Evaluating User Experience in Conversational Recommender Systems: A Systematic Review Across Classical and LLM-Powered Approaches

Raj Mahmud
School of Computer Science
University of Technology Sydney
Sydney, Australia
raj.pgm@outlook.com

Yufeng Wu
School of Computer Science
University of Technology Sydney
Sydney, Australia
yufeng.wu-2@student.uts.edu.au

Abdullah Bin Sawad
Department of CIT
The Applied College - King Abdullaziz University
Jeddah, Saudi Arabia
asawad@kau.edu.sa

Shlomo Berkovsky
Australian Institute of Health Innovation
Macquarie University
Sydney, Australia
shlomo.berkovsky@gmail.com

Mukesh Prasad
Computer Science
University of Technology Sydney
Sydney, Australia
mukesh.prasad@uts.edu.au

A. Baki Kocaballi
School of Computer Science
University of Technology Sydney
Sydney, Australia
baki.kocaballi@uts.edu.au

Abstract

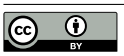
Conversational Recommender Systems (CRSs) are receiving growing research attention across domains, yet their user experience (UX) evaluation remains limited. Existing reviews largely overlook empirical UX studies, particularly in adaptive and large language model (LLM)-based CRSs. To address this gap, we conducted a systematic review following PRISMA guidelines, synthesising 23 empirical studies published between 2017 and 2025. We analysed how UX has been conceptualised, measured, and shaped by domain, adaptivity, and LLM. Our findings reveal persistent limitations: post hoc surveys dominate, turn-level affective UX constructs are rarely assessed, and adaptive behaviours are seldom linked to UX outcomes. LLM-based CRSs introduce further challenges, including epistemic opacity and verbosity, yet evaluations infrequently address these issues. We contribute a structured synthesis of UX metrics, a comparative analysis of adaptive and nonadaptive systems, and a forward-looking agenda for LLM-aware UX evaluation. These findings support the development of more transparent, engaging, and user-centred CRS evaluation practices.

CCS Concepts

• **Human-centered computing** → **User studies**; • **Information systems** → **Recommender systems**; • **Computing methodologies** → *Natural language processing*.

Keywords

User Experience Evaluation, Conversational Recommender Systems, Systematic Review



OZCHI '25, Sydney, Australia

ACM ISBN 979-8-4007-2016-1/25/11
https://doi.org/10.1145/3764687.3764714

ACM Reference Format:
Raj Mahmud, Yufeng Wu, Abdullah Bin Sawad, Shlomo Berkovsky, Mukesh Prasad, and A. Baki Kocaballi. 2025. Evaluating User Experience in Conversational Recommender Systems: A Systematic Review Across Classical and LLM-Powered Approaches. In *37th Australian Conference on Human-Computer Interaction (HCI) (OZCHI '25), November 29–December 03, 2025, Sydney, Australia.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3764687.3764714

1 Introduction

Conversational Recommender Systems (CRSs) are interactive systems that support users in discovering, refining, and selecting items through natural-language dialogue. Unlike traditional recommenders that generate static suggestions based on past behaviour or item similarity, CRSs enable multi-turn, dynamic conversations in which users can iteratively express preferences, explore alternatives, and resolve ambiguities. Early CRS designs combined handcrafted dialogue flows with content-based or collaborative filtering engines [46], often relying on rule-based interaction patterns and constrained slot-filling strategies [24]. Over time, the field has evolved to incorporate adaptive dialogue policies, mixed-initiative capabilities, and user-modelling techniques [68]. The recent surge of large language models (LLMs), such as GPT-4, has further transformed the design of CRS, enabling open-domain dialogue generation and more expressive, human-like conversational styles [10, 14]. However, these advances have also introduced new challenges in UX design, system transparency, and evaluation methodology.

As CRSs are deployed across domains including e-commerce, media, travel, and health [24], understanding how users perceive, interact with, and respond to these systems has become increasingly important. Several recent surveys have comprehensively reviewed technical progress in conversational recommendation, covering models, architectures, and dialogue policies [14, 35, 36, 67], but these works pay limited attention to UX as a distinct evaluative concern. They largely focus on system-level capabilities such as

natural language understanding, response generation, and recommendation accuracy, rather than the experiential dimensions of CRS interaction or empirical UX evaluation practices. While a few recent reviews, such as Jannach et al. [23], have begun to foreground UX concerns and examine evaluation practices from a user-centric perspective, these works remain primarily conceptual or narrow in scope. A systematic, empirical synthesis of how UX is defined, operationalised, and measured in CRS research is still lacking. In particular, prior work has not fully accounted for domain-specific variation, adaptivity mechanisms, or the emergent UX challenges posed by LLM-powered CRS. Although constructs such as satisfaction, trust, and perceived usefulness are frequently measured, there is little consensus on how to evaluate the temporal, affective, and behavioural dynamics that shape user perceptions during interaction [41]. Generative CRS further complicate evaluation by introducing epistemic opacity, verbosity control issues, and novel interaction risks that standard UX instruments may not capture effectively [10, 69]. To address these gaps, we present a systematic literature review (SLR) of CRS studies with an explicit focus on UX evaluation. We apply the PRISMA methodology (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [49] to identify and synthesise 23 empirical studies published up to May 2025. Our analysis focuses on four key questions:

- **RQ1:** What UX dimensions are evaluated in CRS research, and how do these dimensions vary across application domains?
- **RQ2:** What evaluation methods and tools are used to assess UX in CRS, and what are their affordances and limitations?
- **RQ3:** How do adaptivity and personalisation influence the evaluation and experience of CRS?
- **RQ4:** What are the UX design and evaluation challenges posed by LLM-powered CRS?

Our contributions are fourfold. First, we provide a structured synthesis of the UX dimensions assessed in CRS research, categorised by domain. Second, we analyse the methods, tools, and measurement frameworks used to evaluate UX, highlighting common practices and gaps. Third, we offer a comparative analysis of adaptive versus nonadaptive systems, identifying how design choices affect reported user experiences. Fourth, we examine the emerging class of LLM-powered CRS, revealing distinctive UX challenges and underexplored risks that future research must address.

The remainder of this paper is organised as follows: Section 2 reviews related work on CRS design and evaluation. Section 3 outlines our review methodology. Section 4 presents findings for RQ1–RQ4. Section 5 synthesises implications and themes, and Section 6 concludes with reflections and future directions.

2 Related Work

Research on CRS spans architectural design, user modelling, evaluation methodologies, and more recently, the integration of LLMs. This section groups prior work into four strands: architectural and evaluation-oriented surveys, empirical UX studies, LLM-focused CRS reviews, and a summary that positions our review within this landscape.

2.1 Architectural and Evaluation-Oriented Surveys

Several surveys have mapped the system-level evolution of CRS, often highlighting dialogue strategies, recommendation models, and conversational policies. Gao et al. [14] provide a comprehensive overview of architectural components, including natural language understanding, state tracking, and policy learning. Their taxonomy helps clarify technical bottlenecks in multi-turn CRS design but offers limited treatment of UX dimensions or user-centred evaluation frameworks. Similarly, Lei et al. [35] formalise CRS as a joint optimisation task over dialogue and recommendation quality. They outline evaluation strategies ranging from offline ranking metrics to online A/B testing but do not examine user affect, trust, or satisfaction metrics in depth. Zaidi et al. [67] review a broad spectrum of CRS applications, such as fashion, education, and tourism, highlighting domain-specific implementations and dialogue approaches. Their focus remains descriptive, offering little insight into user perception or temporal engagement. Pramod and Bafna [53] examine techniques and tools for CRS development, including ontology-driven recommendation, critiquing interfaces, and adoption barriers. While they emphasise explainability and interface design, their work lacks a methodological breakdown of evaluation protocols or user-centred metrics. In contrast, Jannach et al. [23] develop a dedicated evaluation framework, distinguishing three modes: offline simulations, lab-based studies, and in-the-wild deployments. They critique the over-reliance on accuracy metrics and advocate for more interaction-focused and affective evaluations. However, their review is broad and does not systematically extract constructs, instruments, or domains from empirical CRS studies. Wang et al. [64] complement this by arguing that static evaluation pipelines do not reflect the dynamic nature of user–CRS interaction. While their proposals for interactive and human-centred benchmarks are conceptually rich, they remain largely untested and omit UX operationalisation practices used in prior studies. Li et al. [36] advocate for a holistic view of CRS, integrating multi-modal signals, user goals, and dialogue context. Although they stress the value of personalisation and interactivity, their review falls short of discussing how these qualities are evaluated in practice. Lian et al. [40] introduce RecAI, a modular toolkit for building LLM-enabled recommender systems. While their architecture includes modules for explanation and dialogue monitoring, their evaluation approach focuses on internal benchmarks, not on empirical UX instruments or user feedback.

2.2 Empirical UX Studies

Numerous empirical studies explore UX in CRS, though typically with narrow or domain-specific scope. Siro et al. [59] apply turn-level annotation to the ReDial dataset, identifying satisfaction-related features such as interest arousal and repetition. Their work is valuable for modelling micro-level interaction signals but does not capture longitudinal satisfaction or user preferences. Ma and Ziegler [44] investigate proactive decision aids in CRS, revealing that system-initiated suggestions improve perceived usefulness, though at the cost of increased cognitive load. Their study highlights the UX trade-offs introduced by initiative strategies but

lacks broader construct coverage. Thom et al. [60] develop a GPT-powered nutrition assistant and evaluate it using CSAT, CES, and NPS scores. While the study presents useful deployment insights, it does not explore behavioural metrics, emotional reactions, or adaptation to user preferences. Kraus et al. [31] conduct a Wizard-of-Oz study of group travel CRS, showing that conversation-leading strategies better align with user expectations when group personalities converge. Yet their work does not assess adaptivity mechanisms or engagement sustainability. These studies collectively point to the need for systematic evaluation of diverse UX constructs, ranging from trust and clarity to surprise and engagement, but no synthesis currently exists to integrate these metrics across domains or system types.

2.3 LLM-Focused CRS Reviews

The introduction of LLMs has prompted new concerns in CRS design and evaluation. Wu et al. [65] categorise LLM applications in recommendation into generation, planning, and reasoning. They identify new risks, including hallucination, response verbosity, and misalignment, but do not detail how such issues affect UX or how they are evaluated. Li et al. [37] similarly position LLM-based recommenders as generative pipelines but focus on system capabilities and challenges rather than user-facing evaluation methods. Deldjoo et al. [10] propose a conceptual taxonomy of evaluation risks introduced by generative models, distinguishing between exacerbated challenges (e.g., verbosity, latency) and novel ones (e.g., prompt brittleness, privacy leakage). While this framework is theoretically valuable, it is not grounded in empirical user studies. Huang et al. [20] envision agentic CRS that proactively support user goals, manage dialogue flow, and adapt to long-term preferences. Their vision points to new UX affordances but remains speculative in the absence of empirical validation. Lian et al. [40], Hou et al. [18], and others also explore how LLMs can function as zero-shot rankers or preference matchers, but these works focus on benchmark accuracy and fail to engage with experiential metrics such as trust, control, or expectation calibration. Across these reviews, the experiential implications of LLM deployment are acknowledged, yet the methods to evaluate such experiences remain vague or absent.

2.4 Summary of Gaps and Our Contribution

Despite significant growth in CRS and generative recommender literature, no prior review offers a structured synthesis of UX evaluation constructs, instruments, and methods across empirical studies. Existing surveys primarily focus on architectural taxonomies, dialogue strategies, or conceptual critiques. Even when UX is discussed, as in Jannach et al. [23] or Deldjoo et al. [10], coverage remains abstract, anecdotal, or limited to high-level dimensions. Critically, established UX instruments such as the User Experience Questionnaire (UEQ), System Usability Scale (SUS), and Recommender Quality Evaluation (ResQue) framework are seldom compared across studies. This lack of comparative analysis obscures how different tools capture affective, cognitive, or behavioural dimensions, and whether they are appropriate for CRS-specific evaluation across contexts. Moreover, domain-adaptive evaluation designs, those that account for setting-specific needs in domains like wellbeing, travel, or education, remain underexplored. This fragmentation hinders the development of cumulative knowledge about UX in CRS and limits reproducibility across systems and contexts. In contrast, this review applies a PRISMA-guided methodology to systematically extract and analyse empirical CRS studies. We focus explicitly on how UX is conceptualised and operationalised across application domains, how adaptivity and LLM integration influence evaluation design, and what methods and instruments are employed. By comparing and categorising UX constructs and tools, we reveal which evaluation practices are prevalent, underutilised, or misaligned with the demands of generative CRS. Our synthesis bridges gaps between conceptual reviews and fragmented empirical work, offering actionable insights for CRS researchers and designers seeking to improve UX measurement and design practices.

3 Methodology

This review follows the PRISMA, aiming to enhance reproducibility, transparency, and methodological rigour in synthesising UX evaluations of CRS. Our goal was to examine how CRS UX is conceptualised, operationalised, and measured across different application domains, adaptivity types, and system architectures, including LLM-based implementations.

3.1 Review Protocol and Inclusion Criteria

We applied inclusive criteria to capture the diversity of CRS research. Studies were included if they: (1) implemented a CRS involving natural language interaction, (2) reported empirical UX findings using self-report, behavioural, or third-party methods, and (3) focused on any application domain or evaluation setting (lab, field, simulated). Exclusion criteria were: (1) studies lacking user evaluation (e.g., offline experiments only), (2) conceptual or design-only papers without empirical testing, and (3) technical implementations without user interaction or feedback. Only peer-reviewed conference and journal articles written in English were considered.

3.2 Literature Search and Screening

We conducted a comprehensive literature search across six major academic databases: ACM Digital Library, IEEE Xplore, ScienceDirect, SpringerLink, Scopus, and Web of Science. These were selected for their broad coverage of human–computer interaction (HCI), recommender systems, and conversational technologies. The initial search was conducted in September 2023 and updated in May 2025. The search string was devised by analysing keywords from prominent CRS papers and adjacent UX/HCI studies to ensure sensitivity to both classical and LLM-powered systems:

```
(convers* OR dialog* OR "speech" OR "voice assistant"
OR "voice-enabled" OR "voice agent" OR "voice-based"
OR "voice-activated" OR "spoken-language" OR chatbot
OR chatterbot OR "large language model" OR LLM)
AND ("User Experience" OR UX) AND recommend*
```

No date restrictions were applied, and Boolean operators were used to maximise recall. While we centred the search strategy on "UX" to ensure conceptual clarity, we acknowledge that related constructs such as trust, satisfaction, or engagement may have been reported in CRS research without being explicitly labelled as UX. This prioritised precision but may have reduced recall. We followed PRISMA

Table 1: Comparison of 13 survey and review papers on conversational and LLM-powered recommender systems. Columns assess UX coverage, empirical synthesis, evaluation methods, and LLM focus.

Study	Primary Focus	UX	Empirical Synthesis	Evaluation	LLM Focus
Gao et al. (2021) [14]	CRS Architectures	No	No	Yes	No
Lei et al. (2020) [35]	CRS Formulation & Eval	No	No	Yes	No
Zaidi et al. (2024) [67]	CRS Techniques & Applications	No	No	Yes	No
Pramod & Bafna (2022) [53]	CRS Tools & Adoption	No	No	Yes	No
Jannach et al. (2023) [23]	Evaluation Framework	Yes	No	Yes	No
Wang et al. (2023) [64]	Evaluation Critique	Yes	No	Yes	No
Li et al. (2023) [36]	Holistic CRS Design	No	No	Yes	No
Lian et al. (2024) [39]	LLM Toolkit (RecAI)	No	No	Yes	Yes
Wu et al. (2024) [65]	LLM-RS Survey	Yes	No	Yes	Yes
Li et al. (2024) [38]	LLM-RS Pipeline	No	No	Yes	Yes
Deldjoo et al. (2024) [10]	LLM-RS Eval Taxonomy	Yes	No	Yes	Yes
Huang et al. (2025) [20]	Agentic RS Vision	No	No	No	Yes
Hou et al. (2024) [19]	LLMs as Rankers	No	No	Yes	Yes

guidelines [49] and employed a two-phase screening procedure. In Phase 1, titles and abstracts were screened against predefined inclusion criteria. In Phase 2, full texts were retrieved and assessed independently by multiple reviewers. Three independent reviewers conducted the screening. For the initial phase (up to September 2023), Cohen's κ ranged from 0.625 to 0.85, indicating substantial to almost perfect agreement [47]. During the updated phase (October 2023–May 2025), $\kappa = 0.66$ for title–abstract screening and $\kappa = 0.88$ for full-text screening were achieved. Disagreements were resolved through consensus discussions. Screening was managed using Rayyan [50], a dedicated tool for blinded decisions and conflict tracking.

3.3 Data Extraction and Coding

A structured data extraction sheet was created and iteratively refined. Each included article was coded for the following fields: author(s), publication year, CRS domain (e.g., movie, music, nutrition), evaluation context (lab, field, remote), presence of adaptivity (adaptive vs. nonadaptive), use of LLMs or generative models, UX constructs evaluated, measurement tools used, and participant sample size. UX constructs were extracted using the authors' terminology, then categorised into broader UX types:

- **Affective:** enjoyment, surprise, trust, empathy
- **Cognitive:** usefulness, clarity, novelty, informativeness
- **Relational:** engagement, agency, control
- **Task-oriented:** satisfaction, efficiency, recommendation quality

Instruments and evaluation methods were coded into:

- **SM (Self-report Measures):** Likert scales, interviews, NPS, UEQ
- **BM (Behavioural Measures):** dialogue length, error rate, task completion
- **EM (External Measures):** third-party annotations

Each study's coding was verified by at least two researchers. Discrepancies were reconciled through iterative checking.

3.4 Analysis and Synthesis Strategy

We adopted a structured narrative synthesis aligned with the four research questions. For RQ1, we summarised and compared the UX constructs reported across studies, grouped by application domain. RQ2 was addressed by mapping evaluation tools and study designs to reveal dominant practices and underutilised methods. For RQ3, we categorised CRS adaptivity features and examined how these were evaluated in relation to UX outcomes. Finally, RQ4 was addressed through targeted analysis of LLM-powered CRS papers, triangulated with recent survey literature to surface unique UX challenges. Due to the heterogeneity of UX constructs, evaluation tools, and study designs, we did not conduct quantitative meta-analysis. Instead, we focused on conceptual synthesis, identifying methodological trends, gaps, and design tensions. Descriptive summaries and tabular mappings supported comparison across studies.

4 Results: Reviewed CRS UX Studies

This section presents findings from the included empirical studies evaluating user experience in CRSs. We have structured this section sequentially around our four research questions. We begin with an overview of all the articles.

4.1 Article Overview

The PRISMA in Figure 1 presents the flow of records from initial identification to final inclusion. After removing duplicates, 356 unique records were screened at the title and abstract level. Of these, 50 articles were selected for full-text review. Following exclusion of 27 papers that did not meet the inclusion criteria (e.g., insufficient UX data, non-conversational systems), 23 articles were included in our final synthesis. These studies span diverse domains, including e-commerce, music, movies, restaurants, nutrition, and domain-independent applications. Table 2 provides a detailed summary of each article, including domain, UX metrics, evaluation methods, and key findings.

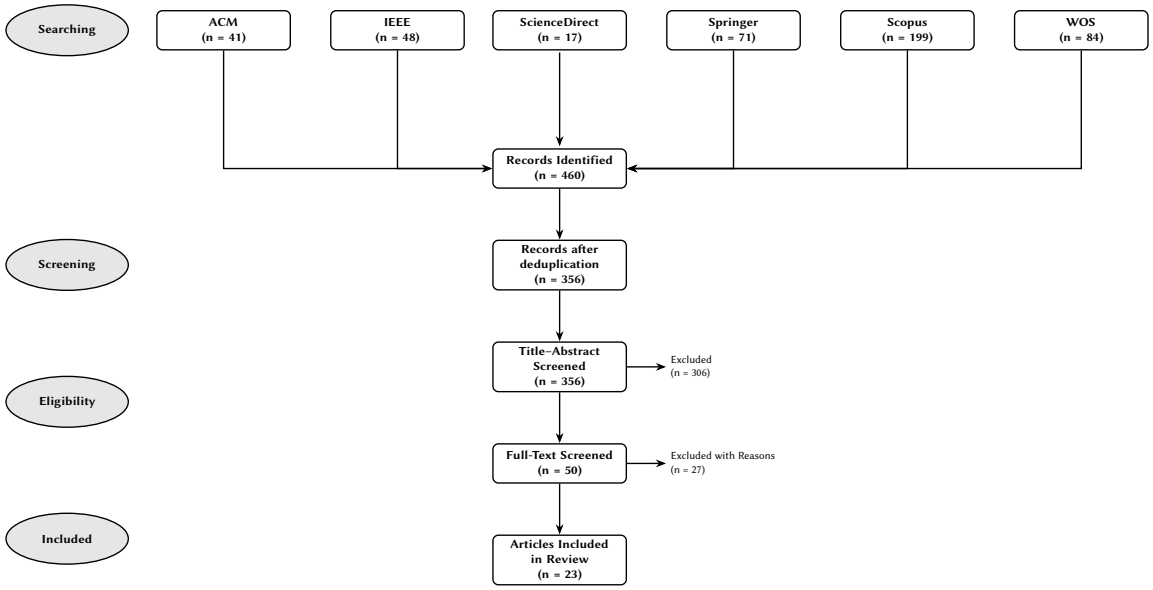


Figure 1: PRISMA flow diagram illustrating the identification, screening, eligibility assessment, and inclusion of studies in this review.

4.2 UX Dimensions and Domain Variation (RQ1)

We examined what UX dimensions were evaluated across the studies and how these dimensions varied by application domain. A total of 71 distinct UX metrics were identified, spanning cognitive, affective, and behavioural dimensions. These metrics were extracted from both self-reported and behavioural measures, with a minority using human annotators who evaluated specific dialogues that occurred between a CRS and other users. We refer to those methods as external measures.

Most Frequently Assessed Metrics. User satisfaction was the most commonly reported UX metric, appearing in 18 of 23 studies. Other frequently evaluated dimensions included personalisation (9 studies), usefulness (7), trust (6), ease of use (6), and recommendation quality (6). Affective and reflective constructs such as curiosity, self-reflection, emotional support, and agency were reported in a smaller number of studies, predominantly those involving LLM-powered systems [17, 66].

Domain-Specific Trends. UX priorities varied systematically across application domains. E-commerce studies (n = 7) often focused on transactional UX factors including purchase intention, initiative, trust, and privacy. Music and movie CRS (n = 10) emphasised hedonic and affective UX such as enjoyment, novelty, serendipity, and transparency. Restaurant-focused studies (n = 3) featured constructs related to interaction quality, perspicuity, and usability. Nutrition systems (n = 2) assessed longer-term UX goals including loyalty, accessibility, and usage frequency. Domain-independent systems (n = 1) prioritised performance-oriented constructs such as interaction cost, recommendation accuracy, and efficiency. We next examined how these UX dimensions were measured and what tools and methodologies were employed to evaluate them across the included studies.

4.3 Evaluation Methods and Tools Used in CRS UX Studies (RQ2)

UX evaluation in CRS research is predominantly conducted using self-reported measures, found in 21 of the 23 studies. These methods typically involved post-interaction survey questionnaires using Likert-type scales (5-point or 7-point), capturing constructs such as satisfaction, usability, trust, novelty, and intention to use. Many studies adapted validated instruments such as UEQ (User Experience Questionnaire), UEQ-S, ResQue, SUS, or domain-specific scales (e.g., CSAT, CES, NPS). Behavioural measures were reported in 12 studies. These included logging the number of dialogue turns, interaction durations, recommendation acceptance rates, moderation actions, and task completion. For example, Ma et al. [42] compared interaction lengths across initiative strategies; Gutierrez et al. [17] measured nDCG@9 and HitRate@9 for CRS versions with and without personalisation. External measures were rarely used. Only two studies adopted third-party annotations or crowdworker ratings. Siro et al. [59] employed Wizard-of-Oz simulations and MTurk evaluations to link dialogue turn quality to satisfaction. Silva et al. [58] annotated politeness and grammar scores externally to benchmark generative responses. Qualitative methods were used in three studies. These included interviews, open-ended reflections, and thematic coding of diary entries. For instance, Yun and Lim [66] conducted a three-week diary study on GPT-based CRS, applying Braun and Clarke's (2006) inductive thematic analysis. Kraus et al. [31] combined content coding with personality correlations to understand group dynamics in multi-party CRS. Regarding tool adoption, 13 studies explicitly cited validated frameworks or tools. The remaining studies either designed novel items or minimally described their instrument sources, reducing replicability. Inferential analyses included t-tests (7 studies), ANOVAs (6), MANOVAs,

Table 2: Summary of 23 empirical studies: UX metrics, evaluation methods, and key findings. Evaluation types: SM = self-report, BM = behavioural, EM = external (non-user rater).

Study	Domain	UX Metrics	Evaluation Methods & Tools	Key Findings
[66]	Music	Agency; Emotional Support; Exploration; Personalisation; Reflection.	**SM**: 3-week diary study ($n = 12$); free-form CRS conversations; daily diaries + post-study interviews. Analysis: inductive thematic + narrative synthesis [3].	CRS enabled exploration, reflection, and emotional support; frustrations from limited memory and generic replies; personalisation linked to anthropomorphism/follow-ups; no standard UX scales or stats.
[43]	E-commerce	Supportive; Easy; Efficient; Clear; Exciting; Interesting; Inventive; Leading Edge.	**SM**: Between-subjects online study ($n = 184$); proactive vs. passive schemes. UX via UEQ-S [57]. Proactive group received rule-based prompts (compare, critique, details). Pre/-post questionnaires: decision style + meta-intents.	No UEQ-S differences overall; proactive ↑ "interesting/exciting" (d=.159, .120); passive ↑ "clear/easy" (d=−.155, −.131). Prompts ↑ decision-aid use ("check details" d=1.45, $p < .001$; "compare in dialog" d=.70, $p < .001$).
[31]	Restaurant	Efficiency; Perspicuity; Attractiveness; Novelty; Stimulation; Dependability.	**BM**: Dialogue logs (duration, moderation acts, utterances, task performance). **SM**: Post-interaction survey ($n = 21$); UEQ (7-pt, long form) + BFI-10. Analysis: Wilcoxon; Bonferroni–Spearman; descriptives; qualitative coding ($\kappa = 0.683$).	No UEQ differences (leader vs. follower, $p > 0.05$). Leader ↓ time (728s vs. 964s, $p = 0.09$), ↑ dominance gap (8.20 vs. 2.87, $p = 0.06$), ↓ moderation ($p = 0.04$); follower ↑ task success (35.6 vs. 32.7). In leader condition, openness ↔ perspicuity ($\rho = 0.65$, $p = 0.01$) and dependability ($\rho = 0.63$, $p = 0.02$).
[60]	Nutrition	Satisfaction; Effort; Loyalty; Usefulness; Accuracy; Accessibility; Frequency of Use.	**SM**: Online survey after demo/independent use ($n = 50$); custom CSAT, CES, NPS, usage metrics (5-pt Likert, frequency). Analysis: formula-based score computation (no inferentials).	CSAT 96% (Mean=4.60); CES 90% (Mean=4.42); NPS 92% (Mean=4.38); Accuracy 94% (Mean=4.56); Use frequency: 5/week; Satisfaction 98%.
[17]	Movie	Enjoyment; Personalisation; Relevance; Satisfaction.	**SM**: Between-subjects ($n = 27$), comparing personalised vs. non-personalised, 5 tasks; post-task 5-pt Likert (satisfaction, experience). **BM**: nDCG@9, HR@9. Analysis: descriptives; Pearson $r = 0.26$ (nDCG@9 ↔ preferences).	Personalised > non-personalised (nDCG@9: 0.4273 > 0.3880; HR@9: 0.78 > 0.74). 40% vs. 10% endorsed all positive items; ∼22% of recommended items not on the platform (fairness concern).
[42]	E-commerce	Information sufficiency; Initiative; Interaction adequacy; Efficiency; Interest; Accuracy; Satisfaction; Usefulness.	**BM**: Interaction count/duration (pre vs. post initiative transfer). **SM**: Pre/post 5-pt Likert (Decision Style, Sense of Agency, Initiative Preference, UEQ-S, ResQue). Analysis: descriptives; t-tests; χ^2; MANOVA.	User-initiated mode ↓ interactions ($t(86.2) = -2.97$, $p = 0.012$) and duration ($t(141) = -1.36$, $p = 0.175$). Initiative transfers ↓ excitement ($F = 7.07$, $p = 0.009$) and interest ($F = 5.25$, $p = 0.023$). Transfer occurrence: $\chi^2(1, N = 143) = 24.40$, $p < 0.001$.
[11]	E-commerce	Personalisation; Sentiment Analysis; User Satisfaction.	**BM**: Sentiment classifiers (F1). **SM**: Post-interaction survey (0–10-pt, custom questions). Analysis: F1 metrics; A/B tests (SA vs. no-SA).	SA ↑ satisfaction (9.13 vs. 8.41). F1: BERT 0.76⟶0.85; GPT 0.72⟶0.78; ELMO 0.73⟶0.82.
[55]	Restaurant	Diversity; Ease of Use; Effectiveness; Intention; Personalisation; Relevance; Satisfaction; Serendipity.	**BM**: Interaction logs (failures, effort, flow, recommendation quality). **SM**: Post-interaction survey (5-pt Likert), custom questions. Analysis: descriptives; one-way t-test; one-sided z-test.	NP-Rec > P-Rec failures (56.1% vs. 50%, $p = 0.039$); NP-Rec ↑ interactions (3.5 vs. 2.8, $p = 0.003$); P-Rec ↑ relevance (85% vs. 78%) and future interest (68% vs. 59%). Ease-of-use κ: P-Rec 0.918; NP-Rec 0.857.
[59]	Movie	Interest arousal; Recommendation quality (precision, recall, F1); Efficiency; User Satisfaction.	**EM**: Wizard-of-Oz dialogue annotations; MTurk ratings (3–5 pt Likert). Analysis: distributions; box plots; correlations; regression; classification.	Turn-level relevance ↔ satisfaction ($\rho = 0.610$). Dialogue-level completion ($\rho = 0.599$) and interest ($\rho = 0.622$) predicted satisfaction. Turn-level prediction: $r = 0.734$; dialogue-level $r = 0.796$. Random forest RMSE=0.768, $\rho = 0.734$.
[27]	E-commerce	Future Intentions; Perceived Friendliness; Product Attitudes; User Satisfaction.	**SM**: Post-interaction surveys (5-pt Likert: satisfaction, friendliness, product attitudes, chatbot intention). Analysis: PROCESS macro (Models 1 & 7); Johnson–Neyman.	Extrovert-matched chatbots ↑ satisfaction and friendliness for extroverts ($F(1, 286) = 52.52$, $p < 0.001$). Mediation: satisfaction ⟶ future intentions and attitudes (Index=0.35, SE=0.13, 95% CI [0.10, 0.62]).
[9]	E-commerce	Purchase Intentions; Psychological Distance; Usability.	**SM**: Post-interaction surveys; scales: anthropomorphism, conversational relevance (custom), psychological distance, usability, purchase intention. Analysis: two-way ANOVA.	High relevance ↑ task proficiency ($F(1, 58) = 56.58$, $p < 0.01$) and perceived intelligence ($F = 24.30$, $p < 0.01$). High anthropomorphism ↑ responsibility ($F = 5.65$, $p = 0.02$). Purchase intentions ↑ with high relevance ($F = 28.74$, $p < 0.01$).
[45]	Movie	Ease of Use; Interaction Adequacy; Intention to Use; Recommendation Accuracy; User Control; User Satisfaction.	**BM**: Logs–duration, HitRate@K, preference count. **SM**: Post-interaction survey (5-pt Likert, ResQue). Analysis: HitRate@1/@3; descriptive statistics.	HitRate@1: Objective 78% vs. Objective+Subjective 88.5%; HitRate@3: 96% vs. 98.1%. Ease 4.24 vs. 4.26; Control 4.28 vs. 4.19; Adequacy 4.28 vs. 4.15; Accuracy 4.30 vs. 4.50; Satisfaction 4.20 vs. 4.40.
[58]	E-commerce	Linguistic Politeness.	**BM**: Automated metrics–Politeness, BLEU, ROUGE, METEOR. **EM**: Dialogue annotations (1–3; binary politeness/-grammar). Analysis: automated and comparative statistics.	Rewrite methods > generative for politeness and content retention. Politeness: RW-Enron 82.24; RW-Fashion 79.76; RW-Fashion-C 78.42; RW-Mixed 80.70. RW-Fashion-C and RW-Mixed ↑ BLEU/ROUGE/METEOR.
[4]	Music	Confidence; Control; Discovery; Diversity; Ease of Use; Interaction Adequacy; Interaction Efficiency; Novelty; Satisfaction; Serendipity; Transparency; Trust.	**BM**: Number of interactions, dialogue turns, duration, number of critiques, accepted recommendations. **SM**: Post-task survey (7-pt Likert, ResQue and Matt et al.). Analysis: SEM; Kruskal–Wallis; Mann–Whitney U; Mixed ANOVA; post-hoc.	Exploration tasks ↑ interaction ($p < 0.001$) and ↓ satisfaction ($p < 0.01$) vs. basic. Cascading critique ↑ diversity ($p < 0.05$); progressive critique ↑ serendipity ($p < 0.01$). Significant differences across critiquing methods for serendipity, diversity, and adequacy ($p < 0.05$).
[56]	Nutrition	Personalisation; Sentiment; Usability; User Satisfaction.	**BM**: Precision/recall/F1 (DIET); sentiment accuracy (VADER). **SM**: Post-interaction surveys: SUS, UEQ, CUQ, SEQ. Analysis: descriptive statistics; t-tests; one-way ANOVA.	SUS=84.7±8.4; CUQ=82.9±8.6. Sentiment accuracy: $\rho = 0.67$, $r = 0.68$, MAE=0.85. Personalised recommendations ↑ satisfaction ($F(1, 58) = 27.45$, $p < 0.001$).
[21]	Domain-independent	Interaction Cost; Recommendation Quality; User Satisfaction.	**BM**: Logs–number of questions, time, query density, accuracy, MAP. **SM**: Post-interaction Likert survey. Analysis: MANOVA.	Natural language interface ⟶ fewer questions but ↑ time vs. button interface. Mixed mode ↑ efficiency and quality ($p < 0.05$); max accuracy=0.63, MAP=0.55.
[1]	E-commerce	Privacy Risk; Social Presence; Trust; Usability.	**BM**: Task time; completion rate. **SM**: Post-interaction surveys and interviews (custom). Analysis: descriptive statistics.	Privacy=6.2/7; Social presence=6.7/7; Usability=94.5/100; Task completion=100%.

Table 2: Summary of Article Characteristics (continued).

Study	Domain	UX Metrics	Evaluation Methods & Tools	Key Findings
[26]	Music	Confidence; Engagement; Humanness; Interaction Adequacy; Novelty; Perceived Ease of Use; Perceived Usefulness; Transparency; Trust; User Control; User Satisfaction.	**BM**: Number of interactions, duration, acceptance rate, task success. **SM**: Post-interaction surveys (7-pt Likert, ResQue plus custom). Analysis: SEM.	SEM: Usefulness ($\beta = 0.52$, $p < 0.01$), Ease ($\beta = 0.45$), Transparency ($\beta = 0.43$) → Satisfaction. Novelty ($\beta = 0.35$, $p < 0.05$), Adequacy ($\beta = 0.32$) → Control. Accuracy=90%; Trust ($\beta = 0.50$), Confidence ($\beta = 0.47$), Satisfaction ($\beta = 0.55$). High engagement and rapport reported.
[13]	Music	Effectiveness; Efficiency; Persuasiveness; Transparency; Trustworthiness; User Satisfaction.	**SM**: Pre/post interaction surveys using standardised questionnaires based on Tintarev. Analysis: descriptive statistics.	Explanation feature ↑ satisfaction, perceived transparency, trustworthiness, effectiveness, and persuasiveness. Recommendation accuracy = 90.48%.
[30]	Restaurant	Cognitive Demand; Habitability; Likeability; Motivation to Interact; Response Accuracy; User Satisfaction.	**SM**: Post-interaction surveys (7-pt Likert) using User Acceptance Scale, SASSI, Motivation Scale. Analysis: one-way ANOVA; post-hoc t-tests.	Both proactive strategies (explicit and implicit) ↑ satisfaction vs. reactive ($F(2, 16) = 3.65$, $p < 0.05$). Implicit strategy ↑ accuracy and habitability ($F(2, 16) = 4.81$, $p < 0.05$).
[22]	Movie	Confidence; Ease of Use; Interaction Adequacy; Overall Satisfaction; Recommendation Accuracy; Transparency; Trust; Use Intentions; Usefulness; User Control.	**BM**: Interaction cost, number of questions, query density, accuracy, MAP. **SM**: Post-interaction survey (5-pt Likert, ResQue). Analysis: Wilcoxon signed-rank.	Overall satisfaction non-significant. Accuracy: Smartphone 3.80 vs. Robot 3.65 (MAP non-significant). Robot ↑ enjoyment ($p = 0.046$) and ↑ irritation ($p = 0.043$). Trust: Robot 3.45 < Smartphone 3.75 ($p < 0.05$); Transparency ↑ with Robot ($p < 0.05$).
[51]	Movie	Decision Confidence; Intention to Return; Intention to Watch; Perceived Effort; Perceived Usefulness; Recommendation Quality; Transparency; User Control; User Satisfaction.	**SM**: Post-interaction surveys (5-pt Likert) based on prior works. Analysis: MANOVA; two-way ANOVA.	Social explanations ↑ recommendation quality ($F = 12.45$, $p < 0.01$) and personalised ratings ($F = 15.32$, $p < 0.01$). Decision confidence ↑ ($F = 10.87$, $p < 0.01$); overall satisfaction ↑ ($p < 0.01$). Transparency, perceived effort, and user control ↑ ($p < 0.05$). Return intention ↑ ($p < 0.05$).
[34]	Movie	Intimacy; Intention to Use; Interactional Enjoyment; Trust; User Satisfaction.	**SM**: Post-interaction surveys (7-pt Likert) based on CASA and Uncertainty Reduction Theory. Analysis: two-way ANOVA; PLS regression.	Self-disclosure ($F = 10.99$, $p < 0.01$) and reciprocity ($F = 17.47$, $p < 0.001$) ↑ satisfaction. Perceived trust and enjoyment mediated effects ($p < 0.05$). Reciprocity was a stronger predictor of relationship building than self-disclosure.

Wilcoxon, and Spearman correlations. A smaller subset used structural models (e.g., SEM in 2 studies), classification models, or effect size reporting (e.g., Cohen's d, Pearson's r). However, several recent works, particularly those using generative CRS, reported descriptive insights without statistical testing, limiting the generalisability of their findings. Taken together, the findings suggest that while survey-based self-reporting dominates CRS UX evaluation, a growing number of studies are augmenting this with behavioural logging and qualitative narratives. Having examined how UX was measured, we next investigated how adaptivity and personalisation influenced CRS design and evaluation.

4.4 Adaptive vs. Nonadaptive UX (RQ3)

We examined how CRS studies incorporated adaptivity and personalisation into their system design and UX evaluation strategies. We classified each of the 23 studies based on whether the system modified its behaviour dynamically during interaction. We considered a system adaptive if it adjusted its dialogue flow, initiative, or content generation in response to user input, preferences, or traits. Systems that included personalisation without runtime behavioural change were classified as nonadaptive. We identified seven studies as fully adaptive [30, 31, 42, 43, 45, 51, 66] and one study as partially adaptive [17]. The remaining fifteen studies used fixed interaction logic and did not incorporate behavioural adaptation during runtime [1, 4, 8, 12, 13, 21, 22, 26, 27, 33, 55, 56, 58–60].

Adaptivity Strategies. Several systems supported initiative shifts based on user input or interaction style, including explicit toggling between proactive and reactive dialogue flows [30, 42, 43]. In group decision-making contexts, adaptive logic guided the assignment or alternation of agent roles, such as leader or supporter, based on conversational dynamics and user preference inputs [31]. Other systems used natural language inputs to infer user goals and adjust recommendations dynamically, including narrative-driven content adaptation [45] and affect-sensitive interaction design using goal-oriented versus open-ended prompts [66]. Pecune et al. [51] implemented real-time personalisation of social explanations by aligning system messages with users' Big Five personality traits. These systems employed a range of adaptivity techniques, including policy-based switching, cue-triggered behaviour modulation, and language model-driven personalisation routines. Researchers often examined adaptive control through experimental manipulations of initiative style, cue types, or framing conditions.

Personalisation Depth. Nineteen studies incorporated personalisation, although its function and depth varied widely. Many nonadaptive systems used static personalisation methods, such as filtering content or priming response tone based on user traits [27, 33], sentiment [12], or domain-specific constraints like dietary needs [60]. These systems typically personalised surface-level content without modifying the interaction logic. In contrast, adaptive systems integrated user traits, preferences, or affective signals directly into dialogue planning, modulating recommendation content,

explanation strategies, or initiative dynamics [43, 51, 66]. While several studies acknowledged the distinction between preference-based and behavioural personalisation, few articulated this difference explicitly. Only a minority of studies evaluated the UX impact of personalisation mechanisms, with most relying on between-condition comparisons or post-interaction feedback to assess user alignment.

Evaluation Coverage. Evaluation approaches also differed by adaptivity. Studies featuring adaptive CRS more frequently assessed interactional UX metrics such as perceived control, initiative alignment, and engagement, often through experimental manipulations or mixed methods [30, 43, 45]. Nonadaptive systems tended to rely on post hoc ratings of satisfaction, clarity, or content relevance, often collected using static Likert scales or sentiment annotations [13, 55, 58, 59]. Only a small number of studies explicitly evaluated the adaptivity mechanism itself, typically through comparative designs testing alternate system behaviours. We found that adaptive systems supported more complex evaluation logic and broader UX metric coverage, while nonadaptive systems focused primarily on interface impressions and content quality. Although most studies incorporated some form of personalisation, only a subset used it to influence system behaviour dynamically during interaction. We now turn to the subset of studies that utilised LLMs, revealing emerging implementation trends and critical methodological limitations.

4.5 Evaluating LLM-Powered CRS (RQ4)

Two studies in the review explicitly implemented CRS systems powered by LLMs: Yun et al. [66] and Granada et al. [17]. Both systems leveraged **OpenAI's GPT** architecture for natural language generation, though their underlying recommendation logic and evaluation designs differed. Yun et al. [66] introduced a GPT-based music CRS that supported affective adaptation through manipulation of dialogue modes (goal-directed vs. open-ended prompts). The system was evaluated using a between-subjects experimental design, with outcome variables including user trust, perceived friendliness, and intention to reuse. GPT-based generation was controlled via prompt framing; however, the system lacked structured explanation modules or user-facing transparency mechanisms. Granada et al. [17] deployed a hybrid CRS architecture combining retrieval-based candidate generation with GPT-3.5-based free-form responses. Recommendations were presented via naturalistic dialogue generated from ranked items embedded within GPT prompts. The evaluation consisted of a comparative user study measuring perceived coherence, engagement, and dialogue naturalness. No formalised UX instrument was employed, and evaluation focused on qualitative user impressions. Neither system incorporated turn-level UX instrumentation, explanation transparency controls, or adaptive modulation of verbosity. Both evaluations were limited to single-session interactions, with no longitudinal tracking of usability, satisfaction, or trust dynamics. Table 4 summarises architecture reporting and implementation tooling across all reviewed studies, illustrating the limited methodological coverage in current LLM-powered CRS evaluations.

5 Discussion

This section interprets the empirical findings of our review, drawing connections across domains, evaluation strategies, and system design choices to surface critical insights for the future of CRS user experience research. We synthesise emergent patterns to guide both researchers and practitioners working at the intersection of conversational AI, recommender systems, and human-centred evaluation. The discussion unfolds in five parts: Section 5.1 offers a cross-RQ synthesis of core findings; Section 5.2 surfaces unresolved conceptual and methodological issues; Section 5.3 analyses UX-specific gaps in LLM-powered CRS; Section 5.4 articulates design and research implications; and Section 5.5 outlines limitations and future directions.

5.1 Cross-RQ Synthesis of CRS UX Findings

Our review reveals significant fragmentation in how UX is conceptualised and evaluated in CRS research. Satisfaction and usability remain the dominant constructs, yet affective, relational, and reflective dimensions, such as emotional support, agency, and trust dynamics, are often neglected or inconsistently measured. This narrow focus constrains our ability to capture the full spectrum of experience, particularly in hedonic, educational, or socially-oriented domains. Evaluation strategies similarly exhibit limitations. Most studies rely on post hoc self-report surveys administered after single-session interactions. While efficient, such methods fail to capture temporal dynamics or interactional nuance. Although some studies incorporate behavioural logging or qualitative feedback, these are seldom integrated systematically with user ratings. As discussed further in Section 5.2, this methodological inertia restricts insight into how CRS experiences unfold over time. Adaptivity and personalisation also remain uneven. While some systems claim to personalise content or tone, most rely on static user traits or pre-scripted rules. Few implement real-time adjustments based on evolving user inputs, and even fewer evaluate the UX implications of such adaptations in situ. As outlined in Section 5.4, real-time, user-aware dialogue modulation remains a promising but underdeveloped area. Finally, the emergence of LLM-powered CRS introduces new challenges. These systems offer expressive generation capabilities, but often lack transparency, controllability, and structured evaluation pipelines. Section 5.3 expands on the UX design and instrumentation gaps specific to LLM-based dialogue systems. Taken together, the findings highlight a mismatch between the complexity of conversational recommendation and the limited scope of current UX evaluations. The field would benefit from multimodal, theoretically grounded, and context-sensitive evaluation strategies capable of tracing how user experience emerges across dialogue.

5.2 Unresolved Issues in Conceptualisation and Evaluation

Despite the proliferation of CRS research in recent years, foundational limitations persist in how UX is framed and evaluated. Two major concerns remain: the inconsistent conceptualisation of UX constructs and the continued reliance on static, post-interaction evaluation methods.

Table 3: Summary of 23 reviewed CRS studies: adaptivity and personalisation (ordered by year). Adaptivity = dynamic system behaviour; Personalisation = tailoring of content, recommendations, or dialogue style.

#	Study	Adapt.	Personal.	Evidence Summary
1	Yun et al. (2025) [66]	Yes	Yes	GPT-based music CRS adapted affectively through goal-directed vs. open-ended dialogue.
2	Ma & Ziegler (2024) [43]	Yes	Yes	Trait-sensitive proactive cues; initiative adjusted by decision style.
3	Kraus et al. (2024) [31]	Yes	Yes	Leader/follower role-switch in restaurant CRS; preferences modelled for group decisions.
4	Damian et al. (2024) [60]	No	Yes	NutrIA app personalised nutrition inputs; no adaptive logic.
5	Rana et al. (2024) [55]	No	Yes	Critiquing CRS explored preferences; no live adaptation.
6	Ma & Ziegler (2023) [42]	Yes	Yes	User-controlled initiative switching with trait framing.
7	El-Ansari & Beni-Hssane (2023) [12]	No	Yes	Sentiment-informed tone; no adaptive logic.
8	Granada et al. (2023) [17]	Partial	Yes	VideolandGPT prompts tailored to traits; no adaptive policy.
9	Siro et al. (2023) [59]	No	Yes	Predictive UX model on features; behaviour static.
10	Jin & Eastin (2022) [27]	No	Yes	Personality similarity influenced trust; not adaptive.
11	Chung & Han (2022) [8]	No	Yes	Rule-based e-commerce bot; personalisation via warm/cold persona.
12	Martina et al. (2022) [45]	Yes	Yes	Narrative CRS adapted recommendations from free-text preferences.
13	Silva et al. (2022) [58]	No	No	Politeness agent scripted; no tailoring.
14	Cai et al. (2022) [4]	No	No	Critiquing CRS; lacked user models/adaptation.
15	Samagaio et al. (2021) [56]	No	No	Recipe CRS, slot-based; no user tailoring.
16	Iovine et al. (2021) [21]	No	No	NL vs. button interface tested; employed fixed interaction flow.
17	Anastasia et al. (2021) [1]	No	Yes	Embodied e-commerce agent designed for social alignment; not adaptive.
18	Jin et al. (2021) [26]	No	Yes	CRS-Que used to analyse perceived quality dimensions; not used for adaptive control.
19	Fernando et al. (2021) [13]	No	Yes	Music recommendations explained via personality traits; system remained non-adaptive.
20	Kraus et al. (2020) [30]	Yes	Yes	Dialogue initiative adapted to user input style; mixed-initiative strategy verified.
21	Iovine et al. (2020) [22]	No	Yes	Robot CRS with static preference input; no dynamic personalisation.
22	Pecune et al. (2019) [51]	Yes	Yes	Social explanations adapted in real-time to Big Five traits.
23	Lee & Choi (2017) [33]	No	Yes	Studied self-disclosure/reciprocity; responses non-adaptive.

Inconsistent Conceptual Framing. Across the reviewed literature, "user experience" is invoked frequently but defined inconsistently. Some studies operationalise UX through acceptance-focused constructs such as satisfaction or intention to use, while others examine affective or relational dimensions like trust, empathy, or engagement. However, these constructs are rarely linked to theoretical frameworks or adapted to the *specific demands* of conversational systems. As a result, the field lacks a shared conceptual foundation for interpreting user experience in CRS, undermining comparability across studies and limiting theoretical progress. Prior calls in HCI and recommender systems research for theory-aligned UX measurement remain largely unheeded in CRS contexts [29, 61].

Methodological Inertia. Most evaluations continue to rely on one-off, post-session Likert-scale questionnaires. While such instruments provide quick feedback, they fail to capture the *evolving nature of experience* in multi-turn, adaptive dialogue. Very few studies use turn-level metrics, real-time prompts, or longitudinal designs that reflect how trust, engagement, or satisfaction may shift over time. Mixed-method evaluations are also rare, with minimal triangulation of behavioural, affective, and self-reported data. This limits insight into how specific system behaviours, such as tone shifts, explanation placement, or initiative control, shape users' moment-by-moment impressions.

Implications. Advancing CRS UX research requires both conceptual clarity and methodological innovation. Conceptually, studies should adopt well-defined constructs aligned with the *interactional* and *adaptive* nature of CRS. Methodologically, evaluations must evolve beyond retrospective ratings to include temporally sensitive, context-aware approaches. Real-time instrumentation (e.g., interaction logs, embedded UX probes) and theory-informed designs (e.g., expectation-confirmation models or affective appraisal frameworks) could offer deeper insight into how user experiences are constructed in dialogue. Without such progress, CRS research risks generating evaluations that are consistent but shallow, and systems that are functional but affectively disengaging.

5.3 UX Challenges in LLM-powered CRS

The integration of LLMs into conversational recommender systems presents new opportunities and challenges for user experience design and evaluation. While LLMs offer increased fluency, expressivity, and generative capabilities, these affordances also introduce risks that current CRS evaluation practices are ill-equipped to address. Our review identified only two empirical studies explicitly evaluating LLM-powered CRS [17, 66]. Both relied primarily on post-hoc user ratings and offered limited UX instrumentation. While participants reported high engagement and affective resonance, neither study incorporated fine-grained diagnostics or tracked experiential shifts across dialogue turns. A core challenge is the *epistemic opacity* of LLM-generated recommendations. Users often cannot discern how or why particular outputs were produced, particularly in the absence of explanation interfaces, source attribution, or transparency controls [2, 48]. This undermines user trust, especially when the system produces confident but hallucinated responses [10, 69]. No reviewed system offered mechanisms for real-time explanation or factual verification during conversation. A second issue is the *lack of user control over verbosity and tone.* Unlike rule-based systems with predictable scripting, LLMs may produce overly verbose or stylistically inconsistent responses, frustrating users in goal-oriented tasks [25, 41]. Current systems offer little

Table 4: Implementation Reporting and Tooling across Reviewed CRS Studies (2025–2017). This table summarises the extent of system architecture reporting and the implementation stack used in each study, with a focus on LLM involvement where relevant.

#	Study	Architecture Reporting	Modelling/Implementation Stack
1	Yun et al. (2025) [66]	Reported: GPT-based architecture	OpenAI GPT API, goal-directed vs. open-ended prompting, music domain KB integration
2	Ma & Ziegler (2024) [43]	Reported: Wizard-of-Oz via Telegram	Predefined response dictionary, simulated Telegram interface using pyTelegramBotAPI + MySQL
3	Kraus et al. (2024) [31]	Reported: Wizard-of-Oz with slot-filling	Rule-based Telegram interface, pyTelegramBotAPI, MySQL cloud logging
4	Damian et al. (2024) [60]	Partially reported	Static logic via nutrition DB, Android frontend, rule-based interactions
5	Rana et al. (2024) [55]	Not reported	Study design focused on user critique responses; no CRS implementation described
6	Ma & Ziegler (2023) [42]	Reported: Modular, non-generative CRS	Google Dialogflow (intent classification), modular action-response with UI widget bindings
7	El-Ansari & Beni-Hssane (2023) [11]	Partially reported	Sentiment-informed response variation, rule-based outputs, JavaScript interface
8	Granada et al. (2023) [17]	Reported: GPT-3.5 for response generation	GPT-3.5 used for NLG; candidate reranking fused into prompts
9	Siro et al. (2023) [59]	Reported: Offline UX modelling only	XGBoost trained on post-interaction logs and features, no interactive system
10	Jin & Eastin (2022) [27]	Partially reported	Scripted persona framing in Qualtrics; no real-time CRS logic described
11	Chung & Han (2022) [8]	Reported: Scripted persona variants	Warm/cold framing scripts for e-commerce scenario; non-adaptive
12	Martina et al. (2022) [45]	Reported: Modular NLP pipeline	Dialogflow (intent recognition), CoreNLP (NER, sentiment), Doc2Vec (CB engine), Python backend
13	Silva et al. (2022) [58]	Reported: Scripted dialogue logic	Fixed politeness templates, no adaptive control
14	Cai et al. (2022) [5]	Reported: Static form-based CRS	Slot-based critique flow, template logic, no runtime adaptation
15	Samagaio et al. (2021) [56]	Reported: Rule-based slot-filling	Fixed-response preference chatbot, no ML or NLP used
16	Iovine et al. (2021) [21]	Reported: Scripted dialogue variants	Dialogue variants used to test linguistic framing; no CRS engine
17	Anastasia et al. (2021) [1]	Reported: Embodied scripted agent	Virtual avatar with scenario-driven, scripted interaction
18	Jin et al. (2021) [28]	Not reported	CRS-Que questionnaire development only; no implemented agent
19	Fernando et al. (2021) [13]	Reported: Template-based trait alignment	Static music explanations based on Big Five traits
20	Kraus et al. (2020) [30]	Reported: Alexa Skill implementation	FlaskAsk, Alexa SDK, proactive cue variants in controlled dialogue
21	Iovine et al. (2020) [22]	Reported: Static NAO robot dialogue	Pre-programmed utterances on humanoid robot, MySQL logging
22	Pecune et al. (2019) [52]	Reported: Rule-based trait adaptation	Adaptive social explanations based on Big Five personality traits
23	Lee & Choi (2017) [34]	Partially reported	Scripted chatbot with social framing logic, no live agent described

to no interaction-level modulation of response length, formality, or elaboration, limiting their adaptability to different user preferences or contexts [69, 70]. Evaluation practices remain similarly limited. Constructs such as hallucination tolerance, prompt sensitivity, or conversational agency are rarely operationalised. None of the studies implemented turn-level UX appraisal, adaptive logging, or real-time diagnostics—methods that are increasingly recognised as essential for understanding interactional breakdowns and affective responses in generative dialogue systems. Finally, both studies reviewed were limited to single-session interactions, offering little insight into longitudinal UX effects. As LLM-based CRS move toward deployment in real-world applications, long-term evaluation designs will be critical to assess evolving user expectations, trust calibration, and experiential adaptation [32]. Table 5 summarises UX design and evaluation gaps in LLM-powered CRS, based on synthesis of SLR-included studies and supplementary literature. Addressing these issues will require not only new instrumentation and control mechanisms but also a reconceptualisation of what "good" experience entails in generative, probabilistic dialogue environments.

5.4 UX Design and Research Implications

Our review highlights several implications for improving the design and evaluation of conversational recommender systems. These implications span interaction design, adaptivity, instrumentation, and methodological rigour. Collectively, they advocate for a shift from static, post-hoc evaluations toward dynamic, user-sensitive, and context-aware systems.

Designing for Interactional Responsiveness. Many reviewed CRS implementations used rigid dialogue flows or template-based responses, even when user traits were explicitly collected or inferable. This limits the system's capacity to respond meaningfully to evolving user input, particularly in exploratory or hedonic contexts. Adaptive strategies such as initiative modulation, tone shifting, and contextual prompt framing can improve responsiveness when grounded in user signals [42, 70]. However, adaptivity must remain intelligible to users [62], avoiding hidden decision logic that erodes

Table 5: Summary of UX gaps in LLM-powered conversational recommender systems. Each row highlights a key challenge and associated design priorities.

Challenge Area	Gap / Design Needs
Transparency and Explanation	No explanation interfaces, source attribution, or rationale generation. Future systems should offer real-time source toggles and adaptive explanation levels.
Hallucination and Trust Calibration	Hallucinated content risks undermining trust. UX evaluation must account for how users detect and respond to false or misaligned recommendations. Feedback loops or verification signals are needed.
Control Over Verbosity and Tone	LLMs often generate verbose or inconsistent dialogue. Users need controls to modulate tone, depth, and response length dynamically.
Lack of Turn-Level UX Instrumentation	No system captured per-turn satisfaction, surprise, or trust shifts. Turn-level logs and probes can locate breakdown points.
Short-Term Evaluation Bias	Studies relied on one-off sessions. Longitudinal designs are needed to assess calibration, trust development, and long-term satisfaction.
Absence of Adaptive Dialogue Policies	Despite LLM flexibility, systems did not adapt initiative, verbosity, or content selection to user traits/inputs. Policies should support real-time adaptation.

trust. Emerging work in affect-aware dialogue suggests that subtle interactional cues, such as mirroring affect or responding to hesitation, can enhance conversational alignment [6]. CRS designers should investigate how such responsiveness can be safely and effectively operationalised in real time.

Moving Beyond Trait-Based Personalisation. Although many systems claim to personalise dialogue, most rely on static traits (e.g., personality, demographics) or one-time input. Few demonstrate *real-time adaptation* based on conversational flow or evolving preferences. In contrast, adjacent domains such as intelligent tutoring systems have successfully implemented session-aware models that update adaptively over time [15, 16]. CRS research should adopt similar models that respond to user inputs, engagement patterns, and evolving preferences, shaping not only what is recommended but how it is communicated. Transparent personalisation, where users can understand, contest, or adjust system behaviour, has been shown to foster trust and agency in AI systems [7, 29]. Future CRS interfaces may benefit from lightweight explanation strategies that disclose adaptive logic without overwhelming or distracting the user.

Enhancing UX Instrumentation. UX instrumentation in CRS remains overly reliant on post-task surveys. However, conversational experience unfolds dynamically through micro-events such as turn transitions, delays, clarification requests, and system initiative. These aspects remain invisible to most current measurement approaches. Researchers should incorporate fine-grained instrumentation such as turn-level satisfaction probes, timestamped affect markers, and error recovery logs [41, 54]. These can enable detailed diagnosis of when and why the user experience breaks down—insights that are critical for improving CRS policy design.

Rethinking UX Study Designs. CRS evaluations remain dominated by short-term, lab-based studies. Yet many UX outcomes, such as trust, satisfaction, or perceived utility, are shaped over time. Our review found few longitudinal studies or evaluations in real-world contexts. As CRS systems become embedded in everyday life, evaluation must capture shifting user expectations, adaptation effects, and behavioural drift [63]. Moreover, few studies control for or examine demographic, digital literacy, or accessibility-related factors, limiting the generalisability of UX findings.

Summary. To advance CRS research, evaluation must be embedded into the interaction itself, not merely appended as a post-hoc measure. This includes not only real-time diagnostic tools but also participatory controls that support user agency. System design must move toward transparent adaptivity, and UX research must embrace more nuanced, diverse, and longitudinal methodologies. These are not only technical goals but also ethical imperatives as CRS systems increasingly mediate decisions, identities, and everyday choices.

5.5 Limitations and Future Directions

This review synthesised empirical studies on CRS user experience published up to May 2025, yet several limitations warrant reflection. First, our scope excluded purely technical papers or speculative design proposals lacking UX evaluation, potentially omitting relevant architectural innovations. Future reviews could bridge technical and experiential perspectives by including hybrid studies. Second, construct heterogeneity and inconsistent instrumentation limited comparability. Many studies used bespoke UX items without validation, reducing the cumulative strength of the evidence base. Progress requires shared UX frameworks that support turn-level, affective, and longitudinal evaluation. Third, most evaluations were short-term and conducted in controlled settings. Naturalistic, multi-session deployments remain rare. Long-term field studies are essential to understand sustained user engagement, adaptation, and trust dynamics. Finally, LLM-powered CRS remain under-evaluated. Existing methods are ill-equipped to capture emergent behaviours, hallucination tolerance, or transparency perceptions. As generative CRS proliferate, evaluation approaches must evolve accordingly. Future work must address these methodological gaps by developing validated UX instruments, embracing longitudinal designs, and equipping CRS with tools for interpreting LLM-driven behaviour in diverse, real-world settings.

6 Conclusion

CRSs are poised to transform personalised recommendation, yet their user experience remains poorly understood, especially with the rise of LLMs. This review synthesised evidence from 23 studies (2017–2025) on how UX has been conceptualised, measured, and shaped by adaptivity and generative architectures. We find CRS UX evaluation highly fragmented. Satisfaction and usability

are frequently assessed, but affective, relational, and adaptive dimensions are often neglected. Methods rely largely on post-hoc self-reports, with limited use of behavioural, turn-level, or longitudinal measures. Domain-specific UX patterns highlight the need for context-sensitive evaluation beyond one-size-fits-all metrics. Adaptivity and personalisation appeared in some systems but were usually shallow or unevaluated. Few studies examined how users perceived adaptivity or how it affected trust, satisfaction, or control. LLM-powered CRS promise richer dialogue, yet only two studies exist, both revealing major gaps in transparency, instrumentation, and traceability. This review offers a rigorous synthesis and sets priorities for future work: (1) validated CRS-specific UX instruments for adaptive and multi-turn dynamics; (2) in-situ and longitudinal designs; (3) diagnostic tooling for LLM behaviour; and (4) stratified, inclusive evaluation practices. Advancing CRS UX research requires moving beyond surface-level metrics and static designs toward interactionally sensitive, context-aware, and user-informed evaluation frameworks. Future work can help build CRS that are accurate, efficient, transparent, trustworthy, and engaging.

Acknowledgments

During the preparation of this work the author(s) used ChatGPT 4o in order to improve readability and language of the final draft. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

References

[1] D. Anastasia, Niels van Berkel, and Vassilis Kostakos. 2021. Designing embodied virtual agent in e-commerce system recommendations using conversational design interaction. *International Journal of Human-Computer Studies* 146 (2021).

[2] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).

[3] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101.

[4] Huixian Cai, Weijia Ju, and Sumona R. Chowdhury. 2022. Enhancing user experience with exploration-oriented conversational recommender systems. *User Modeling and User-Adapted Interaction* 32, 1-2 (2022), 43–72.

[5] Wanling Cai, Yucheng Jin, and Li Chen. 2022. Impacts of personal characteristics on user trust in conversational recommender systems. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–14.

[6] Jacky Casas, Timo Spring, Karl Daher, Elena Mugellini, Omar Abou Khaled, and Philippe Cudré-Mauroux. 2021. Enhancing conversational agents with empathic abilities. In *Proceedings of the 21st ACM international conference on intelligent virtual agents*. 41–47.

[7] Li Chen and Pearl Pu. 2012. Critiquing-based recommenders: survey and emerging trends. *User Modeling and User-Adapted Interaction* 22 (2012), 125–150.

[8] Kyung Joon Chung and Heesang Han. 2022. Consumer perception of chatbots and purchase intentions: Anthropomorphism and conversational relevance. *Journal of Retailing and Consumer Services* 64 (2022).

[9] Sooyun Iris Chung and Kwang-Hee Han. 2022. Consumer perception of chatbots and purchase intentions: Anthropomorphism and conversational relevance. *International Journal of Advanced Culture Technology* 10, 1 (2022), 211–229.

[10] Yashar Deldjoo, Z He, J McAuley, A Korikov, S Sanner, A Ramisa, R Vidal, M Sathiamoorthy, A Kasirzadeh, and S Milano. 2024. A Review of Modern Recommender Systems Using Generative Models (Gen-RecSys). ArXiv. 2024. *arXiv preprint arXiv:2404.00579* (2024).

[11] Anas El-Ansari and Abderrahim Beni-Hssane. 2023. Sentiment analysis for personalized chatbots in e-commerce applications. *Wireless Personal Communications* 129, 3 (2023), 1623–1644.

[12] Wael El-Ansari and Abdelali Beni-Hssane. 2023. Sentiment analysis in conversational recommender systems: An empirical study. *Journal of Intelligent Information Systems* 57, 4 (2023), 515–537.

[13] Samantha Fernando, Hyun Lee, and Ji Choi. 2021. Enhancing transparency and user trust in conversational recommender systems. *Expert Systems with Applications* 176 (2021).

[14] Chongming Gao, Wenqiang Lei, Xiangnan He, Maarten De Rijke, and Tat-Seng Chua. 2021. Advances and challenges in conversational recommender systems: A survey. *AI open* 2 (2021), 100–126.

[15] Arthur C Graesser, Patrick Chipman, Brian C Haynes, and Andrew Olney. 2005. AutoTutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions on Education* 48, 4 (2005), 612–618.

[16] Arthur C Graesser, Shulan Lu, George Tanner Jackson, Heather Hite Mitchell, Mathew Ventura, Andrew Olney, and Max M Louwerse. 2004. AutoTutor: A tutor with dialogue in natural language. *Behavior Research Methods, Instruments, & Computers* 36 (2004), 180–192.

[17] Mateo Gutierrez Granada, Dina Zilbershtein, Daan Odijk, and Francesco Barile. 2023. VideolandGPT: A User Study on a Conversational Recommender System. *arXiv e-prints* (2023), arXiv–2309.

[18] Liangwei Hou, Xu Wang, Tong Wu, Jie Xu, Enhong Chen, Yun Xiong, and Lifeng Zhou. 2024. Large Language Models are Zero-Shot Rankers for Recommender Systems. *arXiv preprint arXiv:2402.00852* (2024).

[19] Yikun Hou, Yao Lu, Kun Zhang, Weinan Wang, and Min Li. 2024. Large language models are zero-shot rankers for recommendation. *arXiv preprint arXiv:2402.09686* (2024).

[20] Wenqiang Huang, Xiao Lin, and Yongfeng Zhang. 2025. Towards Agentic Recommender Systems in the Era of Large Language Models. *arXiv preprint arXiv:2403.04550* (2025).

[21] Antonio Iovine, Riccardo Giordano, and Federico Morbidi. 2021. An investigation on the impact of natural language on conversational recommendations. In *Proceedings of the 2021 ACM Conference on Recommender Systems*. 823–827.

[22] Antonio Iovine, Riccardo Giordano, Daniele Nardi, and Federico Morbidi. 2020. Humanoid robots and conversational recommender systems: A preliminary study. In *Proceedings of the 29th IEEE International Conference on Robot & Human Interactive Communication*. 839–844.

[23] Dietmar Jannach. 2023. Evaluating conversational recommender systems: A landscape of research. *Artificial Intelligence Review* 56, 3 (2023), 2365–2400.

[24] Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. 2021. A survey on conversational recommender systems. *ACM Computing Surveys (CSUR)* 54, 5 (2021), 1–36.

[25] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys* 55, 12 (2023), 1–38.

[26] Heekyoung Jin, Yicheng Wang, and Matthew Eastin. 2021. Key qualities of conversational recommender systems: From users perspective. In *Proceedings of the 29th ACM Conference on User Modeling Adaptation and Personalization*. 169–178.

[27] Seung-A. Annie Jin and Matthew S. Eastin. 2022. Birds of a feather flock together: Matched personality effects of product recommendation chatbots and users. *Computers in Human Behavior* 128 (2022).

[28] Yucheng Jin, Li Chen, Wanling Cai, and Pearl Pu. 2021. Key qualities of conversational recommender systems: From users' perspective. In *Proceedings of the 9th International Conference on Human-Agent Interaction*. 93–102.

[29] Bart P Knijnenburg, Martijn C Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. 2012. Explaining the user experience of recommender systems. *User modeling and user-adapted interaction* 22 (2012), 441–504.

[30] Markus Kraus, Stefan Feuerriegel, and Rabia Ozcelebi. 2020. A comparison of explicit and implicit proactive dialogue strategies for conversational recommendation. In *Proceedings of the 28th ACM Conference on User Modeling Adaptation and Personalization*. 151–160.

[31] Matthias Kraus, Stina Klein, Nicolas Wagner, Wolfgang Minker, and Elisabeth André. 2024. A Pilot Study on Multi-Party Conversation Strategies for Group Recommendations. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces*. 1–7.

[32] Sari Kujala, Virpi Roto, Kaisa Väänänen-Vainio-Mattila, Evangelos Karapanos, and Arto Sinnelä. 2011. UX Curve: A method for evaluating long-term user experience. *Interacting with computers* 23, 5 (2011), 473–483.

[33] Jaewon Lee and Ji Choi. 2017. Enhancing user experience with conversational agent for movie recommendation: Effects of self-disclosure and reciprocity. *International Journal of Human-Computer Studies* 103 (2017), 95–106.

[34] SeoYoung Lee and Junho Choi. 2017. Enhancing user experience with conversational agent for movie recommendation: Effects of self-disclosure and reciprocity. *International Journal of Human-Computer Studies* 103 (2017), 95–105.

[35] Wenqiang Lei, Xiangnan He, Maarten de Rijke, and Tat-Seng Chua. 2020. Conversational recommendation: Formulation, methods, and evaluation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2425–2428.

[36] Chuang Li, Hengchang Hu, Yan Zhang, Min-Yen Kan, and Haizhou Li. 2023. A conversation is worth a thousand recommendations: A survey of holistic conversational recommender systems. *arXiv preprint arXiv:2309.07682* (2023).

[37] Yitong Li, Zihan Liu, Changhua Yu, Jiaqi Guo, Yihong Guo, Xiang Chen, and Yongfeng Zhang. 2024. Large Language Models for Generative Recommendation: Review, Opportunities, and Challenges. *arXiv preprint arXiv:2402.10164* (2024).

[38] Zheng Li, Zijian Chen, Defu Chen, Yuying Li, Hongzhi Zhang, and Dawei Yin. 2024. Large Language Models for Generative Recommendation: A Survey. *arXiv preprint arXiv:2402.01294* (2024).

[39] Defu Lian, Xinyang Cao, Defeng Yang, Zhipeng Zhou, Yong Zhang, Xing Xie, and Min Zhang. 2024. RecAI: Leveraging Large Language Models for Next-Generation Recommender Systems. *arXiv preprint arXiv:2402.16124* (2024).

[40] Jie Lian, Xiaoyuan Liang, Xiaopeng Shen, Jie Tang, and Zhiyuan Liu. 2024. RecAI: Leveraging Large Language Models for Next-Generation Recommender Systems. *arXiv preprint arXiv:2403.01889* (2024).

[41] Ewa Luger and Abigail Sellen. 2016. " Like Having a Really Bad PA" The Gulf between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 5286–5297.

[42] Yuan Ma and Jürgen Ziegler. 2023. Initiative transfer in conversational recommender systems. In *Proceedings of the 17th ACM Conference on Recommender Systems*. 978–984.

[43] Yuan Ma and Jürgen Ziegler. 2024. The effect of proactive cues on the use of decision aids in conversational recommender systems. In *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*. 305–315.

[44] Yuan Ma and Jürgen Ziegler. 2024. Investigating meta-intents: user interaction preferences in conversational recommender systems. *User Modeling and User-Adapted Interaction* 34, 5 (2024), 1535–1580.

[45] Simone Martina, Arnaud Muller, and Francesco Ricci. 2022. Narrative recommendations based on natural language preference elicitation for a virtual assistant for the movie domain. *Journal of Artificial Intelligence Research* 74 (2022).

[46] Kevin McCarthy, James Reilly, Lorraine McGinty, and Barry Smyth. 2004. On the dynamic generation of compound critiques in conversational recommender systems. In *Adaptive Hypermedia and Adaptive Web-Based Systems: Third International Conference, AH 2004, Eindhoven, The Netherlands, August 23-26, 2004. Proceedings 3*. Springer, 176–184.

[47] Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica* 22, 3 (2012), 276–282.

[48] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.

[49] David Moher, Alessandro Liberati, Jennifer Tetzlaff, and Douglas G. Altman. 2009. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine* 6, 7 (2009).

[50] Mourad Ouzzani, Hossam Hammady, Zbys Fedorowicz, and Ahmed Elmagarmid. 2016. Rayyan—a web and mobile app for systematic reviews. *Systematic reviews* 5 (2016), 1–10.

[51] Florian Pecune, Shruti Murali, Vivian Tsai, Yoichi Matsuyama, and Justine Cassell. 2019. A Model of Social Explanations for a Conversational Movie Recommendation System. In *Proceedings of the 7th International Conference on Human-Agent Interaction*. ACM, Kyoto Japan, 135–143. doi:10.1145/3349537.3351899

[52] Florian Pecune, Magalie Ochs, and Catherine Pelachaud. 2019. An exploratory study of user experience in a sentiment-aware conversational agent. *International Journal of Human-Computer Studies* 131 (2019).

[53] Dhanya Pramod and Prafulla Bafna. 2022. Conversational recommender systems techniques, tools, acceptance, and adoption: a state of the art review. *Expert Systems with Applications* 203 (2022), 117539.

[54] Amanda Purington, Jessie G Taft, Shruti Sannon, Natalya N Bazarova, and Samuel Hardman Taylor. 2017. " Alexa is my new BFF" social roles, user satisfaction, and personification of the Amazon Echo. In *Proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems*. 2853–2859.

[55] Arpit Rana, Scott Sanner, Mohamed Reda Bouadjenek, Ronald Di Carlantonio, and Gary Farmaner. 2024. User experience and the role of personalization in critiquing-based conversational recommendation. *ACM Transactions on the Web* 18, 4 (2024), 1–21.

[56] Rui Samagaio, David Carneiro, and Paulo Novais. 2021. A chatbot for recipe recommendation and preference modeling. In *Proceedings of the 19th IEEE International Conference on Smart Technologies*. 130–137.

[57] Martin Schrepp, Andreas Hinderks, and Jörg Thomaschewski. 2017. Design and evaluation of a short version of the user experience questionnaire (UEQ-S). *International Journal of Interactive Multimedia and Artificial Intelligence* 4, 6 (2017), 103–108.

[58] Tiago Silva, Diana Gonçalves, and Nelson Guimarães. 2022. Polite task-oriented dialog agent for enhancing user experience. In *Proceedings of the 21st ACM International Conference on Multimodal Interaction*. 243–252.

[59] Clemencia Siro, Mohammad Aliannejadi, and Maarten De Rijke. 2023. Understanding and predicting user satisfaction with conversational recommender systems. *ACM Transactions on Information Systems* 42, 2 (2023), 1–37.

[60] Damian Thom, Jefferson Ortega, and Rosa Felix. 2024. NutrIA: An Out-of-the-Standard Nutritional Recommendation Mobile Application Powered by Artificial Intelligence. In *2024 IEEE 4th International Conference on Advanced Learning Technologies on Education & Research (ICALTER)*. IEEE, 1–4.

[61] Nava Tintarev and Judith Masthoff. 2015. Explaining recommendations: Design and evaluation. In *Recommender systems handbook*. Springer, 353–382.

[62] Chun-Hua Tsai and Peter Brusilovsky. 2021. The effects of controllability and explainability in a social recommender system. *User Modeling and User-Adapted Interaction* 31 (2021), 591–627.

[63] Ruben S Verhagen, Alexandra Marcu, Mark A Neerincx, and Myrthe L Tielman. 2024. The Influence of Interdependence on Trust Calibration in Human-Machine Teams. In *HHAI 2024: Hybrid Human AI Systems for the Social Good*. IOS Press, 300–314.

[64] Qi Wang, Yujian Zhu, Yu Wang, Hua Cheng, Weinan Xu, and Jing Li. 2023. Rethinking the Evaluation for Conversational Recommender Systems. *arXiv preprint arXiv:2305.10793* (2023).

[65] Likang Wu, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, et al. 2024. A survey on large language models for recommendation. *World Wide Web* 27, 5 (2024), 60.

[66] Sojeong Yun and Youn-kyung Lim. 2025. User Experience with LLM-powered Conversational Recommendation Systems: A Case of Music Recommendation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–15.

[67] Subiya Zaidi, Sawan Rai, and Kapil Juneja. 2024. A Review of Existing Conversational Recommendation Systems. In *2024 2nd International Conference on Disruptive Technologies (ICDT)*. IEEE, 22–26.

[68] Yizhe Zhang, Yucheng Jin, Li Chen, and Ting Yang. 2024. Navigating user experience of chatgpt-based conversational recommender systems: The effects of prompt guidance and recommendation domain. *arXiv preprint arXiv:2405.13560* (2024).

[69] Zihuai Zhao, Wenqi Fan, Jiatong Li, Yunqing Liu, Xiaowei Mei, Yiqi Wang, Zhen Wen, Fei Wang, Xiangyu Zhao, Jiliang Tang, et al. 2024. Recommender systems in the era of large language models (llms). *IEEE Transactions on Knowledge and Data Engineering* (2024).

[70] Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics* 46, 1 (2020), 53–93.