LLM Consumer Behavior Theory: Foundations of a Novel Research Field

Anonymous ACL submission

Abstract

Large language models (LLMs) are increasingly deployed as autonomous agents that make consumption decisions on behalf of users. This shift raises fundamental questions for consumer theory, which has traditionally modeled humans as the primary decision-makers. In this paper, we introduce **LLM Consumer Behavior Theory**, a new field of study concerned with analyzing consumer behavior in agentic markets. Drawing on classical and behavioral economics alongside recent advances in Natural Language Processing, we formalize how human preferences are reflected and acted upon by LLM-based agents, and how agent-level decisions aggregate into market demand. We unify previously fragmented literature on LLM decision-making, human behavior simulation, and preference elicitation under a common economic lens, highlighting where assumptions, such as rationality and heterogeneity, may fail in agentic markets. Rather than providing empirical validation, this paper outlines the scope of LLM consumer behavior and identifies open research questions related to alignment, preference representation, and market dynamics.

1 Introduction

AI systems have undergone a shift from passive conversational tools to *agentic* AI systems that increasingly make decisions on our behalf (Gaarlandt et al., 2025; Purdy, 2024; Goli and Singh, 2024). These emerging AI agents can proactively plan, evaluate alternatives, and execute complex tasks across domains (Acharya et al., 2025; Whiting, 2024). From an economic point of view, these systems are a new type of actors increasingly making purchasing decisions on our behalf (Cherep et al., 2025a). This is also highlighted by the anticipated growth of agentic commerce: McKinsey Research, for example, estimates that by 2030 agentic commerce could orchestrate $ 3 to $ 5 trillion globally (Schumacher et al., 2025). Accordingly, classical economic frameworks that solely model individuals as decision-makers must evolve to account for agentic systems acting on behalf of users.

This paper introduces **LLM Consumer Behavior Theory**, a new research field concerned with consumption decisions made by LLM-based agents acting on behalf of human users. By integrating insights from classical consumer behavior theory, behavioral economics, and Natural Language Processing (NLP), we unify previously fragmented literature under a common economic framework, establishing a coherent foundation for studying agentic consumer behavior. Rather than exhaustively characterizing all implications of this emerging domain, we outline the scope and core conceptual structure of LLM consumer behavior, clarify how existing empirical findings fit within it, and identify central open questions that arise when consumer decision-making is delegated to algorithmic agents.

The paper is organized around three core components of LLM Consumer Behavior Theory, summarized in Figure 1. First, we ground **(LLM) consumer preferences** in classical utility theory and behavioral economics, analyzing how choices arise from preferences, constraints, and cognitive biases. Second, we study **the user–agent alignment**, formalizing how consumer preferences are reflected in LLM-based agents through an agent–principal perspective. Third, we turn to **consumer markets**, aggregating agent-level decisions to characterize agentic demand and exploring how heterogeneity and market dynamics might be affected in comparison to human consumer markets, and how alignment and diversification can mitigate these effects.

2 Consumer preferences and utility

This section reviews the principal theories used to model *individual* consumers, and discusses related work on how these frameworks have been applied in autonomous decision-making settings.

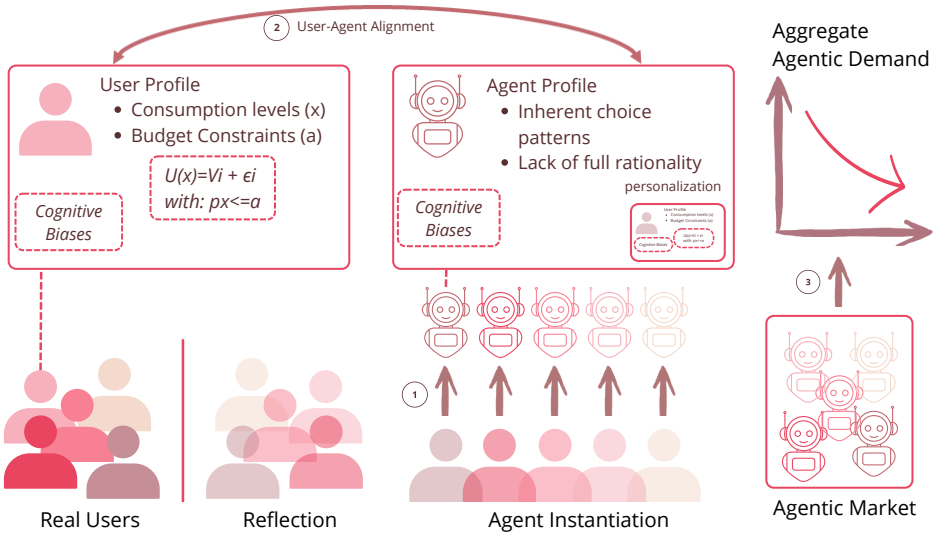

Figure 1: Field of LLM Consumer Behavior Theory. Consumers are modeled via preference structures. These preferences are imperfectly reflected and instantiated in LLM agents. Each agent acts on behalf of a user, producing agent-level choices that aggregate into market demand

2.1 Model of Individual Choice

In classical economic theory, consumers are assumed to choose the option that maximizes utility, represented by $U(x)$, where x is a set of consumption levels, subject to a budget constraint $px \leq a$ where p,where p is a set of prices and a is the income of the consumer (McFadden, 2001). This framework is commonly referred to as the rational choice model (Ooghe et al., 2023). In practice, however, decision-making may also be influenced by random factors, such as noise or bias. Therefore, random utility theory was introduced to analyze choices that maximize utility, while allowing for random elements (McFadden, 2001). Utility is then decomposed into a systematic part V and an error term or unobservable part ϵ (Masiero et al., 2015):

$$U(x) = V + \epsilon \tag{1}$$

Different models can approximate this utility function, such as the multinomial logit model or the mixed logit model (Masiero et al., 2015; McFadden, 2001; Reusens et al., 2026)

Research in NLP has begun to study LLM choice preferences in market-like settings. Cedro et al. (2025) study a series of real-world dilemmas, such as whether to accept a discount in exchange for longer waiting times, and map the resulting preferences of several LLMs. They find that smaller and older models exhibit less consistent preferences, whereas larger and more recent models display sharper and more stable patterns. Related studies consider other forms of LLM decision-making, including preferences over time (Goli and Singh, 2024), risk (Jia et al., 2024), or walking distance (Fulman et al., 2025). From an economic perspective, Reusens et al. (2026) use a discrete choice method to estimate Willingness to Pay for hotel room features.[1].

2.2 Behavioral Economics

Classical economic theory treats the systematic component of human decision-making as rational. Yet even within this component, people are influenced by behavioral biases and do not behave fully rationally. Deviations include loss aversion, anchoring, overconfidence, and framing effects (Tversky and Kahneman, 1979; Kahneman, 2011).

LLM Rationality Jiang (2025) characterizes a rational agent using four main axioms: information grounding, logical consistency, invariance from irrelevant context, and orderability of preferences. Evidence suggests that LLMs do not fully satisfy these criteria, due in part to inconsistent outputs, a bounded knowledge space, and their lack of direct real-world grounding and feedback mechanisms (Jiang et al., 2025). At the same time, some evidence of economic rationality is found, although these are highly sensitive to prompt variations (Chen et al., 2023). A grow-

[1]The maximum willingness to pay signals that the consumer is indifferent between buying or not buying the product (Ooghe et al., 2023)

ing body of work thus investigates how rationality can be improved through, for example multimodal systems, integrated tool use, or additional alignment (Jiang, 2025; Liu et al., 2025a). However, these approaches do not yet succeed in making LLMs fully rational (Jiang, 2025; Liu et al., 2025a).

LLM cognitive biases Beyond violations of rationality, research examines whether LLMs exhibit systematic behavioral deviations similar to those observed in humans (Liu et al., 2025a; Jones and Steinhardt, 2022; Jia et al., 2024). These models apply higher discount rates than humans in intertemporal tasks while simultaneously producing more internally consistent choices in gambling-like settings (Goli and Singh, 2024; Liu et al., 2025b). Jones and Steinhardt (2022) show that prompt designs inspired by human cognitive biases can reliably induce systematic errors. Additionally, both risk and loss aversion are present within LLMs, though the degree is model-dependent (Jia et al., 2024) and LLMs are more susceptible to nudges than humans (Cherep et al., 2025b). In general, LLMs exhibit a range of behaviors between these human biases and more economically rational decisions (Chen et al., 2023; Raman et al., 2024; Ross et al., 2024; Bini et al., 2025; Lou and Sun, 2026).

3 User-Agent Alignment

In agentic markets, LLMs make decisions on behalf of users, requiring alignment with user preferences. Assuming one agent per user, this setting maps onto the classical principal-agent problem, where decision authority is delegated under imperfect information and potentially divergent objectives (Grossman and Hart, 1983). Users delegate decisions to LLMs expecting that these systems act in their best interest. Yet LLMs are shaped by the design choices of their developers, which can lead to misalignment between model behavior and the behavior that a user might seek, as shown by Fan et al. (2024) and Reusens et al. (2025). Additionally, LLMs do not directly observe true user preferences, but rely on proxy representations inferred from prompts, context history, past behavior, and available demographic information (Figure 1).

Persona-based alignment Research studies the use of LLMs to simulate human behavior (Wang et al., 2025b; Yoon et al., 2024; Kim et al., 2026; Horton et al., 2023). A popular approach for improving human alignment is *persona* assignment: instructing an LLM to "*act as*" a user with specific characteristics or preferences (Wang et al., 2025b). Personas can also be based on behavioral rather than demographic attributes (Goethals et al., 2025). While persona assignment can improve alignment in simple tasks (Liu et al., 2025a), it also has important limitations: models may not follow assigned personas consistently (Horton et al., 2023; Reusens et al., 2025), especially when asked to follow uncommon preferences (Fan et al., 2024), and may behave stereotypically (Wang et al., 2025a; Reusens et al., 2025). Lastly, careful persona design is required as it could lead to extreme model behavior (Reusens et al., 2026).

Alignment through few-shot learning and fine-tuning A complementary line of work studies whether LLM-based agents can learn individual preferences from limited interaction data. Several papers use in-context-learning from past choices i.e. (Kim et al., 2026; Singh et al., 2025; Reusens et al., 2026). Other papers finetune a user-specific personalized model i.e. (Thonet et al., 2025). However, full fine-tuning per user can be costly and hard to scale. Therefore, more efficient methods are also proposed, such as LoRe, that models individual preferences as weighted combinations of base reward functions (Bose et al., 2025).

4 Consumer Markets

If everyone uses an agent as their proxy, a fully agentic consumer market would emerge, as shown in Figure 1. In microeconomics, consumer's Willingness to Pay values constitute the demand governed by the *Law of Demand*: as the quantity sold increases, the price must fall to attract additional buyers (Marshall et al., 1961). In agentic markets, aggregating individual agentic willingness to pay values across this population thus yields the agentic market demand curve.

4.1 The Heterogeneous User Demand

Individual human choices are driven by utility, which depends on factors such as preferences and income (Marshall et al., 1961) but also on unobserved dimensions such as tastes and perceptions of the world (McFadden, 2001). As a result, traditional consumer markets are characterized by substantial heterogeneity in preferences reflecting differences in income, taste, and lived experiences. This heterogeneity is a central feature of economic models, as variation across individuals shapes both

market outcomes and welfare implications (McFadden, 1972; Berry et al., 1993).

4.2 The Homogenous Agentic Demand

Heterogeneity may be reduced in agentic markets, since LLMs tend to reproduce average patterns from their training data rather than diverse human preferences (Bender et al., 2021; Matz et al., 2025). This effect may be reinforced by the concentrated LLM ecosystem, in which state-of-the-art models rely on largely overlapping corpora, similar architectures, and related alignment procedures (Bommasani et al., 2021; Intel Market Research, 2024). Although homogenization has been documented in other downstream applications (Weidinger et al., 2021; Shumailov et al., 2024; Goethals and Rhue, 2025), its implications for consumer markets and aggregate demand remain underexplored. Early evidence suggests that agentic demand may be more homogeneous than human demand: LLMs show more uniform out-of-the-box choice patterns (Chen et al., 2023), AI agents can concentrate demand on a few "modal" products (Allouah et al., 2025), and LLM behavior appears less heterogeneous than human behavior (del Rio-Chanona et al., 2025). These preferences may also be unstable, with model updates substantially shifting market shares (Allouah et al., 2025).

4.3 Diversifying the Agentic Demand

Agentic heterogeneity can still arise from both model-side and user-side factors. On the model side, architectural differences across foundation models can lead to different decisions or price boundaries (Reusens et al., 2026; Cedro et al., 2025), and higher temperature settings can also induce response diversity (Chen et al., 2023). On the user side, aligning agents to specific user profiles can further diversify behavior, as discussed in Section 3. For example, LLMs can behave like heterogeneous households when assigned different profiles, occupations, and income levels, generating consumption patterns consistent with macroeconomic regularities (Li et al., 2024).

5 Discussion

A central implication of LLM Consumer Behavior Theory is that consumer analysis can no longer focus solely on human users. When decisions are delegated to LLM-based agents, market outcomes reflect a broader pipeline involving preference reflection, initial agentic preferences, and user–agent alignment (Figure 1). Understanding behavior along this continuum is therefore critical for analyzing emerging consumer markets, motivating several directions for future research.

LLM choice patterns A first key direction concerns LLM choice patterns. What factors influence the choices these models make? How robust and consistent are these choices across models, settings and prompts? To what extent can they be influenced or steered, either through prompting strategies or through modifications to the model itself?

Personalized agents A second direction concerns aligning agents with unique user preferences. Which data sources and alignment mechanisms best capture true user preferences? How do reflection errors or biases propagate from agent decisions to market-level distortions? Can shared agents effectively serve heterogeneous users without disadvantaging minority groups, or does meaningful alignment require individualized models?

Hybrid Markets Rather than an abrupt shift, consumer markets are likely to evolve gradually from human-driven to increasingly agentic systems, creating hybrid market structures. Open questions include the pace of this transition, the dynamics of hybrid markets, and whether shared model weights could induce preference homogenization without collusion. What risks arise when large populations of agents update simultaneously, and how will human consumers adapt to the growing concentration of agentic decision-making? How are prices and equilibria affected in hybrid markets?

6 Conclusion

This paper introduces **LLM Consumer Behavior Theory**, which studies consumption decisions made by LLM-based agents acting on behalf of human users. As consumer decision-making becomes increasingly agentic, market behavior can no longer be understood solely through human preferences. By grounding this in classical and behavioral consumer theory and linking it to recent advances in NLP, we clarify how empirical findings on LLM behavior map to established economic concepts and why agentic consumption raises new challenges of alignment and heterogeneity. We delineate the scope of this emerging field and identify foundational questions for future research.

Limitations

As noted, the primary aim of this paper is to introduce LLM Consumer Behavior Theory as a novel field of study, rather than to provide empirical validation, which we leave to future work. Our analysis is further limited to the demand side of agentic markets; extending these ideas to agentic supply, and to settings in which both supply and demand are partially agentic, raises additional questions for future research. Finally, we do not attempt an exhaustive review of all related literature. Instead, we synthesize insights across economics and NLP to offer a structured perspective on agentic consumer behavior.

Ethical Considerations

Analyzing agentic consumer markets also raises several ethical considerations. A central issue concerns *accountability* when LLMs make decisions on behalf of users. Who is to blame for a poor decision? Currently, there is no global consensus for this question (Schumacher et al., 2025). Individuals should provide informed consent and understand to what extent they remain responsible for agentic choices. This concern extends beyond fully autonomous settings to cases where users explicitly approve recommendations, as approving a suggestion generated by an LLM involves more cognitive engagement and deliberation than making the decision independently. Additional ethical risks arise when agents are misaligned with user preferences and exhibit stereotypical or biased behavior. Finally, the complexity of LLMs limits transparency into their underlying decision-making processes. Addressing these challenges represents an important direction for future research.

References

Deepak Bhaskar Acharya, Karthigeyan Kuppan, and B Divya. 2025. Agentic ai: autonomous intelligence for complex goals - a comprehensive survey. *IEEE Access*.

Amine Allouah, Omar Besbes, Josué Figueroa, Yash Kanoria, and Akshit Kumar. 2025. What is your ai agent buying? evaluation, implications, and emerging questions for agentic e-commerce. *Evaluation, Implications, and Emerging Questions for Agentic E-Commerce (August 04, 2025)*.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Steven T Berry, James A Levinsohn, and Ariel Pakes. 1993. Automobile prices in market equilibrium: Part i and ii.

Pietro Bini, Lin William Cong, Xing Huang, and Lawrence J Jin. 2025. Behavioral economics of ai: Llm biases and corrections. *Available at SSRN 5213130*.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, and 1 others. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Avinandan Bose, Zhihan Xiong, Yuejie Chi, Simon Shaolei Du, Lin Xiao, and Maryam Fazel. 2025. Lore: Personalizing LLMs via low-rank reward modeling. In *Second Conference on Language Modeling*.

Mateusz Cedro, Timour Ichmoukhamedov, Sofie Goethals, Yifan He, James Hinns, and David Martens. 2025. Cash or comfort? how llms value your inconvenience. *arXiv preprint arXiv:2506.17367*.

Yiting Chen, Tracy Xiao Liu, You Shan, and 1 others. 2023. The emergence of economic rationality of GPT. *Proceedings of the National Academy of Sciences*, 120(51).

Manuel Cherep, Chengtian Ma, Abigail Xu, Maya Shaked, Patricia Maes, and Nikhil Singh. 2025a. A framework for studying ai agent behavior: Evidence from consumer choice experiments. In *NeurIPS 2025 Workshop on Bridging Language, Agent, and World Models for Reasoning and Planning*.

Manuel Cherep, Pattie Maes, and Nikhil Singh. 2025b. Llm agents are hypersensitive to nudges. *arXiv preprint arXiv:2505.11584*.

R Maria del Rio-Chanona, Marco Pangallo, and Cars Hommes. 2025. Can generative ai agents behave like humans? evidence from laboratory market experiments. *arXiv preprint arXiv:2505.07457*.

Caoyun Fan, Jindou Chen, Yaohui Jin, and Hao He. 2024. Can large language models serve as rational players in game theory? a systematic analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17960–17967.

Nir Fulman, Abdulkadir Memduhoğlu, and Alexander Zipf. 2025. Utilizing large language models to simulate parking search. *Transportation Research Part A: Policy and Practice*, 199:104542.

Jur Gaarlandt, Wesley Korver, Nathan Furr, and 1 others. 2025. Ai agents are changing how people shop. here's what that means for brands. *Harvard Business Review*.

Sofie Goethals, Johannes Luther, and Sandra Matz. 2025. Words reveal wants: How well can simple llm-based ai agents replicate people's choices based on their social media posts. In *Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization*, pages 126–131.

Sofie Goethals and Lauren Rhue. 2025. One world, one opinion? the superstar effect in llm responses. In *Proceedings of the 3rd Workshop on Cross-Cultural Considerations in NLP (C3NLP 2025)*, pages 89–107.

Ali Goli and Amandeep Singh. 2024. Frontiers: can large language models capture human preferences? *Marketing Science*, 43(4):709–722.

Sanford J. Grossman and Oliver D. Hart. 1983. An analysis of the principal-agent problem. *Econometrica*, 51(1):7–45.

John J Horton, Apostolos Filippas, and Benjamin S Manning. 2023. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research.

Intel Market Research. 2024. Large language model (LLM) market report 2024–2030, by model architecture, geo, tech. Accessed: April 17, 2026.

Jingru Jia, Zehua Yuan, Junhao Pan, Paul E McNamara, and Deming Chen. 2024. Decision-making behavior evaluation framework for llms under uncertain context. *Advances in neural information processing systems*, 37:113360–113382.

Bowen Jiang, Yangxinyu Xie, Xiaomeng Wang, Yuan Yuan, Zhuoqun Hao, Xinyi Bai, Weijie J Su, Camillo Jose Taylor, and Tanwi Mallick. 2025. Towards rationality in language and multimodal agents: A survey. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3656–3675, Albuquerque, New Mexico. Association for Computational Linguistics.

Jingjing Jiang. 2025. Towards human-like dialogue systems: Integrating multimodal emotion recognition and non-verbal cue generation. In *Proceedings of the 21st Workshop of Young Researchers' Roundtable on Spoken Dialogue Systems*, pages 15–17, Avignon, France. Association for Computational Linguistics.

Erik Jones and Jacob Steinhardt. 2022. Capturing failures of large language models via human cognitive biases. *Advances in Neural Information Processing Systems*, 35:11785–11799.

Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Jeongbin Kim, Matthew Kovach, Kyu-Min Lee, Euncheol Shin, and Hector Tzavellas. 2026. Can an llm learn preferences from choice data? *Preprint*, arXiv:2401.07345.

Nian Li, Chen Gao, Mingyu Li, Yong Li, and Qingmin Liao. 2024. EconAgent: Large language model-empowered agents for simulating macroeconomic activities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15523–15536, Bangkok, Thailand. Association for Computational Linguistics.

Jiaxin Liu, Yixuan Tang, Yi Yang, and Kar Yan Tam. 2025a. Evaluating and aligning human economic risk preferences in LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 18174–18188, Suzhou, China. Association for Computational Linguistics.

Ryan Liu, Jiayi Geng, Joshua Peterson, and 1 others. 2025b. Large language models assume people are more rational than we really are. In *The Thirteenth International Conference on Learning Representations*.

Jiaxu Lou and Yifan Sun. 2026. Anchoring bias in large language models: An experimental study. *Journal of Computational Social Science*, 9(1):11.

Alfred Marshall, Claude William Guillebaud, and 1 others. 1961. *Principles of economics*, volume 1. Springer.

Lorenzo Masiero, Cindy Yoonjoung Heo, and Bing Pan. 2015. Determining guests' willingness to pay for hotel room attributes with a discrete choice model. *International Journal of Hospitality Management*, 49:117–124.

Sandra C Matz, C Blaine Horton, and Sofie Goethals. 2025. The basic b*** effect: The use of llm-based agents reduces the distinctiveness and diversity of people's choices. *arXiv preprint arXiv:2509.02910*.

Daniel McFadden. 1972. Conditional logit analysis of qualitative choice behavior.

Daniel McFadden. 2001. Economic choices. *American economic review*, 91(3):351–378.

Erwin Ooghe, Tom Verbeke, Karolien De Bruyne, Simon De Jaeger, Hans Degryse, Johan Eyckmans, Marjan Maes, Nicky Rogge, Patrick Van Cayseele, Stijn Vanormelingen, Dieter Verhaest, and Andreé' Watteyne. 2023. *Economie: Een inleiding*, negende uitgave edition. Acco, Leuven.

Mark Purdy. 2024. What is agentic ai, and how will it change work? *Harvard Business Review*.

Narun Raman, Taylor Lundy, Samuel Amouyal, and 1 others. 2024. Steer: assessing the economic rationality of large language models. *Preprint at https://arxiv.org/pdf/2402.09552*.

Manon Reusens, Bart Baesens, and David Jurgens. 2025. Are economists always more introverted? analyzing consistency in persona-assigned LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 11268–11287, Suzhou, China. Association for Computational Linguistics.

Manon Reusens, Sofie Goethals, Toon Calders, and David Martens. 2026. Would a large language model pay extra for a view? inferring willingness to pay from subjective choices. *Preprint*, arXiv:2602.09802.

Jillian Ross, Yoon Kim, and Andrew Lo. 2024. Llm economicus? mapping the behavioral biases of llms via utility theory. In *First Conference on Language Modeling*.

Katharina Schumacher, Roger Roberts, and Katharina Giebel. 2025. The agentic commerce opportunity: How ai agents are ushering in a new era for consumers and merchants. Report.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759.

Anikait Singh, Sheryl Hsu, Kyle Hsu, Eric Mitchell, Stefano Ermon, Tatsunori Hashimoto, Archit Sharma, and Chelsea Finn. 2025. Fspo: Few-shot preference optimization of synthetic preference data in llms elicits effective personalization to real users. *arXiv preprint arXiv:2502.19312*.

Thibaut Thonet, Germán Kruszewski, Jos Rozen, Pierre Erbacher, and Marc Dymetman. 2025. FaST: Feature-aware sampling and tuning for personalized preference alignment with limited data. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9341–9370, Suzhou, China. Association for Computational Linguistics.

Amos Tversky and Daniel Kahneman. 1979. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–292.

Angelina Wang, Jamie Morgenstern, and John P Dickerson. 2025a. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, 7(3):400–411.

Kuang Wang, Xianfei Li, Shenghao Yang, Li Zhou, Feng Jiang, and Haizhou Li. 2025b. Know you first and be you better: Modeling human-like user simulators via implicit profiles. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21082–21107, Vienna, Austria. Association for Computational Linguistics.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, and 1 others. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Kate Whiting. 2024. The rise of 'ai agents': what they are and how to manage the risks. *World Economic Forum*.

Se-eun Yoon, Zhankui He, Jessica Echterhoff, and Julian McAuley. 2024. Evaluating large language models as generative user simulators for conversational recommendation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1490–1504, Mexico City, Mexico. Association for Computational Linguistics.