AuditAgent: LLM Agent for Risks Auditing in Recommender Systems

Du Su[1*], Zhenxing Chen[2*], Shilong Zhao[1,2], Yuanhao Liu[1,2], Fei Sun[1†], Qi Cao[1], Huawei Shen[1,2]

[1]State Key Laboratory of AI Safety, Institute of Computing Technology, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
sudu@ict.ac.cn, chenzhenxing25z@ict.ac.cn, sunfei@ict.ac.cn

Abstract

Auditing recommendation systems has attracted growing attention due to increasing concerns over filter bubbles, unfairness, and data misuse. A common approach is sock-puppet auditing, where autonomous agents interact with platforms to reveal risks. However, existing approaches rely on hard-coded agents, lacking adaptability to dynamic GUI layouts and generating behaviors far from those of real users, limiting the comprehensiveness and representativeness of assessment. To address these issues, we introduce AuditAgent, an LLM-powered GUI-agent framework for risk auditing. AuditAgent simulates realistic user preferences and performs adaptive, human-like interactions on recommendation platforms. This design enables more thorough and faithful auditing, providing comprehensive assessments across multiple risk dimensions, including filter bubbles, unfairness, and data misuse.

Introduction

While recommendation systems have improved the efficiency of information access, they also pose significant risks, such as filter bubbles (Liu et al. 2025), algorithmic unfairness (Wang et al. 2023), and data misuse (Himeur et al. 2022), underscoring the need for effective risk auditing and governance (Bandy 2021). Autonomous auditing methods, such as sock-puppet auditing, have emerged as promising approaches due to their scalability, adaptability, and reproducibility. For example, Srba et al. (2023) investigated the dynamics of the "filter bubble" on YouTube by deploying scripted agents to watch pre-collected misinformation videos and track changes in homepage recommendations, while Hussein, Juneja, and Mitra (2020) studied how personalization amplifies misinformation using similar methods.

Although they can reveal certain risks, the significant behavioral gap between agents and real users undermines the validity of assessments (Ribeiro, Veselovsky, and West 2023; Haroon et al. 2023). Specifically: (1) **Lack of adaptive GUI interaction.** Existing agents rely on hard-coded scripts tied to fixed UI elements, which fail in dynamic page layouts where element positions change or the UI updates, thereby reducing audit comprehensiveness. (2) **Unrealistic user preference simulation.** Script-driven agents cannot perform personalized actions, such as content selection or feedback, limiting both the representativeness of audits and coverage of users with diverse preferences.

We propose AuditAgent, an auditing framework that leverages LLM-powered GUI agents. By integrating GUI imitation learning with LLM role-playing powered preference simulation, AuditAgent generates human-like interaction trajectories. It also addresses diverse risks, including filter bubbles, unfairness, and data misuse, enhancing the comprehensiveness and representativeness of risk auditing.

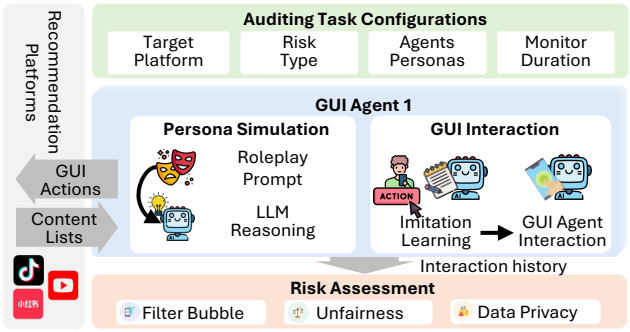

Figure 1: System overview of AuditAgent

System Overview

AuditAgent is an agentic risk auditing framework for recommendation platforms, leveraging LLM-powered, human-like GUI agents to enable representative and comprehensive risk auditing. As shown in Figure 1, the framework comprises the following three core functionalities:

Flexible Configuration of Audit Task

AuditAgent is adaptable across diverse platforms, risk types, and user groups with varying preferences. When configuring an audit task, the auditor specifies the target platform, risk type, the number of GUI agents and their preferences, as well as task parameters such as the audit duration.

GUI Agent Interaction with Platforms

At its core, the system employs LLM-powered GUI agents that interact with recommendation platforms in a human-like

*These authors contributed equally.
†Corresponding author.

manner, which addresses challenges of intelligent interaction and preference simulation, enhancing auditing comprehensiveness and representativeness. Specifically, we leverage two key techniques:

Imitation learning for GUI agent interaction. Agents are trained via imitation learning on fundamental UI actions (e.g., searching, scrolling), enabling them to combine behaviors based on auditing needs and perform flexible, preference-driven actions. Specifically, the auditor first performs fundamental actions, and GUI agents then learn, for each interacted UI element, its on-screen position, path in the UI hierarchy, and appearance from the demonstrations to reproduce the actions, enabling robust adaptation to different page layouts. The learning process is user-friendly, eliminating the need for labor-intensive manual coding.

LLM role-playing for preference simulation. Each agent encodes topics of interest in long-term memory and operates on recommendation platforms accordingly, simulating real user preference. This consistent, preference-driven behavior enhances the representativeness of auditing and enables risk auditing for various users groups.

Comprehensive Risk Assessment Reporting

AuditAgent provides comprehensive coverage of risks, including filter bubbles, unfairness, and data misuse, while preserving detailed interaction histories for risk traceability.

Filter bubble. Filter bubble is assessed using an agent that repeatedly consumes content within topics of interest and then monitors whether subsequent recommendations become overly concentrated on those topics.

Let M^f represent the total number of recommended items during monitoring, and N_1^f, N_2^f denote the number of items from the designated topics before and after the agent's content consumption. We define the increase in recommendations for target topics $\rho^f = {}^{N_2^f}/_{M^f} - {}^{N_1^f}/_{M^f}$ as a risk indicator. A filter bubble is indicated if $\rho^f > \theta^f$.

Unfairness. Unfairness is evaluated using an agent with two topics of interest (e.g., scientific rumors vs. serious technologies). The agent consumes an equal amount of content of the two topics. We then monitor the extent of disproportionate exposure between them.

Let M^u be the total number of items during monitoring, and N_1^u, N_2^u denote the number of items for each topic. Following the classical definition of fairness (Dwork et al. 2012), we denote the exposure difference between the two topics $|\rho_1^u - \rho_2^u|$ as a risk indicator, where $\rho_1^u = {}^{N_1^u}/_{M^u}$, $\rho_2^u = {}^{N_2^u}/_{M^u}$. An unfairness risk is flagged if $|\rho_1^u - \rho_2^u| > \theta^u$.

Data misuse. Data misuse risk arises when a recommendation platform disregards a user's opt-out of data-driven personalization. It is measured by the persistence of topic distribution in recommendations before and after personalization is disabled.

Let $N_1^d, N_2^d, \ldots, N_k^d$ be the counts of items in each topic before opting-out personalization, and $N_1^{d'}, N_2^{d'}, \ldots, N_k^{d'}$ the counts after opt-out, and M^d be the total count. A data misuse risk is flagged if the topic distribution remains similar after the user opts out, $KL(\{\rho_1^d, \ldots, \rho_k^d\} \,\|\, \{\rho_1^{d'}, \ldots, \rho_k^{d'}\}) < \theta^d$, where $\rho_i^d = {}^{N_i^d}/_{M^d}$, $\rho_i^{d'} = {}^{N_i^{d'}}/_{M^d}$, $i = 1, \ldots, k$.

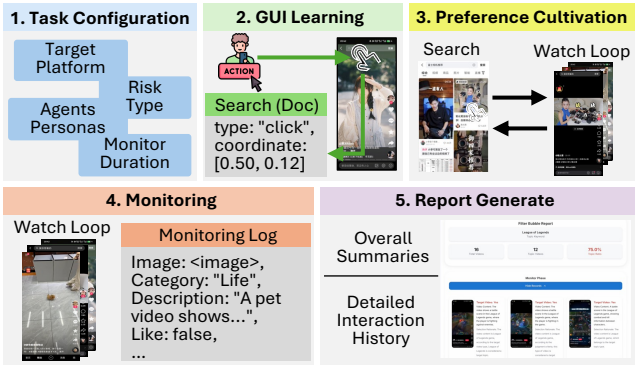

Figure 2: Workflow of audit tasks

Workflow

An audit task follows the workflow illustrated in Figure 2, comprising five major steps.

Task configuration. AuditAgent allows auditors to configure task parameters (platform, duration, risk type), along with the number of GUI agents and their preferences.

GUI learning. AuditAgent enables auditors to train agents to interact through a user-friendly imitation learning framework. Agents learn fundamental actions from auditor demonstrations, such as keyword searches, thumbnail clicks, and swipes, and later on flexibly combine them according to behavior sequences configured for each risk.

Preference cultivation. In this phase, agents start interaction with recommendation platforms to cultivate their personalized user profiles. The agents search for each interested topic and interact with content according to their preferences (e.g., liking or skipping), ensuring personalized recommendations during the subsequent monitoring phase.

Monitoring. After cultivation, agents continue interacting with the platform, consuming content directly from the homepage to mirror real user behavior. They select content of interest and provide preference-driven feedback. Meanwhile, the system logs detailed records of content exposure and interaction decisions for subsequent risk analysis.

Report generation. After monitoring, interaction logs are analyzed using metrics specific to the target risk. The resulting report provides overall summaries and temporal statistics, illustrating risk evolution across task stages and highlighting high-risk content for further safety review.

Conclusion

We introduce AuditAgent, a sock-puppet auditing framework for recommendation systems, leveraging LLM-based GUI agents. AuditAgent addresses two key challenges in agentic auditing: adaptive GUI interaction and realistic preference simulation, which produce human-like trajectories that improve the comprehensiveness and representativeness of risk auditing. It supports auditing across multiple risk dimensions, offering a systematic auditing methodology.

Acknowledgments

This work was supported in part by National Key R&D Program of China under Grant No. 2022YFB3103700 and 2022YFB3103704, the Strategic Priority Research Program of the CAS under Grant No. XDB0680302, the Beijing Natural Science Foundation under Grant No. 4252023, the National Natural Science Foundation of China under Grant No. 62472409.

References

Bandy, J. 2021. Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1).

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226.

Haroon, M.; Wojcieszak, M.; Chhabra, A.; Liu, X.; Mohapatra, P.; and Shafiq, Z. 2023. Auditing YouTube's recommendation system for ideologically congenial, extreme, and problematic recommendations. *Proceedings of the National Academy of Sciences*, 120(50): e2213020120.

Himeur, Y.; Sohail, S. S.; Bensaali, F.; Amira, A.; and Alazab, M. 2022. Latest trends of security and privacy in recommender systems: a comprehensive review and future perspectives. *Computers & Security*, 118: 102746.

Hussein, E.; Juneja, P.; and Mitra, T. 2020. Measuring Misinformation in Video Search Platforms: An Audit Study on YouTube. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW1).

Liu, N.; Hu, X. E.; Savas, Y.; Baum, M. A.; Berinsky, A. J.; Chaney, A. J. B.; Lucas, C.; Mariman, R.; de Benedictis-Kessner, J.; Guess, A. M.; Knox, D.; and Stewart, B. M. 2025. Short-term exposure to filter-bubble recommendation systems has limited polarization effects: Naturalistic experiments on YouTube. *Proceedings of the National Academy of Sciences*, 122(8): e2318127122.

Ribeiro, M. H.; Veselovsky, V.; and West, R. 2023. The amplification paradox in recommender systems. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, 1138–1142.

Srba, I.; Moro, R.; Tomlein, M.; Pecher, B.; Simko, J.; Stefancova, E.; Kompan, M.; Hrckova, A.; Podrouzek, J.; Gavornik, A.; and Bielikova, M. 2023. Auditing YouTube's Recommendation Algorithm for Misinformation Filter Bubbles. *ACM Trans. Recomm. Syst.*, 1(1).

Wang, Y.; Ma, W.; Zhang, M.; Liu, Y.; and Ma, S. 2023. A Survey on the Fairness of Recommender Systems. *ACM Trans. Inf. Syst.*, 41(3).