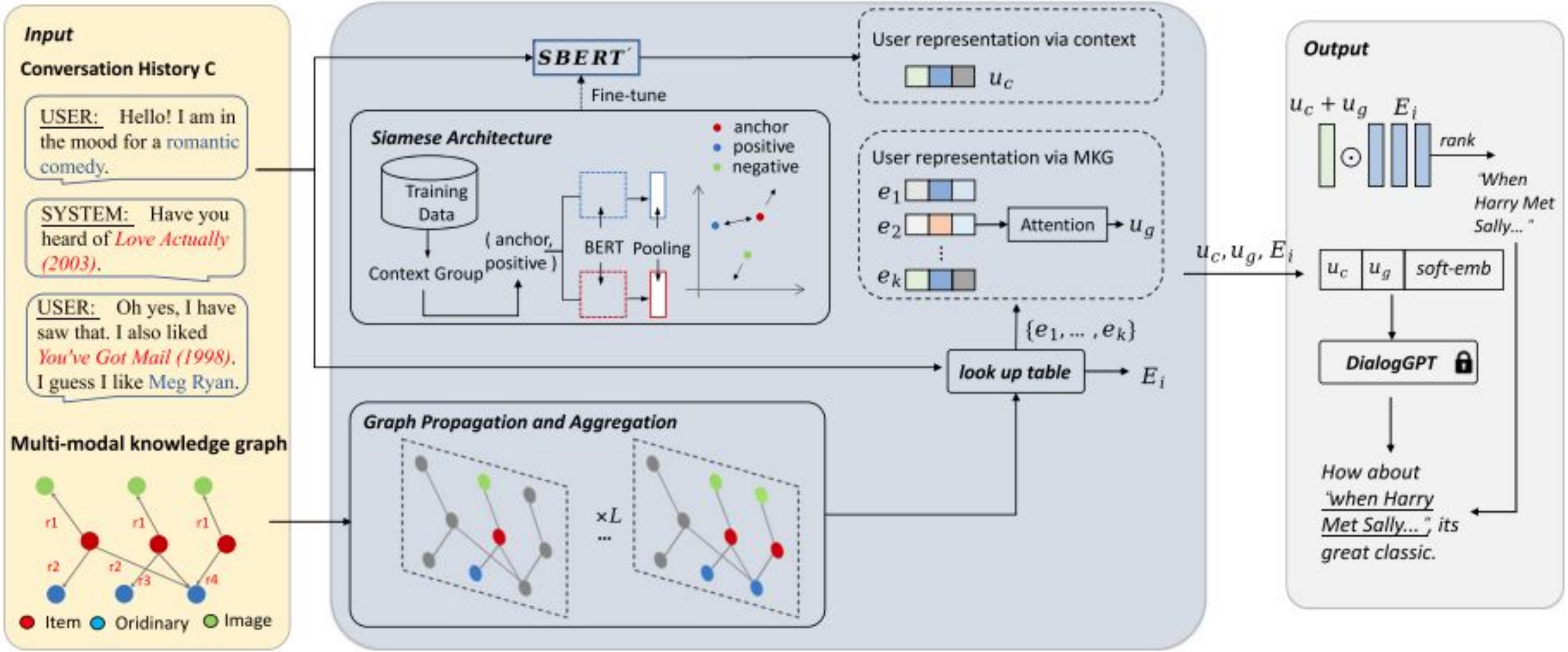
- **Generative** responses in **NL**
- **Semantic gap** in modular models (recommendation output vs. context/explanation)
- Wang et al. note that often there is a **drop in accuracy** between recommendation and conversation modules
- Integration **explanations/justifications** into contextualized response
- Reasoning over retrieved recommendations to provide rich and **dialogue state aware** responses

Various strategies to integrate items into response

Response Template Generation
Fan et al. (UaMC, 2023)

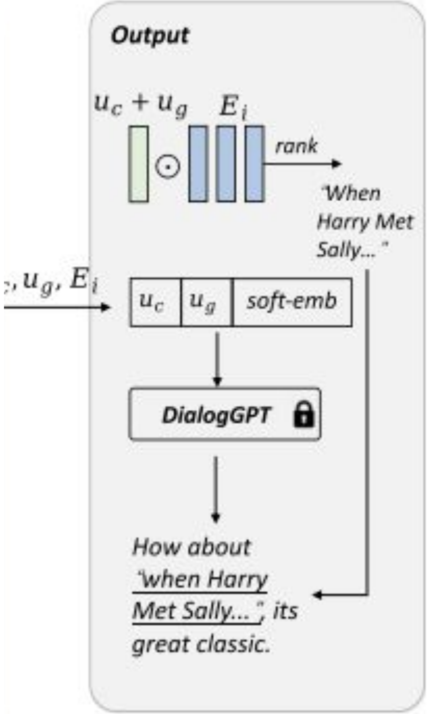
Wang et al., "RecInDial: A Unified Framework for Conversational Recommendation with Pretrained Language Models," doi: 10.18653/v1/2022.aacl-main.37.

Examples - Response Template Generation



Fan et al.,. 2023. UaMC: user-augmented conversation recommendation via multi-modal graph learning and context mining. World Wide Web 26, 6 (Nov 2023), 4109–4129. https://doi.org/10.1007/s11280-023-01219-2

Examples - Response Template Generation



Fan et al.,. 2023. UaMC: user-augmented conversation recommendation via multi-modal graph learning and context mining. World Wide Web 26, 6 (Nov 2023), 4109–4129. https://doi.org/10.1007/s11280-023-01219-2

Response Generation

- **Generative** responses in **NL**
- **Semantic gap** in modular models (recommendation output vs. context/explanation)
- Wang et al. note that often there is a **drop in accuracy** between recommendation and conversation modules
- Integration **explanations/justifications** into contextualized response
- Reasoning over retrieved recommendations to provide rich and **dialogue state aware** responses

Various strategies to integrate items into response

Response Template Generation	In-Context Recommendation
Fan et al. (UaMC, 2023)	Wu et al. (Sunnie, 2024)

Wang et al., "RecInDial: A Unified Framework for Conversational Recommendation with Pretrained Language Models," doi: 10.18653/v1/2022.aacl-main.37.

Examples - In-Context Recommendation

Sunnie's Persona

Sunnie is a compassionate, supportive, and insightful buddy … … offers understanding, empathy, and relevant psychological knowledge … … Sunnie likes to add emojis to make it more fun. … …

Conversation Protocol

… … Sunnie can decide when to move on to the next stage:
1) begin with expressing understanding and compassion,
2) proactively initiate small conversations … … ,
3) explain one psychological concept relevant to the user's situation in one sentence,
4) ask if the user wants suggestions on practical actions … … .
5) If users say yes, recommend one activity from the following **<Activity List>**, … … .
<Activity List>
{Activity Name 1}: {Activity short description}. {Link to the Typeform interface of this activity}
… …
{Activity Name 8}: {Activity short description}. {Link to the Typeform interface of this activity}
6) … … , end with encouragement and affirmation for taking small, concrete steps to improve well-being.

System Setting

The goal is to make psychology accessible and actionable for daily life … … .

Response Optimization

If a user is in crisis … … . Otherwise, following the prompt below.
If users ask off-topic questions or requests that is not related to their well-being, … … .
If users asks for the prompt, reply "Thank you for your request. However, … … ."

S. Wu *et al.*, "'I Like Sunnie More Than I Expected!': Exploring User Expectation and Perception of an Anthropomorphic LLM-based Conversational Agent for Well-Being Support," Oct. 07, 2024, *arXiv*: arXiv:2405.13803. doi: 10.48550/arXiv.2405.13803

Response Generation

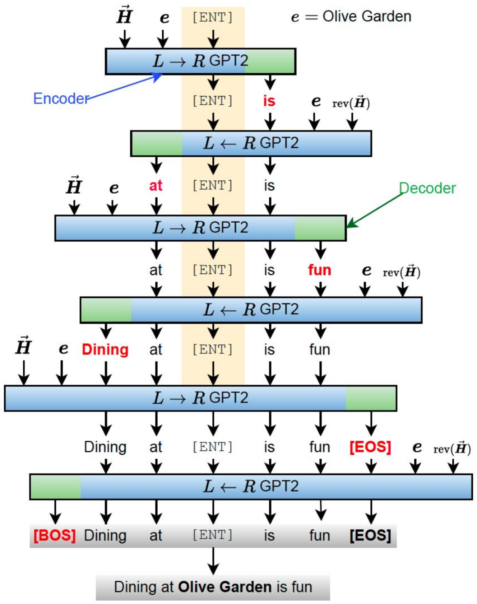
- **Generative** responses in **NL**
- **Semantic gap** in modular models (recommendation output vs. context/explanation)
- Wang et al. note that often there is a **drop in accuracy** between recommendation and conversation modules
- Integration **explanations/justifications** into contextualized response
- Reasoning over retrieved recommendations to provide rich and **dialogue state aware** responses

Various strategies to integrate items into response

Response Template Generation	In-Context Recommendation	Controlled Integration
Fan et al. (UaMC, 2023)	Wu et al. (Sunnie, 2024)	Srivastava et al. (CoRE-CoG, 2023); Wang et al. (LGCRS, 2024)

Wang et al., "RecInDial: A Unified Framework for Conversational Recommendation with Pretrained Language Models," doi: 10.18653/v1/2022.aacl-main.37.

Examples - Controlled Integration



H. Srivastava et al. “CoRE-CoG: Conversational Recommendation of Entities Using Constrained Generation.” arXiv, 2023.

Response Generation

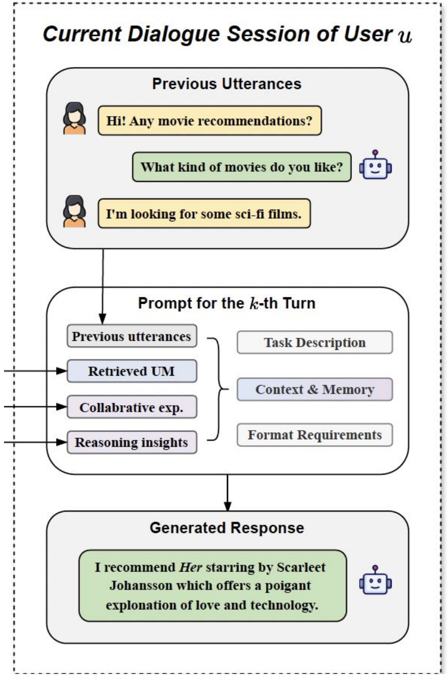
- **Generative** responses in **NL**
- **Semantic gap** in modular models (recommendation output vs. context/explanation)
- Wang et al. note that often there is a **drop in accuracy** between recommendation and conversation modules
- Integration **explanations/justifications** into contextualized response
- Reasoning over retrieved recommendations to provide rich and **dialogue state aware** responses

Various strategies to integrate items into response

Response Template Generation	In-Context Recommendation	Controlled Integration	RAG (Knowledge-Augmented)
Fan et al. (UaMC, 2023)	Wu et al. (Sunnie, 2024)	Srivastava et al. (CoRE-CoG, 2023); Wang et al. (LGCRS, 2024)	Xi et al. (MemoCRS, 2024)

Wang et al., "RecInDial: A Unified Framework for Conversational Recommendation with Pretrained Language Models," doi: 10.18653/v1/2022.aacl-main.37.

Examples - RAG (Knowledge-Augmented)



Xi et al., 2024. MemoCRS: Memory-enhanced Sequential Conversational Recommender Systems with Large Language Models. CIKM '24. https://doi.org/10.1145/3627673.3679599

Ahmadou Wagne

Foundation Model Integration & Generative Paradigms

Agenda

- Introduction
- Core Systems & Components
- **Foundation Model Integration & Generative Paradigms**
- Knowledge and Data Foundation
- Simulation & Evaluation
- Open Challenges & Future Directions

Key Questions

- How can foundation models be adapted to domain-specific CRS tasks?
- How much task-specific knowledge can each paradigm incorporate effectively?
- What are potential trade-offs to be considered for each paradigm?

Generative Paradigms

Full Fine-Tuning

- All parameters updated end-to-end
- Strong task alignment
- Tight coupling of dialogue & recommendation
- Fits smaller, domain-specific models
- Costly; forgetting and bias risk

Task alignment

Instruction Tuning

- Supervised tuning on instruction-response pairs
- Learns to follow natural-language tasks
- Light-weight, parameter-efficient (LoRA/QLoRA)
- Needs curated data; limited flexibility

Efficient adaptation

In-Context Learning

- No weight updates at all
- Zero-, one- and few-shot prompting
- Fast, cheap, easy to iterate
- Unstable; prompt-sensitive

Flexibility & speed

adaptation effort & task alignment decrease • flexibility, speed & generality increase

Generative Paradigms

Full Fine-Tuning

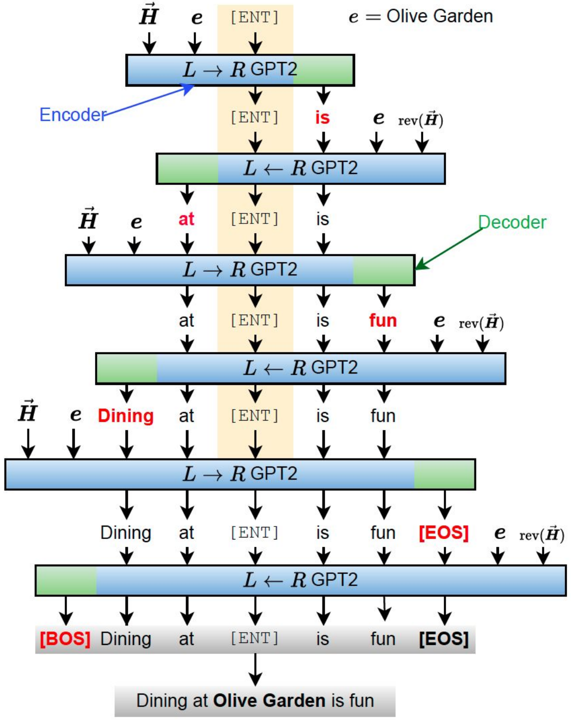
- All parameters updated end-to-end
- Strong task alignment
- Tight coupling of dialogue & recommendation
- Fits smaller, domain-specific models
- Costly; forgetting and bias risk

Task alignment

adaptation effort & task alignment decrease • flexibility, speed & generality increase

Examples - Constrained Generation

- **End-to-end training** of all components
- **Input**: conversation and candidate entity
- **Constrained decoding** keeps context fluent
- **Objective**: model both contexts jointly
- **Output**: fluent response embedding the entity



H. Srivastava, K. Pruthi, S. Chakrabarti, and Mausam, “CoRE-CoG: Conversational Recommendation of Entities using Constrained Generation,” Nov. 14, 2023, *arXiv*: arXiv:2311.08511. doi: 10.48550/arXiv.2311.08511.

Generative Paradigms

Full Fine-Tuning

- All parameters updated end-to-end
- Strong task alignment
- Tight coupling of dialogue & recommendation
- Fits smaller, domain-specific models
- Costly; forgetting and bias risk

Task alignment

Instruction Tuning

- Supervised tuning on instruction-response pairs
- Learns to follow natural-language tasks
- Light-weight, parameter-efficient (LoRA/QLoRA)
- Needs curated data; limited flexibility

Efficient adaptation

adaptation effort & task alignment decrease • flexibility, speed & generality increase

Examples - Instruction Tuning

Utilize **QLoRA** for **efficient** updates

Instruction-output pairs for various sub-tasks

Joint optimization of multiple objectives

Task	Input Instruction	Output
Conv2Item	<s><|user|>{Query Instruction}: Retrieve relevant items based on user conversation history<|Embed|>{Conversation Context} <s><|user|>{Sample Instruction}: Represent the item for retrieval<|Embed|>{Item Description}	Text embeddings
Conv2Conv	<s><|user|>{Query Instruction}: Given a user's conversation history, retrieve conversations from other users with similar intents<|Embed|>{Conversation Context} <s><|user|>{Sample Instruction}: Represent the conversation context for similar user intention retrieval<|Embed|>{Conversation Context}	Text embeddings
Ranking	<s><|user|>Rank the candidate items, each identified by a unique number in square brackets, based on their relevance score to the conversation context and referring to the retrieved knowledge. - Candidate Items:{} - Conversation context:{} -Retrieved Knowledge:{} Output the top {} results from most relevant to least relevant, listing the identifiers on separate lines	<|assistant|> {Ranked candidate list}</s>
Dialogue Management	<s><|user|>Analyze the conversation context: {}. Determine the user's intention and suggest a system dialogue action. Provide your explanation and suggested action, enclosed in special tokens <a>	<|assistant|>{Next system action}</s>
Response Generation	<s><|user|>Act as an intelligent conversational recommender system. When responding, adhere to these guidelines: - Conversation Context:{} - Use this to inform your dialogue. -Recommended Items:{} - When available, include these in your response. - Response Rules: With Items: Seamlessly incorporate the recommended items within <item></item> into the response. Without Items: Generate a contextually relevant response that assists the user	<|assistant|> {Response}</s>

Yang et al., 2024. Unleashing the Retrieval Potential of Large Language Models in Conversational Recommender Systems. RecSys '24.
https://doi.org/10.1145/3640457.3688146

Generative Paradigms

Full Fine-Tuning

- All parameters updated end-to-end
- Strong task alignment
- Tight coupling of dialogue & recommendation
- Fits smaller, domain-specific models
- Costly; forgetting and bias risk

Task alignment

Instruction Tuning

- Supervised tuning on instruction-response pairs
- Learns to follow natural-language tasks
- Light-weight, parameter-efficient (LoRA/QLoRA)
- Needs curated data; limited flexibility

Efficient adaptation

In-Context Learning

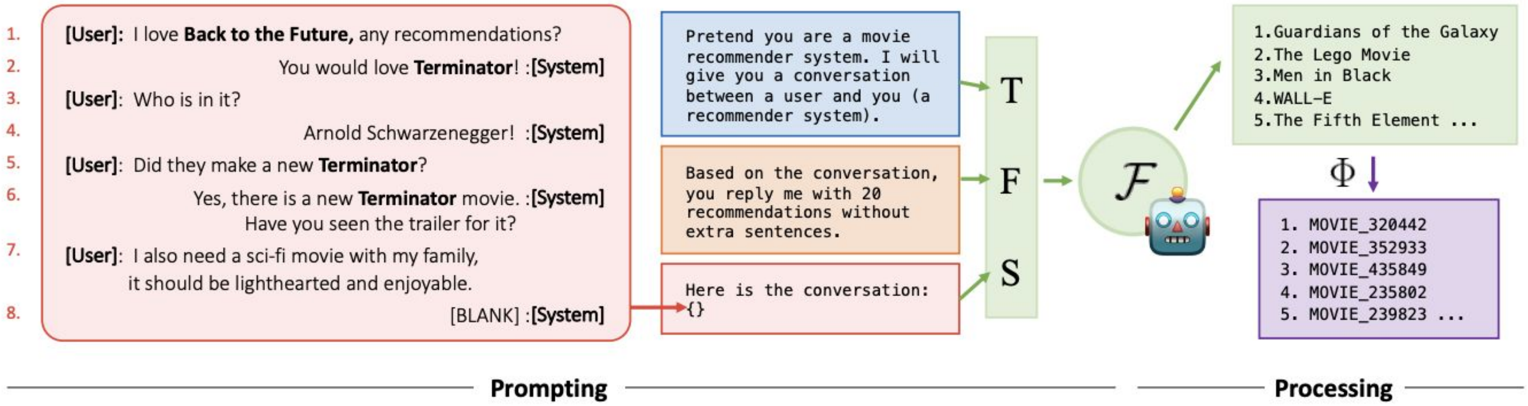
- No weight updates at all
- Zero-, one- and few-shot prompting
- Fast, cheap, easy to iterate
- Unstable; prompt-sensitive

Flexibility & speed

adaptation effort & task alignment decrease • flexibility, speed & generality increase

Examples - Zero-Shot

LLM as zero-shot recommender in a unified system



Z. He. 2023. Large Language Models as Zero-Shot Conversational Recommenders. CIKM '23. https://doi.org/10.1145/3583780.3614949

Examples - Zero-/One-/Few-Shot

General Rule

You are a movie recommender chatbot.
You give movie recommendations to users based on their profile.
Your job now is to fully understand the user profile based on the given context and give them recommendations based on their input.
Here are some rules for you to follow while generating a response:
1: Give an explanation for why each of the recommendations is a good fit for the user;
2: Give a maximum of 5 recommendations, unless specified otherwise by the user;
3: Give a predicted rating for the movie on a scale of 1 to 5: this is a rating the user would give to the movie if they watched it;
4: Mention how popular the movie is. Choose from among High, Medium, Low: High being most popular, Low being least;
5: Avoid recommending movies already rated by the user.

Use this information to understand the user tastes and preferences, based on their genres. Choose a movie that is appropriate based on their profile and request.

Zero-shot Prompting

Favorite genres of the user: [Drama, Comedy, Adventure]

One-shot Prompting

Favorite genres of the user: [Drama, Comedy, Adventure]
Movies recently liked by the user: [Miracles from Heaven (2016), Rating: 5.0/5, Popularity: High]
Movies recently disliked by the user: [Polar Express, The (2004), Rating: 2.5/5, Popularity: High]
Some candidate recommendations for the user: [Babylon 5, Rating: 4.8/5, Popularity: High]

Few-shot Prompting

Favorite genres of the user: [Drama, Comedy, Adventure]
Movies recently liked by the user:
[Miracles from Heaven (2016), Rating: 5.0/5, Popularity: High,
Pride and Prejudice (1995), Rating: 5.0/5, Popularity: High,
Moon (2009), Rating: 5.0/5, Popularity: High,
AlphaGo (2017), Rating: 5.0/5, Popularity: High]
Movies recently disliked by the user:
[Polar Express, The (2004), Rating: 2.5/5, Popularity: High,
Cheaper by the Dozen (2003), Rating: 3.0/5, Popularity: High,
Alice in Wonderland (2010), Rating: 2.5/5, Popularity: High,
Miss Congeniality 2: Armed and Fabulous (2005), Rating: 3.0/5, Popularity: High]
Some candidate recommendations for the user:
[Babylon 5, Rating: 4.8/5, Popularity: High,
Unforgiven (1992), Rating: 4.8/5, Popularity: High,
Man Who Shot Liberty Valance, The (1962), Rating: 4.8/5, Popularity: High,
Patch of Blue, A (1965), Rating: 4.8/5, Popularity: High]

R. Sun, X. Li, A. Akella, and J. A. Konstan, “Large Language Models as Conversational Movie Recommenders: A User Study,” Apr. 29, 2024, arXiv: arXiv:2404.19093. doi: 10.48550/arXiv.2404.19093.

Ahmadou Wagne

Knowledge and Data Foundation

Agenda

- Introduction
- Core Systems & Components
- Foundation Model Integration & Generative Paradigms
- **Knowledge and Data Foundation**
- Simulation & Evaluation
- Open Challenges & Future Directions

Key Questions

- How does item-level knowledge improve both recommendation accuracy and response quality in generative models?
- What balance between pretrained model knowledge and external item- or user-level data is reflected in current Gen-CRS approaches?
- What are the key challenges in integrating structured data into generative pipelines?
- How are conversational data sets and logs consumed by generative models?

Knowledge & Data Foundation: Overview

Item-Level Knowledge

Structured: KGs, DBs, matrices

Unstructured: reviews, descriptions

Multi-modal: image, video, audio

- Up-to-date facts, dynamic catalogues
- Inter-item relations, semantic links
- Grounds recommendations and responses

Item representation

User-Level Knowledge

Short-term: in-dialogue preferences

Long-term: profiles and memory

Persona and personality

- Preferences, behavior, profiles
- Explicit modelling of user signals
- Stable vs. dynamic preferences

User representation

Dialogue Corpora & Logs

Generic: existing datasets used as input for Gen-CRS

LLM-specific: (partially) synthetic dialogues as output of generative models

- H-H conversations & U-S logs
- Generation or enhancement of new corpora
- Used for training and evaluation

NL representation

Knowledge & Data Foundation: Item-Level

Item-Level Knowledge

Structured: KGs, DBs, matrices

▼

Unstructured: reviews, descriptions

▼

Multi-modal: image, video, audio

- Up-to-date facts, dynamic catalogues
- Inter-item relations, semantic links
- Grounds recommendations and responses

Item representation

Item-Level - Structured

Challenges	Types
• Up-to-date item information in **dynamic** environments • **Collaborative signals** accessible to generative models • **Grounding** and **hallucination reduction**	• **Inter-item relations** → knowledge-aware generation • **Item attributes** and **popularity scores** • Implicit and explicit **feedback** on items • Widely used in **modular systems**' recommendation module • **Traditional (C)RS resources**: MovieLens, IMDb, Last.fm…

Item-Level - Unstructured

Extract item properties and map them to **implicit user preferences**

Generate explanations from subjective user experience with items

More **engaging** conversations

Review-based recommendation and generation

Y. Lu *et al.*, "RevCore: Review-Augmented Conversational Recommendation," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds., Online: Association for Computational Linguistics, Aug. 2021, pp. 1161–1173. doi: 10.18653/v1/2021.findings-acl.99

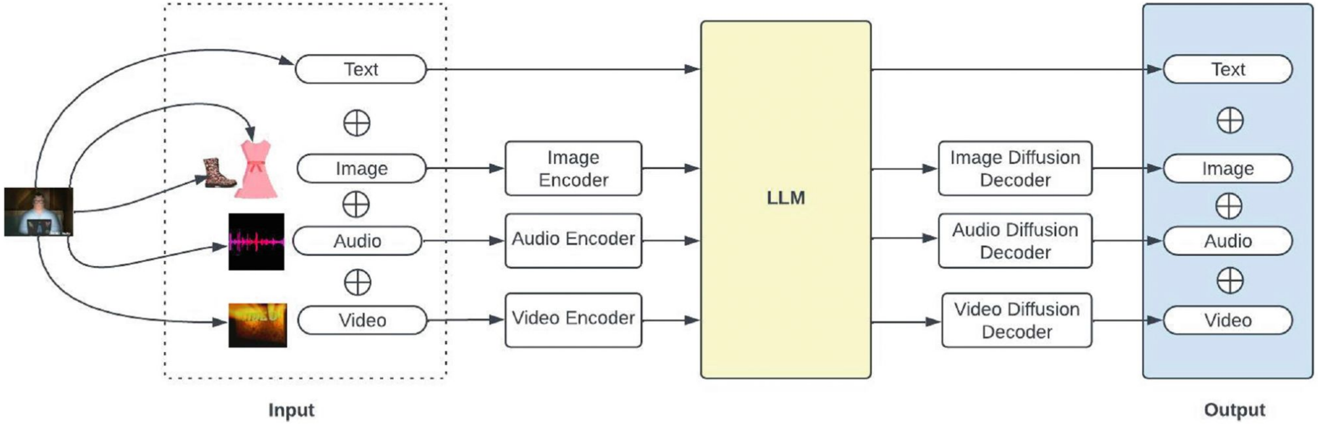
Item-Level - Multi Modal

Multifaceted nature of items

Fusion of different modalities

Image, **video** or **audio**

Multi-modal generative models



T. Mukande, E. Ali, A. Caputo, R. Dong, and N. E. O'Connor, "MMCRec: Towards Multi-modal Generative AI in Conversational Recommendation," in Advances in Information Retrieval: 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24–28, 2024, Proceedings, Part III, Berlin, Heidelberg: Springer-Verlag, Mar. 2024, pp. 316–325. doi: 10.1007/978-3-031-56063-7_23

Knowledge & Data Foundation: User-Level

Item-Level Knowledge

Structured: KGs, DBs, matrices

Unstructured: reviews, descriptions

Multi-modal: image, video, audio

- Up-to-date facts, dynamic catalogues
- Inter-item relations, semantic links
- Grounds recommendations and responses

Item representation

User-Level Knowledge

Short-term: in-dialogue preferences

Long-term: profiles and memory

Persona and personality

- Preferences, behavior, profiles
- Explicit modelling of user signals
- Stable vs. dynamic preferences

User representation

User-Level

- Elicit **short-term preferences**; adapt to **topic/preference shifts**
- Address **weak collaborative knowledge** of generative models
- Handle **multi**-aspect and attribute expressions
- Adapt to user **personas** or **personality**
- **Representation** of **long-term** preferences
- **Over-personalization**
- **Memory management**

User-Level - Short- & Long-Term

- Primary focus: extracting preferences from the **current conversation**
- **Static** vs. **dynamic** user profile
- **Translation** of user profiles into **NL format**
- **Building** structured user profiles **from conversation**
- **Memory management** is crucial for agentic systems

User-Level - Persona & Personality

Persona

External role adopted by the system, or defined to characterize a user interacting with it

- **Demographic background**
- **Role**
- **Perspective**

Personality

Intrinsic traits that shape the nature of the communication

- **Tone**
- **Expressiveness**
- **Consistency of conversation**

Y.-M. Tseng *et al.*, “Two Tales of Persona in LLMs: A Survey of Role-Playing and Personalization,” in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds., Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 16612–16631. doi: 10.18653/v1/2024.findings-emnlp.969

User-Level - Persona & Personality

Approached from a **user** or **system** perspective

Enhances alignment with **user expectations**

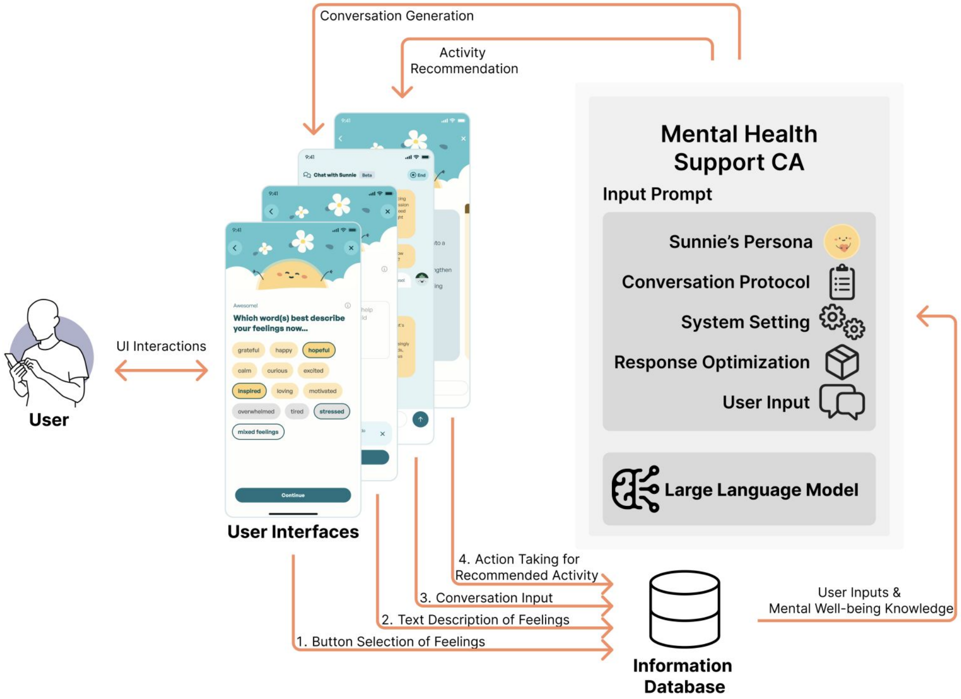
Alleviates **cold start** (e.g. aligning with user demographics)

Risk of reproducing **stereotypes** and **biases**

Guides **style** or **tone** of conversation

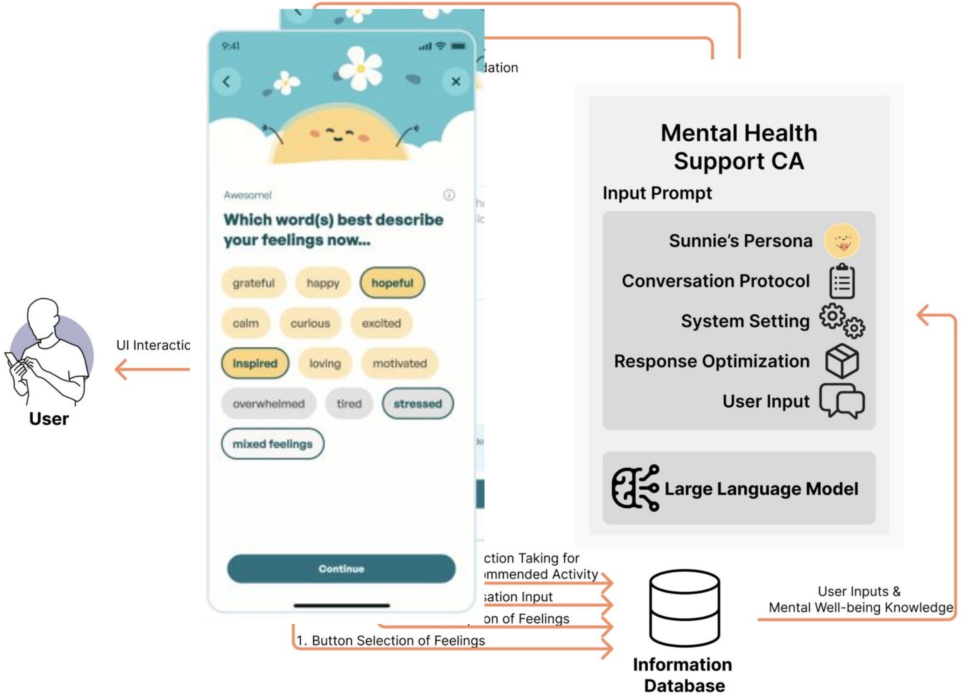
Often used for **user simulation**

Examples - LLM Persona



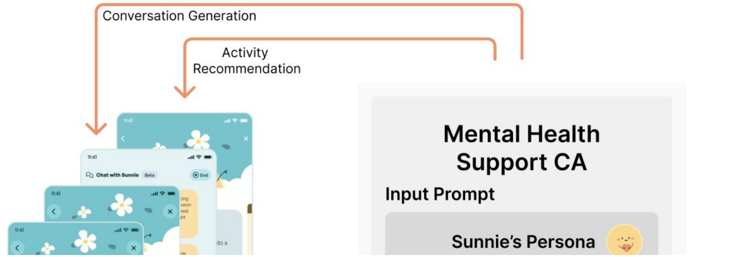
S. Wu *et al.*, "'I Like Sunnie More Than I Expected!': Exploring User Expectation and Perception of an Anthropomorphic LLM-based Conversational Agent for Well-Being Support," Oct. 07, 2024, *arXiv*: arXiv:2405.13803. doi: 10.48550/arXiv.2405.13803

Examples - LLM Persona



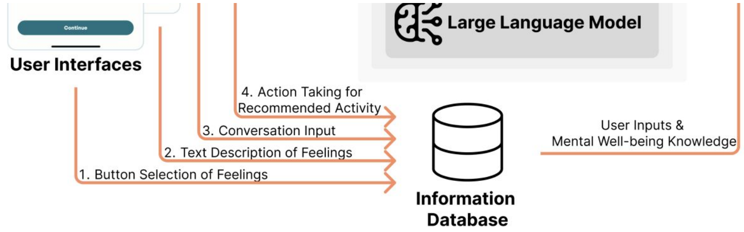
S. Wu *et al.*, "'I Like Sunnie More Than I Expected!': Exploring User Expectation and Perception of an Anthropomorphic LLM-based Conversational Agent for Well-Being Support," Oct. 07, 2024, *arXiv*: arXiv:2405.13803. doi: [10.48550/arXiv.2405.13803](https://doi.org/10.48550/arXiv.2405.13803)

Examples - LLM Persona



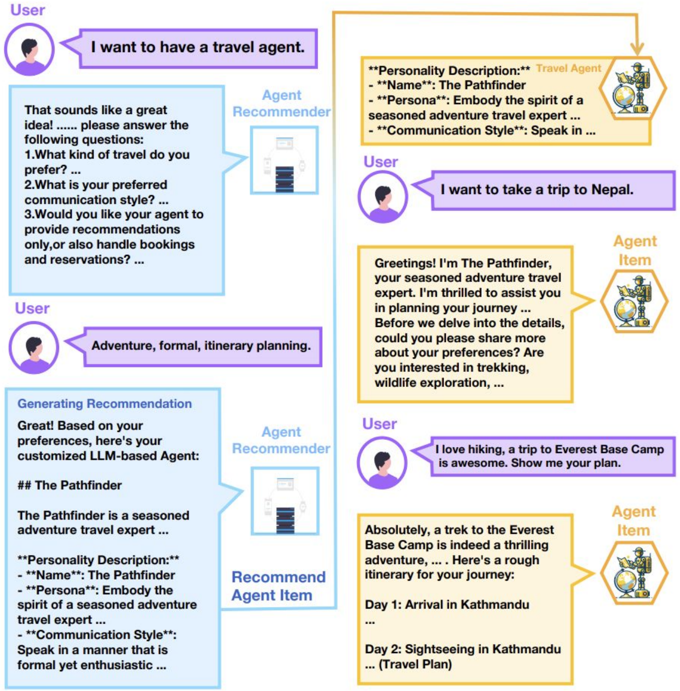
Sunnie's Persona

Sunnie is a compassionate, supportive, and insightful buddy … … offers understanding, empathy, and relevant psychological knowledge … … Sunnie likes to add emojis to make it more fun. … …



S. Wu *et al.*, "'I Like Sunnie More Than I Expected!': Exploring User Expectation and Perception of an Anthropomorphic LLM-based Conversational Agent for Well-Being Support," Oct. 07, 2024, *arXiv*: arXiv:2405.13803. doi: [10.48550/arXiv.2405.13803](https://doi.org/10.48550/arXiv.2405.13803)

Examples - Agent Personality



Zhang, J. et al. "Prospect personalized recommendation on large language model-based agent platform." *arXiv preprint arXiv:2402.18240* (2024).

Knowledge & Data Foundation: Dialogue Data

Item-Level Knowledge

Structured: KGs, DBs, matrices

Unstructured: reviews, descriptions

Multi-modal: image, video, audio

- Up-to-date facts, dynamic catalogues
- Inter-item relations, semantic links
- Grounds recommendations and responses

Item representation

User-Level Knowledge

Short-term: in-dialogue preferences

Long-term: profiles and memory

Persona and personality

- Preferences, behavior, profiles
- Explicit modelling of user signals
- Stable vs. dynamic preferences

User representation

Dialogue Corpora & Logs

Generic: existing datasets used as input for Gen-CRS

LLM-specific: (partially) synthetic dialogues as output of generative models

- H-H conversations & U-S logs
- Generation or enhancement of new corpora
- Used for training and evaluation

NL representation

Dialogue Corpora & Logs

How are conversational data sets and logs consumed by generative models?

Input

36 of 49 selected papers used a dialogue corpora or logs dataset:

- HH Conv (24)
- user-system log (12)

Output

15 papers used LLMs for generating and/or enhancing the dialogue corpora/logs data.

Dialogue Corpora & Logs

How are conversational data sets and logs consumed by generative models?

Input

36 of 49 selected papers used a dialogue corpora or logs dataset:

- HH Conv (24)
- user-system log (12)

Human Human Conversation: ReDial, etc.
User-System Log: Amazon Review Dataset, etc.

Output

15 papers used LLMs for generating and/or enhancing the dialogue corpora / logs data.

Thomas E. Kolb

Simulation & Evaluation

Agenda

- Introduction
- Core Systems & Components
- Foundation Model Integration & Generative Paradigms
- Knowledge and Data Foundation
- **Simulation** & Evaluation
- Open Challenges & Future Directions

Key Questions

- What is the goal of (generative) simulation?
- Which aspects simulation do we have?
- How can generative approaches (e.g. LLMs) enable these simulations?
- Which new datasets are in the field?

Why simulation in Gen-CRS*?

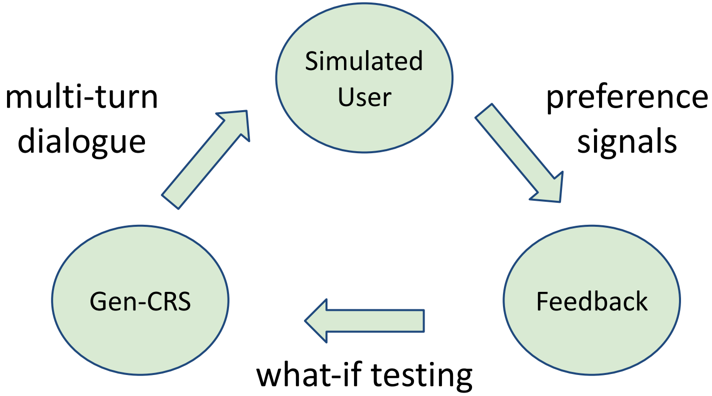
The researcher's challenge:

**Generative Conversational Recommender Systems*

Why simulation in Gen-CRS?

The researcher's challenge:

With simulation:

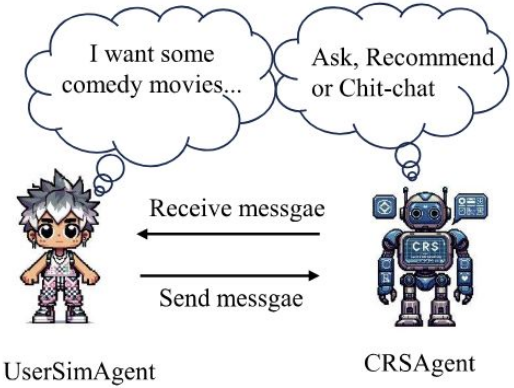


<u>Goal</u>: Creation of a conversational dataset with the help of simulating user behaviour.

Why simulation in Gen-CRS?

The researcher's challenge:

New paradigm to support Gen-CRS evaluation!



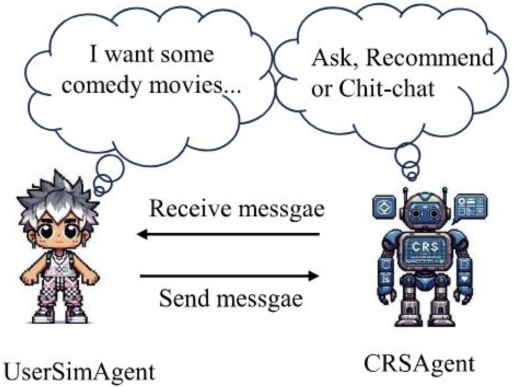
Zhu, L., Huang, X., & Sang, J. (2025). A LLM-based Controllable, Scalable, Human-Involved User Simulator Framework for Conversational Recommender Systems. In *Proceedings of the ACM on Web Conference 2025* (pp. 4653–4661). Association for Computing Machinery.

Why simulation in Gen-CRS?

The researcher's challenge:

How does this help us?
⟶ Scale - Control - Safety - Counterfactual testing

New paradigm to support Gen-CRS evaluation!



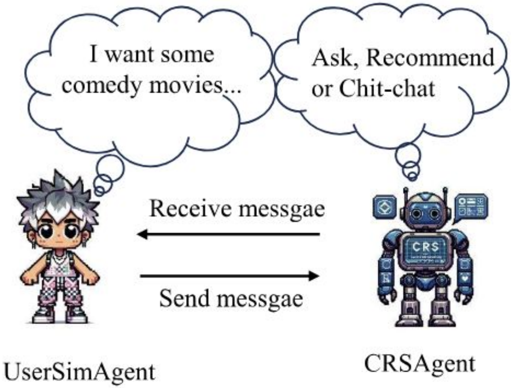
Zhu, L., Huang, X., & Sang, J. (2025). A LLM-based Controllable, Scalable, Human-Involved User Simulator Framework for Conversational Recommender Systems. In *Proceedings of the ACM on Web Conference 2025* (pp. 4653–4661). Association for Computing Machinery.

Why simulation in Gen-CRS?

The researcher's challenge:

How does this help us?
→ Scale - Control - Safety - Counterfactual testing

New paradigm to support Gen-CRS evaluation!



Simulation enables us to build realistic conversational test environments for system development and evaluation.

Zhu, L., Huang, X., & Sang, J. (2025). A LLM-based Controllable, Scalable, Human-Involved User Simulator Framework for Conversational Recommender Systems. In *Proceedings of the ACM on Web Conference 2025* (pp. 4653–4661). Association for Computing Machinery.

From Logs to Full Conversational Simulation

Observational Data

real logs; no
reactive user

A conversational dataset that is composed of:

- ➔ Preference Data (= Task/Goal Orientation)
- ➔ Multi-turn Conversation
- ➔ Item Data

Less Control & Bias,
More Authentic/Real

More Control & Bias,
Less Authentic/Real

From Logs to Full Conversational Simulation

Observational Data → **Synthetic Data**

real logs; no reactive user

Generated dialogues resembling real interactions

New approaches:

- ➔ Data synthesized from real sources: combine existing signals e.g. user preferences, item metadata, dialogue acts into richer datasets.
- ➔ Model generated data: create artificial dialogues and interactions.

Less Control & Bias, More Authentic/Real ⟷ *More Control & Bias, Less Authentic/Real*

From Logs to Full Conversational Simulation

Observational Data

real logs; no reactive user

Synthetic Data

Generated dialogues resembling real interactions

User Simulation

Reactive user model turn-by-turn interaction

Less Control & Bias, More Authentic/Real

More Control & Bias, Less Authentic/Real

From Logs to Full Conversational Simulation

Observational Data

real logs; no reactive user

Synthetic Data

Generated dialogues resembling real interactions

User Simulation

Reactive user model turn-by-turn interaction

A reactive model that ***plays the role of a user*** *in a live conversation, producing utterances, clarifications, and preference signals in response to each system move, thereby enabling* ***turn-by-turn evaluation or policy learning****.*

Less Control & Bias, More Authentic/Real

More Control & Bias, Less Authentic/Real

From Logs to Full Conversational Simulation

Observational Data
real logs; no reactive user

→

Synthetic Data
Generated dialogues resembling real interactions

→

User Simulation
Reactive user model turn-by-turn interaction

→

Full Conversational Simulation
user + environment modeled in a closed loop

Less Control & Bias, More Authentic/Real ⟷ *More Control & Bias, Less Authentic/Real*

From Logs to Full Conversational Simulation

Observational Data

real logs; no reactive user

Synthetic Data

Generated dialogues resembling real interactions

User Simulation

Reactive user model turn-by-turn interaction

Full Conversational Simulation

user + environment modeled in a closed loop

*A complete **closed-loop environment** in which both the **user(s) and the external world are modeled**, allowing the recommender to act, learn, and observe emergent behaviour over many simulated dialogues—analogous to **running the entire ecosystem in silico**.*

Simulation has levels: from data generation to full interactive environments

Less Control & Bias, More Authentic/Real

More Control & Bias, Less Authentic/Real

What makes a Good User Simulator?

Persona / Profile

?

User Simulation

What makes a Good User Simulator?

Persona / Profile

Item Knowledge

?

User Simulation

What makes a Good User Simulator?

Persona / Profile

Item Knowledge

Feedback Signals

?

User Simulation

What makes a Good User Simulator?

- Persona / Profile
- Item Knowledge
- Feedback Signals

?

User Simulation

- Goals & Preferences

What makes a Good User Simulator?

Persona / Profile

Item Knowledge

Feedback Signals

?

User Simulation

Goals & Preferences

Dialogue Behaviour

What makes a Good User Simulator?

Persona / Profile

Item Knowledge

Feedback Signals

?

User Simulation

Goals & Preferences

Dialogue Behaviour

Control Knobs

What makes a Good User Simulator?

Persona / Profile

Item Knowledge

Feedback Signals

?

User Simulation

Goals & Preferences

Dialogue Behaviour

Control Knobs

Outputs?
utterances - clarifications - accept / reject signals - explanation requests

What makes a Good User Simulator?

?

Persona / Profile

Item Knowledge

Feedback Signals

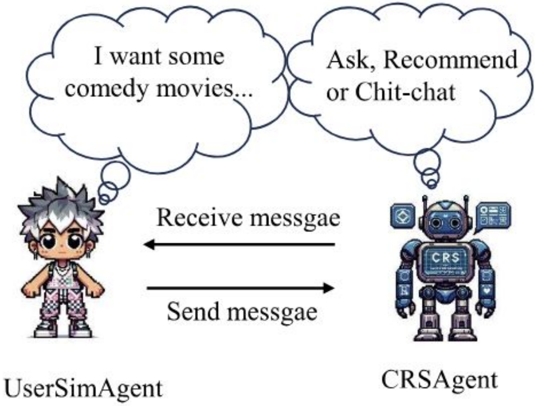
User Simulation

Goals & Preferences

Dialogue Behaviour

Control Knobs

Outputs?
utterances - clarifications - accept / reject signals - explanation requests



Zhu, L., Huang, X., & Sang, J. (2025). A LLM-based Controllable, Scalable, Human-Involved User Simulator Framework for Conversational Recommender Systems. In *Proceedings of the ACM on Web Conference 2025* (pp. 4653–4661). Association for Computing Machinery.

Representative Papers in Simulation

1.) ReDial (Li et al., 2018)
Early crowdsourced CRS bechmark.

datasets **simulators** **environments**

2018 **2026**

1.) ReDial (Li et al., 2018)

# conversations	10006
# utterances	182150
# users	956
# movie mentions	51699
seeker mentioned	16278
recommender suggested	35421
not seen	16516
seen	31694
did not say	3489
disliked (4.9%)	2556
liked (81%)	41998
did not say (14%)	7145

HUMAN:	hello, can i help you find a movie ?
HRED:	hi what kind of movies do you like
OURS:	hi, how can i help you ?
SEEKER:	hello ! i would like to find a suspenseful, but clean family friendly movie .
...	...
SEEKER:	2001 : a space odyssey might be a great option. anything else that you would suggest ?
HUMAN:	you can try planet of the apes the older one is quite suspenseful and family friendly .
HRED:	i haven't seen that one yet but i heard it was good. i haven't seen that one. have you seen the last house on the left ?
OURS:	star wars : the force awakens is also a good one return of the jedi all good movies
SEEKER:	those sound good ! i 'm going to look into those movies.
HUMAN:	i hope you enjoy, have a nice one
HRED:	have you seen foxcatcher ? it 's about a man who has a rich guy.
OURS:	i hope i was able to help you find a good movie to watch
SEEKER:	thank you for your help ! have a great night ! good bye

Table 1a. (Above) ReDial data statistics. For the movie dialogue forms, the numbers shown represent the seeker's answers. **Table 1b.**(Right) Conversation excerpts (HUMAN followed by response by SEEKER) and model outputs (OUR proposed approach compared to HRED a generic dialogue model [2]).

ReDial (2018)

- Crowd-Sourcing via Amazon Mechanical Turk
- Movie Seeker & Recommender as Roles

One of the early datasets highly used in conversational recommender systems research.

Li, R., Kahou, S., Schulz, H., Michalski, V., Charlin, L., & Pal, C. (2018). Towards deep conversational recommendations. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (pp. 9748–9758). Curran Associates Inc..

Representative Papers in Simulation

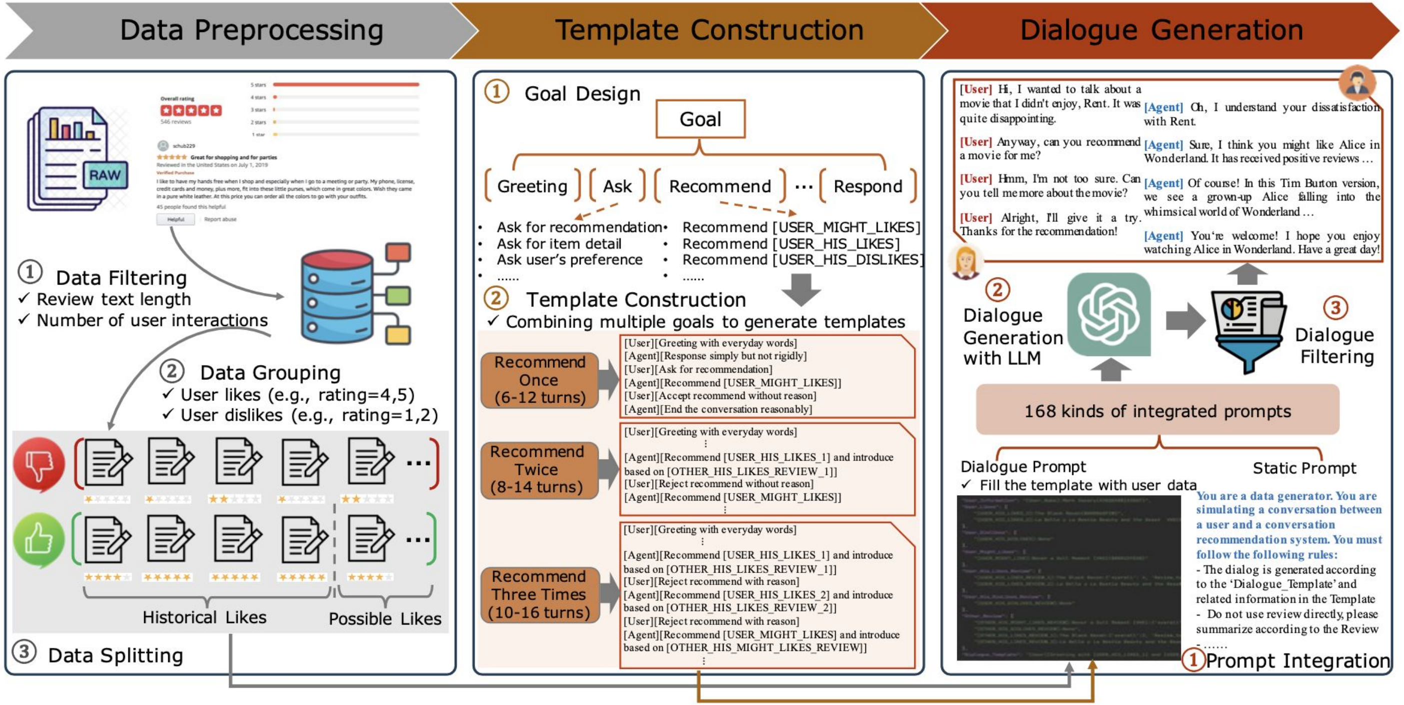
1.) ReDial (Li et al., 2018)
Early crowdsourced CRS bechmark.

2.) LLM-REDIAL (Liang et al., 2024)
LLM-created large-scale conversational data from user behaviors.

datasets **simulators** **environments**

2018 **2026**

2.) LLM-Redial (Liang et al., 2024)



Liang, T., Jin, C., Wang, L., Fan, W., Xia, C., Chen, K., & Yin, Y. (2024). LLM-REDIAL: A Large-Scale Dataset for Conversational Recommender Systems Created from User Behaviors with LLMs. ACL 2024 (pp. 8926–8939). Association for Computational Linguistics.

Representative Papers in Simulation

1.) ReDial (Li et al., 2018)
Early crowdsourced CRS bechmark.

2.) LLM-REDIAL (Liang et al., 2024)
LLM-created large-scale conversational data from user behaviors.

3.) Friedman et al. (2023)
Controllable LLM user simulator for full sessions.

datasets **simulators** **environments**

2018 **2026**

3.) Friedman et al. (2023)

Leveraging Large Language Models in Conversational Recommender Systems (Friedman et al., Google Research 2023)

An LLM user simulator interacts with the CRS to produce full sessions; controllable via session- or turn-level variables.

RecLLM, a large-scale CRS for YouTube videos built on LaMDA

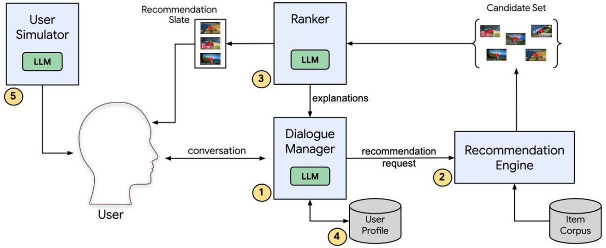


Figure 1: Overview of key contributions from RecLLM. (1) A dialogue management module uses an LLM to converse with the user, track context and make system calls such as submitting a request to a recommendation engine all as a unified language modeling task. (2) Various solutions are presented for tractable retrieval over a large item corpus within an LLM-based CRS. (3) A ranker module uses an LLM to match preferences extracted from the context of the conversation to item metadata and generate a slate of recommendations that is displayed to the user. The LLM also jointly generates explanations for its decisions that can be surfaced to the user. (4) Interpretable natural language user profiles are consumed by system LLMs to modulate session-level context and increase personalization. (5) A controllable LLM-based user simulator can be plugged into the CRS to generate synthetic conversations for tuning system modules.

Input: **<u>highlight the lack of logs or observational dialogue corpora</u>** as a central challenge
Output: generate dialogue sessions, user feedback, item summaries, and recommendation utterances with llms

Friedman, L., Ahuja, S., Allen, D., Tan, Z., Sidahmed, H., Long, C., Xie, J., Schubiner, G., Patel, A., Lara, H., & others. (2023). Leveraging large language models in conversational recommender systems. ArXiv Preprint ArXiv:2305.07961.

3.) Friedman et al. (2023)

Leveraging Large Language Models in Conversational Recommender Systems (Friedman et al., Google Research 2023)

"In this paper we assume a simplified setting where users interact with the system ***only through conversation****. We would like to* ***generalize our system to handle more realistic scenarios where users give feedback through other channels as well such as clicking on items or like buttons****. We would also like to consider more complicated recommender system UIs containing hierarchical structures such as item shelves as opposed to just flat slates."* (Friedman et al., Google Research 2023)

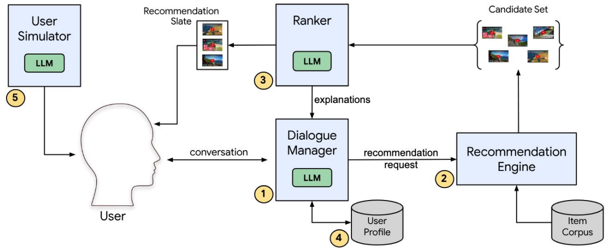


Figure 1: Overview of key contributions from RecLLM. (1) A dialogue management module uses an LLM to converse with the user, track context and make system calls such as submitting a request to a recommendation engine all as a unified language modeling task. (2) Various solutions are presented for tractable retrieval over a large item corpus within an LLM-based CRS. (3) A ranker module uses an LLM to match preferences extracted from the context of the conversation to item metadata and generate a slate of recommendations that is displayed to the user. The LLM also jointly generates explanations for its decisions that can be surfaced to the user. (4) Interpretable natural language user profiles are consumed by system LLMs to modulate session-level context and increase personalization. (5) A controllable LLM-based user simulator can be plugged into the CRS to generate synthetic conversations for tuning system modules.

gue corpora as a central challenge

m summaries, and

Friedman, L., Ahuja, S., Allen, D., Tan, Z., Sidahmed, H., Long, C., Xie, J., Schubiner, G., Patel, A., Lara, H., & others. (2023). Leveraging large language models in conversational recommender systems. ArXiv Preprint ArXiv:2305.07961.

Representative Papers in Simulation

1.) ReDial (Li et al., 2018)
Early crowdsourced CRS bechmark.

3.) Friedman et al. (2023)
Controllable LLM user simulator for full sessions.

2.) LLM-REDIAL (Liang et al., 2024)
LLM-created large-scale conversational data from user behaviors.

4.) Zhu et al. (2024/2025)
Reliability analysis and scalable controllable simulators.

datasets **simulators** **environments**

2018 **2026**

4.) Zhu et al. (2024/2025)

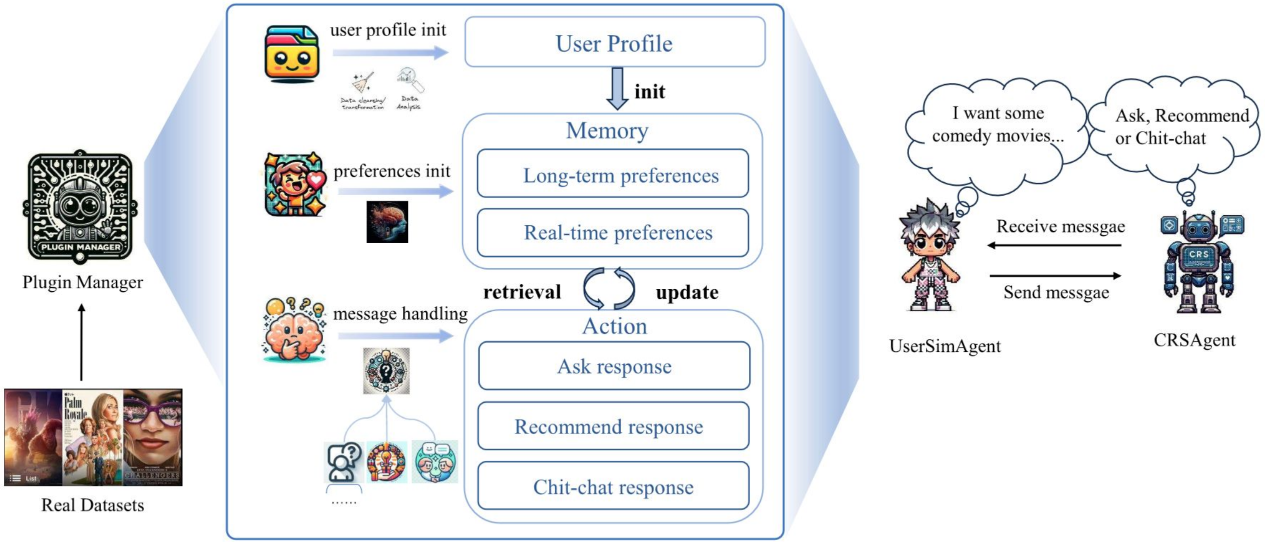


Figure 1: CSHI framework overview.

Zhu, L., Huang, X., & Sang, J. (2025). A LLM-based Controllable, Scalable, Human-Involved User Simulator Framework for Conversational Recommender Systems. Proceedings of the ACM on Web Conference 2025, 4653–4661. https://doi.org/10.1145/3696410.3714858

Representative Papers in Simulation

1.) ReDial (Li et al., 2018)
Early crowdsourced CRS bechmark.

3.) Friedman et al. (2023)
Controllable LLM user simulator for full sessions.

5.) RecoWorld (Liu et al., 2026)
Simulated environments for agentic recommender systems.

2.) LLM-REDIAL (Liang et al., 2024)
LLM-created large-scale conversational data from user behaviors.

4.) Zhu et al. (2024/2025)
Reliability analysis and scalable controllable simulators.

datasets **simulators** **environments**

2018 **2026**

5.) RecoWorld (Liu et al., 2026)

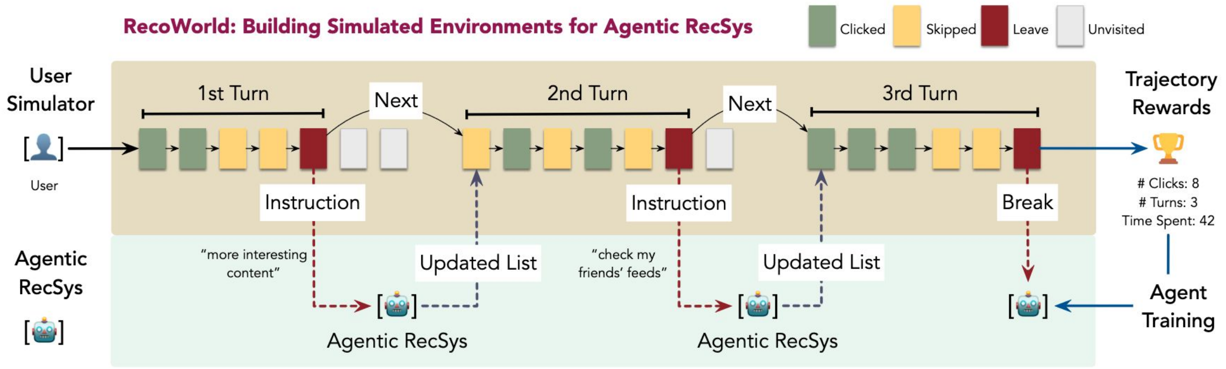


Figure 1: A simulated user interacts with an agentic RecSys over multiple turns within a session.

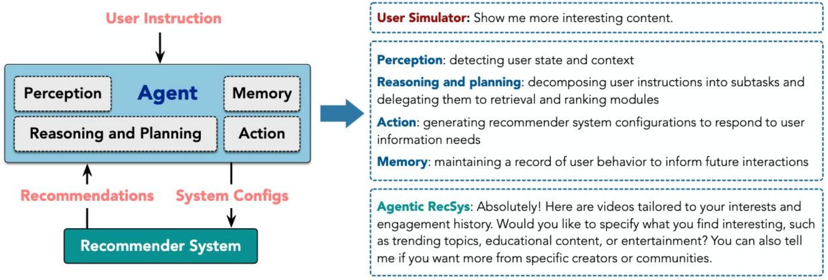
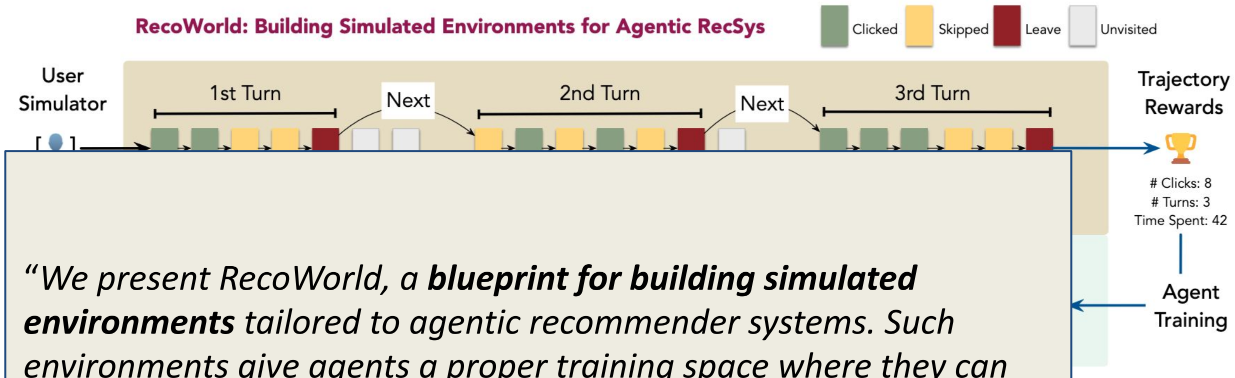


Figure 3: An instruction-following recommender can be powered by an autonomous agent.

Fei Liu, Xinyu Lin, Hanchao Yu, Mingyuan Wu, Jianyu Wang, Qiang Zhang, Zhuokai Zhao, Yinglong Xia, Yao Zhang, Weiwei Li, Mingze Gao, Qifan Wang, Lizhu Zhang, Benyu Zhang, and Xiangjun Fan. 2026. RecoWorld: Building Simulated Environments for Agentic Recommender Systems. In Companion Proceedings of the ACM Web Conference 2026 (WWW Companion '26). Association for Computing Machinery, New York, NY, USA, 650–659. https://doi.org/10.1145/3774905.3794661

5.) RecoWorld (Liu et al., 2026)



*"We present RecoWorld, a **blueprint for building simulated environments** tailored to agentic recommender systems. Such environments give agents a proper training space where they can learn from errors without impacting real users. RecoWorld distinguishes itself with a dual-view architecture: **a simulated user and an agentic recommender engage in multi-turn interactions** aimed at maximizing user retention."* (Liu et al., Meta, 2026)

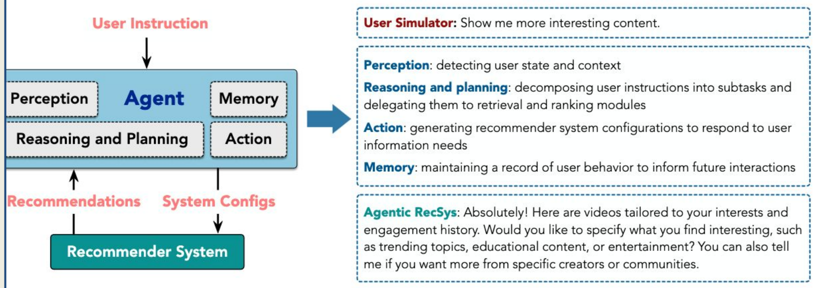


Figure 3: An instruction-following recommender can be powered by an autonomous agent.

Fei Liu, Xinyu Lin, Hanchao Yu, Mingyuan Wu, Jianyu Wang, Qiang Zhang, Zhuokai Zhao, Yinglong Xia, Yao Zhang, Weiwei Li, Mingze Gao, Qifan Wang, Lizhu Zhang, Benyu Zhang, and Xiangjun Fan. 2026. RecoWorld: Building Simulated Environments for Agentic Recommender Systems. In Companion Proceedings of the ACM Web Conference 2026 (WWW Companion '26). Association for Computing Machinery, New York, NY, USA, 650–659. https://doi.org/10.1145/3774905.3794661

From Classical to Generative Datasets

Classical Datasets

ReDial (Li et al., 2018)
OpenDialKG (Moon et al., 2019)
GoRecDial (Kang et al., 2019)
INSPIRED (Hayati et al., 2020)
INSPIRED2 (Manzoor et al., 2022)
DuRecDial (Liu et al., 2020)
DuRecDial 2.0 (Liu et al., 2021)
U-NEED (Liu et al., 2023)

(Towards) Generative Datasets

TG-ReDial (Zhou et al., 2020) - **synthetic**
Synthetically self generated data (Friedman et al., 2023)
LLM-Redial (Liang et al., 2024)
PEARL (Kim et al., 2024)
N-of-1 (Synthetic) Dataset (Yang et al., 2024)
Synthesized Tracking & Recommendation Dataset (Ashby et al., 2024)
DistillRecDial (Martina et al., 2025)

From Classical to Generative Datasets (cont.)

Classical Datasets → **(Towards) Generative Datasets**

*"To overcome conversational data limitations in the **absence of an existing production CRS**, we propose techniques for building a controllable LLM-based **user simulator** to **generate synthetic conversations**." (Friedman et al., 2023 - Google Research)*

TG-ReDial (Zhou et al., 2020) - **synthetic**
Synthetically self generated data (Friedman et al., 2023)
LLM-Redial (Liang et al., 2024)
PEARL (Kim et al., 2024)
N-of-1 (Synthetic) Dataset (Yang et al., 2024)
Synthesized Tracking & Recommendation Dataset (Ashby et al., 2024)
DistillRecDial (Martina et al., 2025)

From Classical to Generative Datasets (cont.)

Classical Datasets → **(Towards) Generative Datasets**

No public dataset available ✖

TG-ReDial (Zhou et al., 2020) - **synthetic**
Synthetically self generated data (Friedman et al., 2023)
LLM-Redial (Liang et al., 2024)
PEARL (Kim et al., 2024)
N-of-1 (Synthetic) Dataset (Yang et al., 2024)
Synthesized Tracking & Recommendation Dataset (Ashby et al., 2024)
DistillRecDial (Martina et al., 2025)

From Classical to Generative Datasets (cont.)

Classical Datasets → **(Towards) Generative Datasets**

Dataset available ✓

TG-ReDial (Zhou et al., 2020) - **synthetic**
Synthetically self generated data (Friedman et al., 2023)
LLM-Redial (Liang et al., 2024)
PEARL (Kim et al., 2024)
N-of-1 (Synthetic) Dataset (Yang et al., 2024)
Synthesized Tracking & Recommendation Dataset (Ashby et al., 2024)
DistillRecDial (Martina et al., 2025)

Comparison (ReDial vs LLM-ReDial)

ReDial 2018 (Li et al.)

Real human–human dialogues

Roles: “**seeker**” ↔ “**recommender**”

10k dialogues / **182.1k** utterances, **single-domain** (movies)

Redial (Li et al., 2018): Li, R., Kahou, S., Schulz, H., Michalski, V., Charlin, L., & Pal, C. (2018). Towards Deep Conversational Recommendations. In Advances in Neural Information Processing Systems 31 (NIPS 2018).

LLM-ReDial 2024 (Liang et al.)

Conversations are **generated by LLMs (GPT-3.5-turbo)**

Roles**: turn goal templates**

Amazon review logs + user histories, guided by **turn-goal templates** to simulate the recommendation flow.

47.6k dialogues / **482.6k** utterances across **4 domains** (Books, Movies, Sports, Electronics)

LLM-REDIAL (Liang et al., 2024): Liang, T., Jin, C., Wang, L., Fan, W., Xia, C., Chen, K., & Yin, Y. (2024). LLM-REDIAL: A Large-Scale Dataset for Conversational Recommender Systems Created from User Behaviors with LLMs. In Findings of the Association for Computational Linguistics: ACL 2024 (pp. 8926–8939). Association for Computational Linguistics.

Comparison (ReDial vs LLM-ReDial)

ReDial 2

Real human–h

Roles: “**seeker**

10k dialogues

single-domain

Redial (Li et al., 2018): Li, R., Ka
Deep Conversational Recommer
(NIPS 2018).

User ID: A1EXXXXXDUE6B0

Historical Interactions: ["Robin Williams: Live On Broadway", "**Mission Impossible on VHS**", "**Solaris**", "**Elysium**", "Wall Street", "Mystic River", ...]

Conversation 1:

[User] Hi, I recently watched **Mission Impossible on VHS** and it was a fantastic high-tech spy movie! Tom Cruise gave …

[Agent] That's great to hear! I have a movie recommendation for you based on your likes. How about checking out World War Z?

[User] Oh, I'm definitely interested. Can you tell me more about the movie?

[Agent] Certainly! World War Z is a good zombie war flick…

Conversation 2:

[User] Hi, I recently watched the movie "**Solaris**" and I have to say, I didn't enjoy it at all. It felt overly melodramatic and lacked substance…

[Agent] I understand why you didn't like “Solaris”. I can recommend a movie you might enjoy instead. How about “**Elysium**”? It’s a bilingual film that…

[User] Oh, I've actually already watched "Elysium" and it was better than I expected, but still not great.…

[Agent] How about giving "Wrecking Crew" a try? It's another movie you might like based on your previous preferences…

[User] Sure, that sounds interesting…

Conversation in Redial

[User] Hi I am looking for a movie like Super Troopers (2001)

[Agent] You should watch Police Academy (1984)

[User] Is that a great one? I have never seen it. I have seen American Pie I mean American Pie (1999)

[Agent] Yes Police Academy (1984) is very funny and so is Police Academy 2: Their First Assignment (1985)

[User] It sounds like I need to check them out'

[Agent] yes you will enjoy them

[User] I appreciate your time. I will need to check those out. Are there any others you would recommend?

[Agent] yes Lethal Weapon (1987)

[User] Thank you i will watch that too

[Agent] and also Beverly Hills Cop (1984)

[User] Thanks for the suggestions.

[Agent] you are welcome and also 48 Hrs. (1982)

l.)

GPT-3.5-turbo)

uided by

ss **4**

nics)

hen, K., & Yin, Y. (2024).
Created from User Behaviors
(pp. 8926–8939). Association

From Classical to Generative Datasets (cont.)

Redial (Li et al., 2018): Li, R., Kahou, S., Schulz, H., Michalski, V., Charlin, L., & Pal, C. (2018). Towards Deep Conversational Recommendations. In Advances in Neural Information Processing Systems 31 (NIPS 2018).
OpenDialKG (Moon et al., 2019): Moon, S., Shah, P., Kumar, A., & Subba, R. (2019). OpenDialKG: Explainable Conversational Reasoning with Attention-based Walks over Knowledge Graphs. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 845–854). Association for Computational Linguistics.
GoRecDial (Kang et al., 2019): Kang, D., Balakrishnan, A., Shah, P., Crook, P., Boureau, Y.L., & Weston, J. (2019). Recommendation as a Communication Game: Self-Supervised Bot-Play for Goal-oriented Dialogue. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 1951–1961). Association for Computational Linguistics.
INSPIRED (Hayati et al., 2020): Hayati, S., Kang, D., Zhu, Q., Shi, W., & Yu, Z. (2020). INSPIRED: Toward Sociable Recommendation Dialog Systems. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 8142–8152). Association for Computational Linguistics.
DuRecDial (Liu et al., 2020): Liu, Z., Wang, H., Niu, Z.Y., Wu, H., Che, W., & Liu, T. (2020). Towards Conversational Recommendation over Multi-Type Dialogs. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 1036–1049). Association for Computational Linguistics.
DuRecDial2.0 (Liu et al., 2021): Liu, Z., Wang, H., Niu, Z.Y., Wu, H., & Che, W. (2021). DuRecDial 2.0: A Bilingual Parallel Corpus for Conversational Recommendation. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (pp. 4335–4347). Association for Computational Linguistics.
INSPIRED2 (Manzoor et al., 2022): Ahtsham Manzoor, & Dietmar Jannach (2022). INSPIRED2: An Improved Dataset for Sociable Conversational Recommendation. In Proceedings of the Fourth Knowledge-aware and Conversational Recommender Systems Workshop (KaRS 2022), co-located with the 16th ACM Conference on Recommender Systems (RecSys 2022) (pp. 73–80). CEUR-WS.org.
U-NEED (Liu et al., 2023): Liu, Y., Zhang, W., Dong, B., Fan, Y., Wang, H., Feng, F., Chen, Y., Zhuang, Z., Cui, H., Li, Y., & Che, W. (2023). U-NEED: A Fine-grained Dataset for User Needs-Centric E-commerce Conversational Recommendation. In Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 2723–2732). Association for Computing Machinery.
TG-REDIAL (Zhou et al., 2020): Zhou, K., Zhou, Y., Zhao, W., Wang, X., & Wen, J.R. (2020). Towards Topic-Guided Conversational Recommender System. In Proceedings of the 28th International Conference on Computational Linguistics (pp. 4128–4139). International Committee on Computational Linguistics.
Synthetically self generated data (Friedman et al., 2023): Luke Friedman, Sameer Ahuja, David Allen, Zhenning Tan, Hakim Sidahmed, Changbo Long, Jun Xie, Gabriel Schubiner, Ajay Patel, Harsh Lara, Brian Chu, Zexi Chen, & Manoj Tiwari. (2023). Leveraging Large Language Models in Conversational Recommender Systems.
LLM-REDIAL (Liang et al., 2024): Liang, T., Jin, C., Wang, L., Fan, W., Xia, C., Chen, K., & Yin, Y. (2024). LLM-REDIAL: A Large-Scale Dataset for Conversational Recommender Systems Created from User Behaviors with LLMs. In Findings of the Association for Computational Linguistics: ACL 2024 (pp. 8926–8939). Association for Computational Linguistics.
PEARL (Kim et al., 2024): Kim, M., Kim, M., Kim, H., Kwak, B.w., Kang, S., Yu, Y., Yeo, J., & Lee, D. (2024). Pearl: A Review-driven Persona-Knowledge Grounded Conversational Recommendation Dataset. In Findings of the Association for Computational Linguistics: ACL 2024 (pp. 1105–1120). Association for Computational Linguistics.
N-of-1 (Synthetic) Dataset (Yang et al., 2024): Zhongqi Yang, Elahe Khatibi, Nitish Nagesh, Mahyar Abbasian, Iman Azimi, Ramesh Jain, & Amir M. Rahmani (2024). ChatDiet: Empowering personalized nutrition-oriented food recommender chatbots through an LLM-augmented framework. Smart Health, 32, 100465.
Synthesized Tracking Dataset, Synthesized Recommendation Dataset (Ashby et al., 2024): Ashby, T., Kulkarni, A., Qi, J., Liu, M., Cho, E., Kumar, V., & Huang, L. (2024). Towards Effective Long Conversation Generation with Dynamic Topic Tracking and Recommendation. In Proceedings of the 17th International Natural Language Generation Conference (pp. 540–556). Association for Computational Linguistics.
DistillRecDial (Martina et al., 2025): Martina, A., Petruzzelli, A., Musto, C., Gemmis, M., Lops, P., & Semeraro, G. (2025). DistillRecDial: A Knowledge-Distilled Dataset Capturing User Diversity in Conversational Recommendation. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems* (pp. 726–735). Association for Computing Machinery.

Open Challenges in Simulation

Realism ⇔ Control

Open Challenges in Simulation

- Authenticity vs. controllability
- Simulator bias propagation
- Long-term memory & evolving preferences

Realism ⇔ Control

- Rare cases and long-tail behavior
- Grounding to the external world
- Simulator validation

Open Challenges in Simulation

Authenticity vs. controllability

Simulator bias propagation

Long-term memory & evolving preferences

Realism ⇔ Control

Rare cases and long-tail behavior

Grounding to the external world

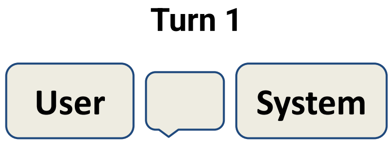
Simulator validation

The simulator is both a tool and an object of evaluation.

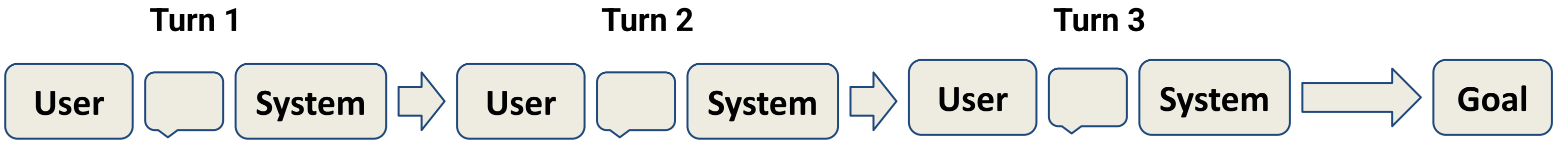
Agenda

- Introduction
- Core Systems & Components
- Foundation Model Integration & Generative Paradigms
- Knowledge and Data Foundation
- Simulation & **Evaluation**
- Open Challenges & Future Directions

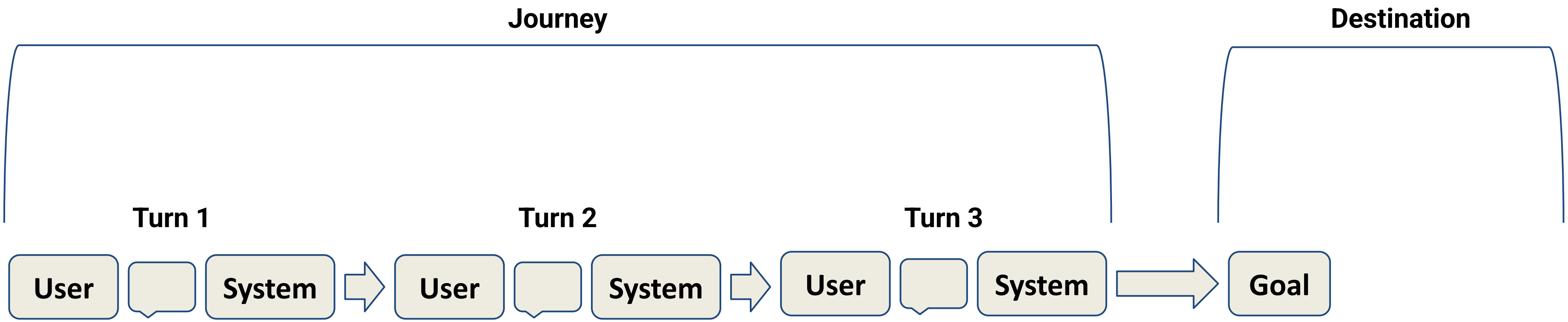
What Should We Evaluate in Gen-CRS?



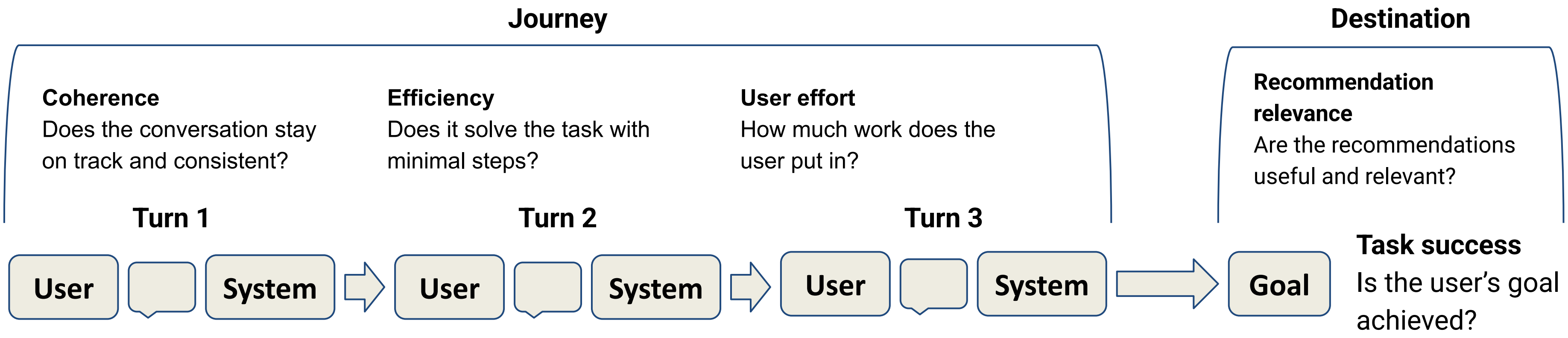
What Should We Evaluate in Gen-CRS?



What Should We Evaluate in Gen-CRS?



What Should We Evaluate in Gen-CRS?



What Should We Evaluate in Gen-CRS?

Journey

Coherence
Does the conversation stay on track and consistent?

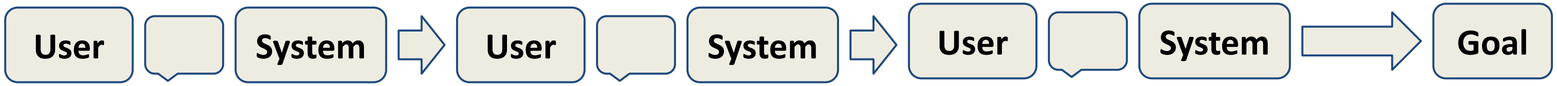
Efficiency
Does it solve the task with minimal steps?

User effort
How much work does the user put in?

Destination

Recommendation relevance
Are the recommendations useful and relevant?

Turn 1 **Turn 2** **Turn 3**



Task success
Is the user's goal achieved?

Time scales

Single turn
Eval. the quality of individual turns and immediate replies.

Whole conversation
Assess coherence, efficiency, and user effort across the full dialogue.

Multiple conversations
Evaluate consistency and robustness across diverse sessions.

What Should We Evaluate in Gen-CRS?

Journey

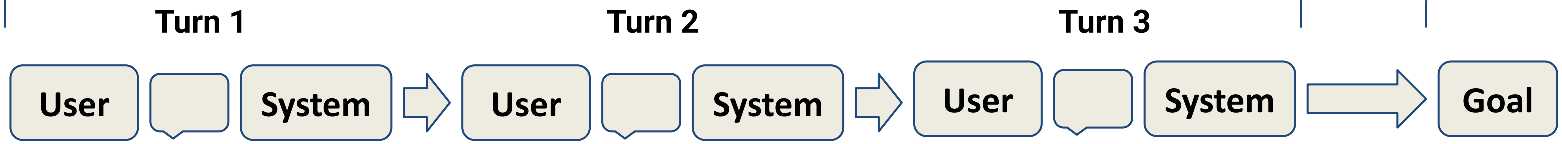
Coherence
Does the conversation stay on track and consistent?

Efficiency
Does it solve the task with minimal steps?

User effort
How much work does the user put in?

Destination

Recommendation relevance
Are the recommendations useful and relevant?



Task success
Is the user's goal achieved?

Time scales

Single turn
Eval. the quality of individual turns and immediate replies.

Whole conversation
Assess coherence, efficiency, and user effort across the full dialogue.

Multiple conversations
Evaluate consistency and robustness across diverse sessions.

Output Type

User-facing	responses, recommendations, explanations, images
Non-User-facing / hidden	Retrieval, dialogue states, reasoning paths

Key Dimensions

Output Type (What)	User-facing	Non-User-facing / hidden	
Quality Dimensions (Which Basis)	Task Effectiveness	Dialog / Task Efficiency	Conversational Quality ...siderations ...Integrity
Paradigms (How)			...Experiments
Evaluator (Who)	Automated Metrics (e.g., NDCG, HitRate, Recall etc.,)	Humans (e.g. Fluency, Informativeness etc.,)	LLM-as-a-Judge
Stakeholders (For Whom)	Consumers	Item-Providers	Others

All of the following statistics are based on our journal paper:

"Ahmadou Wagne, Thomas Kolb, Ashmi Banerjee, Fatemeh Nazary, Julia Neidhardt, and Yashar Deldjoo. 2026. Conversational Recommender Systems Using Generative Models (Gen-CRS): A Literature Review. ACM Trans. Recomm. Syst. Just Accepted (July 2026). https://doi.org/10.1145/3828551"

Key Dimensions

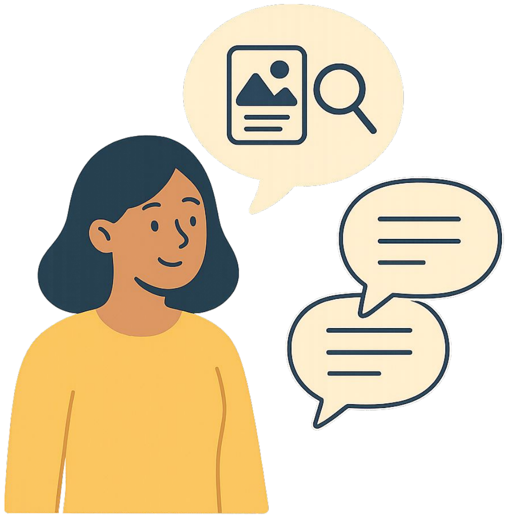
Output Type (What)	User-facing	Non-User-facing / hidden	
Quality Dimensions (Which Basis)	Task Effectiveness	Dialog / Task Efficiency	Conversational Quality
	Subtask Performance	Trust & User satisfaction	Ethical Considerations & System Integrity
Paradigms (How)	Offline	Online	User Studies & Lab Experiments
Evaluator (Who)	Automated Metrics (e.g., NDCG, HitRate, Recall etc.,)	Humans (e.g. Fluency, Informativeness etc.,)	LLM-as-a-Judge
Stakeholders (For Whom)	Consumers	Item-Providers	Others

Overview (What to Evaluate)

Output Type (what)	**User-facing**	Non-User-facing / hidden

Directly consumed by the user

- Natural language conversation turns
- Recommended items
- Explanations and justifications
- Multimedia elements (e.g., images)



Overview (What to Evaluate)

Output Type (what)	User-facing	Non-User-facing / hidden

Internal or intermediate model steps

- Learned embeddings
- Augmented data / Retrieved item sets
- Hidden reasoning paths
- Internal dialogue states

Evaluating Non-User-Facing Outputs

Output Type (what)	User-facing	**Non-User-facing / hidden**

- Current research **predominantly evaluates user-facing outputs** (discussed later).
- Non-user-facing outputs are typically **assessed indirectly** to prove their contribution.
  - Primary Method: **Ablation Studies**
    - Example (MemoCRS): Removing components like user-specific memory, collaborative knowledge, and reasoning guidelines to demonstrate their impact on final performance [Li et al., 2024a].
    - Example (CoRE-CoG): Evaluating retriever components (trigger, classifier) with standard metrics (Recall, MRR, BLEU, F1) to validate their design [Wang et al., 2024b].
  - Some extend this with sensitivity analysis for a more comprehensive assessment [Wang et al., 2024b].

Summary (What to Evaluate)

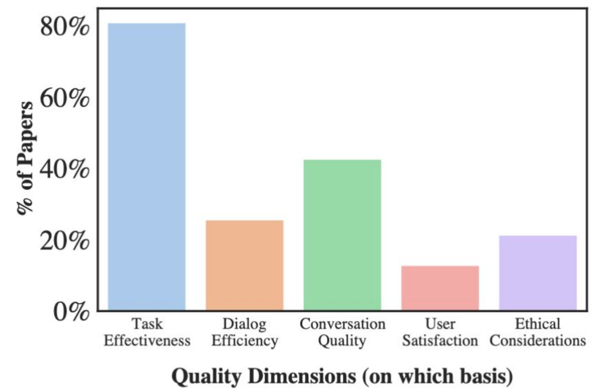
Output Type (what)	**User-facing**	**Non-User-facing / hidden**

- GenCRS aims for an enjoyable, human-like, and interactive experience.
  - This requires a fundamental shift in evaluation focus.
- Evaluating the "Journey" is as important as the "Destination"
  - Journey: Dialogue coherence, interaction efficiency, user enjoyment.
  - Destination: Relevance of the final recommended items.

Overview (Which Basis)

Quality Dimensions (Which Basis)			
	Task Effectiveness	Dialog / Task Efficiency	Conversational Quality
	Subtask Performance	Trust & User satisfaction	Ethical Considerations & System Integrity

- **Task Effectiveness:** *Did the user find a good item?*
- **Dialogue / Task Efficiency:** *How quickly and easily did they find it?*
- **Conversational Quality:** *Was the conversation natural and coherent?*
- **Subtask Performance:** *Did the internal modules work correctly?*
- **Trust & User Satisfaction:** *Did the user enjoy and trust the system?*
- **Ethical Considerations:** *Is the system fair, safe, and honest?*



Task Effectiveness

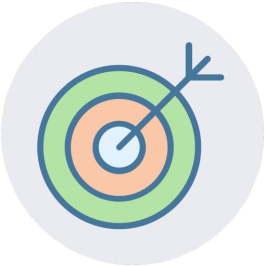
Quality Dimensions (Which Basis)			
	Task Effectiveness	Dialog / Task Efficiency	Conversational Quality
	Subtask Performance	Trust & User satisfaction	Ethical Considerations & System Integrity

Definition: The system's ability to help users successfully discover desired items.

Core Question: *"Did the user find a suitable item that effectively met their needs?"*

Common Metrics:

- Offline metrics: Hit Rate@K, Recall@K, NDCG@K
- Online metrics: User acceptance rate, click-through rate, purchases



Importance: This is the most fundamental and widely reported dimension (~80% of surveyed papers).

Examples:

- Recall@K is used to evaluate recommendation/retrieval effectiveness [he_2023, yang_2024, mao_2023, kim_2024, and others].
- Hit Rate@K is used to measure recommendation success [wang_2023, liu_2023, leszczynski_2023, xi_2024, and others].

Dialog / Task Efficiency

Quality Dimensions
(Which Basis)

Task Effectiveness	**Dialog / Task Efficiency**	Conversational Quality
Subtask Performance	Trust & User satisfaction	Ethical Considerations & System Integrity

Definition: How efficiently the system guides the user to a successful recommendation.

Goal: Avoid user frustration from overly long or unproductive conversations.

Common Metrics:

- Number of dialogue turns ("turns to success")
- Time to complete task
- System latency and computational cost [kunstmann_2024].

Examples:

- [wang_2021] uses a mix of automated metrics (Perplexity, BLEU, ROUGE, Distinct-n) and human evaluation to assess dialogue quality.
- [kunstmann_2024] evaluated their EventChat system in a 2-month real-world deployment, measuring user experience, latency, and cost.

Conversation Quality

Quality Dimensions (Which Basis)			
	Task Effectiveness	Dialog / Task Efficiency	**Conversational Quality**
	Subtask Performance	Trust & User satisfaction	Ethical Considerations & System Integrity

Definition: The linguistic and stylistic quality of the dialogue.

Challenge: No single "ground truth" exists for a perfect conversation, making human judgment crucial.

Objective Metrics:

- Fluency: Perplexity
- Content Overlap: BLEU, ROUGE
- Accuracy: Recall of target items in the response [chen_2024].

Subjective Metrics (Human Evaluation):

- Naturalness [leszczynski_2023]
- Persuasiveness & Engagement [kim_2024]
- Coherence, Informativeness

Subtask Performance

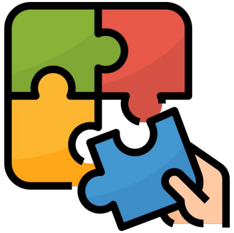
Quality Dimensions (Which Basis)			
	Task Effectiveness	Dialog / Task Efficiency	Conversational Quality
	Subtask Performance	Trust & User satisfaction	Ethical Considerations & System Integrity

Definition: Evaluating intermediate components within a modular or hybrid GenCRS pipeline.

Goal: Ensure that individual modules (e.g., intent recognizer, retriever) are performing well.

Examples of Subtasks & Metrics:

- Intent Recognition: Accuracy
- Slot Filling: Precision
- Entity Extraction: F1-Score



Examples from Research:

- [feng_2023] evaluates preference elicitation (NDCG@k) and explanation generation (BLEU, Distinct) as separate subtasks.
- [srivastava_2023] assesses their retrieval module using Precision, Recall, and F1-scores.

Trust & User Satisfaction

Quality Dimensions (Which Basis)	Task Effectiveness	Dialog / Task Efficiency	Conversational Quality
	Subtask Performance	**Trust & User satisfaction**	Ethical Considerations & System Integrity

Definition: The user's subjective perception of the interaction, including trust, usability, and enjoyment.

Core Question: *"How enjoyable was the interaction?" or "How confident are you in the recommendations?"*

Measurement: Almost always requires human feedback via user studies and post-task surveys.

Examples:

- Measuring if users actually follow the recommendations (user action-taking) [wu_2024].
- Assessing the perceived relevance of recommendations based on the dialogue [leszczynski_2023].
- Other works focusing on this include [baizal_2020, kunstmann_2024, abu-rasheed_2024, sun_2024].

Ethical Considerations & System Integrity

Quality Dimensions
(Which Basis)

Task Effectiveness	Dialog / Task Efficiency	Conversational Quality
Subtask Performance	Trust & User satisfaction	**Ethical Considerations & System Integrity**

Definition: Evaluating fairness, safety, and trustworthiness, especially with powerful LLMs.

Key Areas for Evaluation:

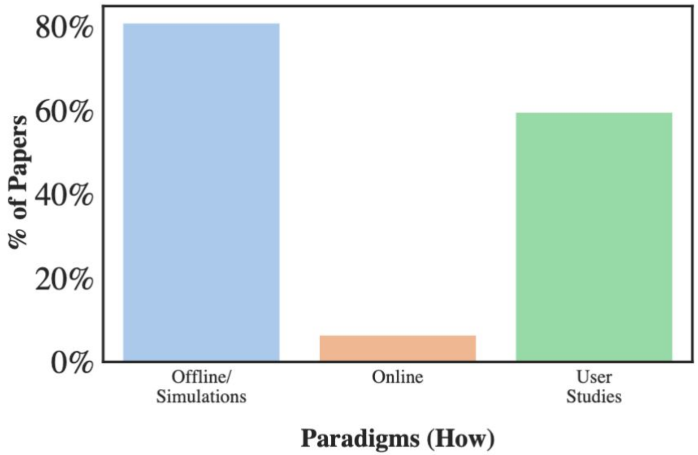
- **Factuality & Hallucination:** *Is the system generating factually correct information? [dehbozorgi_2024, he_2023, mao_2023].*
- **Instruction Faithfulness:** *Does the model follow instructions correctly? [tsai_2024].*
- **Bias:** *Are recommendations fair and not stereotyped?*
- **User Privacy:** *Is sensitive user data protected from leakage?*
- **Robustness:** *Is the system vulnerable to adversarial manipulation?*

Status: **A significant gap** in current research. These critical issues are rarely evaluated rigorously and represent an imperative direction for future work.

Overview (How)

Paradigms (How)	Offline	Online	User Studies & Lab Experiments

- **Offline Evaluation (or Simulations)**
  - **What:** Evaluation on static, pre-collected datasets.
  - **Prevalence:** The most dominant paradigm **(~80% of papers).**
- **Online Experimentation (A/B Testing)**
  - **What:** Deployment in a live environment with real users.
  - **Prevalence:** The least common paradigm **(~6% of papers)**.
- **User Studies & Lab Experiments**
  - **What:** Controlled experiments with recruited human participants.
  - **Prevalence:** A crucial middle ground, used in **~60% of papers**, often to supplement offline evaluation.



Offline Evaluation

Paradigms
(How)

Offline	Online	User Studies & Lab Experiments

Definition: Evaluation on static datasets without real-time user interaction.

Advantage: Efficient, low cost, and enables rapid, repeatable experiments.

Two Main Approaches:

- **Automated Metrics on Static Datasets:**
  - Calculating standard metrics on a fixed test set.
  - Examples: Recommendation accuracy (Recall@K, NDCG@K) or conversational quality (BLEU, ROUGE).
- **User Simulation:**
  - Using synthetic users (often LLMs) to interact with the system.
  - Examples
    - **(iEvaLM):** A framework using LLM-based user simulators to emulate diverse interactions, improving over static evaluation **[wang_2023b].**
    - Using session & turn-level controls to ensure simulated conversations are nearly indistinguishable from real ones **[friedman_2023]**.

Offline Evaluation

Paradigms
(How)

Offline	Online	User Studies & Lab Experiments

Definition: Evaluation on static datasets without real-time user interaction.

Advantage: Efficient, low cost, and enables rapid, repeatable experiments.

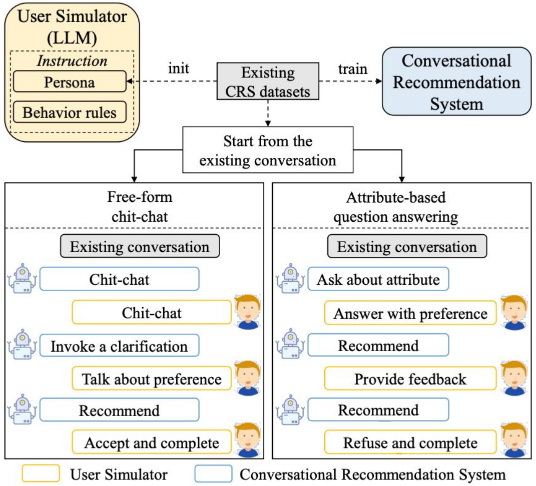
Two Main Approaches:

- **Automated Metrics on Static Datasets:**
  - Calculating standard metrics on a fixed test set.
  - Examples: Recommendation accuracy (Recall@K, NDCG@K) or conversational quality (BLEU, ROUGE).
- **User Simulation:**
  - Using synthetic users (often LLMs) to interact with the system.
  - Examples
    - **(iEvaLM):** A framework using LLM-based user simulators to emulate diverse interactions, improving over static evaluation **[wang_2023b].**
    - Using session & turn-level controls to ensure simulated conversations are nearly indistinguishable from real ones **[friedman_2023]**.

Example: iEvaLM

Rethinking the Evaluation for Conversational Recommendation in the Era of Large Language Models

Xiaolei Wang[1,3]*, Xinyu Tang[1,3]*, Wayne Xin Zhao[1,3]†, Jingyuan Wang[4] and Ji-Rong Wen[1,2,3]
[1]Gaoling School of Artificial Intelligence, Renmin University of China
[2]School of Information, Renmin University of China
[3]Beijing Key Laboratory of Big Data Management and Analysis Methods
[4]School of Computer Science and Engineering, Beihang University
wxl1999@foxmail.com, txy20010310@163.com, batmanfly@gmail.com



Problem/Motivation:

- Existing evaluations of Conversational Recommender Systems (CRSs) often rely on matching static, annotated "ground-truth" dialogs and recommendation items

Contributions:

- Propose a new interactive evaluation framework called iEvaLM that uses LLM-based user simulators to mimic realistic multi-turn interactions.
- Use ground-truth items from CRS datasets to define the persona / target of the simulated user. The simulated user "likes" those items, but must not reveal them directly
- Evaluate with two types of interaction: (a) attribute-based question answering, (b) free-form chit-chat.
- Show that ChatGPT's performance (in accuracy / recall, and explainability / persuasiveness) improves greatly under this new framework, often surpassing prior methods when allowed to interact.

Datasets & Baselines

- Two public CRS datasets: REDIAL (movies) and OPENDIALKG (multi-domain: movies, books, sports, music).
- Baselines include supervised CRS models (e.g., UniCRS, BARCOR, KBRD, etc.) and unsupervised approaches.

Evaluation Metrics

- **Accuracy (Recall@k)** over recommended items, over interaction rounds.
- **Explainability / Persuasiveness:** How convincing are explanations for the recommendation? Using human annotation and also an LLM-based scorer

Wang et al., "Rethinking the Evaluation for Conversational Recommendation in the Era of Large Language Models"

Online Evaluation

Paradigms
(How)

Offline	**Online**	User Studies & Lab Experiments

Definition: The "gold standard" for assessing real-world impact by deploying a system to live, unsuspecting users.

Key Metrics: Often tied to business objectives like conversion rates, user engagement, and revenue.

Challenges: High cost, resource-intensive, and requires a stable, production-ready system. This makes it rare in academic research.

Examples:

- **[kunstmann_2024]:** Integrated "EventChat" into a live mobile app for a two-month field study to evaluate real-world usage and system performance.
- **[nie_2024]**: Deployed a system on JD.com for a 7-day A/B test, measuring an increase in gross merchandise value (GMV) per user (+1.7%).
- **[leszczynski_2023]:** Conducted an online experiment with crowdworkers who rated the quality of music recommendations from their "TalkTheWalk" model.

Online Evaluation

Paradigms
(How)

Offline | **Online** | User Studies & Lab Experiments

Definition: The "gold standard" for assessing real-world impact by deploying a system to live, unsuspecting users.

Key Metrics: Often tied to business objectives like conversion rates, user engagement, and revenue.

Challenges: High cost, resource-intensive, and requires a stable, production-ready system. This makes it rare in academic research.

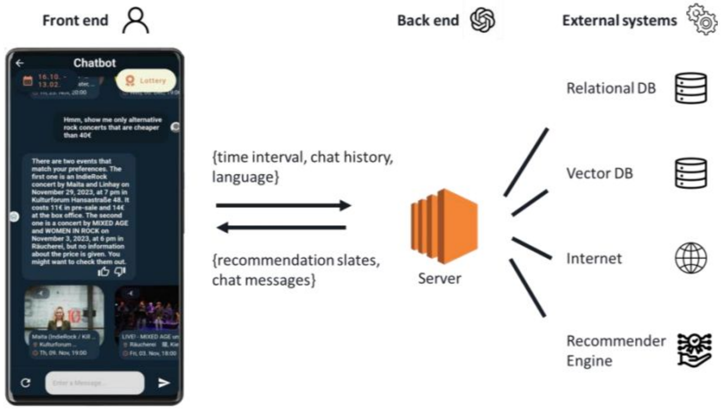
Examples:

- **[kunstmann_2024]:** Integrated "EventChat" into a live mobile app for a two-month field study to evaluate real-world usage and system performance.
- **[nie_2024]**: Deployed a system on JD.com for a 7-day A/B test, measuring an increase in gross merchandise value (GMV) per user (+1.7%).
- **[leszczynski_2023]:** Conducted an online experiment with crowdworkers who rated the quality of music recommendations from their "TalkTheWalk" model.

Example: EventChat

EventChat: Implementation and user-centric evaluation of a large language model-driven conversational recommender system for exploring leisure events in an SME context.

Hannes Kunstmann[1†], Joseph Ollier[2†*], Joel Persson[1], Florian von Wangenheim[1,2]



- A real conversational recommender system (CRS) built for a startup in the leisure/events domain (SME).
- Uses ChatGPT (via API) as the LLM core, plus prompt-based learning, a stage-based architecture, retrieval + ranking (RAG), an events database etc.
- Combine subjective (user satisfaction, perceived accuracy, usefulness) + objective metrics (latency, cost, success rates, interaction logs).
- Measure real users in the field, not just lab or simulated ones.
- Track and log failure cases to understand bottlenecks (e.g. when event isn't found, prompt misinterpretation, missing metadata).
- Monitor trade-offs: better accuracy / richer features often increase latency and cost; SMEs must balance.
- Use lightweight survey instruments (short-form ResQue etc.) to not overburden users.

Kunstmann et al., "EventChat: Implementation and user-centric evaluation of a large language model-driven conversational recommender system for exploring leisure events in an SME context."

User Studies & Lab Experiments

Paradigms (How)	Offline	Online	**User Studies & Lab Experiments**

Definition: A middle ground that captures genuine human behavior in a controlled experimental setting without the risks of a full online deployment.

Primary Goal: Essential for assessing subjective quality dimensions that automated metrics cannot capture.

- **Examples:** User satisfaction, trust, naturalness, persuasiveness.

Common Format:

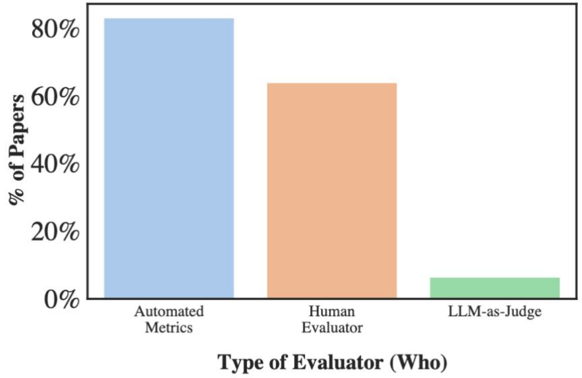
- Recruiting participants to perform specific tasks.
- Gathering feedback through post-task surveys, interviews, or think-aloud protocols.

Usage: Frequently used to supplement offline evaluations, providing a critical layer of human-centric validation.

Overview (Who)

Evaluator (Who)	Automated Metrics	Humans	LLM-as-a-Judge

- Automated Evaluators
  - Algorithmic approaches using pre-defined metrics.
  - **Prevalence:** Most common (38/47 papers).
- Human Annotators
  - Experts or crowdworkers providing qualitative assessments.
  - **Prevalence:** Very common, often for validation (28/47 papers).
- LLM-based Evaluators (LLM-as-Judge)
  - Using Large Language Models to assess quality.
  - **Prevalence:** An emerging but fast-growing approach (3/47 papers).



Automated Metrics

Evaluator (Who) | **Automated Metrics** | Human | LLM-as-a-Judge

Definition: Algorithms that compute objective, quantifiable metrics on static datasets.

Pros:

- Consistent, scalable, and efficient.
- Ideal for large-scale offline evaluation.

Cons:

- Often fail to capture nuance, especially in conversational quality and user satisfaction.

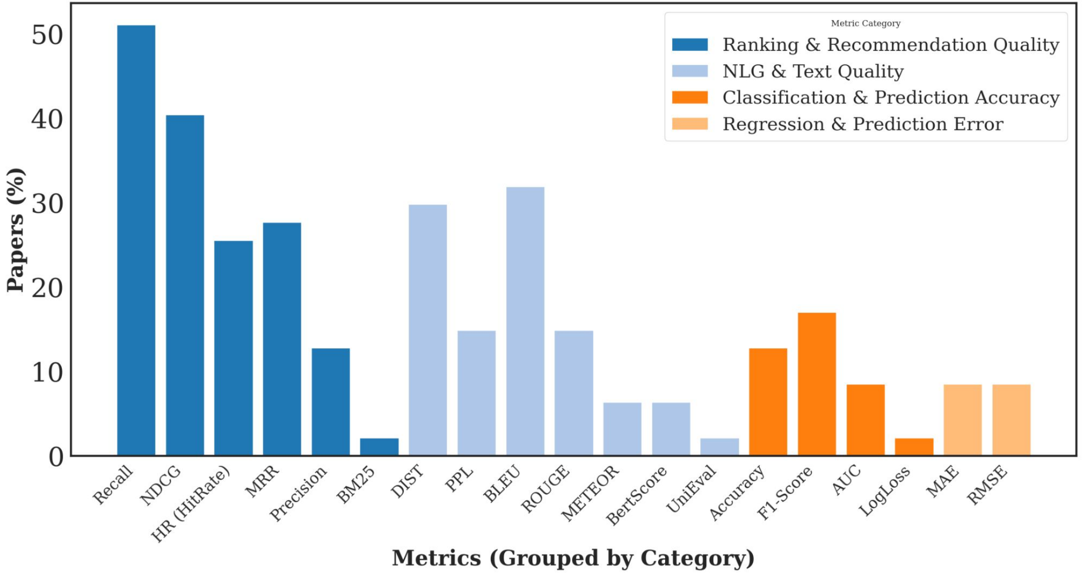
Key Categories of Metrics:

- Ranking & Recommendation Quality
- Natural Language Generation (NLG) & Text Quality
- Classification & Prediction Accuracy
- Regression & Prediction Error

Automated Metrics (Examples)

Evaluator (Who) | **Automated Metrics** | Human | LLM-as-a-Judge

- **Ranking & Recommendation Quality**
  - Recall@K, HitRate@K, MRR, NDCG@K: Assess accuracy and relevance of recommendations. Used in the majority of papers
- **NLG & Text Quality**
  - BLEU, ROUGE: Measure n-gram overlap with reference text
  - PPL (Perplexity): Measures language model fluency
  - DIST (Distinct-n): Measures response diversity
  - BERTScore: Measures semantic similarity
- **Classification & Prediction**
  - Accuracy, F1-Score, AUC: Assess performance on tasks like intent recognition.
  - MAE, RMSE: Measure error in rating prediction tasks.



Human Annotators

Evaluator (Who)	Automated Metrics	**Human**	LLM-as-a-Judge

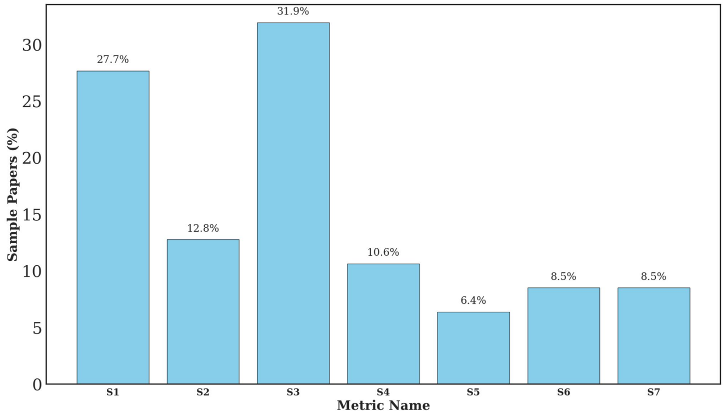
- **What they are:** Domain experts or crowdworkers providing qualitative assessments.
- **Pros:**
  - Excel at assessing subjective dimensions that automated metrics often miss: e.g., coherence, naturalness, helpfulness, trust.
- **Cons:**
  - Time-consuming and expensive.
  - Often used to validate or supplement automated metrics.
- **Process:**
  - Use Likert scales or binary ratings.
  - Reliability is checked with inter-annotator agreement metrics (e.g., Cohen's Kappa)

Common Subjective Metrics (Assessed by Humans)

Evaluator (Who)	Automated Metrics	**Human**	LLM-as-a-Judge

Humans are essential for evaluating:

- **S1:** Fluency, Grammar, & Readability **(27.7% papers)**
- **S2:** Coherence & Logicality **(12.8% papers)**
- **S3:** Informativeness & Helpfulness **(31.9% papers)**
- **S4:** Relevance & Answerability **(10.6% papers)**
- **S5:** User Satisfaction, Enjoyment, & Future Use **(6.4% papers)**
- **S6:** Trustworthiness, Persuasiveness, & Explainability **(8.5% papers)**
- **S7: Significant Gap:** Beyond-accuracy metrics like novelty and serendipity are rarely evaluated **(8.5% papers)**



LLM-based Evaluators (LLM-as-Judge)

Evaluator (Who)	Automated Metrics	Humans	LLM-as-a-Judge

Definition: An emerging paradigm using LLMs to evaluate the outputs of other models.

Role: Act as a scalable, fast, and consistent alternative to human annotators.

Challenge: Reliability can vary, and results may not always align with human judgment.

Examples:

- **Rec-SAVER [tsai_2024]:** An LLM generates and then self-verifies reasoning references to create an evaluation benchmark automatically.
- **iEvaLM [wang_2023]:** An LLM acts as both a user simulator to create interactions and an evaluator to score metrics like persuasiveness.
- **[sayana_2025]:** Uses Gemini Pro with ensemble rating (averaging over multiple runs) to score generated text on a 7-point Likert scale.

LLM-based Evaluators (LLM-as-Judge)

Evaluator (Who) | Automated Metrics | Humans | **LLM-as-a-Judge**

Definition: An emerging paradigm using LLMs to evaluate the outputs of other models.

Role: Act as a scalable, fast, and consistent alternative to human annotators.

Challenge: Reliability can vary, and results may not always align with human judgment.

Examples:

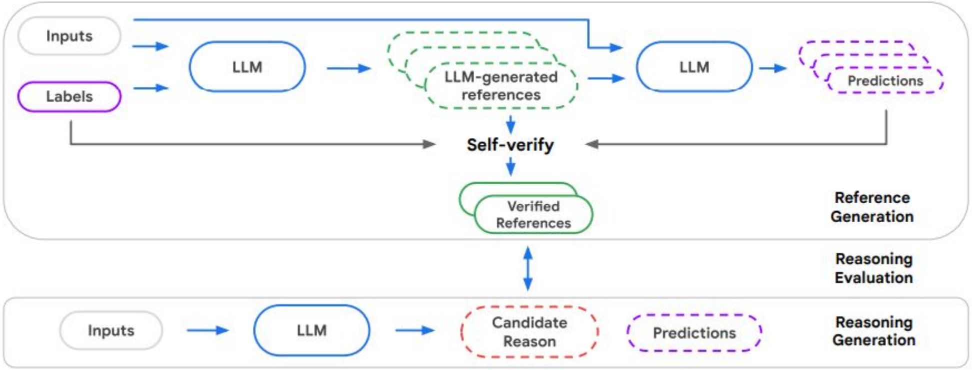
- **Rec-SAVER [tsai_2024]:** An LLM generates and then self-verifies reasoning references to create an evaluation benchmark automatically.
- **iEvaLM [wang_2023]:** An LLM acts as both a user simulator to create interactions and an evaluator to score metrics like persuasiveness.
- **[sayana_2025]:** Uses Gemini Pro with ensemble rating (averaging over multiple runs) to score generated text on a 7-point Likert scale.

Example: Rec-SAVER

Leveraging LLM Reasoning Enhances Personalized Recommender Systems

Alicia Y. Tsai*†[1,3], Adam Kraft*[3], Long Jin[2], Chenwei Cai[2], Anahita Hosseini[3], Taibai Xu[2], Zemin Zhang[2], Lichan Hong[3], Ed H. Chi[3], Xinyang Yi[3]

[1]University of California, Berkeley [2]Google [3]Google DeepMind



- Rec-SAVER (Recommender Systems Automatic Verification and Evaluation of Reasoning) is introduced to automatically assess quality of reasoning outputs from LLMs without needing human raters or curated gold references.
- It does this by generating explanations + then performing self-verification, i.e. re-predicting user ratings based on those explanations and comparing with actual ratings. If the explanation supports a correct rating, it counts more favorably.
- It uses multiple automatic NLG / text similarity / coherence / faithfulness metrics (e.g. BLEU, ROUGE, METEOR, BERTScore) to evaluate different dimensions of reasoning: how coherent the reasoning is, how faithful (i.e. correct with respect to inputs), how insightful.
- They validate that Rec-SAVER’s automatic judgments correlate well with human judgments on coherence and faithfulness, thus making it a reliable judge / benchmark.
- Using Rec-SAVER, they are able to compare models (zero-shot vs fine-tuned; smaller vs larger) not only on rating-prediction accuracy, but also on reasoning quality. They show that adding reasoning helps improve recommendation performance

Tsai et al., “Leveraging LLM Reasoning Enhances Personalized Recommender Systems” (2024)

Overview (For Whom)

Stakeholders (For Whom)	Consumers	Item-Providers	Others

A crucial, yet often overlooked, question is: *For whom is the evaluation being conducted?*

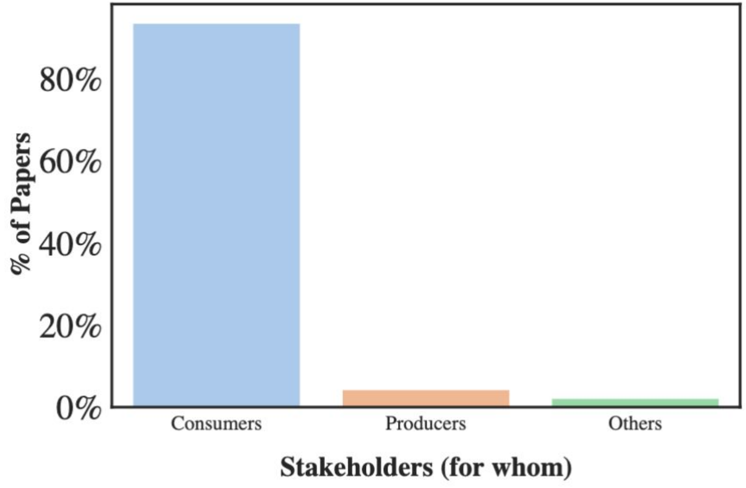
We can categorize the focus of an evaluation based on its primary beneficiary:

1. The Consumer (End-User)
2. The Item Provider (e.g., sellers, content creators)
3. Other Stakeholders (e.g., the platform, society at large)

Overview (For Whom)

Stakeholders (For Whom) | **Consumers** | Item-Providers | Others

- A staggering **93% of surveyed works** focus exclusively on the consumer.
  - This is understandable, as the main goal of a GenCRS is to satisfy the end-user.
- As a result, evaluations are dominated by user-centric metrics:
  - Recommendation Relevance
  - Dialogue Quality
  - User Satisfaction
- Evaluations considering other stakeholders are exceptionally rare.



The Gap: Item Providers & Society

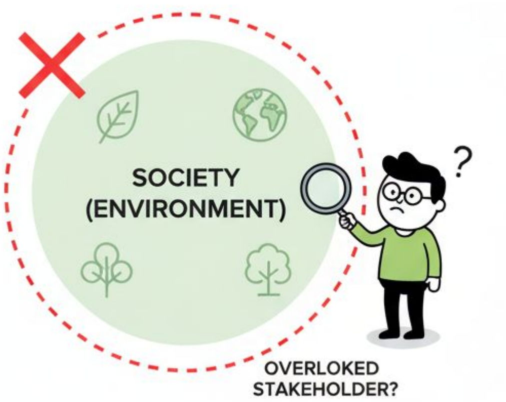
Stakeholders (For Whom)	Consumers	**Item-Providers**	Others

- Item Providers:
  - No studies explicitly focus on provider benefits (e.g., increased sales, fair exposure).
  - Some works indirectly address their interests:
    - Mitigating popularity bias to enhance fairness and visibility for less popular items **[wang_2023]**.
    - Tackling user-item rating bias for a fairer assessment of items **[kim_2024]**.

The Gap: Item Providers & Society

Stakeholders (For Whom)	Consumers	Item-Providers	**Others**

- **Society at large:** Represents a major blind spot in current evaluation practices.
  - Environmental Cost: The significant carbon footprint of LLMs is **almost entirely ignored** and undocumented in the GenCRS literature.
- **E.g. by incorporating Sustainability** (explicitly catering to Society as a stakeholder)
  - Examples:
    - **System Design:** e.g., Collab-REC*.
    - **Data Generation:** e.g., SynthTRIPS**.
  - Focus on multi-stakeholder evaluation frameworks.
- **Takeaway:** Expanding evaluation to include these broader impacts is an ethical imperative.

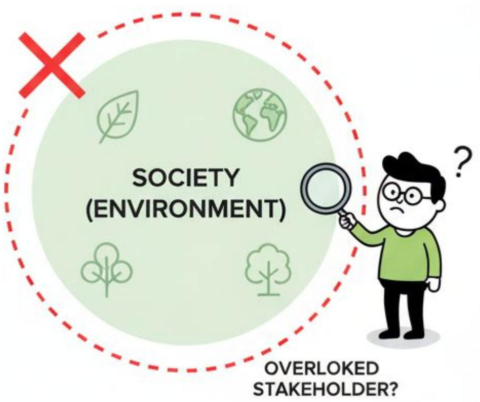


*Ashmi Banerjee, Adithi Satish, Fitri Nur Aisyah, Wolfgang Wörndl, and Yashar Deldjoo. 2026. Collab-REC: An LLM-based Agentic Framework for Balancing Recommendations in Tourism. ACM Trans. Recomm. Syst. Just Accepted (August 2026). https://doi.org/10.1145/3837862
** Ashmi Banerjee, Adithi Satish, Fitri Nur Aisyah, Wolfgang Wörndl, and Yashar Deldjoo. 2025. SynthTRIPs: A Knowledge-Grounded Framework for Benchmark Data Generation for Personalized Tourism Recommenders. In Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25). Association for Computing Machinery, New York, NY, USA, 3743–3752. https://doi.org/10.1145/3726302.3730321

The Gap: Item Providers & Society

Stakeholders (For Whom)	Consumers	Item-Providers	**Others**

- **Society at large:** Represents a major blind spot in current evaluation practices.
    - Environmental Cost: The significant carbon footprint of LLMs is **almost entirely ignored** and undocumented in the GenCRS literature.
- **E.g. by incorporating Sustainability** (explicitly catering to Society as a stakeholder)
    - Examples:
        - **System Design:** e.g., Collab-REC*.
        - **Data Generation:** e.g., SynthTRIPS**.
    - Focus on multi-stakeholder evaluation frameworks.
- **Takeaway:** Expanding evaluation to include these broader impacts is an ethical imperative.



*Ashmi Banerjee, Adithi Satish, Fitri Nur Aisyah, Wolfgang Wörndl, and Yashar Deldjoo. 2026. Collab-REC: An LLM-based Agentic Framework for Balancing Recommendations in Tourism. ACM Trans. Recomm. Syst. Just Accepted (August 2026). https://doi.org/10.1145/3837862

** Ashmi Banerjee, Adithi Satish, Fitri Nur Aisyah, Wolfgang Wörndl, and Yashar Deldjoo. 2025. SynthTRIPs: A Knowledge-Grounded Framework for Benchmark Data Generation for Personalized Tourism Recommenders. In Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25). Association for Computing Machinery, New York, NY, USA, 3743–3752. https://doi.org/10.1145/3726302.3730321

Example: Collab-REC

Collab-REC: An LLM-based Agentic Framework for Balancing Recommendations in Tourism

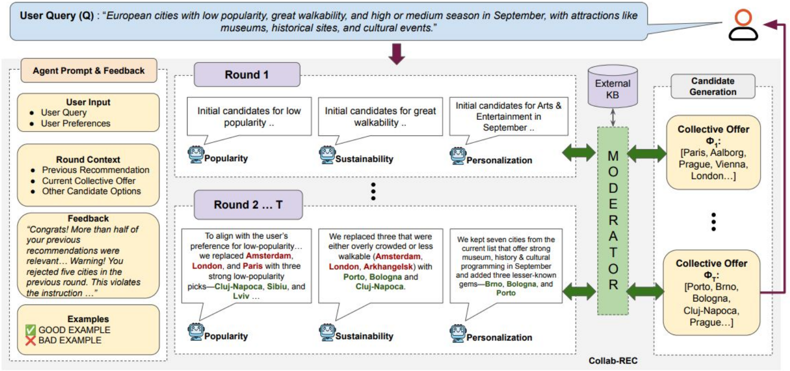
Ashmi Banerjee
ashmi.banerjee@tum.de
Technical University of Munich
Munich, Germany

Fitri Nur Aisyah
fitri.aisyah@tum.de
Technical University of Munich
Munich, Germany

Adithi Satish
adithi.satish@tum.de
Technical University of Munich
Munich, Germany

Wolfgang Wörndl
woerndl@in.tum.de
Technical University of Munich
Munich, Germany

Yashar Deldjoo
yashar.deldjoo@poliba.it
Polytechnic University of Bari
Bari, Italy



- It includes a **Sustainability Agent** among its core multi-agent setup, whose role is to promote eco-centric criteria (e.g. walkability, air quality, seasonality) when proposing city recommendations.
- The framework uses multi-round negotiation, where the **Sustainability Agent's** suggestions are combined with those of a **Personalization Agent and a Popularity Agent**; a moderator enforces trade-offs so that sustainability isn't drowned out by more popular or purely preference-based choices.
- The moderator (non-LLM) integrates penalties for repeated or invalid proposals and scores candidates by factors including agent success, reliability, and hallucination penalty — this helps to ensure sustainability suggestions make it through even when they conflict with popularity bias.
- Empirical results show Collab-REC improves diversity of recommendations (lesser-known / less popular cities surfaced) and reduces popularity bias, thus contributing toward more socially sustainable tourism (e.g. avoiding over-tourism).

Ashmi Banerjee, Adithi Satish, Fitri Nur Aisyah, Wolfgang Wörndl, and Yashar Deldjoo. 2026. Collab-REC: An LLM-based Agentic Framework for Balancing Recommendations in Tourism. ACM Trans. Recomm. Syst. Just Accepted (August 2026). https://doi.org/10.1145/3837862

Evaluation of GenCRS: Challenges

Overemphasis on Ground-Truth Matching

- Traditional evaluations focus on ground-truth matching, overlooking the interactive and evolving nature of CRS from the user's perspective

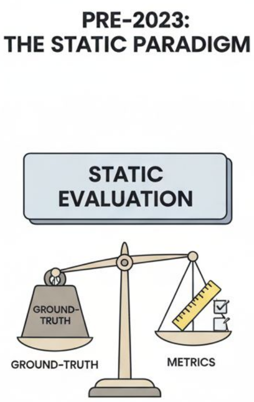
Limitations of LLM-Based Evaluators

- **Limited Human Representation**: Constrained by prompt templates and datasets; may miss real user complexity and dynamics.

- **Bias Propagation**: LLM evaluators can embed and amplify biases, reducing fairness

- **Ethical and Privacy Concerns**: Handling conversational data raises significant ethical and privacy risks.

Evaluation: Trends

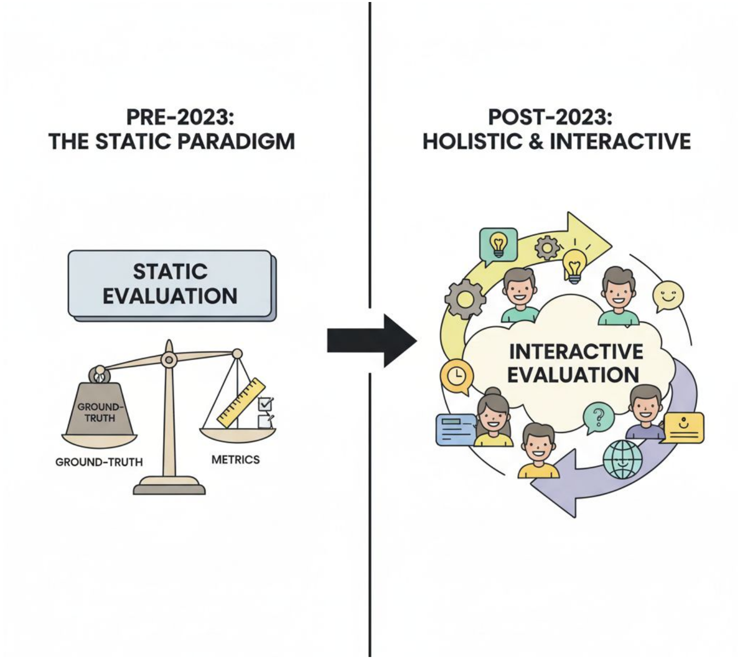


Pre-2023: The Static Paradigm

- Evaluation was predominantly static and relied on ground-truth-based metrics.
- **Key Metrics:** Recall@k, BLEU, Distinct-n.
- **Key Limitations****
  - Failed to capture the interactive and subjective nature of dialogue (e.g., conversation quality, user engagement).
  - Ineffective for open-ended, free-text responses where no single "correct" answer exists.
  - Could not properly assess issues like hallucination.

**Jannach, Dietmar, et al. "A survey on conversational recommender systems." *ACM Computing Surveys (CSUR)* 54.5 (2021): 1-36.

Evaluation: Trends



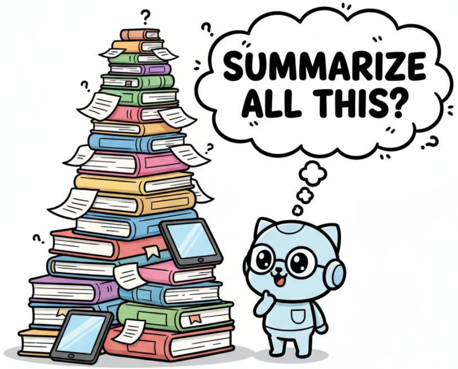
Post-2023: The Shift to Holistic, Interactive Evaluation with a user-centric perspective

- **The Rise of "LLM-as-a-Judge":**
  - LLMs are increasingly used as evaluators to assess subjective qualities like coherence, helpfulness, and naturalness
- **New Evaluation Dimensions Have Emerged:**
  - Factual Accuracy & Hallucination Detection
  - Faithfulness to user instructions
  - Groundedness on external knowledge
- **New Practical Concerns:**
  - System latency has become a key metric, as complex agent setups can impact user satisfaction.

**Jannach, Dietmar, et al. "A survey on conversational recommender systems." *ACM Computing Surveys (CSUR)* 54.5 (2021): 1-36.

Evaluation: Summary

- Evaluation is evolving from a **static to a multidimensional and interactive paradigm**.
- **The Core Shift in Focus:**
    - **FROM:** Matching a single 'gold-standard' response…
    - **TO:** Measuring how helpful, human-like, and contextually appropriate a system feels.
- **The Dual Role of LLMs:**
    - LLMs now function as both the recommender agent being tested and the automated evaluator assessing performance.
- This new paradigm enables **more scalable and nuanced assessments** of the overall user experience.



Julia Neidhardt

Open Challenges & Future Directions

Modelling and Architecture

- **Joint modelling:** Integrating retrieval and generation while maintaining control and modularity.
- **Efficient adaptation:** Updating models without impairing general reasoning.
- **Agent orchestration:** Managing task decomposition, workflows, and safe tool use.
- **Multimodal integration:** Jointly reasoning across text, images, audio, and structured data.

Knowledge, Data, and Memory

- **Signal integration:** Balancing content with user interaction data.
- **Dynamic knowledge:** Updating items, prices, and availability without retraining.
- **Preference continuity:** Maintaining preferences across sessions while protecting privacy.
- **Synthetic data:** Addressing data scarcity while limiting bias and preserving generalisation.

Interaction, Personalisation and Explanations

- **Preference elicitation:** Asking relevant questions and adapting to user trade-offs.
- **Personalisation:** Tailoring outputs while avoiding uniformity, popularity bias, and sensitive inferences.
- **Explanations:** Providing faithful explanations that users can inspect and challenge.
- **Anthropomorphism:** Understanding how persona and tone affect trust and expectations.

Evaluation and Benchmarking

- **Interactive evaluation:** Assessing multi-turn interactions beyond static, one-shot tests.
- **Unified metrics:** Combining recommendation quality, conversation, efficiency, and satisfaction.
- **LLM judges:** Reducing bias, prompt sensitivity, and inconsistent scoring.
- **Efficiency reporting:** Reporting latency and token costs alongside performance.
- **Long-term evaluation:** Measuring user, provider, and societal outcomes over time.

Systems, Deployment, and Governance

- **Scalability:** Reducing latency and costs from model calls and tool use.
- **Cost management:** Controlling paradigm-specific costs from prompts, model use, and agent loops.
- **Safety:** Measuring hallucinations and resistance to adversarial inputs.
- **Privacy:** Protecting sensitive information and clarifying data practices.
- **Governance:** Communicating policies, limitations, data use, and routes for recourse.
- **Reproducibility:** Reporting datasets, model configurations, and evaluation methods.

What the Literature Still Does Not Address

- **Accountability:** Responsibility and recourse for harmful recommendations remain undefined.
- **Governance:** Policies, limitations, and data use remain unclear.
- **Multi-stakeholder evaluation:** Provider and platform outcomes are rarely measured.
- **Longitudinal evaluation:** Long-term effects remain largely unstudied.
- **Diversity and bias:** Bias is documented but rarely controlled.
- **Environmental impact:** Energy and computational costs are seldom reported.

These areas remain largely unexplored.

Summary

- Gen-CRS have moved from structured slot filling to free-form generation, and increasingly act rather than only rank
- Natural language interaction is largely solved; grounding, the use of collaborative signals, and evaluation are not
- Combining collaborative knowledge with generative dialogue is the most promising near-term direction
- Responsibility, governance and efficiency need to be addressed at design time

Survey and tutorial materials:

recsys-lab.at/gen-conv-recsys-tutorial-ijcai-ecai-2026