# Cleo: A Transparent and Controllable Chatbot for Conversational Commerce

Kevin Schott
kevin.schott@gesis.org
GESIS – Leibniz Institute for the Social Sciences
Cologne, Germany

Jan Lattenkamp
jan.lattenkamp@gesis.org
GESIS – Leibniz Institute for the Social Sciences
Cologne, Germany

Daniel Hienert
daniel.hienert@gesis.org
GESIS – Leibniz Institute for the Social Sciences
Cologne, Germany

Dagmar Kern
dagmar.kern@gesis.org
GESIS – Leibniz Institute for the Social Sciences
Cologne, Germany

## Abstract

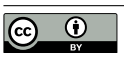
We demonstrate Cleo, a transparent and controllable conversational product advisor that addresses the challenges of opacity, unpredictability of LLMs, and the complexity of comparisons in conversational commerce. With our chatbot system, we make four contributions: First, we introduce transparency by prompting the LLM to reflect on interpreted user needs, while an auditable ranking mechanism reveals loss values per attribute, explaining ranking decisions. Second, we propose controllability through a hybrid architecture separating deterministic ranking from language generation. A ranker applies categorical filters and numeric loss functions over 3,638 product specifications. Meanwhile, a constrained LLM generates grounded descriptions constrained to catalog evidence, thus mitigating the risk of hallucinated or persuasive content. Third, we provide decision support in the form of natural-language comparisons and a highlights feature. These aim to reduce mental workload by contextualizing specifications relative to user needs. Fourth, we contribute an extensible experimental system for IR and HCI researchers, as well as practitioners of conversational search and recommendation. Unlike traditional faceted search or opaque LLM-only recommenders, our approach allows for fluid conversation while maintaining algorithmic transparency. In a live demonstration, attendees will experience information needs elicitation and reflection, conversational refinement with real-time re-ranking, inspection of per-attribute loss explanations, and AI-generated multi-item comparisons. The system aims to advance the design of transparent and controllable conversational systems that provide support for decision-making during online product search.

## CCS Concepts

• **Information systems** → **Users and interactive retrieval**; **Recommender systems**; • **Human-centered computing** → **Natural language interfaces**; • **Applied computing** → *Online shopping*.

## Keywords

Conversational search, conversational commerce, transparent ranking, hybrid LLM systems, explainable recommendations, decision support





## 1 Introduction & Contribution

Conversational commerce, the buying activity conducted via a digital assistant, aims to provide chat- or voice-based product advice that mimics a human shop assistant: the assistant elicits needs and preferences based on which it recommends suitable items [2, 13, 22, 23]. Since these assistants act as interactive, socially present decision-making aids, they can foster trust in e-commerce platforms, which is a prerequisite for purchase intentions [6, 24].

The process of online shopping with a digital assistant combines search and recommendation [16, 18, 28]. Contemporary conversational recommenders promise multi-turn preference elicitation beyond one-shot results lists, yet dialogic answers often hide ranking logic and underlying evidence [5, 8]. For conversational search, Radlinski and Craswell [16] introduced the concept of *revealment*: helping users express or discover their needs (user revealment) and exposing the system's capabilities and corpus (system revealment). We adapt their notion of system revealment to also encompass surfacing internal decision-making, allowing users to assess system capabilities. Łajewska et al. [9] show that in conversational information systems, where details like source attribution and model confidence are often concealed, users struggle to detect factual errors and biases in responses.

Many shopping tasks are faceted and multi-attribute, with users often uncertain which attributes matter or how to reference them [16]. Mixed-initiative conversation supports piecewise elicitation and lightweight teaching [16], while critiquing enables attribute-level refinement by the user through natural-language feedback [4]. In addition, clarifying questions from the assistant can reduce ambiguity before showing results, improving retrieval effectiveness [1].

However, comparing and evaluating retrieved products remains challenging. Users frequently open multiple browser tabs for side-by-side comparison [19], while existing comparison features typically present specification tables requiring manual interpretation of technical details and trade-offs. Personalized natural-language comparisons provided by conversational assistants offer the potential to reduce mental demand and support informed decision-making.

These benefits are undermined when recommenders remain black boxes, reducing scrutability and potentially affecting trust [21]. In conversational commerce, disclosing what was understood has been shown to increase perceived competence and engagement [15], while natural-language explanations linking user utterances to

product attributes can improve perceived transparency [18]. However, explanation quality matters: Łajewska et al. [9] demonstrate that while high-quality explanations of sources, model confidence, and limitations tend to raise perceived response usefulness and explanation ratings, inaccurate explanations can significantly reduce perceived usefulness. For e-commerce systems, Tsagkias et al. [22] point out that explanations can increase transparency by enabling users to understand what data from their input is being processed and how the search or recommendation mechanism works.

Large language models (LLMs) introduce additional complexity. Supervised fine-tuning leaves models prone to undesired behavior, while reinforcement learning from human feedback (RLHF) substantially improves but does not guarantee instruction-following [14, 25]. LLM-based agents can exhibit overconfidence [20, 26] or employ persuasive tactics [12, 17]. Lin et al. [11] advocate hybrid approaches, noting that LLM-only recommenders without domain-specific fine-tuning can lag in performance and introduce latency. While retrieval-augmented generation (RAG) provides grounding [7, 27], challenges such as hallucinations, biases, and incorrect instruction-following persist [3, 10, 29].

We present a chatbot, acting as a conversational product advisor, addressing these challenges through four contributions:

**(1) Controllability:** We demonstrate a hybrid architecture separating deterministic ranking from language generation, mitigating hallucinated or persuasive content.

**(2) Transparency:** We operationalize system revealment through transparent and verifiable ranking.

**(3) Decision support:** We provide constrained natural-language comparisons and a highlights feature designed to reduce cognitive load by contextualizing specifications in relation to user needs.

**(4) Experimental system:** We contribute an open-source extensible system designed to support research on transparency, control, and decision support in conversational search and recommendation[1].

This approach provides a practical balance between fully automated LLM-based recommenders and traditional faceted search. It reduces comparison complexity through transparency and improves interpretability with verifiable rankings. It furthermore limits LLM unpredictability through constrained text generation.

## 2 User Interaction

We implemented and instantiated Cleo as a conversational agent (chatbot) designed to support users in finding a suitable laptop. While the current implementation focuses on laptops, the system can easily be adapted to other product domains characterized by a set of structured attributes. We selected laptops as our use case because they represent a complex yet familiar product category, involving multiple comparable features such as performance, portability, and price.

Cleo supports both text-based and voice-based input, allowing users to express their information need through their preferred modality. The system includes text-to-speech functionality to vocalize the chatbot's responses, and allows users to edit their most recent message. Through natural conversational interaction, users

[1]The current, work-in-progress version of Cleo can be tested via https://multiweb.gesis.org/vacos5/.

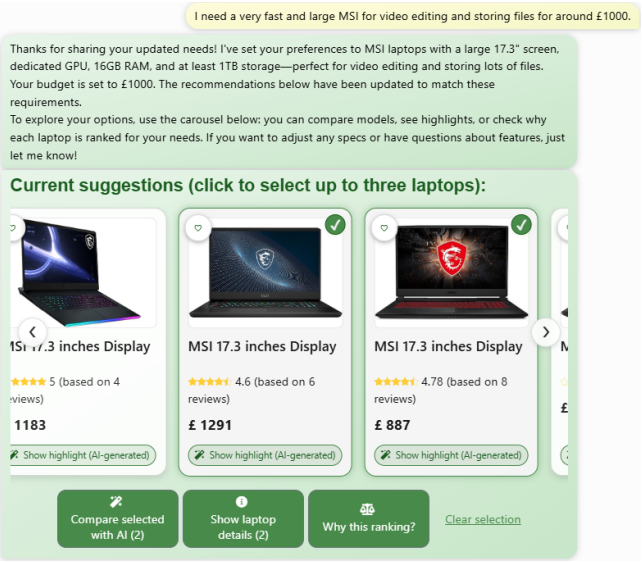



**Figure 1: Product suggestions: If a user expresses new or refined needs, Cleo renders a suggestions carousel presenting the nine top-ranked items.**

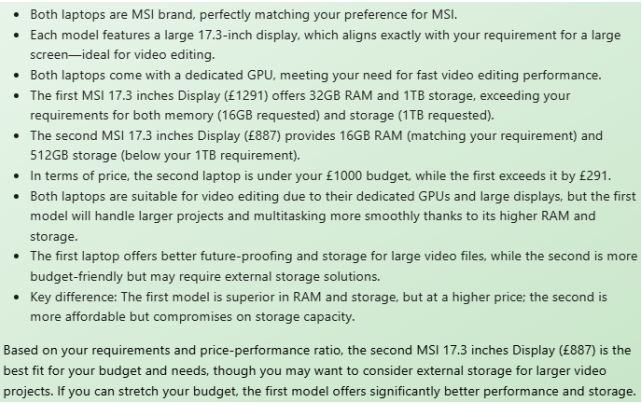



**Figure 2: Multi-item analysis: selecting 2–3 cards enables the "Compare selected with AI" feature, which returns a bullet-pointed comparison plus a brief personalized recommendation.**

engage in a mixed-initiative dialogue, iteratively refining recommendations by stating requirements such as "the laptop should support video editing," "smaller screen," or "must have an integrated GPU."

With each conversational turn that expresses new or refined requirements, the system re-ranks the product catalog and presents an updated product carousel within the chat history—always allowing users to revisit previous results.

### 2.1 Product Suggestions & Highlights

The nine top-ranked laptops are presented in a carousel view (see Figure 1), displaying a product image, star rating, and price. Each

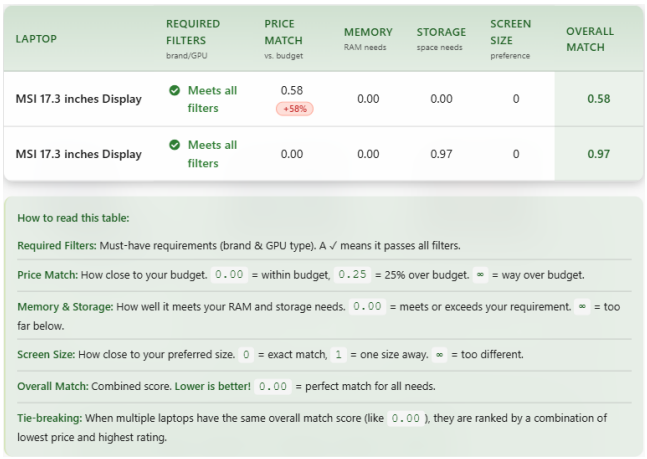



**Figure 3: Explanations: The "Why this ranking?" modal displays required filters (brand/GPU), per-attribute loss values (price/RAM/storage/screen size), and an overall score (lower is better). It also shows a legend explaining the loss values.**

product card supports three actions: (i) selecting the item to view product specifications or enable AI-generated comparison, (ii) adding the item to a wishlist (heart icon) for later review, and (iii) triggering an AI-generated product highlight that explains how and why the product matches the user's information need, for example: "This Lenovo laptop matches your preferred 14-inch portable size, offers more RAM (20 GB) and storage (1 TB) than you requested, and includes a dedicated GPU for enhanced performance."

## 2.2 Multi-item Comparison & Details

Users can select up to three laptops for comparison. Once at least two items are selected, the "Compare selected with AI" button generates a bullet-pointed comparison that takes the user's initial information need into account, followed by a concise personalized recommendation (see Figure 2). A complementary "Show laptop details" button opens a modal window containing a side-by-side product specifications table. The wishlist view replicates these features for saved items.

## 2.3 Explanations: "Why this ranking?"

The "Why this ranking?" button opens a modal window explaining the system's ranking rationale (see Figure 3). The modal displays per-item required filters (brand and GPU type, indicated by "✓" or "✗" symbols with human-readable reasons such as "Meets all filters") and numeric loss values for price, RAM, storage, and screen size, culminating in an overall score where lower values indicate better matches. The modal includes a legend explaining how to interpret each column.

# 3 System Overview

Cleo is implemented with a hybrid architecture that separates decision policy from language generation, addressing opacity and unpredictability in conversational search and recommendation. The system combines a deterministic, auditable ranking core with constrained LLM generation, enabling both conversational fluidity and algorithmic transparency.

## 3.1 Architecture

Figure 4 illustrates the backend architecture enforcing separation of tasks through two layers. The backend provides a REST API supporting message handling, product highlights and comparisons, and ranking explanations. The *management layer* orchestrates dialogue flow through **Conversation Manager**, which invokes **AI Manager** and **Ranker** in sequence, and maintains the conversation state. The *functionality layer* comprises two independent modules: **AI Manager** and **Ranker**. The **AI Manager** module operates in two modes: *structured extraction* (anticipated user requirements in JSON format) and *natural language generation* (responses, highlights, comparisons). Meanwhile, the **Ranker** implements a deterministic ranking, ensuring AI-generated conversational responses remain decoupled from predictable ranking decisions.

## 3.2 Data Flow

A typical interaction follows a four-stage process coordinated by **Conversation Manager**: (1) requirement extraction with (**AI Manager**), (2) deterministic re-ranking based on updated requirements (**Ranker**), (3) response generation using updated requirements and newly ranked laptops (**AI Manager**), (4) return of response and top-ranked items to the frontend.

Product comparisons and highlights use a simplified flow: hard-coded rules compute specification comparisons against requirements and across items, then the LLM formats them into bullet points through **AI Manager**. Ranking explanations use identical loss computations as the **Ranker** to ensure model-intrinsic transparency. User-chatbot conversations can optionally be logged.

## 3.3 User requirements extraction

A prompt instructs an LLM how the user's input can be mapped to structured JSON output with few-shot examples, e.g., "casual gaming" is mapped to `{"gpu": "dedicated", "ram": 32}`. It is supplemented by explicit scope instructions limiting generation to the laptop domain, and update policies that preserve previous values unless explicitly changed by the user. Following extraction, a sanitizer function corrects potential mistakes in the LLM's output through (1) unit conversion (e.g., converting terabytes to gigabytes or averaging price ranges), (2) aligning to catalog constraints (e.g., storage values to discrete tiers [256, 512, 1000, 2000 GB]), and (3) filtering invalid updates (e.g., if the LLM extracted a brand that is not in the catalog). This ensures extracted requirements map directly to queryable product attributes.

## 3.4 Deterministic Ranking

The **Ranker** module operationalizes system transparency through deterministic and inspectable ranking. An exemplary catalog of 3,638 laptop specifications is indexed for efficient computation. The ranking algorithm proceeds in three steps: Categorical filtering first removes devices that violate user requirements for brand and GPU type from the result list. Numeric loss functions then compute deviations for price, RAM, storage, and screen size. Finally, laptops

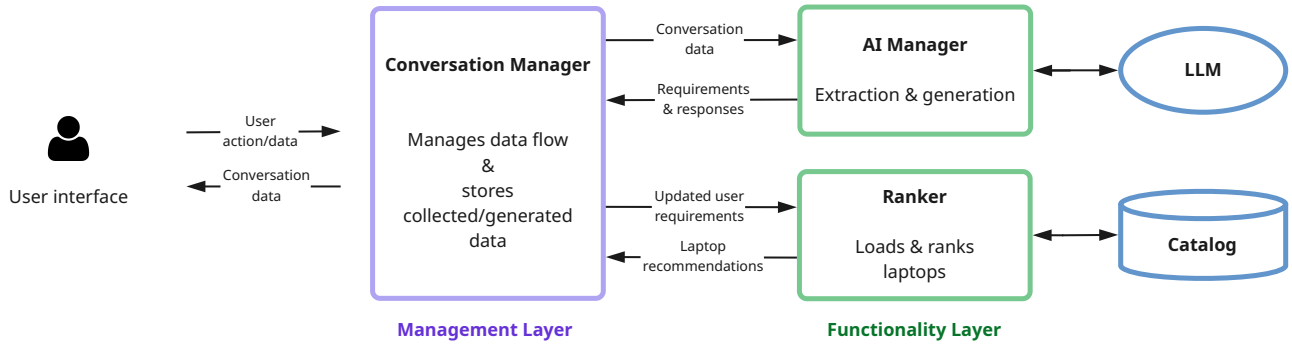



**Figure 4: System architecture implementing our hybrid approach: the *Functionality Layer* separates deterministic ranking (Ranker) from constrained language generation (AI Manager), while the *Management Layer* (Conversation Manager) orchestrates their interaction.**

are ranked by ascending total loss, considering both higher rating and lower price in case of loss value ties. Loss components are exposed through the ranking explanation endpoint, making every ranking decision auditable.

### 3.5 Response Generation

Another prompt instructs the system to articulate how vague user needs are mapped to concrete attribute values (e.g., "For video editing, I've updated to a dedicated GPU"). Following RAG principles, responses are grounded in current user requirements and laptop recommendations to mitigate hallucinations. Few-shot examples demonstrate appropriate conversational style and handling of different user message types, e.g., for feedback, questions, or clarifications.

### 3.6 Highlights and Comparisons

For single-item product attributes, deterministic rules compare specifications to requirements (e.g., "32 GB RAM exceeds 16 GB requirement"), while the LLM converts these into natural language descriptions. For multi-item comparisons, hard-coded rules first compare specifications across selected laptops and against requirements, then the LLM formats these as structured bullet points and adds a brief personalized recommendation based on price/performance trade-offs. This division aims to mitigate hallucinated specifications and over-persuasive content while allowing natural language comparison and contextual advice.

## 4 Conclusion

We demonstrated a transparent and controllable conversational product advisor addressing comparison complexity, opacity, and LLM unpredictability through a hybrid architecture. Our system provides transparency through auditable rankings and requirement reflection, controllability by separating deterministic ranking from language generation, and decision support through grounded natural-language comparisons and highlights. We contribute an extensible experimental system targeting IR/HCI researchers and practitioners seeking inspiration regarding the design of conversational search and recommendation systems.

Our approach differs from traditional faceted search (which lacks conversational interaction), end-to-end LLM recommenders (which typically hide decision logic), and existing comparison tools (which tend to present uncontextualized specifications). During the live demonstration, CHIIR attendees will be able to test their own search scenarios and provide feedback.

This work provides an adaptable pattern for building transparent and controllable conversational search systems in different domains. Current limitations include rigid loss weights and potential mental workload from the ranking explanations. We propose three potential user studies with Cleo: (1) comparing hybrid, LLM-only, and traditional faceted interfaces; (2) evaluating the explanation and decision support features to measure impact on decision confidence, trust, and time-to-decision; (3) exploring user preferences for guidance versus autonomy within a conversational search/recommendation system. Future work includes personalizing loss weights, expanding the product attributes taken into account, integrating customer reviews into item comparisons, and adding counterfactual steering controls.

We invite the CHIIR community to adapt this experimental system to advance transparency, control, and decision support in conversational search and recommendation.[2]

## Acknowledgments

This work was supported by the German Research Foundation (DFG) as part of the "VACOS 2" project (no. 388815326).

## References

[1] Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W. Bruce Croft. 2019. Asking Clarifying Questions in Open-Domain Information-Seeking Conversations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Paris, France) *(SIGIR'19)*. Association for Computing Machinery, New York, NY, USA, 475–484. doi:10.1145/3331184.3331265

[2] Janarthanan Balakrishnan and Yogesh K. Dwivedi. 2021. Conversational commerce: entering the next stage of AI-powered digital assistants. *Annals of Operations Research* (12 Apr 2021). doi:10.1007/s10479-021-04049-5

[3] Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference*

[2]The source code will be made publicly available upon publication at https://osf.io/scxpv/overview?view_only=3dbccde002a846a5a13ab2001cd507b9.

*on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence (AAAI'24/IAAI'24/EAAI'24)*, Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan (Eds.). AAAI Press, Article 1980, 9 pages. doi:10.1609/aaai.v38i16.29728

[4] Li Chen and Pearl Pu. 2012. Critiquing-based recommenders: survey and emerging trends. *User Model. User-adapt Interact.* 22, 1-2 (April 2012), 125–150.

[5] Chongming Gao, Wenqiang Lei, Xiangnan He, Maarten de Rijke, and Tat-Seng Chua. 2021. Advances and challenges in conversational recommender systems: A survey. *AI Open* 2 (2021), 100–126. doi:10.1016/j.aiopen.2021.06.002

[6] David Gefen, Elena Karahanna, and Detmar W. Straub. 2003. Trust and TAM in Online Shopping: An Integrated Model. *MIS Quarterly* 27, 1 (2003), 51–90. http://www.jstor.org/stable/30036519

[7] Trey Grainger, Doug Turnbull, and Max Irwin. 2025. *AI-Powered Search*. Simon and Schuster.

[8] Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. 2021. A Survey on Conversational Recommender Systems. *ACM Comput. Surv.* 54, 5, Article 105 (May 2021), 36 pages. doi:10.1145/3453154

[9] Weronika Łajewska, Damiano Spina, Johanne Trippas, and Krisztian Balog. 2024. Explainability for Transparent Conversational Information-Seeking. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 1040–1050. doi:10.1145/3626772.3657768

[10] Jingling Li, Zeyu Tang, Xiaoyu Liu, Peter Spirtes, Kun Zhang, Liu Leqi, and Yang Liu. 2024. Steering LLMs Towards Unbiased Responses: A Causality-Guided Debiasing Framework. *arXiv preprint arXiv:2403.08743* (2024). doi:10.48550/arXiv.2403.08743

[11] Jianghao Lin, Xinyi Dai, Yunjia Xi, Weiwen Liu, Bo Chen, Hao Zhang, Yong Liu, Chuhan Wu, Xiangyang Li, Chenxu Zhu, Huifeng Guo, Yong Yu, Ruiming Tang, and Weinan Zhang. 2025. How Can Recommender Systems Benefit from Large Language Models: A Survey. *ACM Trans. Inf. Syst.* 43, 2, Article 28 (Jan. 2025), 47 pages. doi:10.1145/3678004

[12] Minqian Liu, Zhiyang Xu, Xinyi Zhang, Heajun An, Sarvech Qadir, Qi Zhang, Pamela J. Wisniewski, Jin-Hee Cho, Sang Won Lee, Ruoxi Jia, and Lifu Huang. 2025. LLM Can be a Dangerous Persuader: Empirical Study of Persuasion Safety in Large Language Models. arXiv:2504.10430 [cs.CL] https://arxiv.org/abs/2504.10430

[13] Chris Messina. 2024. Conversational commerce. When the point of sale comes to the… | by Chris Messina | Chris Messina | Medium. https://medium.com/chris-messina/conversational-commerce-92e0bccfc3ff. Accessed: 2024-02-15.

[14] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(NIPS '22)*. Curran Associates Inc., Red Hook, NY, USA, Article 2011, 15 pages.

[15] Andrea Papenmeier and Elin Anna Topp. 2023. Ah, Alright, Okay! Communicating Understanding in Conversational Product Search. In *Proceedings of the 5th International Conference on Conversational User Interfaces* (Eindhoven, Netherlands) *(CUI '23)*. Association for Computing Machinery, New York, NY, USA, Article 49, 5 pages. doi:10.1145/3571884.3604318

[16] Filip Radlinski and Nick Craswell. 2017. A Theoretical Framework for Conversational Search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval* (Oslo, Norway) *(CHIIR '17)*. Association for Computing Machinery, New York, NY, USA, 117–126. doi:10.1145/3020165.3020183

[17] Alexander Rogiers, Sander Noels, Maarten Buyl, and Tijl De Bie. 2024. Persuasion with Large Language Models: a Survey. arXiv:2411.06837 [cs.CL] https://arxiv.org/abs/2411.06837

[18] Kevin Schott, Andrea Papenmeier, Daniel Hienert, and Dagmar Kern. 2024. What Did I Say Again? Relating User Needs to Search Outcomes in Conversational Commerce. In *Proceedings of Mensch Und Computer 2024* (Karlsruhe, Germany) *(MuC '24)*. Association for Computing Machinery, New York, NY, USA, 129–139. doi:10.1145/3670653.3670680

[19] Kevin Schott, Andrea Papenmeier, Daniel Hienert, and Dagmar Kern. 2025. Click-Click-Add – Product Search Strategies in Online Shopping. *Proceedings of the Association for Information Science and Technology* 62, 1 (2025), 621–633. arXiv:https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/pra2.1283 doi:10.1002/pra2.1283

[20] Fengfei Sun, Ningke Li, Kailong Wang, and Lorenz Goette. 2025. Large Language Models are overconfident and amplify human bias. arXiv:2505.02151 [cs.SE] https://arxiv.org/abs/2505.02151

[21] Nava Tintarev and Judith Masthoff. 2012. Evaluating the effectiveness of explanations for recommender systems. *User Model. User-adapt Interact.* 22, 4-5 (Oct. 2012), 399–439. doi:10.1007/s11257-011-9117-5

[22] Manos Tsagkias, Tracy Holloway King, Surya Kallumadi, Vanessa Murdock, and Maarten de Rijke. 2021. Challenges and research opportunities in eCommerce search and recommendations. *SIGIR Forum* 54, 1, Article 2 (feb 2021), 23 pages. doi:10.1145/3451964.3451966

[23] Sven Tuzovic and Stefanie Paluch. 2018. *Conversational Commerce – A New Era for Service Business Development?* Springer Fachmedien Wiesbaden, Wiesbaden, 81–100. doi:10.1007/978-3-658-22426-4_4

[24] Preeti Virdi, Arti D. Kalro, and Dinesh Sharma. 2020. Online decision aids: the role of decision-making styles and decision-making stages. *International Journal of Retail & Distribution Management* 48, 6 (01 Jan 2020), 555–574. doi:10.1108/IJRDM-02-2019-0068

[25] Zhichao Wang, Bin Bi, Shiva Kumar Pentyala, Kiran Ramnath, Sougata Chaudhuri, Shubham Mehrotra, Zixu, Zhu, Xiang-Bo Mao, Sitaram Asur, Na, and Cheng. 2024. A Comprehensive Survey of LLM Alignment Techniques: RLHF, RLAIF, PPO, DPO and More. arXiv:2407.16216 [cs.CL] https://arxiv.org/abs/2407.16216

[26] Bingbing Wen, Chenjun Xu, Bin HAN, Robert Wolfe, Lucy Lu Wang, and Bill Howe. 2024. From Human to Model Overconfidence: Evaluating Confidence Dynamics in Large Language Models. In *NeurIPS 2024 Workshop on Behavioral Machine Learning*. https://openreview.net/forum?id=y9UdO5cmHs

[27] Tessa Withorn. 2025. Google AI Overviews Are Here to Stay: A Call to Teach AI Literacy. *College & Research Libraries News* 86, 5 (2025), 214. doi:10.5860/crln.86.5.214

[28] Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W. Bruce Croft. 2018. Towards Conversational Search and Recommendation: System Ask, User Respond. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management* (Torino, Italy) *(CIKM '18)*. Association for Computing Machinery, New York, NY, USA, 177–186. doi:10.1145/3269206.3271776

[29] Yang Zhang, Hanlei Jin, Dan Meng, Jun Wang, and Jinghua Tan. 2025. A Comprehensive Survey on Process-Oriented Automatic Text Summarization with Exploration of LLM-Based Methods. arXiv:2403.02901 [cs.AI] https://arxiv.org/abs/2403.02901