Understanding and Improving Decision-Making UX in Conversational E-Commerce Agents via Evidence Selection

Tianqi Liu[1]

[1]University of Delaware, Newark, DE, USA

Abstract

Conversational e-commerce agents are changing how users interpret review evidence and form purchase judgments, yet existing systems often optimize for "answering questions" or "summarizing reviews," with less direct focus on users' decision-making experience. We formulate this as a decision-making UX problem and introduce a decision-making UX gap, where the system can produce fluent answers, but users still struggle to judge whether an answer is trustworthy, covers key risks, and directly supports a decision. To narrow this gap, we propose DEEP-Agent (Decision-Enhancement Evidence-Processing Agent). Instead of optimizing a single model-level metric in isolation, we treat UX principles as constraints on evidence selection and presentation—through intent recognition, risk-coverage constraints, evidence anchoring, and confidence expression—to improve decision usability for consumers. We conduct systematic evaluation on Amazon Handmade review and Q&A data and a controlled exploratory user study. Results show that DEEP-Agent yields gains on trust, decision satisfaction, decision confidence, and time-to-decision.

Keywords

Conversational commerce, Decision-making UX, E-commerce agent, Review-based QA, Evidence selection, Trust, Cognitive load

1. Introduction

With the growth of conversational commerce, AI shopping agents (e.g., Amazon Rufus) have become a major entry point for product information and purchase decisions. Users no longer only search for product attributes; they ask questions closer to real shopping situations, e.g., "Is this suitable as a gift?" "Is the quality reliable?" [1]

Such questions look like QA on the surface but are in essence decision support questions [2]. Users expect the AI not merely to give an answer, but to help them understand review perspectives, weigh pros and cons, identify risks, and reach an actionable judgment.

However, real-world AI shopping assistants still show clear limitations. On one hand, audit-based evaluation finds that commercial shopping agents vary substantially under different linguistic formulations and input noise, hurting quality-of-service [3]. With LLM-driven assistants deployed in production, work also reports instability and inconsistent behavior across contexts [3]. On the other hand, industry observation suggests some assistant answers read more like marketing summaries than user-centered, objective decision support. These issues weaken perceived helpfulness, because even fluent answers may send users back to raw reviews to verify.

We summarize this as a decision-making UX gap in conversational e-commerce agents. In short, the system can perform UGC-based information retrieval [4] or summarize UGC, yet users still lack a low-burden, verifiable interaction experience that supports purchase judgment.

We argue that improving answer correctness alone is insufficient for conversational commerce; ease of making a purchase judgment should be treated as a primary optimization target alongside fluency.

We address three research questions. First, in review-grounded shopping QA, what factors shape users' decision-making experience? Second, if we strengthen evidence selection on the input side using decision usefulness in UGC scenarios, and include risk cues and confidence expression on the output

Jul 24, 2026, Melbourne, Australia

anneqiqi@udel.edu (T. Liu)

side, can we reduce cognitive burden and increase user trust? Third, how do users' interaction behaviors differ under different answer formats?

Our contributions are to define a decision-making UX gap; to propose DEEP-Agent as a framework for controllable, interpretable evidence selection and presentation aligned with UX goals; and to validate effects on decision-oriented experience through offline evaluation and a user study.

2. Related Work

2.1. Conversational IR, Conversational Recommendation, and Review-Based QA

Conversational information retrieval emphasizes collaboration between the system and the user to achieve retrieval goals [5]; neural conversational IR further organizes modeling directions in the deep learning era [6]. On this basis, user-generated content (UGC) is widely used for information retrieval and question answering in e-commerce. Early work formalizes review-based QA as retrieving relevant snippets from reviews and answer generation to respond to queries [7]. For example, RAGE (Retrieval-Augmented Generation) uses a retrieval + generation pipeline to extract relevant sentences and produce fluent answers with broader coverage [8]; the general RAG paradigm is also used for knowledge-intensive NLP tasks [9]. AmazonQA scales real user Q&A paired with reviews [10]. To exploit subjective signals, opinion-aware models incorporate sentiment and preference into generation [11].

Conversational recommender systems (CRS) typically follow intent understanding, candidate retrieval, and multi-turn feedback; surveys and evaluation perspectives appear in [12, 13, 14], and multi-turn interactive recommendation further extends user–system collaboration [15]; a user-centric framework for evaluating CRS is in [16]. Yet much of this line still optimizes system metrics such as relevance, faithfulness, and answer quality, with less explicit modeling of how users use answers to complete purchase decisions. In other words, the system may produce a correct summary without making users more confident to decide.

2.2. Information Overload, Cognitive Load, Trust, and Human–AI Collaboration

Information overload and its consequences have a long history in information science and user studies [17, 18]; cognitive load theory provides a psychological basis for how structured external representations reduce integration cost [19]. In modeling UGC helpfulness (e.g., review helpfulness, RHP), usefulness is often assessed along information richness, cognitive burden, and veracity [20, 21]. In shopping, these are not independent objectives but trade off. Richer information may cover more user-relevant facets but increase reading burden; shorter answers are easier to skim but may hide important risks; clearer AI conclusions may aid decisions, but without evidence users may not trust them.

Accuracy and related factors also affect trust calibration for machine learning models [22]. HCI work proposes guidelines for human–AI interaction [23], and surveys on human–AI collaboration summarize research agendas [24]. Thus the core challenge for e-commerce agents is not merely maximizing information, but balancing information richness, cognitive load, and trustworthiness for user experience.

3. Problem Definition

3.1. Task Definition

Given review set $R = \{r_1, \ldots, r_n\}$ for target product a and user question q, the system produces a decision-support answer y. Unlike conventional review-based QA, we optimize not only "whether the answer is relevant" but also decision usability, formalized as

$$y^* = \arg\max_y \mathcal{U}(y|q, R), \quad \mathcal{U} = f(\text{Richness}, -\text{CognitiveLoad}, \text{Trust})$$

Here Richness is coverage of key decision facets; CognitiveLoad is comprehension cost; Trust is evidence verifiability and risk transparency.

3.2. Decision-Making UX Gap

On product pages with massive reviews, users face information overload. In traditional browsing, users may manually filter high-vote, low-star, and recent reviews to form a judgment. AI shopping assistants should lower this cost, yet generic summaries still leave decision uncertainty, evidence opacity, and cognitive switching cost, which together constitute a decision-making UX gap.

We summarize the mismatch between typical assistants and user needs in Table 1.

Table 1
User needs vs. typical assistant outputs (UX gap).

User need	Typical system output	UX gap
Judge fit to one's scenario	Generic product summary	Weak scenario alignment
Understand main pros and risks	Mostly positive framing	Insufficient risk signals
Judge whether the answer is trustworthy	No or weak sourcing	Weak evidence anchoring
Decide quickly	Long unstructured prose	Unclear decision structure

Accordingly, our design goal for the Agent is to help users understand review evidence and form purchase judgments faster, with less effort, and with greater confidence.

4. DEEP-Agent: An Evidence-Selection Framework for Decision-Oriented UX

DEEP-Agent (Decision-Enhancement Evidence-Processing Agent) frames review QA as optimizing practical decision support for shoppers. Inputs are target ASIN, UGC-grounded natural-language question q, and the product's review set. Output is not a single summary but a structured, decision-oriented answer with overall judgment, core evidence, pros and cons, risk notes, review citations, and confidence expression.

DEEP-Agent has three parts. At retrieval, we rank evidence by decision usefulness in UGC scenarios, where high-usefulness UGC tends to be helpful, trustworthy, intent-aligned, and to carry necessary risk signals. At reasoning, we generate from ranked evidence and extract theme coverage, supporting points, concerns, and representative quotes. At generation, we produce information-rich text and confidence expression.

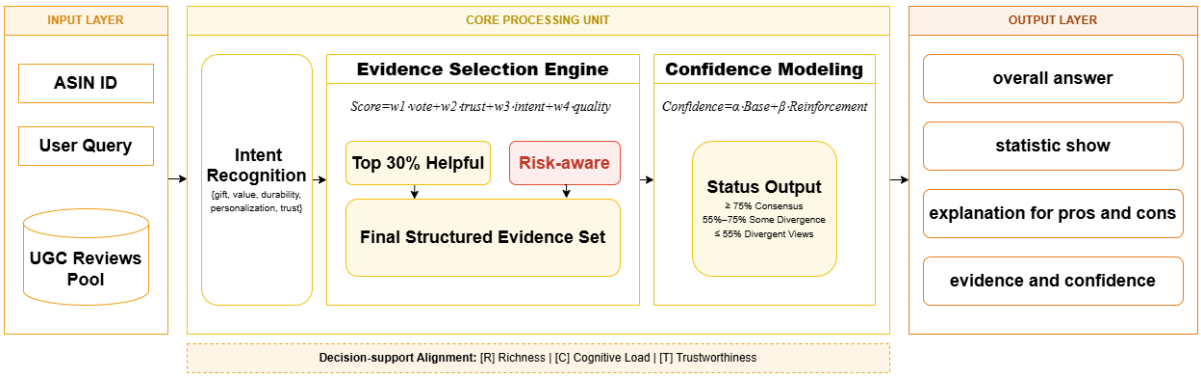

Figure 1: Overview of the DEEP-Agent pipeline: intent recognition, BM25 retrieval with decision-weighted re-ranking and risk-aware retention, confidence estimation, and structured answer assembly (judgment, pros and cons, anchored evidence).

Conversational recommender systems (CRS) often build an Agent pipeline through intent understanding, candidate retrieval, intent-conditioned reasoning, and response generation [12, 13, 14, 15].

4.1. Evidence Selection and Reasoning

At evidence selection, we first retrieve relevant reviews with BM25, then rank reviews with an aggregate decision-helpfulness score. The score is defined as follows.

$$\text{score}(r) = \sum_{k=1}^{K} w_k \cdot x_k(r), \quad w \in \mathbb{R}^K_{\geq 0},\ \sum_{k=1}^{K} w_k = 1$$

$$x(r) = \left[\tilde{p}(r), t(r), m(r), i(r), \ell(r)\right]$$

Per-review features are defined as follows. General helpfulness $\tilde{p}$ is strength from `helpful_vote`. Trust cue t indicates verified purchase or not. Intent match m is relevance to current decision intent (BM25). Multimedia cue i marks image or image URL present. Risk signal (keyword proxy) captures mentions of damage, disappointment, shipping, quality, and related issues. Length as an information proxy ℓ is review text length.

To ensure a neutral and unbiased baseline, we instantiate the model with a uniform weighting scheme. Specifically,

$$w_k = \frac{1}{K}, \quad \forall k$$

This removes manual preference over feature importance and allows us to evaluate the intrinsic effectiveness of the feature set and downstream decision-making performance.

The weighting vector w lies in a continuous parameter space, so in principle it supports learning-based instantiation. In an extended setting, w can be optimized using a linear utility model or pairwise learning-to-rank objective [25], allowing the scoring function to adapt to observed user preferences or decision signals.

To avoid selecting only high-scoring positive reviews, we add risk review supplementation. We take top-n by score and top-n_r by risk, merge, deduplicate, re-rank, and truncate to the target size, preserving both supportive and counter-evidence under a fixed display budget.

Unlike standard RAG pipelines that optimize for relevance, DEEP-Agent explicitly optimizes for decision usability, where evidence selection is constrained by risk coverage, scenario alignment, and cognitive budget, among other factors. Reasoning on these enriched cues highlights the most decision-relevant information for the user's question; extracted pros, cons, and quotes align better with decision needs than a shallow summary over all reviews. Users receive richer, more trustworthy information for deciding, improving shopping UX.

4.2. Confidence Modeling and Risk Expression

In purchase decisions users care not only what the answer is, but how reliable it is. DEEP-Agent aggregates helpfulness, trustworthiness, risk, and intent-match signals on evidence subset E into a confidence estimate. When presenting confidence to users, we may draw on discussions of neural network calibration [26] to turn numbers into understandable uncertainty wording. Tunable mixing weights $\alpha_1, \alpha_2, \alpha_3 \geq 0$ with $\alpha_1 + \alpha_2 + \alpha_3 = 1$, $\beta_1, \beta_2, \beta_3 \geq 0$ with $\beta_1 + \beta_2 + \beta_3 = 1$, and $\gamma \in [0, 1]$ are fixed in our experiments.

$$\text{base} = \text{clip}\big(\alpha_1\, \overline{\tilde{p}} + \alpha_2\, \overline{t} + \alpha_3\, (1 - \overline{r}), 0, 1\big)$$

$$\text{reinf} = \text{clip}\big(\beta_1\, \overline{m} + \beta_2\, \overline{t} + \beta_3\, (1 - \overline{r}), 0, 1\big)$$

$$\text{conf} = \text{clip}\big(\gamma\, \text{base} + (1 - \gamma)\, \text{reinf}, 0, 1\big)$$

In user-facing text we omit formulas and rely on natural language, for example by stating that review evidence is fairly consistent so the conclusion is relatively reliable, that reviews disagree and size

and shipping feedback should be checked closely before buying, or that most ratings are positive but negatives cluster on packaging damage.

The aim is to communicate uncertainty without overconfident phrasing.

Building on the above, the final answer extends a clear-answer-and-explanation backbone with explicit risks and counterpoints under a "things to be aware of" rubric, representative quotes or anchored evidence as customer evidence, and a confidence line that states evidence consistency and suggested use.

For example, for "Is this a good gift for someone who likes handmade crafts according to the reviews?", a structured answer covers gift fit, quality, personalization, shipping risk, and sizing—not only "it is a good gift." Users locate what matters faster and decide whether to read raw reviews. When reviews strongly disagree on a facet, the Agent flags disagreement and asks users to judge, rather than sticking to a marketing tone—improving perceived trustworthiness.

5. User Study

We conducted a comparative user study to assess how alternative agent answers affect decision-making UX. $N = 150$ participants (81 female, 69 male; mean age 21.3, SD 3.1), from campus volunteers and online recruitment. We use a within-subject design in which each participant completes multiple decision tasks (Baselines A–D) in randomized order, with scales after each task and time-to-decision logged. We aim not for population-level generalization but for consistent UX effects across tasks and conditions.

5.1. Task

Participants complete realistic shopping-decision tasks. Each includes product context, one user question, and one AI answer. Questions follow real product Q&A style, for example whether an item is a good gift for someone who likes handmade crafts, whether it would last long according to reviews, whether personalization quality is reliable, or whether it is worth buying for the price.

After reading, participants indicate whether they would buy, not buy, or inspect more reviews based on the answer.

5.2. Conditions

The design remains within-subject across four conditions. Baseline A is a Rufus-style summary from Amazon's in-production shopping assistant. It is a black-box, often marketing-leaning system with a clear answer and rationale, review citations, and risks and counterpoints rare in roughly one out of twelve cases, and it provides no confidence explanation. Responses for Baseline A were manually queried on Amazon for each experimental ASIN and matching question, saved as text, and used as fixed stimuli. Baseline B is a relevance-optimized summary based on classic BM25 RAG with ranking by relevance score, clear answer and rationale, citations, risks and counterpoints present, and no confidence. Baseline C is a decision-score-optimized summary after BM25 retrieval with ranking by decision-score, clear answer and rationale, high-value citations, risks and counterpoints, and no confidence. Treatment D is DEEP-Agent, which after BM25 retrieval uses decision-score ranking, a clear answer and rationale, high-value citations, risks and counterpoints, and includes a confidence explanation.

5.3. Measures

Table 2
User study measures.

Category	Metric	Measurement	Interpretation
Subjective	Answer satisfaction	1–5 Likert	Overall satisfaction with the answer
Subjective	Cognitive load	1–5 Likert	Difficulty understanding and processing
Subjective	Trust perception	1–5 Likert	Perceived credibility and objectivity
Behavioral	Decision time	seconds	Time to decision
Behavioral	Verification behavior	binary (click or no click)	Opens raw reviews for more checking

5.4. Hypotheses

We expect decision-score ranking (C and D) compared with relevance ranking (B) and black-box summary (A) to improve decision satisfaction. We further expect decision-score systems (C and D) compared with relevance (B) and black-box (A), with better evidence and risk retention, to improve trust. We expect explicit confidence (D) compared with no confidence (A–C) to help trust and satisfaction, while verification and decision time may trade off with shorter or simpler summaries, which we leave to empirical results.

5.5. Ethics and Data Handling

We collect only anonymized task ratings and durations, no identifiable personal data; participants may withdraw without penalty. Instructions state answers come from different systems, tasks are for research only, not purchase advice. Statistics are reported in aggregate by condition.

6. Experiments

6.1. Dataset

We build a large review corpus from Amazon Handmade (664,162 reviews with rating, verified purchase, text, and image links) [27]. With existing Q&A data [28], we keep 150 ASINs with ≥ 50 reviews and ≥ 20 Q&As.

We drop questions not answerable from a single ASIN's reviews (cross-product, out-of-scope). We embed questions per ASIN, K-means into five clusters for coverage, then sample 20 questions per ASIN, forming 3000 $(\text{ASIN}, \text{question})$ pairs.

6.2. Setup

We implement the pipeline with GPT-5 as the backbone LLM for generative baselines and our method. Evaluation mixes human annotation and automation; human judgments are primary for correctness and usefulness. Gemini 2.5 Flash serves as an auxiliary LLM judge for scalable, consistent auto-scoring.

6.3. Offline Evaluation

Relevance is embedding similarity between answer and question. Faithfulness uses an LLM judge to assess whether claims are supported by reviews. Coverage uses an LLM judge to summarize key review facets, then scores answer coverage.

7. Results and User Effects

7.1. System-Level Results

In most runs, DEEP-Agent links conclusion, evidence, risk, and confidence in its outputs, which yields comparatively strong gains on justification quality and trust-related metrics in our offline evaluation,

although we occasionally observe weaker grounding when review signals are sparse or stylistically repetitive.

Table 3
Offline metrics (higher is better for all columns).

Method	Relevance	Faithfulness	Coverage
Baseline B: relevance-optimized summary	0.86	0.91	0.71
Baseline C: decision-score-optimized summary	0.74	0.83	0.76
Baseline A: Rufus-style summary	0.82	0.75	0.69
DEEP-Agent	0.84	0.80	0.81

DEEP-Agent is not necessarily highest on raw relevance or faithfulness, but coverage is strongest, with the other two still competitive—consistent with prioritizing decision-relevant evidence coverage. The margin on coverage is not always large across ASIN subsets, so we do not claim uniform dominance on every product–question pair.

7.2. User Study Results

Table 4
User study outcomes by condition (satisfaction and trust: higher is better; cognitive load, decision time, and verification rate: lower is better).

Condition	Sat.	Load	Trust	Time (s)	Verif.
Baseline A: Rufus-style	3.8	2.9	3.6	44.2	61%
Baseline B: Relevance RAG	4.1	3.9	4.0	61.4	59%
Baseline C: Decision-score RAG	4.6	3.4	4.3	55.7	56%
Treatment D: DEEP-Agent	4.8	3.5	4.6	58.3	58%

On subjective experience and trust, DEEP-Agent (D) scores highest on answer satisfaction (4.8) and trust (4.6); decision-score RAG (C) is next; Rufus-style (A) lowest—aligned with H1 and H2 on decision-oriented evidence and structured expression. On cognitive load and speed, A has lowest load (2.9) and shortest time (44.2s); D and C are mid (D: 3.5, 58.3s); B slowest and heaviest. Short, marketing-leaning summaries still win on "read fast, think less"; denser structured answers show a time–quality tradeoff. In optional remarks after tasks, several participants reported that they "skimmed and clicked quickly" under Baseline A, often without checking cited evidence, whereas others said Condition D "took longer to read" but made risks easier to weigh. On verification, C has lowest verification rate (56%); D (58%) beats A (61%) and B (59%) but not C overall. D's edge is higher trust and satisfaction, not dominance on every behavioral metric.

8. Discussion

8.1. Why a UX-Driven Framing?

Viewed purely as a system architecture, the work is evidence selection paired with generation; the practical question in e-commerce is whether users can decide with less friction. We therefore frame the contribution in terms of shopping decision experience, not only nominal answer quality.

This perspective links IR, dialogue, user behavior, and purchase decisions. IR finds relevant evidence in reviews, LLMs organize and phrase that evidence, and UX evaluation tests whether users trust more, strain less, and decide faster.

8.2. Alignment with User Results: Strengths and Tradeoffs

Table 4 shows DEEP-Agent leads on satisfaction and trust, matching our emphasis on anchoring, risk, and confidence. Rufus-style (A) is best on cognitive load and decision time, consistent with "less text,

smoother skim and thus faster processing"; that need not mean better deliberation—participants may rely on heuristics and click earlier. A few participants explicitly contrasted the same product under A versus D, noting that A "felt faster for a first impression" while D "slowed them down but reduced guesswork." Decision-score RAG (C) reaches strong satisfaction and lowest verification without explicit confidence text—better ranking alone partly eases "need to open reviews"; D's gain over C is mainly trust and satisfaction, with slightly worse verification.

Our claim is not that we are faster than every baseline everywhere. Rather, within acceptable reading time we prioritize credibility and overall satisfaction, and if the product goal is minimal latency or ultra-light UI, one should explicitly trade off against short-summary paradigms or adapt length.

8.3. Limitations

User-study scale is limited ($n = 150$); it is exploratory evidence, not population inference—larger, cross-platform replication is needed. Tasks are controlled, not real checkout; ecological validity needs online controlled experiments and longitudinal logs. Load, time, and verification depend on UI, instructions, and participant strategy; D is not uniformly "fastest and easiest," so do not generalize to all journeys. Helpful votes, verified purchase, and images are imperfect trust proxies for fraud or platform bias. Rufus-style is described as a representative industry pattern rather than a full system reimplementation; comparisons reflect interaction paradigm differences, not absolute ranking of a commercial product.

9. Conclusion

We recast review-grounded conversational e-commerce QA as decision-making UX. Current AI shopping assistants produce fluent summaries but still show weak scenario fit, opaque evidence, insufficient risk, and limited trust.

We propose DEEP-Agent—intent recognition, review evidence selection, pros and cons, risk notes, and confidence—to turn raw reviews into structured answers for decisions; the goal is to instantiate UX principles, not to chase a single "best model" alone (§4). User results show DEEP-Agent beats baselines on answer satisfaction and trust; shortest Rufus-style summaries may still win on cognitive load and decision time; decision-score RAG is slightly better on verification rate than DEEP-Agent—illustrating tradeoffs across experience dimensions. Evaluating shopping agents should report satisfaction, trust, burden, speed, and verification alongside offline coverage and faithfulness, and avoid overrating or dismissing a method on one behavioral metric alone.

Overall, in conversational commerce, improving answer correctness alone is not enough; decision usability deserves first-order status alongside traditional relevance. Future evaluation should place decision-centered UX metrics next to classical IR metrics.

Future work includes personalized decision preferences, multimodal review evidence, multi-turn clarification, and long-term behavioral effects on real platforms.

Declaration on Generative AI

This paper involves generative AI for e-commerce decision support. AI-assisted tools were used for drafting and language refinement, while the research design, experiments, and conclusions were fully conducted and validated by the authors.

References

[1] J. Trippas, et al., Conversational search: Current research and open questions, ACM SIGIR Forum (2020).

[2] Anonymous, Shopping with a platform AI assistant: Who adopts, when in the journey, and what for, arXiv preprint arXiv:2603.24947 (2026).

[3] E. Harvey, R. F. Kizilcec, A. Koenecke, A framework for auditing chatbots for dialect-based quality-of-service harms, arXiv preprint arXiv:2506.04419 (2025).
[4] S. Zhang, et al., Product question answering with customer reviews, in: Proceedings of the Web Conference (WWW), 2018.
[5] F. Radlinski, N. Craswell, Cooperative and conversational information retrieval, in: Proceedings of SIGIR, 2017.
[6] J. Gao, M. Galley, L. Li, Neural approaches to conversational information retrieval, Foundations and Trends in Information Retrieval 13 (2018) 127–298.
[7] J. McAuley, A. Yang, Addressing complex and subjective product-related queries with customer reviews, in: Proceedings of the Web Conference (WWW), 2016.
[8] Q. Chen, Z. Lin, Y. He, RAGE: Retrieval-augmented generation for product question answering, in: Proceedings of WSDM, 2019.
[9] P. Lewis, et al., Retrieval-augmented generation for knowledge-intensive NLP tasks, in: Advances in Neural Information Processing Systems (NeurIPS), 2020.
[10] R. Gupta, et al., AmazonQA: A review-based question answering dataset, arXiv preprint arXiv:1908.04364 (2019).
[11] Y. Deng, Y. Shen, et al., Opinion-aware neural networks for review-based question answering, in: Proceedings of CIKM, 2020.
[12] D. Jannach, A. Manzoor, W. Cai, L. Chen, A survey on conversational recommender systems, arXiv preprint arXiv:2004.00646 (2020).
[13] Y. Zhang, et al., Conversational recommender systems: A survey, ACM Computing Surveys (2021).
[14] Y. Sun, S. Zhang, et al., Conversational recommender systems: A survey of the state of the art, ACM Computing Surveys (2023).
[15] S. Zhang, L. Yao, et al., Interactive recommendation with multi-round feedback, in: Proceedings of IJCAI, 2020.
[16] T. Chen, R. Zhang, et al., Evaluating conversational recommender systems: A user-centric perspective, in: Proceedings of RecSys, 2023.
[17] M. J. Eppler, J. Mengis, The concept of information overload: A review of literature, The Information Society 20 (2004) 325–344.
[18] D. Bawden, L. Robinson, The dark side of information: Overload, anxiety and other paradoxes, Journal of Information Science 35 (2009) 180–191.
[19] J. Sweller, Cognitive load theory, in: Psychology of Learning and Motivation, volume 55, 2011, pp. 37–76.
[20] X. Li, Y. Zhang, T. Chen, CRCNet: Consistency-aware multimodal review helpfulness prediction framework, IEEE Access 10 (2022) 45678–45689. doi:10.1109/ACCESS.2022.3185678.
[21] A. Ghose, P. G. Ipeirotis, Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics, IEEE Transactions on Knowledge and Data Engineering 23 (2011) 1498–1512.
[22] M. Yin, J. Wortman Vaughan, Understanding the effect of accuracy on trust in machine learning models, in: Proceedings of CHI, 2019.
[23] S. Amershi, D. Weld, et al., Guidelines for human-AI interaction, in: Proceedings of CHI, 2019.
[24] V. Lai, C. Tan, Human-AI collaboration: A review and research agenda, arXiv preprint arXiv:2001.09990 (2020).
[25] C. J. C. Burges, From RankNet to LambdaRank to LambdaMART: An overview, in: Learning to Rank for Information Retrieval Workshop, 2010.
[26] C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, On calibration of modern neural networks, in: Proceedings of ICML, 2017.
[27] J. Ni, J. Li, J. McAuley, Justifying recommendations using distantly-labeled reviews and fine-grained aspects, in: Proceedings of EMNLP, 2019, pp. 188–197.
[28] Y. Hou, J. Li, Z. He, A. Yan, X. Chen, J. McAuley, Bridging language and items for retrieval and recommendation, arXiv preprint arXiv:2403.03952 (2024).