Recommender Systems Meet Large Language Model Agents: A Survey

Other titles in Foundations and Trends® in Privacy and Security

Trustworthy Machine Learning: From Data to Models
Bo Han, Jiangchao Yao, Tongliang Liu, Bo Li, Sanmi Koyejo and Feng Liu
ISBN: 978-1-63828-548-9

Advances in Secure IoT Data Sharing
Phu Nguyen, Arda Goknil, Gencer Erdogan, Shukun Tokas, Nicolas Ferry and Thanh Thao Thi Tran
ISBN: 978-1-63828-422-2

Navigating the Soundscape of Deception: A Comprehensive Survey on Audio Deepfake Generation, Detection, and Future Horizons
Taiba Majid Wani, Syed Asif Ahmad Qadri, Farooq Ahmad Wani and Irene Amerini
ISBN: 978-1-63828-492-5

Reverse Engineering of Deceptions on Machine- and Human-Centric Attacks
Yuguang Yao, Xiao Guo, Vishal Asnani, Yifan Gong, Jiancheng Liu, Xue Lin, Xiaoming Liu and Sijia Liu
ISBN: 978-1-63828-340-9

Identifying and Mitigating the Security Risks of Generative AI
Clark Barrett *et al.*
ISBN: 978-1-63828-312-6

Cybersecurity for Modern Smart Grid Against Emerging Threats
Daisuke Mashima, Yao Chen, Muhammad M. Roomi, Subhash Lakshminarayana and Deming Chen
ISBN: 978-1-63828-294-5

Recommender Systems Meet Large Language Model Agents: A Survey

Xi Zhu
Rutgers University

Yu Wang
Netflix

Hang Gao
Rutgers University

Wujiang Xu
Rutgers University

Chen Wang
University of Illinois Chicago

Zhiwei Liu
Salesforce AI Research

Kun Wang
Squirrel Ai Learning

Mingyu Jin
Rutgers University

Linsey Pang
Salesforce

Qingsong Wen
Squirrel Ai Learning

Philip S. Yu
University of Illinois Chicago

Yongfeng Zhang
Rutgers University

now
the essence of knowledge
Boston — Delft

Foundations and Trends® in Privacy and Security

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

X. Zhu *et al.*. *Recommender Systems Meet Large Language Model Agents: A Survey.* Foundations and Trends® in Privacy and Security, vol. 7, no. 4, pp. 247–396, 2025.

ISBN: 978-1-63828-565-6

Full text available at: http://dx.doi.org/10.1561/3300000050

Foundations and Trends® in Privacy and Security

Volume 7, Issue 4, 2025

Editorial Board

Editorial Scope

Foundations and Trends® in Privacy and Security publishes survey and tutorial articles in the following topics:

- Access control
- Accountability
- Anonymity
- Application security
- Artifical intelligence methods in security and privacy
- Authentication
- Big data analytics and privacy
- Cloud security
- Cyber-physical systems security and privacy
- Distributed systems security and privacy
- Embedded systems security and privacy
- Forensics
- Hardware security
- Human factors in security and privacy
- Information flow
- Intrusion detection
- Malware
- Metrics
- Mobile security and privacy
- Language-based security and privacy
- Network security
- Privacy-preserving systems
- Protocol security
- Security and privacy policies
- Security architectures
- System security
- Web security and privacy

Information for Librarians

Foundations and Trends® in Privacy and Security, 2025, Volume 7, 4 issues. ISSN paper version 2474-1558. ISSN online version 2474-1566. Also available as a combined paper and online subscription.

Contents

Recommender Systems Meet Large Language Model Agents: A Survey

Xi Zhu[1*], Yu Wang[2*], Hang Gao[1*], Wujiang Xu[1*], Chen Wang[3], Zhiwei Liu[4], Kun Wang[5], Mingyu Jin[1], Linsey Pang[6], Qingsong Wen[5], Philip S. Yu[3] and Yongfeng Zhang[1]

[1] *Rutgers University, USA*
[2] *Netflix, USA*
[3] *University of Illinois Chicago, USA*
[4] *Salesforce AI Research, USA*
[5] *Squirrel Ai Learning, USA*
[6] *Salesforce, USA*

ABSTRACT

In recent years, the integration of Large Language Models (LLMs) and Recommender Systems (RS) has revolutionized the way personalized and intelligent user experiences are delivered. This survey provides an extensive review of critical challenges, current landscape, and future directions in the collaboration between LLM-based AI agents (LLM Agent) and recommender systems. We begin with an introduction to the foundational knowledge, exploring the components of LLM agents and the applications of LLMs in recommender systems. The survey then delves into the symbiotic relationship between LLM agents and recommender systems, illustrating how LLM agents enhance

*Xi Zhu, Yu Wang, Hang Gao, and Wujiang Xu are co-first authors of this work.

Xi Zhu, Yu Wang, Hang Gao, Wujiang Xu, Chen Wang, Zhiwei Liu, Kun Wang, Mingyu Jin, Linsey Pang, Qingsong Wen, Philip S. Yu and Yongfeng Zhang (2025), "Recommender Systems Meet Large Language Model Agents: A Survey", Foundations and Trends® in Privacy and Security: Vol. 7, No. 4, pp 247–396. DOI: 10.1561/3300000050.

recommender systems and how recommender systems support better LLM agents. Specifically, we discuss the overall architectures for designing LLM agents for recommendation, encompassing profile, memory, planning, and action components, along with multi-agent collaboration. Conversely, we investigate how recommender systems contribute to LLM agents, focusing on areas such as memory recommendation, plan recommendation, tool recommendation, agent recommendation, and personalized LLMs and LLM agents. Furthermore, a critical evaluation of trustworthy AI agents and recommender systems follows, addressing key issues of safety, explainability, fairness, and privacy. Finally, we propose potential future research directions, highlighting emerging trends and opportunities in the intersection of AI agents and recommender systems. This survey concludes by summarizing the key insights of current research and outlining promising avenues for future exploration in this rapidly evolving field. A curated collection of relevant papers for this survey is available in the GitHub repository: https://github.com/agiresearch/AgentRecSys.

1

Introduction

The integration of Large Language Model (LLM) and Recommender Systems (RS) has marked a transformative shift in how personalized recommendations are generated and delivered. Recommender systems, designed to predict user preferences and suggest relevant items, are ubiquitous in applications ranging from e-commerce to entertainment and social media. Historically, these systems have relied on techniques such as collaborative filtering, content-based filtering, and hybrid approaches. However, the advent of LLMs and AI agents has introduced new paradigms, significantly enhancing the capabilities and performance of recommender systems.

This survey seeks to thoroughly explore the interplay between LLM-based AI Agents (LLM agents) and recommender systems. It explores how LLM agents can enhance the functionality and effectiveness of recommender systems and, conversely, how recommender systems can optimize the performance and utility of LLM agents. By delving into these interconnections, we aim to shed light on the current state of research, highlight key challenges, and outline future directions in this fast-developing field. The importance of this survey is underscored by the growing sophistication and prevalence of LLM agents in various domains. As LLM agents continue to advance, their potential to enhance the accuracy, efficiency, and user experience of recommender systems grows increasingly impactful. Understanding the dynamic relationship between LLM agents and recommender systems is crucial for researchers and practitioners aiming to leverage AI technologies to develop next-generation recommender systems.

First, we introduce the foundational concepts necessary for understanding the integration of LLM agents into recommender systems in Section 2. This includes an overview of the evolution and capabilities of LLM-based AI agents and the application of LLMs in enhancing recommender systems. Additionally, we highlight the symbiotic relationship between LLM agents and recommender systems, which motivates us to organize the subsequent sections.

Then, we explore various approaches through which LLM agents can benefit recommender systems in Section 3. Specifically, we begin by discussing the limitations of existing recommender systems and how LLM agents address them, followed by the challenges of developing LLM agent-based recommender systems. Next, we explore the overall architecture and key components including memory, planning, and action that are essential for designing LLM agent recommender systems, along with the details of relevant technologies. Furthermore, we discuss how multiple agents collaborate to support more complex and effective recommender systems.

Conversely, we also investigate how recommender systems can enhance the functionality of LLM agents in Section 4. Specifically, we begin by analyzing the motivations, benefits, and challenges associated with applying recommender systems to LLM agents. Furthermore, we examine research on memory recommendation, plan recommendation for agents, tool recommendation, agent recommendation, and personalized agent configurations in the context of LLM agents. This section further highlights the bidirectional relationship, emphasizing the mutual benefits of integrating recommender systems with LLM agents.

Furthermore, as discussed in Section 5, the deployment of LLM agents in recommender systems raises critical issues related to trustworthiness. We address key challenges such as safety, explainability, fairness, and privacy of LLM agents within recommender systems. Ensuring that these systems are trustworthy, reliable, and robust is essential for their widespread adoption and effectiveness.

Finally, we explore potential future research directions in Section 6, highlighting emerging trends and opportunities at the intersection of LLM agents and recommender systems. We conclude this survey by highlighting our main contributions and the promising future of this field in Section 7.

This survey is timely and crucial due to the rapid advancements in LLM agents and the increasing need for sophisticated recommender systems. By exploring the intersection of these two fields, this survey provides a comprehensive understanding of recent advancements and future possibilities, offering valuable insights into how LLM agents can enhance recommendation capabilities and how recommender systems can, in turn, optimize LLM agents. What distinguishes this survey from existing literature is its holistic approach. To the best of our knowledge, this is the first survey to thoroughly detail the interaction between LLM agents and recommender systems, while other surveys might focus on specific aspects of LLM agents or recommender systems. Our survey encompasses the full spectrum of the interaction of LLM agents and recommender systems, covering key aspects such as definitions, motivations, current advancements, methodologies, and techniques, as well as future challenges and opportunities within each branch of research. Additionally, we address the critical issue of trustworthiness in the context of LLM agents and recommender systems, which is often overlooked in other surveys. In conclusion, our comprehensive analysis and forward-looking perspective make this survey a valuable resource for anyone interested in cutting-edge developments at the intersection of LLM agents and recommender systems.

Full text available at: http://dx.doi.org/10.1561/3300000050

References

Abdollahpouri, H., R. Burke, and B. Mobasher. (2017). "Controlling popularity bias in learning-to-rank recommendation". In: *Proceedings of the eleventh ACM conference on recommender systems (RecSys)*. 42–46.

Abdollahpouri, H. and M. Mansoury. (2020). "Multi-sided exposure bias in recommendation". *arXiv preprint arXiv:2006.15772*.

Abdollahpouri, H., M. Mansoury, R. Burke, and B. Mobasher. (2019). "The unfairness of popularity bias in recommendation". *arXiv preprint arXiv:1907.13286*.

Abdollahpouri, H., M. Mansoury, R. Burke, and B. Mobasher. (2020). "The connection between popularity bias, calibration, and fairness in recommendation". In: *Proceedings of the 14th ACM conference on recommender systems*. 726–731.

Abnar, S. and W. Zuidema. (2020). "Quantifying attention flow in transformers". *arXiv preprint arXiv:2005.00928*.

Achiam, J., S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, *et al.* (2023a). "Gpt-4 technical report". *arXiv preprint arXiv:2303.08774*.

Achiam, J., S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, *et al.* (2023b). "Gpt-4 technical report". *arXiv preprint arXiv:2303.08774*.

Ai, Q., V. Azizi, X. Chen, and Y. Zhang. (2018). “Learning heterogeneous knowledge base embeddings for explainable recommendation”. *Algorithms*. 11(9): 137.

Aïmeur, E., G. Brassard, J. M. Fernandez, and F. S. Mani Onana. (2008). “Alambic: a privacy-preserving recommender system for electronic commerce”. *International Journal of Information Security*. 7(5): 307–334.

Alayrac, J.-B., J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, *et al.* (2022). “Flamingo: a visual language model for few-shot learning”. *Advances in neural information processing systems*. 35: 23716–23736.

Alfrink, K., I. Keller, N. Doorn, and G. Kortuem. (2022). “Tensions in transparent urban AI: designing a smart electric vehicle charge point”. *AI & SOCIETY*: 1–17.

Alon, G. and M. Kamfonas. (2023). “Detecting language model attacks with perplexity”. *arXiv preprint arXiv:2308.14132*.

Amayuelas, A., X. Yang, A. Antoniades, W. Hua, L. Pan, and W. Wang. (2024). “Multiagent collaboration attack: Investigating adversarial attacks in large language model collaborations via debate”. *arXiv preprint arXiv:2406.14711*.

Andric, M., I. Ivanova, and F. Ricci. (2021). “Climbing Route Difficulty Grade Prediction and Explanation”. In: *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. 285–292.

Anelli, V. W., Y. Deldjoo, T. Di Noia, D. Malitesta, and F. A. Merra. (2021). “A study of defensive methods to protect visual recommendation against adversarial manipulation of images”. In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1094–1103.

Atanasova, P. (2024). “A diagnostic study of explainability techniques for text classification”. In: *Accountable and Explainable Methods for Complex Reasoning over Text*. Springer. 155–187.

Awad, N. F. and M. S. Krishnan. (2006). “The personalization privacy paradox: an empirical evaluation of information transparency and the willingness to be profiled online for personalization”. *MIS quarterly*: 13–28.

Badsha, S., X. Yi, and I. Khalil. (2016). “A practical privacy-preserving recommender system”. *Data Science and Engineering*. 1(3): 161–177.

Bagdasaryan, E. and V. Shmatikov. (2022). “Spinning Language Models: Risks of Propaganda-As-A-Service and Countermeasures”. In: *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE. DOI: 10.1109/sp46214.2022.9833572.

Bagdasaryan, E., R. Yi, S. Ghalebikesabi, P. Kairouz, M. Gruteser, S. Oh, B. Balle, and D. Ramage. (2024). “Air Gap: Protecting Privacy-Conscious Conversational Agents”. *arXiv preprint arXiv:2405.05175*.

Bansal, R. (2022). “A survey on bias and fairness in natural language processing”. *arXiv preprint arXiv:2204.09591*.

Bao, K., J. Zhang, Y. Zhang, W. Wang, F. Feng, and X. He. (2023). “Tallrec: An effective and efficient tuning framework to align large language model with recommendation”. In: *Proceedings of the 17th ACM Conference on Recommender Systems*. 1007–1014.

Barria Pineda, J. and P. Brusilovsky. (2019). “Making educational recommendations transparent through a fine-grained open learner model”. In: *Proceedings of Workshop on Intelligent User Interfaces for Algorithmic Transparency in Emerging Technologies at the 24th ACM Conference on Intelligent User Interfaces, IUI 2019, Los Angeles, USA, March 20, 2019*. Vol. 2327.

Bauman, K., B. Liu, and A. Tuzhilin. (2017). “Aspect based recommendations: Recommending items with the most valuable aspects based on user reviews”. In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 717–725.

Beigi, G., A. Mosallanezhad, R. Guo, H. Alvari, A. Nou, and H. Liu. (2020). “Privacy-aware recommendation with private-attribute protection using adversarial learning”. In: *Proceedings of the 13th International Conference on Web Search and Data Mining*. 34–42.

Belrose, N., Z. Furman, L. Smith, D. Halawi, I. Ostrovsky, L. McKinney, S. Biderman, and J. Steinhardt. (2023). “Eliciting latent predictions from transformers with the tuned lens”. *arXiv preprint arXiv:2303.08112*.

Bender, E. M., T. Gebru, A. McMillan-Major, and S. 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*. 610–623.

Beutel, A., J. Chen, T. Doshi, H. Qian, L. Wei, Y. Wu, L. Heldt, Z. Zhao, L. Hong, E. H. Chi, *et al.* (2019). “Fairness in recommendation ranking through pairwise comparisons”. In: *Proceedings of the 25th ACM SIGKDD*.

Bhagat, S., I. Rozenbaum, and G. Cormode. (2007). “Applying link-based classification to label blogs”. In: *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*. 92–101.

Bhardwaj, R. and S. Poria. (2023). “Red-teaming large language models using chain of utterances for safety-alignment”. *arXiv preprint arXiv:2308.09662*.

Bilgic, M. and R. J. Mooney. (2005). “Explaining recommendations: Satisfaction vs. promotion”. In: *Beyond personalization workshop, IUI*. Vol. 5. 153.

Blodgett, S. L., S. Barocas, H. Daumé III, and H. Wallach. (2020). “Language (technology) is power: A critical survey of" bias" in nlp”. *arXiv preprint arXiv:2005.14050*.

Bocklisch, T., J. Faulkner, N. Pawlowski, and A. Nichol. (2017). “Rasa: Open source language understanding and dialogue management”. *arXiv preprint arXiv:1712.05181*.

Bommasani, R., D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, *et al.* (2021). “On the opportunities and risks of foundation models”. *arXiv preprint arXiv:2108.07258*.

Bonawitz, K., V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth. (2017). “Practical secure aggregation for privacy-preserving machine learning”. In: *proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. 1175–1191.

Borgeaud, S., A. Mensch, J. Hoffmann, T. Cai, E. Rutherford, K. Millican, G. B. Van Den Driessche, J.-B. Lespiau, B. Damoc, A. Clark, *et al.* (2022). “Improving language models by retrieving from trillions of tokens”. In: *International conference on machine learning*. PMLR. 2206–2240.

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.* (2020). “Language models are few-shot learners”. *Advances in neural information processing systems*. 33: 1877–1901.

Bubeck, S., V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, *et al.* (2023). “Sparks of artificial general intelligence: Early experiments with gpt-4”. *arXiv preprint arXiv:2303.12712*.

Burke, R., B. Mobasher, and R. Bhaumik. (2005a). “Limited knowledge shilling attacks in collaborative filtering systems”. In: *Proceedings of 3rd international workshop on intelligent techniques for web personalization (ITWP 2005), 19th international joint conference on artificial intelligence (IJCAI 2005)*. 17–24.

Burke, R., B. Mobasher, R. Bhaumik, and C. Williams. (2005b). “Segment-based injection attacks against collaborative filtering recommender systems”. In: *Fifth IEEE International Conference on Data Mining (ICDM’05)*. IEEE. 4–pp.

Burke, R., B. Mobasher, C. Williams, and R. Bhaumik. (2006). “Classification features for attack detection in collaborative recommender systems”. In: *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. 542–547.

Burke, R., M. P. O’Mahony, and N. J. Hurley. (2015). “Robust collaborative recommendation”. *Recommender systems handbook*: 961–995.

Calandrino, J. A., A. Kilzer, A. Narayanan, E. W. Felten, and V. Shmatikov. (2011). “" You might also like:" Privacy risks of collaborative filtering”. In: *2011 IEEE symposium on security and privacy*. IEEE. 231–246.

Canny, J. (2002). “Collaborative filtering with privacy”. In: *Proceedings 2002 IEEE Symposium on Security and Privacy*. IEEE. 45–57.

Carbonell, J. and J. Goldstein. (1998). “The use of MMR, diversity-based reranking for reordering documents and producing summaries”. In: *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR ’98.* Melbourne, Australia: Association for Computing Machinery. 335–336. DOI: 10.1145/290941.291025.

Casper, S., J. Lin, J. Kwon, G. Culp, and D. Hadfield-Menell. (2023). “Explore, Establish, Exploit: Red Teaming Language Models from Scratch”. arXiv: 2306.09442 [cs.CL].

Celis, L. E., S. Kapoor, F. Salehi, and N. Vishnoi. (2019). “Controlling polarization in personalization: An algorithmic framework”. In: *Proceedings of the conference on fairness, accountability, and transparency*. 160–169.

Chaabane, A., G. Acs, M. A. Kaafar, *et al.* (2012). “You are what you like! information leakage through users’ interests”. In: *Proceedings of the 19th annual network & distributed system security symposium (NDSS).* Citeseer.

Chai, D., L. Wang, K. Chen, and Q. Yang. (2020). “Secure federated matrix factorization”. *IEEE Intelligent Systems*. 36(5): 11–20.

Chang, Y., X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, *et al.* (2024). “A survey on evaluation of large language models”. *ACM Transactions on Intelligent Systems and Technology*. 15(3): 1–45.

Chao, P., A. Robey, E. Dobriban, H. Hassani, G. J. Pappas, and E. Wong. (2023). “Jailbreaking Black Box Large Language Models in Twenty Queries”. arXiv: 2310.08419 [cs.LG].

Chen, H., G. Zheng, and Y. Ji. (2020). “Generating hierarchical explanations on text classification via feature interaction detection”. *arXiv preprint arXiv:2004.02015.*

Chen, H., X. Chen, S. Shi, and Y. Zhang. (2019a). “Generate natural language explanations for recommendation”. In: *Proceedings of the SIGIR 2019 Workshop on ExplainAble Recommendation and Search.*

Chen, H., Y. Li, S. Shi, S. Liu, H. Zhu, and Y. Zhang. (2022a). “Graph collaborative reasoning”. In: *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. 75–84.

Chen, H., S. Shi, Y. Li, and Y. Zhang. (2021a). "Neural collaborative reasoning". In: *Proceedings of the World Wide Web Conference 2021*. 1516–1527.

Chen, H., H. Chen, M. Yan, W. Xu, X. Gao, W. Shen, X. Quan, C. Li, J. Zhang, F. Huang, *et al.* (2024). "RoleInteract: Evaluating the Social Interaction of Role-Playing Agents". *arXiv preprint arXiv:2403.13679*.

Chen, H. and J. Li. (2019). "Adversarial tensor factorization for context-aware recommendation". In: *Proceedings of the 13th ACM Conference on Recommender Systems (RecSys)*. 363–367.

Chen, J., H. Dong, X. Wang, F. Feng, M. Wang, and X. He. (2023). "Bias and debias in recommender system: A survey and future directions". *ACM Transactions on Information Systems*. 41(3): 1–39.

Chen, J., W. Fan, G. Zhu, X. Zhao, C. Yuan, Q. Li, and Y. Huang. (2022b). "Knowledge-enhanced black-box attacks for recommendations". In: *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 108–117.

Chen, M., J. Tworek, H. Jun, Q. Yuan, H. P. D. O. Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, *et al.* (2021b). "Evaluating large language models trained on code". *arXiv preprint arXiv:2107.03374*.

Chen, N., Y. Wang, H. Jiang, D. Cai, Y. Li, Z. Chen, L. Wang, and J. Li. (2022c). "Large Language Models Meet Harry Potter: A Bilingual Dataset for Aligning Dialogue Agents with Characters". *arXiv preprint arXiv:2211.06869*.

Chen, X. and V. Huang. (2012). "Privacy preserving data publishing for recommender system". In: *2012 IEEE 36th Annual Computer Software and Applications Conference Workshops*. IEEE. 128–133.

Chen, X., H. Chen, H. Xu, Y. Zhang, Y. Cao, Z. Qin, and H. Zha. (2019b). "Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation". In: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 765–774.

Chen, X., Y. Zhang, and Z. Qin. (2019c). “Dynamic explainable recommendation based on neural attentive models”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. No. 01. 53–60.

Chen, X., Y. Zhang, H. Xu, Y. Cao, Z. Qin, and H. Zha. (2018). “Visually explainable recommendation”. *arXiv preprint arXiv:1801.10288*.

Cheng, J., C. Danescu-Niculescu-Mizil, and J. Leskovec. (2015). “Antisocial behavior in online discussion communities”. In: *Proceedings of the international aaai conference on web and social media*. Vol. 9. No. 1. 61–70.

Cheng, M., E. Durmus, and D. Jurafsky. (2023). “Marked personas: Using natural language prompts to measure stereotypes in language models”. *arXiv preprint arXiv:2305.18189*.

Cheng, Z., X. Chang, L. Zhu, R. C. Kanjirathinkal, and M. Kankanhalli. (2019). “MMALFM: Explainable recommendation by leveraging reviews and images”. *ACM Transactions on Information Systems (TOIS)*. 37(2): 1–28.

Chiang, W.-L., Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, and E. P. Xing. (2023). “Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality”. URL: https://lmsys.org/blog/2023-03-30-vicuna/.

Chowdhery, A., S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, *et al.* (2022). “Palm: Scaling language modeling with pathways”. *arXiv preprint arXiv:2204.02311*.

Christakopoulou, K. and A. Banerjee. (2018). “Adversarial recommendation: Attack of the learned fake users”. *arXiv preprint arXiv:1809.08336*.

Christakopoulou, K. and A. Banerjee. (2019). “Adversarial attacks on an oblivious recommender”. In: *Proceedings of the 13th ACM Conference on Recommender Systems*. 322–330.

Christakopoulou, K., A. Lalama, C. Adams, I. Qu, Y. Amir, S. Chucri, P. Vollucci, F. Soldo, D. Bseiso, S. Scodel, *et al.* (2023). “Large language models for user interest journeys”. *arXiv preprint arXiv:2305.15498*.

Chu, Z., Z. Wang, and W. Zhang. (2024). “Fairness in large language models: A taxonomic survey”. *ACM SIGKDD explorations newsletter*. 26(1): 34–48.

Chughtai, B., L. Chan, and N. Nanda. (2023). “A toy model of universality: Reverse engineering how networks learn group operations”. In: *International Conference on Machine Learning*. PMLR. 6243–6267.

Cissée, R. and S. Albayrak. (2007). “An agent-based approach for privacy-preserving recommender systems”. In: *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems*. 1–8.

Collins, A., D. Tkaczyk, A. Aizawa, and J. Beel. (2018). “A study of position bias in digital library recommender systems”. *arXiv preprint arXiv:1802.06565*.

Corecco, N., G. Piatti, L. A. Lanzendörfer, F. X. Fan, and R. Wattenhofer. (2024). “An LLM-based Recommender System Environment”. *arXiv preprint arXiv:2406.01631*.

Costa-jussà, M. R., J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, *et al.* (2022). “No language left behind: Scaling human-centered machine translation”. *arXiv preprint arXiv:2207.04672*.

Crocco, M. S., A. Segall, A.-L. Halvorsen, A. Stamm, and R. Jacobsen. (2020). ““It’s not like they’re selling your data to dangerous people”: Internet privacy, teens, and (non-) controversial public issues”. *The Journal of Social Studies Research*. 44(1): 21–33.

Cui, S., Z. Zhang, Y. Chen, W. Zhang, T. Liu, S. Wang, and T. Liu. (2023). “FFT: Towards Harmlessness Evaluation and Analysis for LLMs with Factuality, Fairness, Toxicity”. arXiv: 2311.18580 [cs.CL].

Cui, Z., J. Ma, C. Zhou, J. Zhou, and H. Yang. (2022). “M6-Rec: Generative Pretrained Language Models are Open-Ended Recommender Systems”. *arXiv preprint arXiv:2205.08084*.

Dafoe, A., E. Hughes, Y. Bachrach, T. Collins, K. R. McKee, J. Z. Leibo, K. Larson, and T. Graepel. (2020). “Open problems in cooperative ai”. *arXiv preprint arXiv:2012.08630*.

Dai, G., W. Zhang, J. Li, S. Yang, S. Rao, A. Caetano, M. Sra, *et al.* (2024a). "Artificial Leviathan: Exploring Social Evolution of LLM Agents Through the Lens of Hobbesian Social Contract Theory". *arXiv preprint arXiv:2406.14373*.

Dai, Y., H. Hu, L. Wang, S. Jin, X. Chen, and Z. Lu. (2024b). "MM-Role: A Comprehensive Framework for Developing and Evaluating Multimodal Role-Playing Agents". *arXiv preprint arXiv:2408.04203*.

Dai, Z., Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov. (2019). "Transformer-xl: Attentive language models beyond a fixed-length context". *arXiv preprint arXiv:1901.02860*.

Deng, Y., W. Zhang, W. Xu, W. Lei, T.-S. Chua, and W. Lam. (2023). "A unified multi-task learning framework for multi-goal conversational recommender systems". *ACM Transactions on Information Systems*. 41(3): 1–25.

Dettmers, T., A. Pagnoni, A. Holtzman, and L. Zettlemoyer. (2023). "Qlora: Efficient finetuning of quantized llms". *Advances in neural information processing systems*. 36: 10088–10115.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding". *arXiv preprint arXiv:1810.04805*.

Dey, R., C. Tang, K. Ross, and N. Saxena. (2012). "Estimating age privacy leakage in online social networks". In: *2012 proceedings ieee infocom*. IEEE. 2836–2840.

Dong, Q., L. Li, D. Dai, C. Zheng, J. Ma, R. Li, H. Xia, J. Xu, Z. Wu, T. Liu, *et al.* (2022). "A survey on in-context learning". *arXiv preprint arXiv:2301.00234*.

Dong, Z., Z. Zhou, C. Yang, J. Shao, and Y. Qiao. (2024). "Attacks, defenses and evaluations for llm conversation safety: A survey". *arXiv preprint arXiv:2402.09283*.

Al-Doulat, A. (2021). "FIRST: Finding Interesting StoRies about STudents-An Interactive Narrative Approach to Explainable Learning Analytics". *PhD thesis*. The University of North Carolina at Charlotte.

Edemacu, K. and X. Wu. (2024). "Privacy preserving prompt engineering: A survey". *arXiv preprint arXiv:2404.06001*.

Edge, D., H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, and J. Larson. (2024). "From local to global: A graph rag approach to query-focused summarization". *arXiv preprint arXiv:2404.16130*.

Ekstrand, M. D., M. Tian, I. M. Azpiazu, J. D. Ekstrand, O. Anuyah, D. McNeill, and M. S. Pera. (2018). "All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness". In: *Conference on fairness, accountability and transparency*. PMLR. 172–186.

Enguehard, J. (2023). "Sequential Integrated Gradients: a simple but effective method for explaining language models". *arXiv preprint arXiv:2305.15853*.

Erkin, Z., T. Veugen, T. Toft, and R. L. Lagendijk. (2012). "Generating private recommendations efficiently using homomorphic encryption and data packing". *IEEE transactions on information forensics and security*. 7(3): 1053–1066.

Esiobu, D., X. Tan, S. Hosseini, M. Ung, Y. Zhang, J. Fernandes, J. Dwivedi-Yu, E. Presani, A. Williams, and E. Smith. (2023). "ROBBIE: Robust bias evaluation of large generative language models". In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 3764–3814.

Fan, A., M. Lewis, and Y. Dauphin. (2018). "Hierarchical Neural Story Generation". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 889–898.

Fan, W., T. Derr, X. Zhao, Y. Ma, H. Liu, J. Wang, J. Tang, and Q. Li. (2021). "Attacking black-box recommendations via copying cross-domain user profiles". In: *2021 IEEE 37th international conference on data engineering (ICDE)*. IEEE. 1583–1594.

Fan, W., X. Zhao, X. Chen, J. Su, J. Gao, L. Wang, Q. Liu, Y. Wang, H. Xu, L. Chen, and Q. Li. (2022). "A Comprehensive Survey on Trustworthy Recommender Systems". arXiv: 2209.10117 [cs.IR]. URL: https://arxiv.org/abs/2209.10117.

Fang, J., S. Gao, P. Ren, X. Chen, S. Verberne, and Z. Ren. (2024). "A multi-agent conversational recommender system". *arXiv preprint arXiv:2402.01135*.

Fang, M., N. Z. Gong, and J. Liu. (2020). "Influence function based data poisoning attacks to top-n recommender systems". In: *Proceedings of the World Wide Web Conference 2020*. 3019–3025.

Fang, M., G. Yang, N. Z. Gong, and J. Liu. (2018). "Poisoning attacks to graph-based recommender systems". In: *Proceedings of the 34th Annual Computer Security Applications Conference*. 381–392.

Fedus, W., B. Zoph, and N. Shazeer. (2022). "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity". *Journal of Machine Learning Research*. 23(120): 1–39.

Feng, S., E. Wallace, A. Grissom II, M. Iyyer, P. Rodriguez, and J. Boyd-Graber. (2018). "Pathologies of neural models make interpretations difficult". *arXiv preprint arXiv:1804.07781*.

Feng, Y., S. Liu, Z. Xue, Q. Cai, L. Hu, P. Jiang, K. Gai, and F. Sun. (2023). "A large language model enhanced conversational recommender system". *arXiv preprint arXiv:2308.06212*.

Ferrando, J., G. I. Gállego, and M. R. Costa-Jussà. (2022). "Measuring the mixing of contextual information in the transformer". *arXiv preprint arXiv:2203.04212*.

Fore, M., S. Singh, and D. Stamoulis. (2024). "GeckOpt: LLM System Efficiency via Intent-Based Tool Selection". In: *Proceedings of the Great Lakes Symposium on VLSI 2024*. 353–354.

Fredrikson, M., S. Jha, and T. Ristenpart. (2015). "Model inversion attacks that exploit confidence information and basic countermeasures". In: *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*. 1322–1333.

Friedman, L., S. Ahuja, D. Allen, Z. Tan, H. Sidahmed, C. Long, J. Xie, G. Schubiner, A. Patel, H. Lara, *et al.* (2023). "Leveraging large language models in conversational recommender systems". *arXiv preprint arXiv:2305.07961*.

Gade, P., S. Lermen, C. Rogers-Smith, and J. Ladish. (2023). "BadLlama: cheaply removing safety fine-tuning from Llama 2-Chat 13B". arXiv: 2311.00117 [cs.CL].

Ganguli, D., L. Lovitt, J. Kernion, A. Askell, Y. Bai, S. Kadavath, B. Mann, E. Perez, N. Schiefer, K. Ndousse, A. Jones, S. Bowman, A. Chen, T. Conerly, N. DasSarma, D. Drain, N. Elhage, S. El-Showk, S. Fort, Z. Hatfield-Dodds, T. Henighan, D. Hernandez, T. Hume, J. Jacobson, S. Johnston, S. Kravec, C. Olsson, S. Ringer, E. Tran-Johnson, D. Amodei, T. Brown, N. Joseph, S. McCandlish, C. Olah, J. Kaplan, and J. Clark. (2022). "Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned". arXiv: 2209.07858 [cs.CL].

Ganta, S. R., S. P. Kasiviswanathan, and A. Smith. (2008). "Composition attacks and auxiliary information in data privacy". In: *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. 265–273.

Gao, H. and Y. Zhang. (2024a). "Memory Sharing for Large Language Model based Agents". *arXiv preprint arXiv:2404.09982*.

Gao, H. and Y. Zhang. (2024b). "PTR: Precision-Driven Tool Recommendation for Large Language Models". *arXiv preprint arXiv:2411.09613*.

Gao, H. and Y. Zhang. (2024c). "VRSD: Rethinking Similarity and Diversity for Retrieval in Large Language Models". arXiv: 2407.04573 [cs.IR]. URL: https://arxiv.org/abs/2407.04573.

Gao, J., L. Qi, H. Huang, and C. Sha. (2020a). "Shilling attack detection scheme in collaborative filtering recommendation system based on recurrent neural network". In: *Advances in Information and Communication: Proceedings of the 2020 Future of Information and Communication Conference (FICC), Volume 1*. Springer. 634–644.

Gao, J., B. Chen, X. Zhao, W. Liu, X. Li, Y. Wang, Z. Zhang, W. Wang, Y. Ye, S. Lin, *et al.* (2024a). "LLM-enhanced Reranking in Recommender Systems". *arXiv preprint arXiv:2406.12433*.

Gao, L., A. Madaan, S. Zhou, U. Alon, P. Liu, Y. Yang, J. Callan, and G. Neubig. (2023a). "Pal: Program-aided language models". In: *International Conference on Machine Learning*. PMLR. 10764–10799.

Gao, R. and C. Shah. (2021). “Addressing bias and fairness in search systems”. In: *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*. 2643–2646.

Gao, S., Z. Shi, M. Zhu, B. Fang, X. Xin, P. Ren, Z. Chen, J. Ma, and Z. Ren. (2024b). “Confucius: Iterative tool learning from introspection feedback by easy-to-difficult curriculum”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. No. 16. 18030–18038.

Gao, S., J. Dwivedi-Yu, P. Yu, X. E. Tan, R. Pasunuru, O. Golovneva, K. Sinha, A. Celikyilmaz, A. Bosselut, and T. Wang. (2024c). “Efficient Tool Use with Chain-of-Abstraction Reasoning”. *arXiv preprint arXiv:2401.17464*.

Gao, T., A. Fisch, and D. Chen. (2020b). “Making pre-trained language models better few-shot learners”. *arXiv preprint arXiv:2012.15723*.

Gao, Y., T. Sheng, Y. Xiang, Y. Xiong, H. Wang, and J. Zhang. (2023b). “Chat-rec: Towards interactive and explainable llms-augmented recommender system”. *arXiv preprint arXiv:2303.14524*.

Gao, Y., Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, and H. Wang. (2023c). “Retrieval-augmented generation for large language models: A survey”. *arXiv preprint arXiv:2312.10997*.

Garrido-Muñoz, I., A. Montejo-Ráez, F. Martínez-Santiago, and L. A. Ureña-López. (2021). “A survey on bias in deep NLP”. *Applied Sciences*. 11(7): 3184.

Gawlikowski, J., C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher, *et al.* (2023). “A survey of uncertainty in deep neural networks”. *Artificial Intelligence Review*. 56(Suppl 1): 1513–1589.

Ge, Y., W. Hua, K. Mei, J. Tan, S. Xu, Z. Li, Y. Zhang, *et al.* (2024). “Openagi: When llm meets domain experts”. *Advances in Neural Information Processing Systems*. 36.

Ge, Y., S. Liu, Z. Fu, J. Tan, Z. Li, S. Xu, Y. Li, Y. Xian, and Y. Zhang. (2022a). “A survey on trustworthy recommender systems”. *ACM Transactions on Recommender Systems*.

Ge, Y., S. Liu, R. Gao, Y. Xian, Y. Li, X. Zhao, C. Pei, F. Sun, J. Ge, W. Ou, and Y. Zhang. (2021). "Towards Long-term Fairness in Recommendation". In: *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 445–453.

Ge, Y., Y. Ren, W. Hua, S. Xu, J. Tan, and Y. Zhang. (2023). "Llm as os (llmao), agents as apps: Envisioning aios, agents and the aios-agent ecosystem". *arXiv preprint arXiv:2312.03815*.

Ge, Y., X. Zhao, L. Yu, S. Paul, D. Hu, C.-C. Hsieh, and Y. Zhang. (2022b). "Toward Pareto Efficient Fairness-Utility Trade-off in Recommendation through Reinforcement Learning". In: *Proceedings of the 15th ACM International Conference on Web Search and Data Mining*.

Gehman, S., S. Gururangan, M. Sap, Y. Choi, and N. A. Smith. (2020). "Realtoxicityprompts: Evaluating neural toxic degeneration in language models". *arXiv preprint arXiv:2009.11462*.

Geng, S., S. Liu, Z. Fu, Y. Ge, and Y. Zhang. (2022). "Recommendation as Language Processing (RLP): A Unified Pretrain, Personalized Prompt & Predict Paradigm (P5)". In: *Proceedings of the 16th ACM Conference on Recommender Systems (RecSys)*.

Geva, M., A. Caciularu, K. R. Wang, and Y. Goldberg. (2022). "Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space". *arXiv preprint arXiv:2203.14680*.

Geyik, S. C., S. Ambler, and K. Kenthapadi. (2019). "Fairness-Aware Ranking in Search & Recommendation Systems with Application to LinkedIn Talent Search". In: *Proceedings of SIGKDD*. ACM. 2221–2231.

Ghazimatin, A., O. Balalau, R. Saha Roy, and G. Weikum. (2020). "PRINCE: Provider-side interpretability with counterfactual explanations in recommender systems". In: *Proceedings of the 13th International Conference on Web Search and Data Mining*. 196–204.

Glavic, B., A. Meliou, and S. Roy. (2021). "Trends in explanations: Understanding and debugging data-driven systems". *Foundations and Trends® in Databases*. 11(3).

Gong, P., J. Li, and J. Mao. (2024). "CoSearchAgent: A Lightweight Collaborative Search Agent with Large Language Models". In: *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 2729–2733.

Graves, A., G. Wayne, M. Reynolds, T. Harley, I. Danihelka, A. Grabska-Barwińska, S. G. Colmenarejo, E. Grefenstette, T. Ramalho, J. Agapiou, *et al.* (2016). "Hybrid computing using a neural network with dynamic external memory". *Nature.* 538(7626): 471–476.

Gruver, N., M. Finzi, S. Qiu, and A. G. Wilson. (2024). "Large language models are zero-shot time series forecasters". *Advances in Neural Information Processing Systems.* 36.

Gu, Z., X. Zhu, H. Guo, L. Zhang, Y. Cai, H. Shen, J. Chen, Z. Ye, Y. Dai, Y. Gao, *et al.* (2024). "Agent Group Chat: An Interactive Group Chat Simulacra For Better Eliciting Collective Emergent Behavior". *arXiv preprint arXiv:2403.13433.*

Gupta, A., E. Johnson, J. Payan, A. K. Roy, A. Kobren, S. Panda, J.-B. Tristan, and M. Wick. (2021). "Online post-processing in rankings for fair utility maximization". In: *Proceedings of the 14th ACM International Conference on Web Search and Data Mining.* 454–462.

Gupta, M., C. Akiri, K. Aryal, E. Parker, and L. Praharaj. (2023). "From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy". arXiv: 2307.00691 [cs.CR]. URL: https://arxiv.org/abs/2307.00691.

Gururangan, S., A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith. (2020). "Don't stop pretraining: Adapt language models to domains and tasks". *arXiv preprint arXiv:2004.10964.*

Guu, K., K. Lee, Z. Tung, P. Pasupat, and M. Chang. (2020). "Retrieval augmented language model pre-training". In: *International conference on machine learning.* PMLR. 3929–3938.

Ha, M., X. Tao, W. Lin, Q. Ma, W. Xu, and L. Chen. (2024). "Fine-Grained Dynamic Framework for Bias-Variance Joint Optimization on Data Missing Not at Random". *arXiv preprint arXiv:2405.15403.*

Hada, D. V. and S. K. Shevade. (2021). "ReXPlug: Explainable Recommendation using Plug-and-Play Language Model". In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 81–91.

Hadi, M. U., R. Qureshi, A. Shah, M. Irfan, A. Zafar, M. B. Shaikh, N. Akhtar, J. Wu, S. Mirjalili, *et al.* (2023). "A survey on large language models: Applications, challenges, limitations, and practical usage". *Authorea Preprints.*

Halawi, D., J.-S. Denain, and J. Steinhardt. (2023). "Overthinking the truth: Understanding how language models process false demonstrations". *arXiv preprint arXiv:2307.09476.*

Halder, K., M.-Y. Kan, and K. Sugiyama. (2017). "Health forum thread recommendation using an interest aware topic model". In: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM).* 1589–1598.

Hao, S., T. Liu, Z. Wang, and Z. Hu. (2024). "Toolkengpt: Augmenting frozen language models with massive tools via tool embeddings". *Advances in neural information processing systems.* 36.

Hardt, M., E. Price, and N. Srebro. (2016). "Equality of opportunity in supervised learning". In: *NeurIPS.* 3315–3323.

Hartvigsen, T., S. Gabriel, H. Palangi, M. Sap, D. Ray, and E. Kamar. (2022). "Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection". *arXiv preprint arXiv:2203.09509.*

He, J., W. W. Chu, and Z. V. Liu. (2006). "Inferring privacy information from social networks". In: *International Conference on Intelligence and Security Informatics.* Springer. 154–165.

He, P., J. Gao, and W. Chen. (2023). "DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing". arXiv: 2111.09543 [cs.CL].

He, X., T. Chen, M.-Y. Kan, and X. Chen. (2015). "Trirank: Review-aware explainable recommendation by modeling aspects". In: *Proceedings of the ACM International Conference on Information & Knowledge Management (CIKM).*

He, X., K. Deng, X. Wang, Y. Li, Y. Zhang, and M. Wang. (2020). "Lightgcn: Simplifying and powering graph convolution network for recommendation". In: *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*. 639–648.

He, X., Z. He, X. Du, and T.-S. Chua. (2018). "Adversarial Personalized Ranking for Recommendation". In: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM. 355–364.

He, X., L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua. (2017). "Neural Collaborative Filtering". In: *Proceedings of the World Wide Web Conference*. 173–182.

Herlocker, J. L., J. A. Konstan, and J. Riedl. (2000). "Explaining collaborative filtering recommendations". In: *Proceedings of the 2000 ACM conference on Computer supported cooperative work*. 241–250.

Hernandez, E., B. Z. Li, and J. Andreas. (2023). "Inspecting and editing knowledge representations in language models". *arXiv preprint arXiv:2304.00740*.

Hewitt, J. and C. D. Manning. (2019). "A structural probe for finding syntax in word representations". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4129–4138.

Hidano, S., T. Murakami, S. Katsumata, S. Kiyomoto, and G. Hanaoka. (2017). "Model inversion attacks for prediction systems: Without knowledge of non-sensitive attributes". In: *2017 15th Annual Conference on Privacy, Security and Trust (PST)*. IEEE. 115–11509.

Himeur, Y., A. Alsalemi, A. Al-Kababji, F. Bensaali, A. Amira, C. Sardianos, G. Dimitrakopoulos, and I. Varlamis. (2021a). "A survey of recommender systems for energy efficiency in buildings: Principles, challenges and prospects". *Information Fusion*. 72: 1–21.

Himeur, Y., K. Ghanem, A. Alsalemi, F. Bensaali, and A. Amira. (2021b). "Artificial intelligence based anomaly detection of energy consumption in buildings: A review, current trends and new perspectives". *Applied Energy*. 287: 116601.

Hogan, A., E. Blomqvist, M. Cochez, C. d'Amato, G. D. Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, *et al.* (2021). "Knowledge graphs". *ACM Computing Surveys (Csur)*. 54(4): 1–37.

Houlsby, N., A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly. (2019). "Parameter-efficient transfer learning for NLP". In: *International conference on machine learning*. PMLR. 2790–2799.

Hu, E. J., Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, *et al.* (2022). "Lora: Low-rank adaptation of large language models." *ICLR*. 1(2): 3.

Hu, J., W. Liu, and M. Du. (2024). "Strategic Demonstration Selection for Improved Fairness in LLM In-Context Learning". *arXiv preprint arXiv:2408.09757*.

Hua, W., L. Fan, L. Li, K. Mei, J. Ji, Y. Ge, L. Hemphill, and Y. Zhang. (2023). "War and peace (waragent): Large language model-based multi-agent simulation of world wars". *arXiv preprint arXiv:2311.17227*.

Hua, W., Y. Ge, S. Xu, J. Ji, and Y. Zhang. (2024a). "UP5: Unbiased Foundation Model for Fairness-aware Recommendation". *EACL*.

Hua, W., X. Yang, M. Jin, Z. Li, W. Cheng, R. Tang, and Y. Zhang. (2024b). "Trustagent: Towards safe and trustworthy llm-based agents through agent constitution". In: *Trustworthy Multi-modal Foundation Models and AI Agents (TiFA)*.

Huang, F., Z. Yang, J. Jiang, Y. Bei, Y. Zhang, and H. Chen. (2024). "Large Language Model Interaction Simulator for Cold-Start Item Recommendation". *arXiv preprint arXiv:2402.09176*.

Huang, J. and K. C.-C. Chang. (2022). "Towards reasoning in large language models: A survey". *arXiv preprint arXiv:2212.10403*.

Huang, X., J. Lian, Y. Lei, J. Yao, D. Lian, and X. Xie. (2023). "Recommender ai agent: Integrating large language models for interactive recommendations". *arXiv preprint arXiv:2308.16505*.

Islam, R., K. N. Keya, Z. Zeng, S. Pan, and J. Foulds. (2021). "Debiasing career recommendations with neural fair collaborative filtering". In: *Proceedings of the Web Conference 2021*. 3779–3790.

Izacard, G., M. Caron, L. Hosseini, S. Riedel, P. Bojanowski, A. Joulin, and E. Grave. (2021). "Unsupervised dense information retrieval with contrastive learning". *arXiv preprint arXiv:2112.09118*.

Izacard, G. and É. Grave. (2021). "Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 874–880.

Ji, J., Y. Chen, M. Jin, W. Xu, W. Hua, and Y. Zhang. (2024). "MoralBench: Moral Evaluation of LLMs". *arXiv preprint arXiv:2406.04428*.

Ji, S., S. Pan, E. Cambria, P. Marttinen, and S. Y. Philip. (2021). "A survey on knowledge graphs: Representation, acquisition, and applications". *IEEE transactions on neural networks and learning systems*. 33(2): 494–514.

Jia, J. and N. Z. Gong. (2018). "{AttriGuard}: A Practical Defense Against Attribute Inference Attacks via Adversarial Machine Learning". In: *27th USENIX Security Symposium (USENIX Security 18)*. 513–529.

Jiang, J.-Y., C.-T. Li, and S.-D. Lin. (2019). "Towards a more reliable privacy-preserving recommender system". *Information Sciences*. 482: 248–265.

Jin, C., H. Peng, A. Zhang, N. Chen, J. Zhao, X. Xie, K. Li, S. Feng, K. Zhong, C. Ding, *et al.* (2025a). "RankFlow: A Multi-Role Collaborative Reranking Workflow Utilizing Large Language Models". *arXiv preprint arXiv:2502.00709*.

Jin, C., H. Peng, S. Zhao, Z. Wang, W. Xu, L. Han, J. Zhao, K. Zhong, S. Rajasekaran, and D. N. Metaxas. (2024a). "Apeer: Automatic prompt engineering enhances large language model reranking". *arXiv preprint arXiv:2406.14449*.

Jin, M., S. Wang, L. Ma, Z. Chu, J. Y. Zhang, X. Shi, P.-Y. Chen, Y. Liang, Y.-F. Li, S. Pan, *et al.* (2023). "Time-llm: Time series forecasting by reprogramming large language models". *arXiv preprint arXiv:2310.01728*.

Jin, M., K. Mei, W. Xu, M. Sun, R. Tang, M. Du, Z. Liu, and Y. Zhang. (2025b). “Massive Values in Self-Attention Modules are the Key to Contextual Knowledge Understanding”. *arXiv preprint arXiv:2502.01563*.

Jin, M., Q. Yu, J. Huang, Q. Zeng, Z. Wang, W. Hua, H. Zhao, K. Mei, Y. Meng, K. Ding, *et al.* (2024b). “Exploring Concept Depth: How Large Language Models Acquire Knowledge at Different Layers?” *arXiv preprint arXiv:2404.07066*.

Jin, M., Q. Yu, D. Shu, H. Zhao, W. Hua, Y. Meng, Y. Zhang, and M. Du. (2024c). “The impact of reasoning step length on large language models”. *arXiv preprint arXiv:2401.04925*.

Jin, M., S. Zhu, B. Wang, Z. Zhou, C. Zhang, Y. Zhang, *et al.* (2024d). “Attackeval: How to evaluate the effectiveness of jailbreak attacking on large language models”. *arXiv preprint arXiv:2401.09002*.

Joachims, T., L. Granka, B. Pan, H. Hembrooke, and G. Gay. (2017). “Accurately interpreting clickthrough data as implicit feedback”. In: *Acm Sigir Forum*. Vol. 51. No. 1. Acm New York, NY, USA. 4–11.

Joachims, T., L. Granka, B. Pan, H. Hembrooke, F. Radlinski, and G. Gay. (2007). “Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search”. *ACM Transactions on Information Systems (TOIS)*. 25(2): 7–es.

Johnson, J., M. Douze, and H. Jégou. (2019). “Billion-scale similarity search with GPUs”. *IEEE Transactions on Big Data*. 7(3): 535–547.

Jones, E., A. Dragan, A. Raghunathan, and J. Steinhardt. (2023). “Automatically Auditing Large Language Models via Discrete Optimization”. arXiv: 2303.04381 [cs.LG].

Kaneko, M., D. Bollegala, N. Okazaki, and T. Baldwin. (2024). “Evaluating Gender Bias in Large Language Models via Chain-of-Thought Prompting”. arXiv: 2401.15585 [cs.CL]. URL: https://arxiv.org/abs/2401.15585.

Kang, D., X. Li, I. Stoica, C. Guestrin, M. Zaharia, and T. Hashimoto. (2024). “Exploiting programmatic behavior of llms: Dual-use through standard security attacks”. In: *2024 IEEE Security and Privacy Workshops (SPW)*. IEEE. 132–143.

Kang, W.-C. and J. McAuley. (2018). "Self-attentive sequential recommendation". In: *2018 IEEE international conference on data mining (ICDM)*. IEEE. 197–206.

Kapoor, S. (2018). "Multi-agent reinforcement learning: A report on challenges and approaches". *arXiv preprint arXiv:1807.09427*.

Karpas, E., O. Abend, Y. Belinkov, B. Lenz, O. Lieber, N. Ratner, Y. Shoham, H. Bata, Y. Levine, K. Leyton-Brown, *et al.* (2022). "MRKL Systems: A modular, neuro-symbolic architecture that combines large language models, external knowledge sources and discrete reasoning". *arXiv preprint arXiv:2205.00445*.

Karthikeyan, P., S. T. Selvi, G. Neeraja, R. Deepika, A. Vincent, and V. Abinaya. (2017). "Prevention of shilling attack in recommender systems using discrete wavelet transform and support vector machine". In: *2016 eighth international conference on Advanced Computing (ICoAC)*. IEEE. 99–104.

Katz, D. M., M. J. Bommarito, S. Gao, and P. Arredondo. (2024). "Gpt-4 passes the bar exam". *Philosophical Transactions of the Royal Society A*. 382(2270): 20230254.

Khandelwal, U., O. Levy, D. Jurafsky, L. Zettlemoyer, and M. Lewis. (2019). "Generalization through memorization: Nearest neighbor language models". *arXiv preprint arXiv:1911.00172*.

Khashabi, D., S. Min, T. Khot, A. Sabharwal, O. Tafjord, P. Clark, and H. Hajishirzi. (2020). "UNIFIEDQA: Crossing Format Boundaries with a Single QA System". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. 1896–1907.

Kindermans, P.-J., S. Hooker, J. Adebayo, M. Alber, K. T. Schütt, S. Dähne, D. Erhan, and B. Kim. (2019). "The (un) reliability of saliency methods". *Explainable AI: Interpreting, explaining and visualizing deep learning*: 267–280.

Kindermans, P.-J., K. Schütt, K.-R. Müller, and S. Dähne. (2016). "Investigating the influence of noise and distractors on the interpretation of neural networks". *arXiv preprint arXiv:1611.07270*.

Kojima, T., S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. (2022). "Large language models are zero-shot reasoners". *Advances in neural information processing systems*. 35: 22199–22213.

Kong, Y., J. Ruan, Y. Chen, B. Zhang, T. Bao, S. Shi, G. Du, X. Hu, H. Mao, Z. Li, *et al.* (2023). "Tptu-v2: Boosting task planning and tool usage of large language model-based agents in real-world systems". *arXiv preprint arXiv:2311.11315*.

Krishnamurthy, B. and C. E. Wills. (2009). "On the leakage of personally identifiable information via online social networks". In: *Proceedings of the 2nd ACM workshop on Online social networks*. 7–12.

Kumar, S., V. Balachandran, L. Njoo, A. Anastasopoulos, and Y. Tsvetkov. (2022). "Language generation models can cause harm: so what can we do about it". *An actionable survey. CoRR abs/2210.07700*.

Lahoti, P., K. P. Gummadi, and G. Weikum. (2019). "ifair: Learning individually fair data representations for algorithmic decision making". In: *2019 ieee 35th international conference on data engineering (icde)*. IEEE. 1334–1345.

Lam, S. K. and J. Riedl. (2004). "Shilling recommender systems for fun and profit". In: *Proceedings of the World Wide Web Conference*. 393–402.

Lambert, N., L. Castricato, L. von Werra, and A. Havrilla. (2022). "Illustrating Reinforcement Learning from Human Feedback (RLHF)". *Hugging Face Blog*.

Lampinen, A. K., I. Dasgupta, S. C. Chan, K. Matthewson, M. H. Tessler, A. Creswell, J. L. McClelland, J. X. Wang, and F. Hill. (2022). "Can language models learn from explanations in context?" *arXiv preprint arXiv:2204.02329*.

Lee, J.-S. and D. Zhu. (2012). "Shilling attack detection—a new approach for a trustworthy recommender system". *INFORMS Journal on Computing*. 24(1): 117–131.

Lepikhin, D., H. Lee, Y. Xu, D. Chen, O. Firat, Y. Huang, M. Krikun, N. Shazeer, and Z. Chen. (2020). "Gshard: Scaling giant models with conditional computation and automatic sharding". *arXiv preprint arXiv:2006.16668*.

Lermen, S., C. Rogers-Smith, and J. Ladish. (2023). "LoRA Fine-tuning Efficiently Undoes Safety Training in Llama 2-Chat 70B". arXiv: 2310.20624 [cs.LG].

Lewis, P., E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, *et al.* (2020). "Retrieval-augmented generation for knowledge-intensive nlp tasks". *Advances in Neural Information Processing Systems*. 33: 9459–9474.

Li, B., Y. Wang, A. Singh, and Y. Vorobeychik. (2016). "Data poisoning attacks on factorization-based collaborative filtering". *Advances in neural information processing systems*. 29.

Li, H., Q. Chen, H. Zhu, D. Ma, H. Wen, and X. S. Shen. (2017). "Privacy leakage via de-anonymization and aggregation in heterogeneous social networks". *IEEE Transactions on Dependable and Secure Computing*. 17(2): 350–362.

Li, J., W. Zhang, T. Wang, G. Xiong, A. Lu, and G. Medioni. (2023a). "GPT4Rec: A generative framework for personalized recommendation and user interests interpretation". *arXiv preprint arXiv:2304.03879*.

Li, J., S. Wang, M. Zhang, W. Li, Y. Lai, X. Kang, W. Ma, and Y. Liu. (2024a). "Agent hospital: A simulacrum of hospital with evolvable medical agents". *arXiv preprint arXiv:2405.02957*.

Li, K., A. K. Hopkins, D. Bau, F. Viégas, H. Pfister, and M. Wattenberg. (2022a). "Emergent world representations: Exploring a sequence model trained on a synthetic task". *arXiv preprint arXiv:2210.13382*.

Li, K., O. Patel, F. Viégas, H. Pfister, and M. Wattenberg. (2024b). "Inference-time intervention: Eliciting truthful answers from a language model". *Advances in Neural Information Processing Systems*. 36.

Li, L., Y. Zhang, and L. Chen. (2021a). "Personalized Transformer for Explainable Recommendation". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 4947–4957.

Li, L., L. Fan, S. Atreja, and L. Hemphill. (2024c). ""HOT" ChatGPT: The Promise of ChatGPT in Detecting and Discriminating Hateful, Offensive, and Toxic Comments on Social Media". *ACM Trans. Web*. 18(2). DOI: 10.1145/3643829.

Li, M., Y. Zhao, B. Yu, F. Song, H. Li, H. Yu, Z. Li, F. Huang, and Y. Li. (2023b). "Api-bank: A comprehensive benchmark for tool-augmented llms". *arXiv preprint arXiv:2304.08244*.

Li, N., T. Li, and S. Venkatasubramanian. (2007). "t-closeness: Privacy beyond k-anonymity and l-diversity". In: *2007 IEEE 23rd international conference on data engineering*. IEEE. 106–115.

Li, T. and T. Unger. (2012). "Willing to pay for quality personalization? Trade-off between quality and privacy". *European Journal of Information Systems*. 21(6): 621–642.

Li, X., Z. Zhou, J. Zhu, J. Yao, T. Liu, and B. Han. (2023c). "DeepInception: Hypnotize Large Language Model to Be Jailbreaker". arXiv: 2311.03191 [cs.LG].

Li, Y., M. Du, R. Song, X. Wang, and Y. Wang. (2023d). "A survey on fairness in large language models". *arXiv preprint arXiv:2308.10149*.

Li, Y., H. Chen, Z. Fu, Y. Ge, and Y. Zhang. (2021b). "User-oriented Fairness in Recommendation". In: *Proceedings of the World Wide Web Conference 2021*. 624–632.

Li, Y., H. Chen, S. Xu, Y. Ge, J. Tan, S. Liu, and Y. Zhang. (2022b). "Fairness in Recommendation: A Survey". *arXiv preprint arXiv:2205.13619*.

Li, Y., L. Zhang, and Y. Zhang. (2023e). "Fairness of chatgpt". *arXiv preprint arXiv:2305.18569*.

Lian, J., Y. Lei, X. Huang, J. Yao, W. Xu, and X. Xie. (2024). "RecAI: Leveraging Large Language Models for Next-Generation Recommender Systems". In: *Companion Proceedings of the ACM on Web Conference 2024*. 1031–1034.

Liang, P. P., C. Wu, L.-P. Morency, and R. Salakhutdinov. (2021). "Towards understanding and mitigating social biases in language models". In: *International Conference on Machine Learning*. PMLR. 6565–6576.

Liao, J., S. Li, Z. Yang, J. Wu, Y. Yuan, and X. Wang. (2023). "Llara: Aligning large language models with sequential recommenders". *CoRR*.

Liao, Z., L. Mo, C. Xu, M. Kang, J. Zhang, C. Xiao, Y. Tian, B. Li, and H. Sun. (2024). "Eia: Environmental injection attack on generalist web agents for privacy leakage". *arXiv preprint arXiv:2409.11295*.

Lin, C., S. Chen, H. Li, Y. Xiao, L. Li, and Q. Yang. (2020). "Attacking recommender systems with augmented user profiles". In: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management (CIKM)*. 855–864.

Lin, C.-Y. (2004). "Rouge: A package for automatic evaluation of summaries". In: *Text summarization branches out*. 74–81.

Lin, F., X. Zhu, Z. Zhao, D. Huang, Y. Yu, X. Li, T. Xu, and E. Chen. (2024a). "Knowledge Graph Pruning for Recommendation". *arXiv preprint arXiv:2405.11531*.

Lin, G., W. Hua, and Y. Zhang. (2024b). "Promptcrypt: Prompt encryption for secure communication with large language models". *arXiv preprint arXiv:2402.05868*.

Lin, S., J. Hilton, and O. Evans. (2021). "Truthfulqa: Measuring how models mimic human falsehoods". *arXiv preprint arXiv:2109.07958*.

Lin, X., W. Wang, Y. Li, F. Feng, S.-K. Ng, and T.-S. Chua. (2024c). "Bridging items and language: A transition paradigm for large language model-based recommendation". In: *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1816–1826.

Liu, D., P. Cheng, Z. Dong, X. He, W. Pan, and Z. Ming. (2020). "A general knowledge distillation framework for counterfactual recommendation via uniform data". In: *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*. 831–840.

Liu, J., Y. Zhu, S. Wang, X. Wei, E. Min, Y. Lu, S. Wang, D. Yin, and Z. Dou. (2024a). "LLMs+ Persona-Plug= Personalized LLMs". *arXiv preprint arXiv:2409.11901*.

Liu, J., C. Liu, P. Zhou, R. Lv, K. Zhou, and Y. Zhang. (2023a). "Is chatgpt a good recommender? a preliminary study". *arXiv preprint arXiv:2304.10149*.

Liu, L., X. Yang, Y. Shen, B. Hu, Z. Zhang, J. Gu, and G. Zhang. (2023b). "Think-in-memory: Recalling and post-thinking enable llms with long-term memory". *arXiv preprint arXiv:2311.08719*.

Liu, Q., Y. Zeng, R. Mokhosi, and H. Zhang. (2018). "STAMP: short-term attention/memory priority model for session-based recommendation". In: *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 1831–1839.

Liu, Q., N. Chen, T. Sakai, and X.-M. Wu. (2023c). "A first look at llm-powered generative news recommendation". *arXiv preprint arXiv:2305.06566*.

Liu, X., Z. Peng, X. Yi, X. Xie, L. Xiang, Y. Liu, and D. Xu. (2024b). "ToolNet: Connecting large language models with massive tools via tool graph". *arXiv preprint arXiv:2403.00839*.

Liu, Z., Y. Chen, J. Li, P. S. Yu, J. McAuley, and C. Xiong. (2021). "Contrastive self-supervised sequential recommendation with robust augmentation". *arXiv preprint arXiv:2108.06479*.

Liu, Z., L. Yang, Z. Fan, H. Peng, and P. S. Yu. (2022). "Federated social recommendation with graph neural network". *ACM Transactions on Intelligent Systems and Technology (TIST)*. 13(4): 1–24.

Liu, Z., W. Yao, J. Zhang, R. Murthy, L. Yang, Z. Liu, T. Lan, M. Zhu, J. Tan, S. Kokane, *et al.* (2024c). "PRACT: Optimizing Principled Reasoning and Acting of LLM Agent". *arXiv preprint arXiv:2410.18528*.

Liu, Z., W. Yao, J. Zhang, L. Xue, S. Heinecke, R. Murthy, Y. Feng, Z. Chen, J. C. Niebles, D. Arpit, *et al.* (2023d). "Bolaa: Benchmarking and orchestrating llm-augmented autonomous agents". *arXiv preprint arXiv:2308.05960*.

Liu, Z., W. Yao, J. Zhang, L. Yang, Z. Liu, J. Tan, P. K. Choubey, T. Lan, J. Wu, H. Wang, *et al.* (2024d). "AgentLite: A Lightweight Library for Building and Advancing Task-Oriented LLM Agent System". *arXiv preprint arXiv:2402.15538*.

Liu, Z., T. Hoang, J. Zhang, M. Zhu, T. Lan, S. Kokane, J. Tan, W. Yao, Z. Liu, Y. Feng, *et al.* (2024e). "Apigen: Automated pipeline for generating verifiable and diverse function-calling datasets". *arXiv preprint arXiv:2406.18518*.

Lundberg, S. (2017). "A unified approach to interpreting model predictions". *arXiv preprint arXiv:1705.07874*.

Luo, P., X. Zhu, T. Xu, Y. Zheng, and E. Chen. (2024). "Semantic Interaction Matching Network for Few-Shot Knowledge Graph Completion". *ACM Trans. Web.* 18(2). DOI: 10.1145/3589557.

Ma, Y., Z. Gou, J. Hao, R. Xu, S. Wang, L. Pan, Y. Yang, Y. Cao, and A. Sun. (2024). "SciAgent: Tool-augmented Language Models for Scientific Reasoning". *arXiv preprint arXiv:2402.11451*.

Machanavajjhala, A., D. Kifer, J. Gehrke, and M. Venkitasubramaniam. (2007). "l-diversity: Privacy beyond k-anonymity". *ACM Transactions on Knowledge Discovery from Data (TKDD)*. 1(1): 3–es.

Magister, L. C., J. Mallinson, J. Adamek, E. Malmi, and A. Severyn. (2022). "Teaching small language models to reason". *arXiv preprint arXiv:2212.08410*.

Mao, J., C. Gan, P. Kohli, J. B. Tenenbaum, and J. Wu. (2019). "The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision". *arXiv preprint arXiv:1904.12584*.

Marlin, B., R. S. Zemel, S. Roweis, and M. Slaney. (2012). "Collaborative filtering and the missing at random assumption". *arXiv preprint arXiv:1206.5267*.

Mazeika, M., L. Phan, X. Yin, A. Zou, Z. Wang, N. Mu, E. Sakhaee, N. Li, S. Basart, B. Li, *et al.* (2024). "Harmbench: A standardized evaluation framework for automated red teaming and robust refusal". *arXiv preprint arXiv:2402.04249*.

Mehnaz, S., S. V. Dibbo, E. Kabir, N. Li, and E. Bertino. (2022). "Are Your Sensitive Attributes Private? Novel Model Inversion Attribute Inference Attacks on Classification Models". Aug.

Mehrabi, N., P. Goyal, C. Dupuy, Q. Hu, S. Ghosh, R. Zemel, K.-W. Chang, A. Galstyan, and R. Gupta. (2023). "FLIRT: Feedback Loop In-context Red Teaming". arXiv: 2308.04265 [cs.AI].

Mehrabi, N., F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. (2021). "A survey on bias and fairness in machine learning". *ACM Computing Surveys (CSUR)*. 54(6): 1–35.

Mehrotra, A., M. Zampetakis, P. Kassianik, B. Nelson, H. Anderson, Y. Singer, and A. Karbasi. (2023). "Tree of Attacks: Jailbreaking Black-Box LLMs Automatically". arXiv: 2312.02119 [cs.LG].

Mehta, B. (2007). "Unsupervised shilling detection for collaborative filtering". In: *AAAI*. 1402–1407.

Mehta, B. and W. Nejdl. (2009). "Unsupervised strategies for shilling detection and robust collaborative filtering". *User Modeling and User-Adapted Interaction*. 19(1): 65–97.

Mei, K., W. Xu, S. Lin, and Y. Zhang. (2025). "ECCOS: Efficient Capability and Cost Coordinated Scheduling for Multi-LLM Serving". *arXiv preprint arXiv:2502.20576*.

Mei, K. and Y. Zhang. (2023). "LightLM: a lightweight deep and narrow language model for generative recommendation". *arXiv preprint arXiv:2310.17488*.

Mei, K., X. Zhu, W. Xu, W. Hua, M. Jin, Z. Li, S. Xu, R. Ye, Y. Ge, and Y. Zhang. (2024). "AIOS: LLM agent operating system". *arXiv e-prints, pp. arXiv–2403*.

Mekala, D., J. Weston, J. Lanchantin, R. Raileanu, M. Lomeli, J. Shang, and J. Dwivedi-Yu. (2024). "TOOLVERIFIER: Generalization to New Tools via Self-Verification". *arXiv preprint arXiv:2402.14158*.

Meng, K., D. Bau, A. Andonian, and Y. Belinkov. (2022). "Locating and editing factual associations in GPT". *Advances in Neural Information Processing Systems*. 35: 17359–17372.

Mobasher, B., R. Burke, R. Bhaumik, and C. Williams. (2007). "Toward trustworthy recommender systems: An analysis of attack models and algorithm robustness". *ACM Transactions on Internet Technology (TOIT)*. 7(4): 23–es.

Modarressi, A., M. Fayyaz, E. Aghazadeh, Y. Yaghoobzadeh, and M. T. Pilehvar. (2023). "DecompX: Explaining transformers decisions by propagating token decomposition". *arXiv preprint arXiv:2306.02873*.

Modarressi, A., M. Fayyaz, Y. Yaghoobzadeh, and M. T. Pilehvar. (2022). "GlobEnc: Quantifying global token attribution by incorporating the whole encoder layer in transformers". *arXiv preprint arXiv:2205.03286*.

Mozes, M., X. He, B. Kleinberg, and L. D. Griffin. (2023). "Use of llms for illicit purposes: Threats, prevention measures, and vulnerabilities". *arXiv preprint arXiv:2308.12833*.

Muhammad, K., Q. Wang, D. O'Reilly-Morgan, E. Tragos, B. Smyth, N. Hurley, J. Geraci, and A. Lawlor. (2020). "Fedfast: Going beyond average for faster training of federated recommender systems". In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1234–1242.

Narayanan, A. and V. Shmatikov. (2008). “Robust de-anonymization of large sparse datasets”. In: *2008 IEEE Symposium on Security and Privacy (sp 2008).* IEEE. 111–125.

Narayanan, D., M. Shoeybi, J. Casper, P. LeGresley, M. Patwary, V. Korthikanti, D. Vainbrand, P. Kashinkunti, J. Bernauer, B. Catanzaro, *et al.* (2021). “Efficient large-scale language model training on gpu clusters using megatron-lm”. In: *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis.* 1–15.

Nikolaenko, V., S. Ioannidis, U. Weinsberg, M. Joye, N. Taft, and D. Boneh. (2013). “Privacy-preserving matrix factorization”. In: *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security.* 801–812.

Ning, X., W. Xu, X. Liu, M. Ha, Q. Ma, Y. Li, L. Chen, and Y. Zhang. (2024). “Information maximization via variational autoencoders for cross-domain recommendation”. *arXiv preprint arXiv:2405.20710.*

Nobata, C., J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang. (2016). “Abusive language detection in online user content”. In: *Proceedings of the 25th international conference on world wide web.* 145–153.

Nori, H., N. King, S. M. McKinney, D. Carignan, and E. Horvitz. (2023). “Capabilities of gpt-4 on medical challenge problems”. *arXiv preprint arXiv:2303.13375.*

Nye, M., A. J. Andreassen, G. Gur-Ari, H. Michalewski, J. Austin, D. Bieber, D. Dohan, A. Lewkowycz, M. Bosma, D. Luan, *et al.* (2021). “Show your work: Scratchpads for intermediate computation with language models”. *arXiv preprint arXiv:2112.00114.*

O’Brien, M. and M. T. Keane. (2006). “Modeling result-list searching in the World Wide Web: The role of relevance topologies and trust bias”. In: *Proceedings of the 28th annual conference of the cognitive science society.* Vol. 28. Citeseer. 1881–1886.

Ohm, P. (2009). “Broken promises of privacy: Responding to the surprising failure of anonymization”. *UCLA l. Rev.* 57: 1701.

Ooge, J., S. Kato, and K. Verbert. (2022). “Explaining Recommendations in E-Learning: Effects on Adolescents’ Trust”. In: *27th International Conference on Intelligent User Interfaces.* 93–105.

Ouyang, L., J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, *et al.* (2022). "Training language models to follow instructions with human feedback". *Advances in Neural Information Processing Systems*. 35: 27730–27744.

Ovaisi, Z., R. Ahsan, Y. Zhang, K. Vasilaky, and E. Zheleva. (2020). "Correcting for selection bias in learning-to-rank systems". In: *Proceedings of The Web Conference 2020*. 1863–1873.

Palato, M. (2021). "Federated Variational Autoencoder for Collaborative Filtering". In: *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 1–8.

Pan, S., D. Li, H. Gu, T. Lu, X. Luo, and N. Gu. (2022). "Accurate and Explainable Recommendation via Review Rationalization". In: *Proceedings of the World Wide Web Conference 2022*. 3092–3101.

Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu. (2002). "Bleu: a method for automatic evaluation of machine translation". In: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 311–318.

Paranjape, B., S. Lundberg, S. Singh, H. Hajishirzi, L. Zettlemoyer, and M. T. Ribeiro. (2023). "Art: Automatic multi-step reasoning and tool-use for large language models". *arXiv preprint arXiv:2303.09014*.

Park, J. S., J. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein. (2023). "Generative agents: Interactive simulacra of human behavior". In: *Proceedings of the 36th annual acm symposium on user interface software and technology*. 1–22.

Patro, G. K., A. Biswas, N. Ganguly, K. P. Gummadi, and A. Chakraborty. (2020). "Fairrec: Two-sided fairness for personalized recommendations in two-sided platforms". In: *Proceedings of the web conference 2020*. 1194–1204.

Peng, H., X. Wang, S. Hu, H. Jin, L. Hou, J. Li, Z. Liu, and Q. Liu. (2022). "Copen: Probing conceptual knowledge in pre-trained language models". *arXiv preprint arXiv:2211.04079*.

Perez, E., S. Huang, F. Song, T. Cai, R. Ring, J. Aslanides, A. Glaese, N. McAleese, and G. Irving. (2022a). "Red Teaming Language Models with Language Models". arXiv: 2202.03286 [cs.CL].

Perez, E., S. Ringer, K. Lukošiūtė, K. Nguyen, E. Chen, S. Heiner, C. Pettit, C. Olsson, S. Kundu, S. Kadavath, A. Jones, A. Chen, B. Mann, B. Israel, B. Seethor, C. McKinnon, C. Olah, D. Yan, D. Amodei, D. Amodei, D. Drain, D. Li, E. Tran-Johnson, G. Khundadze, J. Kernion, J. Landis, J. Kerr, J. Mueller, J. Hyun, J. Landau, K. Ndousse, L. Goldberg, L. Lovitt, M. Lucas, M. Sellitto, M. Zhang, N. Kingsland, N. Elhage, N. Joseph, N. Mercado, N. DasSarma, O. Rausch, R. Larson, S. McCandlish, S. Johnston, S. Kravec, S. E. Showk, T. Lanham, T. Telleen-Lawton, T. Brown, T. Henighan, T. Hume, Y. Bai, Z. Hatfield-Dodds, J. Clark, S. R. Bowman, A. Askell, R. Grosse, D. Hernandez, D. Ganguli, E. Hubinger, N. Schiefer, and J. Kaplan. (2022b). "Discovering Language Model Behaviors with Model-Written Evaluations". arXiv: 2212.09251 [cs.CL].

Perez, F. and I. Ribeiro. (2022). "Ignore Previous Prompt: Attack Techniques For Language Models". arXiv: 2211.09527 [cs.CL]. URL: https://arxiv.org/abs/2211.09527.

Petroni, F., T. Rocktäschel, P. Lewis, A. Bakhtin, Y. Wu, A. H. Miller, and S. Riedel. (2019). "Language models as knowledge bases?" *arXiv preprint arXiv:1909.01066*.

Phute, M., A. Helbling, M. Hull, S. Peng, S. Szyller, C. Cornelius, and D. H. Chau. (2023). "LLM Self Defense: By Self Examination, LLMs Know They Are Being Tricked". arXiv: 2308.07308 [cs.CL].

Polat, H. and W. Du. (2003). "Privacy-preserving collaborative filtering using randomized perturbation techniques". In: *Third IEEE International Conference on Data Mining*. IEEE. 625–628.

Porat, T., R. Nyrup, R. A. Calvo, P. Paudyal, and E. Ford. (2020). "Public health and risk communication during COVID-19—enhancing psychological needs to promote sustainable behavior change". *Frontiers in public health*: 637.

Press, O., M. Zhang, S. Min, L. Schmidt, N. A. Smith, and M. Lewis. (2022). "Measuring and narrowing the compositionality gap in language models". *arXiv preprint arXiv:2210.03350*.

Pu, P. and L. Chen. (2006). "Trust building with explanation interfaces". In: *Proceedings of the 11th international conference on Intelligent user interfaces*. 93–100.

Qiao, S., H. Gui, C. Lv, Q. Jia, H. Chen, and N. Zhang. (2024). “Making Language Models Better Tool Learners with Execution Feedback”. In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 3550–3568.

Qin, Y., S. Liang, Y. Ye, K. Zhu, L. Yan, Y. Lu, Y. Lin, X. Cong, X. Tang, B. Qian, *et al.* (2023). “Toolllm: Facilitating large language models to master 16000+ real-world apis”. *arXiv preprint arXiv:2307.16789*.

Qiu, H., S. Zhang, A. Li, H. He, and Z. Lan. (2023). “Latent Jailbreak: A Benchmark for Evaluating Text Safety and Output Robustness of Large Language Models”. arXiv: 2307.08487 [cs.CL].

Qiu, X., T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang. (2020). “Pretrained models for natural language processing: A survey”. *Science China technological sciences*. 63(10): 1872–1897.

Qu, C., S. Dai, X. Wei, H. Cai, S. Wang, D. Yin, J. Xu, and J.-R. Wen. (2024). “COLT: Towards Completeness-Oriented Tool Retrieval for Large Language Models”. *arXiv preprint arXiv:2405.16089*.

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.* (2019). “Language models are unsupervised multitask learners”. *OpenAI blog*. 1(8): 9.

Rae, J. W., S. Borgeaud, T. Cai, K. Millican, J. Hoffmann, F. Song, J. Aslanides, S. Henderson, R. Ring, S. Young, *et al.* (2021). “Scaling language models: Methods, analysis & insights from training gopher”. *arXiv preprint arXiv:2112.11446*.

Rae, J. W., A. Potapenko, S. M. Jayakumar, and T. P. Lillicrap. (2019). “Compressive transformers for long-range sequence modelling”. *arXiv preprint arXiv:1911.05507*.

Rafailov, R., A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn. (2023). “Direct preference optimization: Your language model is secretly a reward model”. *arXiv preprint arXiv:2305.18290*.

Rastegarpanah, B., K. P. Gummadi, and M. Crovella. (2019). “Fighting fire with fire: Using antidote data to improve polarization and fairness of recommender systems”. In: *Proceedings of the twelfth ACM international conference on web search and data mining*. 231–239.

Razeghi, Y., R. L. Logan IV, M. Gardner, and S. Singh. (2022). "Impact of pretraining term frequencies on few-shot reasoning". *arXiv preprint arXiv:2202.07206*.

Reimers, N. (2019). "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". *arXiv preprint arXiv:1908.10084*.

Ribeiro, M. T., S. Singh, and C. Guestrin. (2016). "" Why should i trust you?" Explaining the predictions of any classifier". In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

Richardson, C., Y. Zhang, K. Gillespie, S. Kar, A. Singh, Z. Raeesy, O. Z. Khan, and A. Sethy. (2023). "Integrating summarization and retrieval for enhanced personalization via large language models". *arXiv preprint arXiv:2310.20081*.

Robertson, S., H. Zaragoza, *et al.* (2009). "The probabilistic relevance framework: BM25 and beyond". *Foundations and Trends® in Information Retrieval*. 3(4): 333–389.

Salecha, A., M. E. Ireland, S. Subrahmanya, J. Sedoc, L. H. Ungar, and J. C. Eichstaedt. (2024). "Large Language Models Show Human-like Social Desirability Biases in Survey Responses". *arXiv preprint arXiv:2405.06058*.

Salemi, A., S. Kallumadi, and H. Zamani. (2024). "Optimization methods for personalizing large language models through retrieval augmentation". In: *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 752–762.

Salemi, A., S. Mysore, M. Bendersky, and H. Zamani. (2023). "Lamp: When large language models meet personalization". *arXiv preprint arXiv:2304.11406*.

Sanner, S., K. Balog, F. Radlinski, B. Wedin, and L. Dixon. (2023). "Large language models are competitive near cold-start recommenders for language-and item-based preferences". In: *Proceedings of the 17th ACM conference on recommender systems*. 890–896.

Santoro, A., S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap. (2016). "Meta-learning with memory-augmented neural networks". In: *International conference on machine learning*. PMLR. 1842–1850.

Santy, S., J. Liang, R. Le Bras, K. Reinecke, and M. Sap. (2023). "NLPositionality: Characterizing Design Biases of Datasets and Models". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by A. Rogers, J. Boyd-Graber, and N. Okazaki. Toronto, Canada: Association for Computational Linguistics. 9080–9102. DOI: 10.18653/v1/2023.acl-long.505.

Sardianos, C., I. Varlamis, C. Chronis, G. Dimitrakopoulos, A. Alsalemi, Y. Himeur, F. Bensaali, and A. Amira. (2021). "The emergence of explainability of intelligent systems: Delivering explainable and personalized recommendations for energy efficiency". *International Journal of Intelligent Systems*. 36(2): 656–680.

Schick, T., J. Dwivedi-Yu, R. Dessı`, R. Raileanu, M. Lomeli, E. Hambro, L. Zettlemoyer, N. Cancedda, and T. Scialom. (2024). "Toolformer: Language models can teach themselves to use tools". *Advances in Neural Information Processing Systems*. 36.

Shah, D. S., H. A. Schwartz, and D. Hovy. (2020). "Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault. Online: Association for Computational Linguistics. 5248–5264. DOI: 10.18653/v1/2020.acl-main.468.

Shah, R., Q. Feuillade–Montixi, S. Pour, A. Tagade, S. Casper, and J. Rando. (2023). "Scalable and Transferable Black-Box Jailbreaks for Language Models via Persona Modulation". arXiv: 2311.03348 [cs.CL].

Shahrasbi, B., V. Mani, A. R. Arrabothu, D. Sharma, K. Achan, and S. Kumar. (2020). "On detecting data pollution attacks on recommender systems using sequential gans". *arXiv preprint arXiv:2012.02509*.

Shanahan, M., K. McDonell, and L. Reynolds. (2023). "Role play with large language models". *Nature*. 623(7987): 493–498.

Sharma, A. and D. Cosley. (2013). "Do social explanations work? Studying and modeling the effects of social explanations in recommender systems". In: *Proceedings of the World Wide Web Conference*. 1133–1144.

Shen, X., Z. Chen, M. Backes, Y. Shen, and Y. Zhang. (2024a). “"Do Anything Now": Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models”. arXiv: 2308.03825 [cs.CR]. URL: https://arxiv.org/abs/2308.03825.

Shen, Y., K. Song, X. Tan, D. Li, W. Lu, and Y. Zhuang. (2024b). “Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face”. *Advances in Neural Information Processing Systems*. 36.

Shi, S., H. Chen, W. Ma, J. Mao, M. Zhang, and Y. Zhang. (2020). “Neural logic reasoning”. In: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management (CIKM)*. 1365–1374.

Shi, W., X. He, Y. Zhang, C. Gao, X. Li, J. Zhang, Q. Wang, and F. Feng. (2024a). “Large language models are learnable planners for long-term recommendation”. In: *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1893–1903.

Shi, Y., W. Xu, M. Jin, H. Zhang, Q. Wu, Y. Zhang, and M. Xu. (2024b). “Beyond KAN: Introducing KarSein for Adaptive High-Order Feature Interaction Modeling in CTR Prediction”. *arXiv preprint arXiv:2408.08713*.

Shi, Y., W. Xu, Z. Zhang, X. Zi, Q. Wu, and M. Xu. (2025). “PersonaX: A Recommendation Agent Oriented User Modeling Framework for Long Behavior Sequence”. *arXiv preprint arXiv:2503.02398*.

Shi, Z., K. Mei, M. Jin, Y. Su, C. Zuo, W. Hua, W. Xu, Y. Ren, Z. Liu, M. Du, *et al.* (2024c). “From Commands to Prompts: LLM-based Semantic File System for AIOS”. *arXiv preprint arXiv:2410.11843*.

Shin, T., Y. Razeghi, R. L. L. I. au2, E. Wallace, and S. Singh. (2020). “AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts”. arXiv: 2010.15980 [cs.CL].

Shinn, N., F. Cassano, A. Gopinath, K. Narasimhan, and S. Yao. (2024). “Reflexion: Language agents with verbal reinforcement learning”. *Advances in Neural Information Processing Systems*. 36.

Shokri, R., M. Stronati, C. Song, and V. Shmatikov. (2017). “Membership inference attacks against machine learning models”. In: *2017 IEEE symposium on security and privacy (SP)*. IEEE. 3–18.

Shu, Y., H. Zhang, H. Gu, P. Zhang, T. Lu, D. Li, and N. Gu. (2024). "RAH! RecSys–Assistant–Human: A Human-Centered Recommendation Framework With LLM Agents". *IEEE Transactions on Computational Social Systems*.

Sikdar, S., P. Bhattacharya, and K. Heese. (2021). "Integrated directional gradients: Feature interaction attribution for neural NLP models". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 865–878.

Singh, A. and T. Joachims. (2018). "Fairness of Exposure in Rankings". In: *Proceedings of the 24th ACM SIGKDD*. London, United Kingdom.

Sobitha Ahila, S. and K. Shunmuganathan. (2016). "Role of agent technology in web usage mining: homomorphic encryption based recommendation for e-commerce applications". *Wireless Personal Communications*. 87(2): 499–512.

Song, J., Z. Li, Z. Hu, Y. Wu, Z. Li, J. Li, and J. Gao. (2020). "Poisonrec: an adaptive data poisoning framework for attacking black-box recommender systems". In: *2020 IEEE 36th International Conference on Data Engineering (ICDE)*. IEEE. 157–168.

Song, Y., W. Xiong, D. Zhu, W. Wu, H. Qian, M. Song, H. Huang, C. Li, K. Wang, R. Yao, *et al.* (2023). "RestGPT: Connecting Large Language Models with Real-World RESTful APIs". *arXiv preprint arXiv:2306.06624*.

Sood, S. O., E. F. Churchill, and J. Antin. (2012). "Automatic identification of personal insults on social news sites". *Journal of the American Society for Information Science and Technology*. 63(2): 270–285.

Sparck Jones, K. (1972). "A statistical interpretation of term specificity and its application in retrieval". *Journal of documentation*. 28(1): 11–21.

Stiennon, N., L. Ouyang, J. Wu, D. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. F. Christiano. (2020). "Learning to summarize with human feedback". *Advances in Neural Information Processing Systems*. 33: 3008–3021.

Sun, F., J. Liu, J. Wu, C. Pei, X. Lin, W. Ou, and P. Jiang. (2019a). "BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer". In: *Proceedings of the 28th ACM international conference on information and knowledge management (CIKM)*. 1441–1450.

Sun, T., A. Gaut, S. Tang, Y. Huang, M. ElSherief, J. Zhao, D. Mirza, E. Belding, K.-W. Chang, and W. Y. Wang. (2019b). "Mitigating Gender Bias in Natural Language Processing: Literature Review". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 1630–1640.

Sundararajan, M., A. Taly, and Q. Yan. (2017). "Axiomatic attribution for deep networks". In: *International conference on machine learning*. PMLR. 3319–3328.

Sweeney, L. (2002). "k-anonymity: A model for protecting privacy". *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*. 10(05): 557–570.

Takami, K., Y. Dai, B. Flanagan, and H. Ogata. (2022). "Educational Explainable Recommender Usage and its Effectiveness in High School Summer Vacation Assignment". In: *LAK22: 12th International Learning Analytics and Knowledge Conference*. 458–464.

Tan, J., S. Geng, Z. Fu, Y. Ge, S. Xu, Y. Li, and Y. Zhang. (2022). "Learning and Evaluating Graph Neural Network Explanations based on Counterfactual and Factual Reasoning". In: *Proceedings of the World Wide Web Conference 2022*. 1018–1027.

Tan, J., S. Xu, Y. Ge, Y. Li, X. Chen, and Y. Zhang. (2021). "Counterfactual explainable recommendation". In: *Proceedings of the 30th ACM International Conference on Information & Knowledge Management (CIKM)*. 1784–1793.

Tan, Z., Z. Liu, and M. Jiang. (2024a). "Personalized Pieces: Efficient Personalized Large Language Models through Collaborative Efforts". *arXiv preprint arXiv:2406.10471*.

Tan, Z., Q. Zeng, Y. Tian, Z. Liu, B. Yin, and M. Jiang. (2024b). "Democratizing large language models via personalized parameter-efficient fine-tuning". *arXiv preprint arXiv:2402.04401*.

Tang, H., L. Cheng, N. Liu, and M. Du. (2023a). “A Theoretical Approach to Characterize the Accuracy-Fairness Trade-off Pareto Frontier”. *arXiv preprint arXiv:2310.12785*.

Tang, H., C. Zhang, M. Jin, Q. Yu, Z. Wang, X. Jin, Y. Zhang, and M. Du. (2024). “Time series forecasting with llms: Understanding and enhancing model capabilities”. *arXiv preprint arXiv:2402.10835*.

Tang, J., H. Wen, and K. Wang. (2020). “Revisiting adversarially learned injection attacks against recommender systems”. In: *Proceedings of the 14th ACM Conference on Recommender Systems*. 318–327.

Tang, J., X. Du, X. He, F. Yuan, Q. Tian, and T.-S. Chua. (2019). “Adversarial training towards robust multimedia recommender system”. *IEEE Transactions on Knowledge and Data Engineering*. 32(5): 855–867.

Tang, Q., Z. Deng, H. Lin, X. Han, Q. Liang, B. Cao, and L. Sun. (2023b). “Toolalpaca: Generalized tool learning for language models with 3000 simulated cases”. *arXiv preprint arXiv:2306.05301*.

Team, G., R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, *et al.* (2023). “Gemini: a family of highly capable multimodal models”. *arXiv preprint arXiv:2312.11805*.

Toroghi, A., W. Guo, M. M. A. Pour, and S. Sanner. (2024). “Right for Right Reasons: Large Language Models for Verifiable Commonsense Knowledge Graph Question Answering”. *arXiv preprint arXiv:2403.01390*.

Touvron, H., T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, *et al.* (2023a). “Llama: Open and efficient foundation language models”. *arXiv preprint arXiv:2302.13971*.

Touvron, H., L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom. (2023b). "Llama 2: Open Foundation and Fine-Tuned Chat Models". arXiv: 2307.09288 [cs.CL]. URL: https://arxiv.org/abs/2307.09288.

Tran, K. H., A. Ghazimatin, and R. Saha Roy. (2021). "Counterfactual Explanations for Neural Recommenders". In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1627–1631.

Tu, Q., S. Fan, Z. Tian, and R. Yan. (2024). "Charactereval: A chinese benchmark for role-playing conversational agent evaluation". *arXiv preprint arXiv:2401.01275*.

Umemoto, K., T. Milo, and M. Kitsuregawa. (2020). "Toward recommendation for upskilling: Modeling skill improvement and item difficulty in action sequences". In: *2020 IEEE 36th International Conference on Data Engineering (ICDE)*. IEEE. 169–180.

Vasile, F., E. Smirnova, and A. Conneau. (2016). "Meta-prod2vec: Product embeddings using side-information for recommendation". In: *Proceedings of the 10th ACM conference on recommender systems (RecSys)*. 225–232.

Vaswani, A. (2017). "Attention is all you need". *Advances in Neural Information Processing Systems*.

Voigt, P. and A. Von dem Bussche. (2017). "The eu general data protection regulation (gdpr)". *A Practical Guide, 1st Ed., Cham: Springer International Publishing*. 10(3152676): 10–5555.

Wallace, E., P. Rodriguez, S. Feng, I. Yamada, and J. Boyd-Graber. (2019). "Trick Me If You Can: Human-in-the-loop Generation of Adversarial Examples for Question Answering". arXiv: 1809.02701 [cs.CL].

Wang, C., L. Yang, Z. Liu, X. Liu, M. Yang, Y. Liang, and P. S. Yu. (2024a). "Collaborative Alignment for Recommendation". In: *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 2315–2325.

Wang, J. and P. Han. (2019). "Adversarial training-based mean Bayesian personalized ranking for recommender system". *IEEE Access*. 8: 7958–7968.

Wang, K., A. Variengien, A. Conmy, B. Shlegeris, and J. Steinhardt. (2022a). "Interpretability in the wild: a circuit for indirect object identification in gpt-2 small". *arXiv preprint arXiv:2211.00593*.

Wang, L. and E.-P. Lim. (2023). "Zero-shot next-item recommendation using large pretrained language models". *arXiv preprint arXiv:2304.03153*.

Wang, L., C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin, *et al.* (2024b). "A survey on large language model based autonomous agents". *Frontiers of Computer Science*. 18(6): 186345.

Wang, L., J. Zhang, H. Yang, Z. Chen, J. Tang, Z. Zhang, X. Chen, Y. Lin, R. Song, W. X. Zhao, *et al.* (2023a). "User behavior simulation with large language model based agents". *arXiv preprint arXiv:2306.02552*.

Wang, L. and H. Zhong. (2024). "LLM-SAP: Large Language Models Situational Awareness-Based Planning". In: *2024 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*. IEEE. 1–6.

Wang, P., D. Zhang, L. Li, C. Tan, X. Wang, K. Ren, B. Jiang, and X. Qiu. (2024c). "Inferaligner: Inference-time alignment for harmlessness through cross-model guidance". *arXiv preprint arXiv:2401.11206*.

Wang, S., L. Hu, L. Cao, X. Huang, D. Lian, and W. Liu. (2018a). "Attention-based transactional context embedding for next-item recommendation". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. No. 1.

Wang, X., X. He, Y. Cao, M. Liu, and T.-S. Chua. (2019a). "Kgat: Knowledge graph attention network for recommendation". In: *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 950–958.

Wang, X., X. He, F. Feng, L. Nie, and T.-S. Chua. (2018b). "Tem: Tree-enhanced embedding model for explainable recommendation". In: *Proceedings of the 2018 World Wide Web Conference*. 1543–1552.

Wang, X., D. Wang, C. Xu, X. He, Y. Cao, and T.-S. Chua. (2019b). "Explainable reasoning over knowledge graphs for recommendation". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. No. 01. 5329–5336.

Wang, X., K. Zhou, J.-R. Wen, and W. X. Zhao. (2022b). "Towards unified conversational recommender systems via knowledge-enhanced prompt learning". In: *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1929–1937.

Wang, X., Y. Chen, J. Yang, L. Wu, Z. Wu, and X. Xie. (2018c). "A reinforcement learning framework for explainable recommendation". In: *2018 IEEE international conference on data mining (ICDM)*. IEEE. 587–596.

Wang, X., J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou. (2023b). "Self-Consistency Improves Chain of Thought Reasoning in Language Models". arXiv: 2203.11171 [cs.CL]. URL: https://arxiv.org/abs/2203.11171.

Wang, Y., Z. Jiang, Z. Chen, F. Yang, Y. Zhou, E. Cho, X. Fan, X. Huang, Y. Lu, and Y. Yang. (2023c). "Recmind: Large language model powered agent for recommendation". *arXiv preprint arXiv:2308.14296*.

Wang, Y., Z. Liu, J. Zhang, W. Yao, S. Heinecke, and P. S. Yu. (2023d). "Drdt: Dynamic reflection with divergent thinking for llm-based sequential recommendation". *arXiv preprint arXiv:2312.11336*.

Wang, Z., Y. Yu, W. Zheng, W. Ma, and M. Zhang. (2024d). "MACRec: A Multi-Agent Collaboration Framework for Recommendation". In: *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2760–2764.

Wang, Z., S. Mao, W. Wu, T. Ge, F. Wei, and H. Ji. (2023e). "Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration". *arXiv preprint arXiv:2307.05300*.

Wang, Z., C. Chen, L. Lyu, D. N. Metaxas, and S. Ma. (2024e). "DIAGNOSIS: Detecting Unauthorized Data Usages in Text-to-image Diffusion Models". In: *The Twelfth International Conference on Learning Representations*. URL: https://openreview.net/forum?id=f8S3aLm0Vp.

Wang, Z., C. Chen, V. Sehwag, M. Pan, and L. Lyu. (2024f). "Evaluating and Mitigating IP Infringement in Visual Generative AI". *arXiv preprint arXiv:2406.04662*.

Wei, J., M. Bosma, V. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le. (2022a). "Finetuned Language Models are Zero-Shot Learners". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=gEZrGCozdqR.

Wei, J., X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, *et al.* (2022b). "Chain-of-thought prompting elicits reasoning in large language models". *Advances in neural information processing systems*. 35: 24824–24837.

Wei, W., X. Ren, J. Tang, Q. Wang, L. Su, S. Cheng, J. Wang, D. Yin, and C. Huang. (2024). "Llmrec: Large language models with graph augmentation for recommendation". In: *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*. 806–815.

Weikum, G., X. L. Dong, S. Razniewski, F. Suchanek, *et al.* (2021). "Machine knowledge: Creation and curation of comprehensive knowledge bases". *Foundations and Trends® in Databases*. 10(2-4): 108–490.

Weinsberg, U., S. Bhagat, S. Ioannidis, and N. Taft. (2012). "BlurMe: Inferring and obfuscating user gender based on ratings". In: *Proceedings of the sixth ACM conference on Recommender systems (RecSys)*. 195–202.

Williams, C. and B. Mobasher. (2006). “Profile injection attack detection for securing collaborative recommender systems”. *DePaul University CTI Technical Report*: 1–47.

Woźniak, S., B. Koptyra, A. Janz, P. Kazienko, and J. Kocoń. (2024). “Personalized large language models”. *arXiv preprint arXiv:2402.09269*.

Wu, C.-Y., A. Beutel, A. Ahmed, and A. J. Smola. (2016). “Explaining reviews and ratings with paco: poisson additive co-clustering”. In: *Proceedings of the World Wide Web Conference*.

Wu, C., D. Lian, Y. Ge, Z. Zhu, and E. Chen. (2021a). “Triple adversarial learning for influence based poisoning attack in recommender systems”. In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 1830–1840.

Wu, C., F. Wu, Y. Cao, Y. Huang, and X. Xie. (2021b). “Fedgnn: Federated graph neural network for privacy-preserving recommendation”. *arXiv preprint arXiv:2102.04925*.

Wu, C., F. Wu, T. Qi, J. Lian, Y. Huang, and X. Xie. (2020). “PTUM: Pre-training user model from unlabeled user behaviors via self-supervision”. *arXiv preprint arXiv:2010.01494*.

Wu, J., X. Wang, F. Feng, X. He, L. Chen, J. Lian, and X. Xie. (2021c). “Self-supervised graph learning for recommendation”. In: *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*. 726–735.

Wu, L., L. Chen, P. Shao, R. Hong, X. Wang, and M. Wang. (2021d). “Learning fair representations for recommendation: A graph-based perspective”. In: *Proceedings of the World Wide Web Conference 2021*. 2198–2208.

Wu, Y., R. Xie, Y. Zhu, F. Zhuang, X. Ao, X. Zhang, L. Lin, and Q. He. (2022). “Selective Fairness in Recommendation via Prompts”. *Proceddings of the SIGIR Conference*.

Wu, Y., R. Xie, Y. Zhu, F. Zhuang, X. Zhang, L. Lin, and Q. He. (2024a). “Personalized prompt for sequential recommendation”. *IEEE Transactions on Knowledge and Data Engineering*. 36(7): 3376–3389.

Wu, Z., A. Geiger, T. Icard, C. Potts, and N. Goodman. (2024b). “Interpretability at scale: Identifying causal mechanisms in alpaca”. *Advances in Neural Information Processing Systems*. 36.

Wulczyn, E., N. Thain, and L. Dixon. (2017). “Ex Machina: Personal Attacks Seen at Scale”. arXiv: 1610.08914 [cs.CL].

Xi, Y., W. Liu, J. Lin, X. Cai, H. Zhu, J. Zhu, B. Chen, R. Tang, W. Zhang, and Y. Yu. (2024). “Towards open-world recommendation with knowledge augmentation from large language models”. In: *Proceedings of the 18th ACM Conference on Recommender Systems.* 12–22.

Xian, Y., Z. Fu, Q. Huang, S. Muthukrishnan, and Y. Zhang. (2020a). “Neural-symbolic reasoning over knowledge graph for multi-stage explainable recommendation”. *arXiv preprint arXiv:2007.13207.*

Xian, Y., Z. Fu, S. Muthukrishnan, G. De Melo, and Y. Zhang. (2019). “Reinforcement knowledge graph reasoning for explainable recommendation”. In: *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval.* 285–294.

Xian, Y., Z. Fu, H. Zhao, Y. Ge, X. Chen, Q. Huang, S. Geng, Z. Qin, G. De Melo, S. Muthukrishnan, and Y. Zhang. (2020b). “CAFE: Coarse-to-fine neural symbolic reasoning for explainable recommendation”. In: *Proceedings of the ACM International Conference on Information & Knowledge Management (CIKM).*

Xian, Y., T. Zhao, J. Li, J. Chan, A. Kan, J. Ma, X. L. Dong, C. Faloutsos, G. Karypis, S. Muthukrishnan, and Y. Zhang. (2021). “Ex3: Explainable attribute-aware item-set recommendations”. In: *Fifteenth ACM Conference on Recommender Systems (RecSys).* 484–494.

Xiao, G., J. Tang, J. Zuo, J. Guo, S. Yang, H. Tang, Y. Fu, and S. Han. (2024). “DuoAttention: Efficient Long-Context LLM Inference with Retrieval and Streaming Heads”. *arXiv preprint arXiv:2410.10819.*

Xiao, G., Y. Tian, B. Chen, S. Han, and M. Lewis. (2023). “Efficient streaming language models with attention sinks”. *arXiv preprint arXiv:2309.17453.*

Xiong, L., C. Xiong, Y. Li, K.-F. Tang, J. Liu, P. Bennett, J. Ahmed, and A. Overwijk. (2020). “Approximate nearest neighbor negative contrastive learning for dense text retrieval”. *arXiv preprint arXiv:2007.00808.*

Xu, J., M. D. Ma, F. Wang, C. Xiao, and M. Chen. (2023a). "Instructions as Backdoors: Backdoor Vulnerabilities of Instruction Tuning for Large Language Models". arXiv: 2305.14710 [cs.CL].

Xu, Q., Y. Li, H. Xia, and W. Li. (2024a). "Enhancing Tool Retrieval with Iterative Feedback from Large Language Models". *arXiv preprint arXiv:2406.17465*.

Xu, S., W. Hua, and Y. Zhang. (2023b). "OpenP5: Benchmarking Foundation Models for Recommendation". *arXiv:2306.11134*.

Xu, S., Y. Li, S. Liu, Z. Fu, Y. Ge, X. Chen, and Y. Zhang. (2021). "Learning causal explanations for recommendation". In: *The 1st International Workshop on Causality in Search and Recommendation*.

Xu, W., S. Li, M. Ha, X. Guo, Q. Ma, X. Liu, L. Chen, and Z. Zhu. (2023c). "Neural node matching for multi-target cross domain recommendation". In: *2023 IEEE 39th International Conference on Data Engineering (ICDE)*. IEEE. 2154–2166.

Xu, W., Z. Liang, J. Han, X. Ning, W. Lin, L. Chen, F. Wei, and Y. Zhang. (2024b). "SLMRec: Empowering Small Language Models for Sequential Recommendation". *arXiv preprint arXiv:2405.17890*.

Xu, W., X. Ning, W. Lin, M. Ha, Q. Ma, Q. Liang, X. Tao, L. Chen, B. Han, and M. Luo. (2024c). "Towards open-world cross-domain sequential recommendation: A model-agnostic contrastive denoising approach". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 161–179.

Xu, W., Y. Shi, Z. Liang, X. Ning, K. Mei, K. Wang, X. Zhu, M. Xu, and Y. Zhang. (2025). "Instructagent: Building user controllable recommender via llm agent". *arXiv preprint arXiv:2502.14662*.

Xu, W., Q. Wu, R. Wang, M. Ha, Q. Ma, L. Chen, B. Han, and J. Yan. (2024d). "Rethinking cross-domain sequential recommendation under open-world assumptions". In: *Proceedings of the ACM on Web Conference 2024*. 3173–3184.

Xue, H. and F. D. Salim. (2023). "Promptcast: A new prompt-based learning paradigm for time series forecasting". *IEEE Transactions on Knowledge and Data Engineering*.

Yang, C., X. Wang, Y. Lu, H. Liu, Q. V. Le, D. Zhou, and X. Chen. (2023a). "Large Language Models as Optimizers". arXiv: 2309.03409 [cs.LG].

Yang, F., Z. Chen, Z. Jiang, E. Cho, X. Huang, and Y. Lu. (2023b). "Palr: Personalization aware llms for recommendation". *arXiv preprint arXiv:2305.07622*.

Yang, R., L. Song, Y. Li, S. Zhao, Y. Ge, X. Li, and Y. Shan. (2024). "Gpt4tools: Teaching large language model to use tools via self-instruction". *Advances in Neural Information Processing Systems*. 36.

Yang, S., S. Huang, W. Zou, J. Zhang, X. Dai, and J. Chen. (2023c). "Local interpretation of transformer based on linear decomposition". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 10270–10287.

Yang, T. and Q. Ai. (2021). "Maximizing Marginal Fairness for Dynamic Learning to Rank". In: *Proceedings of the World Wide Web Conference 2021*. 137–145.

Yang, X., X. Wang, Q. Zhang, L. Petzold, W. Y. Wang, X. Zhao, and D. Lin. (2023d). "Shadow Alignment: The Ease of Subverting Safely-Aligned Language Models". arXiv: 2310.02949 [cs.CL].

Yao, S., D. Yu, J. Zhao, I. Shafran, T. Griffiths, Y. Cao, and K. Narasimhan. (2024a). "Tree of thoughts: Deliberate problem solving with large language models". *Advances in Neural Information Processing Systems*. 36.

Yao, S., J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao. (2023). "ReAct: Synergizing Reasoning and Acting in Language Models". In: *International Conference on Learning Representations (ICLR)*.

Yao, Y., J. Duan, K. Xu, Y. Cai, Z. Sun, and Y. Zhang. (2024b). "A survey on large language model (LLM) security and privacy: The Good, The Bad, and The Ugly". *High-Confidence Computing*. 4(2): 100211. DOI: 10.1016/j.hcc.2024.100211.

Ye, R., C. Zhang, R. Wang, S. Xu, and Y. Zhang. (2024). "Language is all a graph needs". In: *Findings of the Association for Computational Linguistics: EACL 2024*. 1955–1973.

Yoon, S.-e., Z. He, J. M. Echterhoff, and J. McAuley. (2024). "Evaluating Large Language Models as Generative User Simulators for Conversational Recommendation". *arXiv preprint arXiv:2403.09738*.

Yu, C., X. Liu, J. Maia, Y. Li, T. Cao, Y. Gao, Y. Song, R. Goutam, H. Zhang, B. Yin, *et al.* (2024). "COSMO: A large-scale e-commerce common sense knowledge generation and serving system at Amazon". In: *Companion of the 2024 International Conference on Management of Data*. 148–160.

Yu, J., X. Lin, and X. Xing. (2023). "Gptfuzzer: Red teaming large language models with auto-generated jailbreak prompts". *arXiv preprint arXiv:2309.10253*.

Yuan, F., L. Yao, and B. Benatallah. (2019). "Adversarial collaborative neural network for robust recommendation". In: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1065–1068.

Yuan, S., K. Song, J. Chen, X. Tan, D. Li, and D. Yang. (2024a). "EvoAgent: Towards Automatic Multi-Agent Generation via Evolutionary Algorithms". *arXiv preprint arXiv:2406.14228*.

Yuan, S., K. Song, J. Chen, X. Tan, Y. Shen, R. Kan, D. Li, and D. Yang. (2024b). "Easytool: Enhancing llm-based agents with concise tool instruction". *arXiv preprint arXiv:2401.06201*.

Yuan, Y., W. Jiao, W. Wang, J.-t. Huang, P. He, S. Shi, and Z. Tu. (2024c). "GPT-4 Is Too Smart To Be Safe: Stealthy Chat with LLMs via Cipher". arXiv: 2308.06463 [cs.CL]. URL: https://arxiv.org/abs/2308.06463.

Zellers, R., A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi. (2020). "Defending Against Neural Fake News". arXiv: 1905.12616 [cs.CL].

Zeng, Q., M. Jin, Q. Yu, Z. Wang, W. Hua, Z. Zhou, G. Sun, Y. Meng, S. Ma, Q. Wang, *et al.* (2024a). "Uncertainty is fragile: Manipulating uncertainty in large language models". *arXiv preprint arXiv:2407.11282*.

Zeng, Y., A. Rajasekharan, P. Padalkar, K. Basu, J. Arias, and G. Gupta. (2024b). "Automated interactive domain-specific conversational agents that understand human dialogs". In: *International Symposium on Practical Aspects of Declarative Languages*. Springer. 204–222.

Zhan, H., L. Li, S. Li, W. Liu, M. Gupta, and A. C. Kot. (2023). "Towards explainable recommendation via bert-guided explanation generator". In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 1–5.

Zhan, J., C.-L. Hsieh, I.-C. Wang, T.-S. Hsu, C.-J. Liau, and D.-W. Wang. (2010). "Privacy-preserving collaborative recommender systems". *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*. 40(4): 472–476.

Zhang, A., Y. Chen, L. Sheng, X. Wang, and T.-S. Chua. (2024a). "On generative agents in recommendation". In: *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1807–1817.

Zhang, B. H., B. Lemoine, and M. Mitchell. (2018). "Mitigating unwanted biases with adversarial learning". In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 335–340.

Zhang, F. and Q. Zhou. (2014). "HHT–SVM: An online method for detecting profile injection attacks in collaborative recommender systems". *Knowledge-Based Systems*. 65: 96–105.

Zhang, H., H. Song, S. Li, M. Zhou, and D. Song. (2022a). "A Survey of Controllable Text Generation using Transformer-based Pre-trained Language Models". *arXiv preprint arXiv:2201.05337*.

Zhang, J., T. Lan, R. Murthy, Z. Liu, W. Yao, J. Tan, T. Hoang, L. Yang, Y. Feng, Z. Liu, *et al.* (2024b). "AgentOhana: Design Unified Data and Training Pipeline for Effective Agent Learning". *arXiv preprint arXiv:2402.15506*.

Zhang, J., T. Lan, M. Zhu, Z. Liu, T. Hoang, S. Kokane, W. Yao, J. Tan, A. Prabhakar, H. Chen, *et al.* (2024c). "xlam: A family of large action models to empower ai agent systems". *arXiv preprint arXiv:2409.03215*.

Zhang, J., X. Xu, and S. Deng. (2023a). "Exploring collaboration mechanisms for llm agents: A social psychology view". *arXiv preprint arXiv:2310.02124*.

Zhang, J., K. Bao, W. Wang, Y. Zhang, W. Shi, W. Xu, F. Feng, and T.-S. Chua. (2024d). "Prospect Personalized Recommendation on Large Language Model-based Agent Platform". *arXiv preprint arXiv:2402.18240*.

Zhang, J., Y. Hou, R. Xie, W. Sun, J. McAuley, W. X. Zhao, L. Lin, and J.-R. Wen. (2024e). "Agentcf: Collaborative learning with autonomous language agents for recommender systems". In: *Proceedings of the ACM on Web Conference 2024*. 3679–3689.

Zhang, J., R. Xie, Y. Hou, X. Zhao, L. Lin, and J.-R. Wen. (2023b). "Recommendation as instruction following: A large language model empowered recommendation approach". *ACM Transactions on Information Systems*.

Zhang, K., L. Qing, Y. Kang, and X. Liu. (2024f). "Personalized LLM Response Generation with Parameterized Memory Injection". *arXiv preprint arXiv:2404.03565*.

Zhang, K., H. Chen, L. Li, and W. Wang. (2023c). "Syntax error-free and generalizable tool use for llms via finite-state decoding". *arXiv preprint arXiv:2310.07075*.

Zhang, K., W. Yao, Z. Liu, Y. Feng, Z. Liu, R. Murthy, T. Lan, L. Li, R. Lou, J. Xu, B. Pang, Y. Zhou, S. Heinecke, S. Savarese, H. Wang, and C. Xiong. (2024g). "Diversity empowers intelligence: Integrating expertise of software engineering agents". *arXiv preprint arXiv:2408.07060*.

Zhang, M., Z. Ren, Z. Wang, P. Ren, Z. Chen, P. Hu, and Y. Zhang. (2021). "Membership Inference Attacks Against Recommender Systems". In: *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*. 864–879.

Zhang, S., E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston. "Personalizing dialogue agents: I have a dog, do you have pets too? arXiv 2018". *arXiv preprint arXiv:1801.07243*.

Zhang, S., H. Yin, T. Chen, Q. V. N. Hung, Z. Huang, and L. Cui. (2020). "Gcn-based user representation learning for unifying robust recommendation and fraudster detection". In: *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*. 689–698.

Zhang, S., L. Yao, A. Sun, and Y. Tay. (2019). "Deep learning based recommender system: A survey and new perspectives". *ACM Computing Surveys (CSUR)*. 52(1): 1–38.

Zhang, W., J. Yan, Z. Wang, and J. Wang. (2022b). “Neuro-Symbolic Interpretable Collaborative Filtering for Attribute-based Recommendation”. In: *Proceedings of the World Wide Web Conference 2022*. 3229–3238.

Zhang, X., H. Xu, Z. Ba, Z. Wang, Y. Hong, J. Liu, Z. Qin, and K. Ren. (2024h). “Privacyasst: Safeguarding user privacy in tool-using large language model agents”. *IEEE Transactions on Dependable and Secure Computing*.

Zhang, Y. and X. Chen. (2020). “Explainable recommendation: A survey and new perspectives”. *Foundations and Trends® in Information Retrieval*. 14(1): 1–101.

Zhang, Y., G. Lai, M. Zhang, Y. Zhang, Y. Liu, and S. Ma. (2014). “Explicit factor models for explainable recommendation based on phrase-level sentiment analysis”. In: *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. 83–92.

Zhang, Z., L. Lei, L. Wu, R. Sun, Y. Huang, C. Long, X. Liu, X. Lei, J. Tang, and M. Huang. (2023d). “SafetyBench: Evaluating the Safety of Large Language Models with Multiple Choice Questions”. arXiv: 2309.07045 [cs.CL].

Zhang, Z., J. Yang, P. Ke, and M. Huang. (2023e). “Defending Large Language Models Against Jailbreaking Attacks Through Goal Prioritization”. arXiv: 2311.09096 [cs.CL].

Zhao, H., H. Chen, F. Yang, N. Liu, H. Deng, H. Cai, S. Wang, D. Yin, and M. Du. (2024a). “Explainability for large language models: A survey”. *ACM Transactions on Intelligent Systems and Technology*. 15(2): 1–38.

Zhao, W. X., K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, *et al.* (2023). “A survey of large language models”. *arXiv preprint arXiv:2303.18223*.

Zhao, Y., J. Wu, X. Wang, W. Tang, D. Wang, and M. de Rijke. (2024b). “Let Me Do It For You: Towards LLM Empowered Recommendation via Tool Learning”. In: *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1796–1806.

Zhao, Z., F. Lin, X. Zhu, Z. Zheng, T. Xu, S. Shen, X. Li, Z. Yin, and E. Chen. (2024c). "DynLLM: When Large Language Models Meet Dynamic Graph Recommendation". *arXiv preprint arXiv:2405.07580*.

Zheng, G., F. Zhang, Z. Zheng, Y. Xiang, N. J. Yuan, X. Xie, and Z. Li. (2018). "DRN: A deep reinforcement learning framework for news recommendation". In: *Proceedings of the 2018 world wide web conference*. 167–176.

Zheng, Y., C. Gao, X. Li, X. He, Y. Li, and D. Jin. (2021). "Disentangling user interest and conformity for recommendation with causal embedding". In: *Proceedings of the Web Conference 2021*. 2980–2991.

Zheng, Y., P. Li, W. Liu, Y. Liu, J. Luan, and B. Wang. (2024). "ToolRerank: Adaptive and Hierarchy-Aware Reranking for Tool Retrieval". *arXiv preprint arXiv:2403.06551*.

Zheng, Z., Z. Qiu, X. Hu, L. Wu, H. Zhu, and H. Xiong. (2023). "Generative job recommendations with large language model". *arXiv preprint arXiv:2307.02157*.

Zhou, C., P. Liu, P. Xu, S. Iyer, J. Sun, Y. Mao, X. Ma, A. Efrat, P. Yu, L. Yu, S. Zhang, G. Ghosh, M. Lewis, L. Zettlemoyer, and O. Levy. (2023). "LIMA: Less Is More for Alignment". arXiv: 2305.11206 [cs.CL].

Zhou, D., N. Schärli, L. Hou, J. Wei, N. Scales, X. Wang, D. Schuurmans, C. Cui, O. Bousquet, Q. Le, *et al.* (2022). "Least-to-most prompting enables complex reasoning in large language models". *arXiv preprint arXiv:2205.10625*.

Zhou, Z., Q. Wang, M. Jin, J. Yao, J. Ye, W. Liu, W. Wang, X. Huang, and K. Huang. (2024). "Mathattack: Attacking large language models towards math solving ability". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. No. 17. 19750–19758.

Zhu, L., X. Huang, and J. Sang. (2024a). "A LLM-based Controllable, Scalable, Human-Involved User Simulator Framework for Conversational Recommender Systems". *arXiv preprint arXiv:2405.08035*.

Zhu, S., R. Zhang, B. An, G. Wu, J. Barrow, Z. Wang, F. Huang, A. Nenkova, and T. Sun. (2023). "AutoDAN: Automatic and Interpretable Adversarial Attacks on Large Language Models". arXiv: 2310.15140 [cs.CR].

Zhu, X., F. Lin, Z. Zhao, T. Xu, X. Zhao, Z. Yin, X. Li, and E. Chen. (2024b). "Multi-Behavior Recommendation with Personalized Directed Acyclic Behavior Graphs". *ACM Transactions on Information Systems*.

Zhu, Y., Y. Xian, Z. Fu, G. de Melo, and Y. Zhang. (2021). "Faithfully explainable recommendation via neural logic reasoning". *arXiv preprint arXiv:2104.07869*.

Zhuang, Y., H. Sun, Y. Yu, Q. Wang, C. Zhang, and B. Dai. (2024). "HYDRA: Model Factorization Framework for Black-Box LLM Personalization". *arXiv preprint arXiv:2406.02888*.

Ziegler, D. M., S. Nix, L. Chan, T. Bauman, P. Schmidt-Nielsen, T. Lin, A. Scherlis, N. Nabeshima, B. Weinstein-Raun, D. de Haas, B. Shlegeris, and N. Thomas. (2022). "Adversarial Training for High-Stakes Reliability". arXiv: 2205.01663 [cs.LG].

Zou, A., Z. Wang, J. Z. Kolter, and M. Fredrikson. (2023). "Universal and Transferable Adversarial Attacks on Aligned Language Models". arXiv: 2307.15043 [cs.CL].

Zucco, C., H. Liang, G. Di Fatta, and M. Cannataro. (2018). "Explainable sentiment analysis with applications in medicine". In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 1740–1747.