

Attacking and Defending Multi-Agent Collaborative Filtering Systems Through Connectivity

Anjun Hu
University of Oxford
Oxford, United Kingdom
anjun.hu@eng.ox.ac.uk

Hanting Xie
Amazon
London, United Kingdom
hanting@amazon.com

Saranya Govindan
Amazon
London, United Kingdom
saranygo@amazon.co.uk

Jas Kandola
Amazon
London, United Kingdom
jaskand@amazon.co.uk

Kurt Cutajar
Amazon
London, United Kingdom
cutajark@amazon.co.uk

Abstract

Multi-agent collaborative filtering systems coordinate autonomous LLM-powered user and item agents through natural-language interaction to refine preferences and generate recommendations. These systems inherit vulnerabilities from both their data-driven nature and multi-agent interactions, which manifest in distinct ways. Understanding how connectivity modulates vulnerability in these systems could facilitate the development of more robust recommendation pipelines. In this work, we adapt attacks and defenses from the general multi-agent systems (MAS) literature to the agent-based CF setting, evaluating them under systematically varied connectivity in the AgentCF framework, where CF connectivity is characterized along two axes: (i) **candidate count** (the number of item candidates per turn per user, measuring user-side interaction density) and (ii) **catalog concentration** (the degree of item catalog overlap across users). Our contributions include: **(1) Adaptation:** we reproduce MAS-inspired attacks and defenses in the agentic CF domain, confirming partial transferability of original observations. **(2) Characterization:** we characterize how the two aspects of connectivity shape attack and defense outcomes, revealing role asymmetries between user and item agents, non-monotonic temporal dynamics in attack efficacy, and divergent patterns across dissemination and extraction attack goals. Additionally, as an exploratory extension, we assess the applicability of epidemic-inspired static metrics in ranking CF configurations by expected attack outcome, potentially enabling cost-efficient robustness assessment. Implementation is available at <https://github.com/anjunhu/ConnACF>.

CCS Concepts

• **Security and privacy** → *Domain-specific security and privacy architectures; Usability in security and privacy.*

Keywords

Responsible AI, Multi-Agent Systems, Collaborative Filtering

ACM Reference Format:

Anjun Hu, Hanting Xie, Saranya Govindan, Jas Kandola, and Kurt Cutajar. 2026. Attacking and Defending Multi-Agent Collaborative Filtering Systems Through Connectivity. In *20th ACM Conference on Recommender Systems (RecSys '26)*, September 27–October 02, 2026, Minneapolis, MN, USA. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3773078.3831748>

1 Introduction

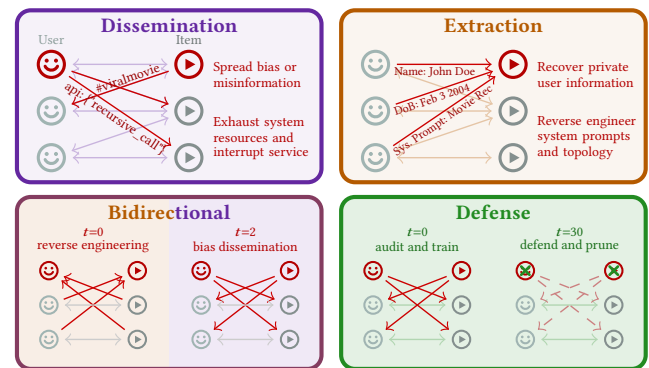


Figure 1: Attack/defense families reproduced on multi-agent CF. Red nodes are attackers; layouts are schematic. How connectivity modulates the outcome is the subject of this work.

Multi-agent collaborative filtering (CF) systems augment traditional recommendation with stateful, autonomous LLM-powered agents that represent users and items [72]. Connectivity in these systems arises from **structural design choices** (communication topology and candidate item count per round) and users' **historical interactions** (overlap in past interactions and preference alignment). These dimensions shape both the evolution of preferences and how vulnerabilities propagate through the system. The latter has received limited attention, and this work aims to characterize the connectivity-robustness relationship in multi-agent CF.

The same inter-agent communication pathways that enable preference refinement could also introduce (non-exhaustively) two types of unwanted information flows: Firstly, unwanted *dissemination* may occur when a compromised agent injects a payload that spreads outward: a poisoned item description persists in a user's memory, shapes future interactions, and reaches peers through opinion exchange. Secondly, unwanted *extraction* may occur when



This work is licensed under a Creative Commons Attribution 4.0 International License. *RecSys '26, Minneapolis, MN, USA*
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2284-4/26/09
<https://doi.org/10.1145/3773078.3831748>

an attacker agent recovers unauthorized proprietary information stored in a victim agent’s memory, with leakage observable on communication channels accumulating across interaction turns. Both directions are potentially governed by the same connectivity variables, resulting in variations in recommendation quality and system safety.

Recommender systems are known to be susceptible to data poisoning and adversarial attacks [3, 30]. Recent red-teaming works increasingly target agentic recommenders through their stateful nature and inter-agent communication channels, devising novel attacks through profile poisoning and memory corruption [32, 64, 73]. These attacks confirm system vulnerability but are tested on fixed system configurations, leaving open whether attack efficacy changes as connectivity varies. On the other hand, general-domain MAS research has shown connectivity to be a significant determinant of attack outcomes [24, 26, 67, 81], but those findings come from systems that differ substantially in scale and structure from CF, and their transferability is not yet established.

Multi-agent CF has structural properties that are distinct from query-driven MAS, making attack transferability non-trivial. (1) CF systems are *bipartite*, with asymmetric user and item roles that have no direct counterpart in role-homogeneous or tool-heterogeneous MAS networks. (2) Connectivity manifests more diversely in CF, originating from both inference system design and historical interaction data, defined more formally in §3.3. (3) CF usually operates at larger scale with agent states persisting over longer interaction horizons, giving rise to intriguing temporal dynamics (F3, §4) not previously reported for smaller task-driven MAS.

In this work, we reproduce topology-vulnerability findings from general MAS in the CF setting, using AgentCF [72] as the primary evaluation platform and incorporating MACF [61]. We scope our threat model to application-layer adversaries who compromise a fraction of user and/or item agents through prompt injection or profile poisoning, with goals spanning dissemination (preference manipulation or resource exhaustion), extraction (privacy breach or reverse engineering), or a combination of both. Our contributions include: **(1) Adaptation and Reproduction.** We adapt MAS attacks and defenses to the multi-agent CF setting and validate the applicability of original methods. **(2) Characterizing the Effect of Connectivity.** We find that candidate count and catalog concentration produce qualitatively different effects on attack success, with direction and magnitude varying across attacker goals and role configurations, revealing user-item partition asymmetries unique to CF systems. Additionally, we assess the applicability of epidemic-inspired static metrics derived from system configuration parameters in ranking CF configurations by expected attack outcome, potentially enabling cost-efficient, *a priori* risk assessment.

2 Related Work

Multi-agent recommender systems (MARS) inherit vulnerabilities from both their data-driven nature and their multi-agent communication structure. Our work sits at the intersection of MAS and RecSys safety, bringing topology awareness from MAS works to MARS and grounding evaluation in CF’s unique structures.

Multi-Agent Recommender Systems. LLM-based recommender

systems have evolved from single-agent interfaces [12, 23, 50, 78] towards multi-agent architectures [10, 11, 28, 37] with diverse topologies and configurations: *centralized* systems route coordination through a hub or orchestrator [56, 58, 66]; *mesh* systems model recommendation as emergent peer-to-peer behavior [27, 61, 71, 72] and facilitate behavioral simulation [47, 69, 82]; and *hybrid* systems combine both [41, 43]. Their safety is receiving increasing attention [14], and our work builds upon this line of research.

LLM and Agentic Recommender System Safety. Classical poisoning and adversarial attacks spanning item-profile manipulation, shilling, and prompt injection [3, 25, 30, 53, 55, 59, 73] remain relevant to LLM and agentic recommenders. Recent work targets agentic recommenders directly via profile poisoning, memory corruption, backdoor implantation, and retrieval poisoning [22, 31, 32, 49, 52, 64, 65, 75]. Indirect prompt injection and self-replicating payloads further extend attack surfaces to inter-agent communication [20, 68]. Control-flow hijacking demonstrates the blast radius of these vulnerabilities in general MAS [15, 45]. Privacy leakage [36, 51, 62, 74], collusion [6, 13], collective manipulation [44], and closed-loop ranking exploitation [2] represent further threat dimensions. We adapt representative attacks from this body of work to the CF setting, and ask how their efficacy changes as connectivity varies.

Evaluating MAS Safety. ASB [70] and TAMAS [17] benchmark attacks and defenses in general MAS; AgentLeak [62] and failure-mode analyses [1] cover privacy leakage and systematic breakdowns. In RecSys, safety tutorials [38, 54], AgentRecBench [39], and agentic evaluation at scale [76] survey the broader landscape. Evaluation has also expanded beyond utility [16] to consider safety and trustworthiness. Our evaluation harness is informed by these benchmarks and extends them with per-partition, per-turn metrics that separate transient from steady-state regimes.

Topology Awareness in MAS Safety. NetSafe [67] demonstrates that lower-connectivity topologies are safer and introduces static graph metrics predicting safety rankings. Stealthy multi-round tampering [63], optimized prompt attacks [18], and information propagation analysis [40] further characterize how topology shapes attack spread. GNN-based defenses [4, 29, 33, 48, 79] operationalize topology awareness on the defense side, several of which we adapt and evaluate in this work. We bring topology awareness to CF by reproducing a dual-source connectivity-driven evaluation, characterizing how candidate count and catalog concentration jointly shape outcomes across user and item partitions.

3 Experimental Setup

We build a unified evaluation harness on top of AgentCF [72], also incorporating MACF [61], equipping both with a common attack, defence and metric tracking surface and exposing two connectivity axes as independent experimental variables.

3.1 Base Systems. AgentCF [72] is a bipartite mesh CF system: despite being orchestrator-moderated, user and item agents can directly impact each others’ memory state. In the canonical setting, at each step an orchestrator selects a (user, positive item, negative item) triple; the user ranks the two items ($k=2$); both item agents update their descriptions; and the user updates its preference memory.

An optional user-user opinion-exchange step lets users share preference summaries with peers. This canonical setting corresponds to $k=2$ and a **medium-overlap catalog**. We generalize this system to other k by changing the binary task to 3-item ranking in case of $k=3$ and binary like/dislike feedback in case of $k=1$; and vary **catalog density** by resampling the dataset (§3.4). MACF [61] is an orchestrator-mediated star topology CF system ($k=0$): all user-item communication is routed through a central hub and no direct U-I exchange triggers immediate memory updates. MACF is included in the linked codebase and available for analysis, but excluded from the plots in §4 since its workflow introduces structural confounds beyond connectivity. Moreover, attacks with extraction components are not applicable to MACF without significant changes due to the lack of immediate peripheral agent contact.

3.2 Memory Management. Each agent’s profile is a rolling episodic memory list: after each interaction the LLM appends a semantic summary, providing a persistent structured surface for injection (CORBA, NetSafe, etc.) and extraction (MAMA, MASLeak, etc.).

3.3 Two Connectivity Axes. The two axes are illustrated in Fig. 2. k is a system configuration set at inference time independently of the dataset; ρ is determined by historical interaction data.

(1) **Candidate count** (k , inference-time communication density). $k \in \{1, 2, 3\}$ is the number of items each user can interact with at each turn. In other words, it is the user’s decision bandwidth. A higher k means each user can immediately impact more items.
 (2) **Catalog concentration** (ρ , interaction history density). We fix $n_U=100$ users and vary item catalog size n_I by re-sampling MovieLens-100K (§3.4) to $n_I \in \{200, 100, 50\}$, giving $\rho \equiv n_U/n_I \in \{0.5, 1, 2\}$. Denser catalogs concentrate the same number of historical interactions on fewer items, making each item more influential.

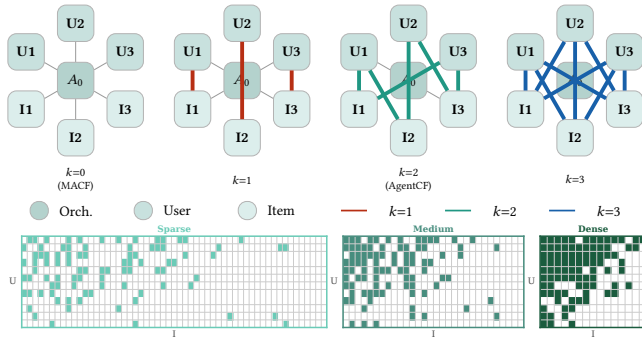


Figure 2: (Top) Candidate count variants: graph edge colors match $k=1/2/3$ line colors used in metric plots. **(Bottom) catalog concentration variants:** $\rho=0.5/1/2$ correspond to low/medium/high overlap across users.

3.4 Dataset and Task. We use the MovieLens-100K dataset [8], subsampling 100 users and varying the item catalog size to produce three **catalog concentration** levels (§3.3). MovieLens-100K is the canonical benchmark of the AgentCF base system we reproduce [72]; keeping to it lets us faithfully replicate the original system and isolate connectivity effects rather than confounding them with

a change of domain. The single-dataset, single-domain scope is a deliberate reproduction choice and a stated limitation (§8). The task is sequential movie rating prediction: at each turn, each user agent takes an action (rate/choose/rank) on $k \in \{1, 2, 3\}$ items, updating its preference memory and the item descriptions accordingly. In parallel to profile optimization, the system task (sequential item ranking prediction) is triggered by separate evaluation queries.

3.5 Threat Models. We consider application-layer adversaries who compromise a fixed fraction α of user agents, item agents, or both (α_U, α_I), with $(0.25, 0.25)$ for main experiments and more combinations as ablations (§6.1-6.2). Compromise is achieved via two surfaces: (i) *system-prompt override* (replacing an agent’s system/task prompt with an adversarial persona) and (ii) *user profile/item description poisoning* (seeding canary content into a user agent’s memory or an item agent’s description). LLM weights and inference infrastructure are assumed secure; the adversary has no access to model internals or the orchestrator’s scheduling logic. Attacker goals fall into two directions:

- **Dissemination** (bias induction, resource exhaustion). Attacker agents inject canary concepts (e.g., esoteric theatrical genres such as Letterist cinema) or self-replicating blocking payloads into their profiles. Success is the *contamination rate*, the fraction of non-attacker agents whose current memory contains these payloads, tracked per turn for the user (U) and item (I) partitions.
- **Extraction** (privacy breach, reverse engineering). Attacker agents elicit private information from victim memories through targeted dialogues. Success is the *leakage rate*: for example, the fraction of seeded PII fields recovered; or the fraction of system-design IP (prompts, instructions, topology) recovered.

3.6 Metrics: Attack Success Rate (ASR) via LLM Judge. For each turn, an LLM judge evaluates every non-attacker agent by comparing its current memory against its original state and the known canary concepts, returning a binary contaminated/clean label. ASR is defined as the ratio between contaminated non-attacker agents and total non-attacker agents, reported separately for user (U) and item (I) partitions per turn. ASR percentages exclude attacker agents, so reported rates reflect spread or leakage among the benign population only. For canary-style dissemination and PII-extraction attacks, where the seeded concept or field is known a priori, contamination and leakage can also be checked deterministically via exact and embedding-based matching; we use these deterministic checks to spot-check the judge’s labels and treat the LLM judge as the primary instrument. We flag full human-annotated calibration of the judge (false-positive/false-negative rates, prompt-sensitivity analysis) as a limitation and a target for extension. For temporal regime-aware analysis, we also distinguish the initial transient stage **slope** (U_{tr}, I_{tr}) from the steady-state **mean** (U_{ss}, I_{ss}) in §4. Attack-specific metrics (e.g., blocking or Denial-of-Service rate, per-category PII leakage, topology recovery rate) are reported alongside ASR where applicable.

3.7 Computational Environment. All experiments were conducted on an AWS EC2 g6.12xlarge instance using the Deep Learning OSS Nvidia Driver AMI image. The instance is equipped with 4 NVIDIA L4 GPUs (23 GB VRAM each), running Ubuntu 24.04.4 LTS (kernel 6.17.0-1010-aws), CUDA 12.9 and PyTorch 2.11. Default agent LLM

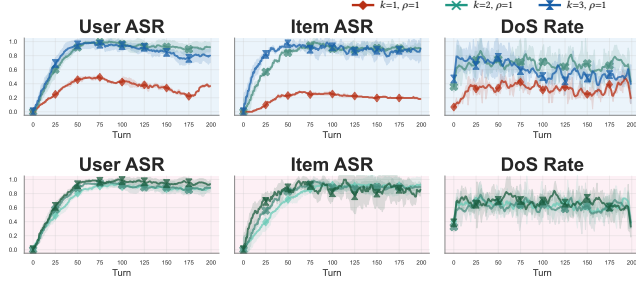


Figure 3: CORBA ASR and DoS rate (y-axis) across turns (x-axis) and connectivity configurations: candidate count (k , top) and catalog concentration (ρ , bottom).

is Qwen3-235B-A22B. Default judge model is Claude Sonnet 4.5. Full conda environment specification and run scripts are included in the repository. LLM inference is handled via Amazon Bedrock.

4 Evaluation

This section presents how we adapt existing MAS evaluation, attack and defenses to the CF setting and reproduced results on connectivity variants. We reproduced red-teaming methods grouped by the direction of unwanted information flow they cause: dissemination-driven (§4.1), extraction-driven (§4.2), and bidirectional (§4.3). Defenses follow in §4.4. For each method, we report: **Adaptation** (original methods, metrics, how we adapt them to CF setting) and **Characterization** (per-turn temporal metrics that demonstrate how the two connectivity axes affect efficacy).

Three cross-cutting patterns are observed across adapted methods: **(F1) ASRs exhibit growth-saturation patterns.** Contamination and leakage curves exhibit rapid early growth that decelerates and saturates, consistent with epidemic spreading dynamics observed in prior MAS red-teaming work.

(F2) Role-based asymmetry. For attacks that are applicable to both User and Item roles, attack efficacy on the two agentic role partitions (U_{**} and I_{**}) and the modulation effects of both connectivity axes are generally not the same and warrant separate tracking.

(F3) Decoupled slope and plateau. A faster early-phase growth rate (higher initial transient slope, U_{tr} , I_{tr}) does not always translate to higher steady-state mean attack performance (U_{ss} , I_{ss}).

4.1 Dissemination-Driven Attacks

We first reproduce two dissemination-driven attacks: CORBA (resource exhaustion via a self-replicating recursive blocking payload) and NetSafe (bias and misinformation spread). Increased connectivity in both k and ρ generally facilitate dissemination, with attack-specific nuances detailed below.

4.1.1 CORBA. Adaptation. CORBA [81] injects a self-replicating recursive blocking payload that combines contagion (pass to neighbours) with recursion (re-inject to oneself), achieving near-complete contamination within 20 turns. Our adapted version seeds the payload into a target item agent’s memory and into the system prompts of a fixed fraction of user agents. Since AgentCF’s backward reflection step can partially wash out the payload, lower absolute rates

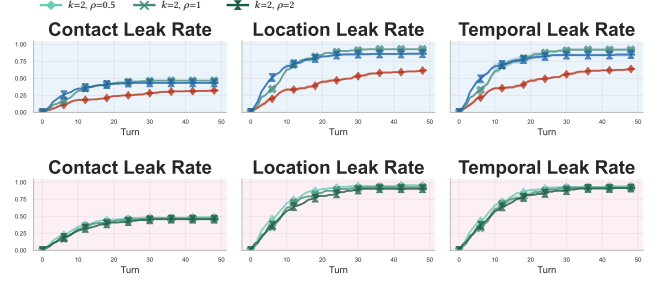


Figure 4: MAMA per-category PII leakage across various k and ρ connectivity configurations.

than the original are expected. We track user and item victim rate separately, along with overall Denial-of-Service (DoS) rate.

Characterization. As shown in Fig. 3, both partitions follow growth-saturation patterns (F1), but respond differently to each axis (F2). Increasing connectivity in k raises both user and item initial growth slopes; at steady state, $k=2$ and $k=3$ curves converge while both remain ahead of $k=1$. Varying ρ leaves user-side non-monotonic throughout. Meanwhile, higher ρ yields higher item-side initial ASR growth slope, but item-side ASR becomes non-monotonic at steady state (F3).

4.1.2 NetSafe. Adaptation. NetSafe [67] replaces the system prompts of a fixed fraction of agents with attackers’ and measures contamination rate across the network, finding that higher connectivity increases adversarial spread and that static attacker-distance metrics predict safety rankings better than classical graph metrics. We adapt the misinformation variant by seeding attacker agents via both surfaces: a poisoned system prompt and profile for user agents, or a poisoned product description for item agents. In addition to the default setting, three ablations (attacker ratio, role-split and style variants, and LLM swap) are reported in §6.

Characterization. As shown in Fig. 9a, both partitions follow growth-saturation patterns (F1) but respond differently to connectivity modulation (F2). User-side contamination is non-monotonic with k across both transient and steady-state regimes. Item-side transient slope is also non-monotonic ($I_{tr}^{k=3} > I_{tr}^{k=1} > I_{tr}^{k=2}$), but a reversal between $k=1$ and $k=2$ eventually resolves: higher k yields higher item-side plateaus (F3). Varying ρ leaves user-side non-monotonic throughout, while higher ρ yields higher item-side growth rates and final contamination levels.

4.2 Extraction-Driven Attacks

We evaluate two Extraction-Driven Attacks, MAMA (Privacy breach measured through PII presence on communication channels) and MASLeak (reverse engineering of system prompts and topologies). This family of attacks is generally more sensitive to variations in k than ρ , where higher candidate counts k facilitate extraction.

4.2.1 MAMA. Adaptation. MAMA [26] seeds synthetic PII into a target agent’s memory and deploys an attacker agent to elicit it via multi-round resonance dialogue, targeting only user agents as victims. The original paper finds that denser topologies and shorter attacker-target distances increase leakage, that leakage rises sharply

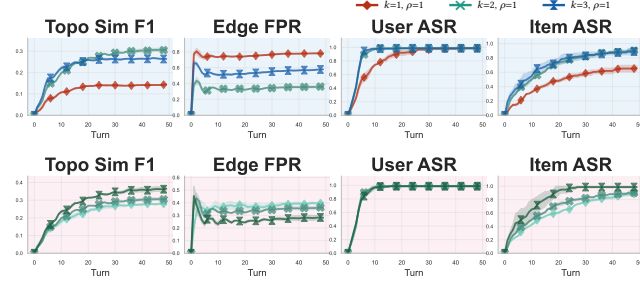


Figure 5: TOMA: Reverse engineering metrics (Topology Similarity F1, Edge FPR), and ASR among User / Item agents.

in early rounds then plateaus, and that model choice shifts absolute rates but preserves topology ordering; temporal and location attributes leak most readily while regulated identifiers remain near zero. We adapt it to CF by seeding 13 PII fields across five categories (identity, contact, location, temporal, regulated) into user agent memories and enabling user-item dialogue as the primary leakage channel. We report per-category leakage rates.

Characterization. As shown in Fig. 4, increasing k raises leakage growth rate in the initial transient regime but steady-state rates are nearly identical for $k=2$ and $k=3$, both of which lead $k=1$ by a significant margin: connectivity accelerates early exposure without lifting the plateau beyond $k=2$. Variations in p produces negligible differences across all variants.

4.2.2 MASLeak. Adaptation. MASLeak [46] uses a two-phase worm to extract system prompts, task instructions, agent count, and communication topology from black-box multi-agent systems, finding that the worm can recover internal IP across diverse frameworks. We relax the reflection prompt to preserve structured data blocks. We track extraction rate (fraction of IP fields recovered), system-prompt and task-instruction recovery rates separately, and topology leak rate (fraction of interaction edges reconstructed).

Characterization. Higher k yields higher IP extraction rate in the transient regime ($U_{tr}^{k=1} < U_{tr}^{k=2} < U_{tr}^{k=3}$), but steady-state rates are non-monotonic: $k=2$ outperforms both $k=1$ and $k=3$, suggesting that faster initial growth does not necessarily translate to a higher final plateau, consistent with (F3). Varying p produces negligible differences in extraction rate across both regimes, indicating that catalog concentration is not a primary modulating factor for this attack. Both axes are consistent with MASLeak’s hypothesis that reachability drives extraction: once the worm reaches enough agents the marginal gain from additional connectivity is small.

4.3 Bidirectional Attacks

We evaluate two Bidirectional Attacks, TOMA and MASTER, both of which begin with a reverse-engineering phase followed by dissemination. The extraction component is measured by two metrics: **Topology Similarity F1** (Jaccard similarity between the inferred and true U-I interaction edge sets: $|E_{pred} \cap E_{true}| / |E_{pred} \cup E_{true}|$) and **Node/Edge FPR** (fraction of predicted nodes or edges that are false positives: $|E_{pred} \setminus E_{true}| / |E_{pred}|$). They have divergent behaviours.

4.3.1 TOMA. Adaptation. TOMA [24] routes malicious payloads toward high-value core agents by exploiting known topology,

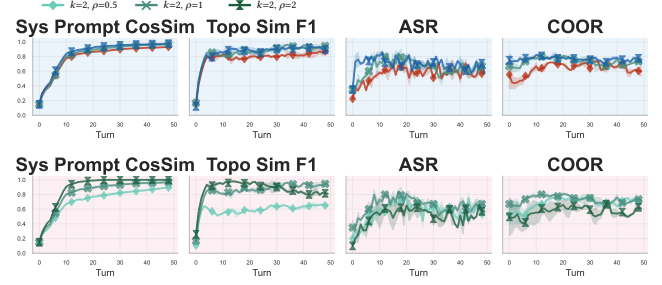


Figure 6: MASTER: Reverse engineering metrics and Dissemination metrics (ASR, Harmful Team Cooperativeness).

finding that centralized topologies (star, tree) are more vulnerable to adversarial spread than decentralized ones (mesh, ring), as payloads routed through hub nodes reach more agents faster. We recast TOMA as a bidirectional attack: the contamination component injects canary concepts into bridge items (items with high connectivity to multiple users, identified by offline topology analysis), while the topology-inference component tracks canary propagation to reconstruct the hidden interaction graph. We track ASR for both user and item partitions for the dissemination component; topology recovery F1 and edge FPR for the extraction component.

Characterization. As shown in Fig. 5, higher k yields higher initial growth rates of topology extraction similarity. This is then reversed ($k=2$ leads $k=3$) at steady state (F3). While higher k leads to higher recall, precision is consistently worse for $k=3$ than $k=2$ (see FPR), causing this reversal in system reverse engineering (extraction side) efficacy. Higher k yields higher ASR for both user and item partitions, but $k=3$ has a negligible advantage over $k=2$. Higher p (denser catalog) leads to higher reverse engineering metrics and item-side steady-state ASR, but steady-state user-side ASR is mostly unaffected by p variation. Initial ASR growth rates for both partitions are non-monotonic with respect to p .

4.3.2 MASTER. Adaptation. MASTER [83] combines topology probing with role-adaptive dark-trait injection [77] (e.g., “I prioritize self-interest over user welfare” for a Machiavellian persona), finding that targeting agents by role and centrality amplifies attack success. MASTER’s probing phase assumes agents respond to direct capability queries; we adapt it to indirect elicitation via canary items. The injection phase maps directly: dark-trait persona strings are appended to attacker user agents’ memory.

Characterization. As shown in Fig. 6, higher k generally leads to faster growth and higher steady state levels of ASR, though the gap across connectivity variants shrinks at steady state. Reverse engineering performance is relatively insensitive to k but higher p yields higher reverse engineering performance throughout: denser catalogs facilitate the extraction of system prompt and agent count.

4.4 Defenses

We reproduce four defenses: two GNN-based observe-then-protect mechanisms (G-Safeguard and BlindGuard), benchmarked against NetSafe but applicable to other attacks, and two tailored defenses designed specifically for the bidirectional attacks (T-Guard for TOMA and M-Guard for MASTER).

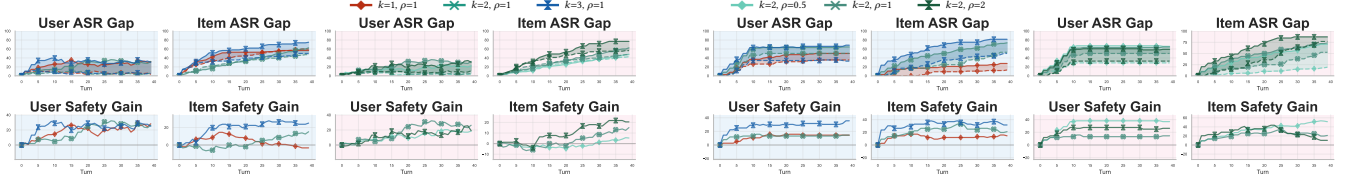


Figure 7: G-Safeguard against NetSafe (left) and T-Guard against TOMA (right). Top: guarded (solid) vs. unguarded (dashed) ASR across k and ρ connectivity configurations, for users and items. Bottom: safety gain $\Delta\text{ASR} = \text{Unguarded} - \text{Guarded}$.

4.4.1 G-Safeguard. Adaptation. G-Safeguard [48] was originally evaluated on role-homogeneous LLM-MAS with random topologies. It builds a discourse graph over agent communications, embeds each agent’s memory as a dense vector, and trains a supervised edge-aware GAT to label attacker nodes; detected nodes are quarantined by removing their outgoing edges. We apply it against NetSafe, reproducing the GNN architecture and re-implement graph construction for bipartite user-item graphs: nodes are user and item agents, and edges carry temporal windows of memory embeddings as attributes. The supervised training procedures (collect labelled data, train GNN, evaluate with active detection) are preserved.

Characterization. G-Safeguard effectively reduces user-side contamination relative to the unguarded baselines in all cases. On item side, sparsest variants ($\rho=0.5$ and $k=1$) seem to perform notably worse with nearly no meaningful safety gain (ASR difference between unguarded and guarded runs). With respect to both k and ρ , safety gain is non-monotonic for users throughout and for items initially, but higher connectivity yields greater item-side safety gain at steady state. (Fig. 7)

4.4.2 BlindGuard. Adaptation. BlindGuard [29] was evaluated on the same attack suite as G-Safeguard. Its key distinction is *unsupervised* training: it requires only normal interaction data and learns to detect anomalies without attacker labels, using a contrastive objective that scores nodes by deviation from the learned normal distribution. We re-implement the encoder as a hierarchical MLP that fuses self, neighbour, and global features, replacing the original pooling-based encoder to better handle the degree asymmetry between user and item agents in the bipartite graph. The corruption-guided contrastive training strategy is preserved.

Characterization. BlindGuard moderately reduces **user**-side contamination in most settings. However, items receive little security benefit apart from the densest $\rho=2$ setting. This may originate from the fact that item agents have highly variable contact degrees whereas users have fixed contact degrees, so the contrastive detector cannot function reliably. Denser catalog concentration ($\rho \uparrow$), for both user and item partitions, leads to a larger drop in ASR.

4.4.3 T-Guard. Adaptation. T-Guard [24] was originally designed for tool-using MAS (Magenti-One [5], LangManus [19], OWL [9]) with visual edge agents, comprising four components: a cross-modal validator (visual-textual consistency), a topology trust evaluator (taint propagation), an access control manager (policy enforcement), and a dynamic policy updater. The cross-modal validator is not applicable to our setting which has no visual modality. We adapt and transfer the rest: the topology trust evaluator runs the taint propagation model on bipartite U-I graphs. The access control manager then enforces the original work’s three-tier policy

based on suspect level: quarantine (blocked from all memory updates), restrict (backward memory update suppressed), or proceed as usual.

Characterization. Contamination growth is delayed and plateau levels are reduced relative to the undefended baseline across all k variants. Higher connectivity in k generally results in greater attack suppression. ASR reduction is comparable between $k=1$ and $k=2$ but substantially larger at $k=3$. On the other hand, ASR reduction is non-monotonic with respect to ρ variations. (Fig. 7)

4.4.4 M-Guard. Adaptation. M-Guard [83] comprises three mechanisms: (i) *prompt leakage defense*: an LLM detector monitors responses for system-prompt disclosure and injects a warning into the next-turn input; (ii) *hierarchical monitoring*: agents are ranked by role criticality and topological position, with high-importance agents monitored more frequently by a supervisory agent; (iii) *preemptive defense*: scenario-specific hardening instructions are prepended to each agent’s system prompt before deployment. We transplant this guardrail by fixing the domain to movie recommendation, and adapting the importance classifier to U/I partitions.

Characterization. Denser catalogs ($\rho \uparrow$) yield greater ASR suppression for both partitions in the transient regime and for users at steady state, while item-side steady-state suppression becomes non-monotonic. Higher k produces steeper slopes in security gain for both partitions, but steady-state ASR drop is not consistently ordered for either.

5 Discussions: Elements of Reproducibility

Closer scrutiny of our empirical results surfaces eight sources of system robustness variability. These findings hint that one single scalar ASR is not sufficiently descriptive or informative about CF robustness under modern agentic attacks and defenses and we encourage practitioners to consider the following elements for reproducibility and completeness of metric reporting:

5.1 Explicit Controlled Variables. (1) Candidate count (k) and **(2) catalog concentration (ρ)** are the primary independent variables swept in all experiments.

5.2 System Structure. (3) Temporal regime. Nearly all attacks and defenses have an early-stage growth or ramp-up regime and reach equilibrium in attack metrics. We find that these two regimes can rank configurations differently. This temporal-phase distinction echoes concurrent observations in feedback-loop analysis [34]. **(4) Victim’s role asymmetry.** The bipartite role asymmetry across users and items is an inherent property peculiar to CF systems, and we observe distinct behaviours in response to attacks and defenses.

Table 1: Effect of increasing k and ρ on attack and defense outcomes, measured by transient slope (tr) and steady-state mean (ss) of user and item ASR (attacks) or Δ ASR (defenses). $\nearrow$: higher connectivity yields higher attack or defense efficacy; $-$: no meaningful difference; $\updownarrow$: curve behaviors differ but lack consistent ordering.

Attack	Class	$\uparrow k$				$\uparrow \rho$			
		U_{tr}	I_{tr}	U_{ss}	I_{ss}	U_{tr}	I_{tr}	U_{ss}	I_{ss}
CORBA	Dissem.	$\nearrow$	$\nearrow$	$\nearrow$	$\nearrow$	$\updownarrow$	$\nearrow$	$\updownarrow$	$\updownarrow$
NetSafe	Dissem.	$\updownarrow$	$\updownarrow$	$\updownarrow$	$\nearrow$	$\updownarrow$	$\nearrow$	$\updownarrow$	$\nearrow$
TOMA	Dissem.	$\nearrow$	$\nearrow$	$\nearrow$	$\nearrow$	$\updownarrow$	$\updownarrow$	$-$	$\nearrow$
MASTER	Dissem.	$\nearrow$	$\nearrow$	$\updownarrow$	$\updownarrow$	$\updownarrow$	$\updownarrow$	$\updownarrow$	$\updownarrow$
MAMA	Extract.	$\nearrow$	$\updownarrow$	$-$	$-$	$-$	$-$	$-$	$-$
MASLeak	Extract.	$\nearrow$	$\updownarrow$	$-$	$-$	$-$	$-$	$-$	$-$
TOMA	Extract.	$\nearrow$	$\updownarrow$	$\nearrow$	$\nearrow$	$\nearrow$	$\nearrow$	$\nearrow$	$\nearrow$
MASTER	Extract.	$-$	$-$	$\nearrow$	$\nearrow$	$\nearrow$	$\nearrow$	$\nearrow$	$\nearrow$
G-Safeguard	Defense	$\updownarrow$	$\updownarrow$	$\updownarrow$	$\updownarrow$	$\updownarrow$	$\updownarrow$	$\updownarrow$	$\updownarrow$
BlindGuard	Defense	$\updownarrow$	$\updownarrow$	$\updownarrow$	$\updownarrow$	$\updownarrow$	$\updownarrow$	$\updownarrow$	$\updownarrow$
T-Guard	Defense	$\nearrow$	$\nearrow$	$\nearrow$	$\nearrow$	$\updownarrow$	$\updownarrow$	$\updownarrow$	$\updownarrow$
M-Guard	Defense	$\nearrow$	$\nearrow$	$\updownarrow$	$\updownarrow$	$\nearrow$	$\nearrow$	$\nearrow$	$\nearrow$

Extraction metrics are either user-only (PII leakage) or role-agnostic (reverse engineering) so U/I columns are merged. Variations in ρ induce no meaningful performance difference in solely extraction attacks.

5.3 Attack Design. (5) Attacker placement. The attacker set is characterised by a tuple (α_U, α_I) describing attacker concentration and user/item role split, ablated in §6.1. **(6) Attacker’s prompt style.** Attack style is ablated in §6.2, where four RecSys-native attacks evaluated through the lens of NetSafe reveal how entry point and payload style shift attack outcomes given the same connectivity.

5.4 Connectivity-Agnostic. (7) LLM safety alignment. The LLM’s safety alignment acts as a *topology-agnostic recovery force*. Higher connectivity drives malicious spread, but also brings more benign interactions that facilitates recovery. We replicate experiments across a few more LLM backbones in §6.3 to demonstrate how the choice of LLM affects this aspect. **(8) Task and system prompt design.** Task and domain affect what injections are plausible, and both are difficult to quantify independently of the LLM and attack payload. We only study classical CF on MovieLens variants, but behaviour could vary across domains (e.g., private datasets the LLM has not seen, niche subject areas), especially when confounded by the LLM’s world knowledge. This dimension is out of scope for this study.

6 Ablations

We test robustness of the connectivity modulation findings across various configurations (attacker concentration, role split, and LLM backbone), using NetSafe unless specified otherwise.

6.1 Attacker Concentration. We repeat NetSafe across $(\alpha_U, \alpha_I) \in \{0.10, 0.25, 0.50, 0.75\}^2$. Three takeaways emerge (Fig 8): (1) *Higher attacker concentration shortens the transient regime*: higher α steepens the initial slope and brings the plateau forward. (2) *Qualitative behaviours remain similar*: users generally exhibit larger variance

and turbulence (recovery through benign interactions) than items regardless of α . (3) *Asymmetric partition sensitivity*: higher α yields higher ASR in both role partitions, but user-side contamination is more sensitive to α changes (larger gap between $\alpha=0.10$ and 0.75), whereas item-side contamination varies less across α values.

6.2 Attacker Style and Role Split. We evaluate three RecSys-native attacks that differ in attacker placement and payload style (Fig. 9) through the lens of NetSafe. Two key observations surface:

Attack	U	I	Payload Style	Injection Surface
NetSafe [67]	✓	✓	canary concept via system prompt	system prompt
CheatAgent [32]	✓	✓	canary concept in user memory	user profile
DrunkAgent [64]	✓	✓	canary concept in item memory	item profile
RecTextAttack [73]	✓	✓	lexical perturbation of item text	item profile

(1) *Semantic attacks substantially outperform lexical attacks.* Under identical attacker placement ($\alpha_U=0, \alpha_I>0$) and connectivity, DrunkAgent (semantic, Fig. 9c) outperforms RecTextAttack (lexical, Fig. 9d) by a significant margin. The CF memory update step rewrites profiles as coherent natural-language summaries, silently correcting lexically perturbed text while preserving semantically injected content, making the latter more durable. (2) *Role asymmetry in CF propagation.* DrunkAgent (item-side entry, Fig. 9c) contaminates users directly and propagates back to items effectively. In contrast, CheatAgent (user-side entry, Fig. 9b) reaches items directly, but struggles to spread among users via a two-hop path through items. Users seem to be more resilient to higher-order, multi-hop dissemination.

6.3 Cross-LLM Consistency. We replicate NetSafe experiments with alternative LLMs in Fig. 10. Notably, systems built upon newer models are harder to attack, likely due to better alignment, extensive world knowledge, and thus, more chances for reflection-driven self-correction of bias and misinformation during profile updates. More detailed analysis is available in the linked codebase.

7 Predictive Metrics

Simulations of large-scale CF systems are expensive, and it is desirable to rank attack severity across system configurations from graph-structural quantities alone, without running full experiments. Our empirical observations suggest that a useful predictor would satisfy the following: (i) it is defined *per role partition* (user-side vs. item-side), since the two agent populations exhibit distinct contamination dynamics; (ii) it captures *expected per-turn contact frequency* between attacker and victim agents; and (iii) it accounts for *LLM alignment-driven recovery*, or the tendency of non-attacker victim agents to self-correct through subsequent benign interactions and reflections; (iv) the transient growth slope and the steady-state contamination level should be treated as separate targets.

Such patterns are qualitatively consistent with epidemic spreading on bipartite networks [35]: recovery confers no immunity, so agents re-enter the susceptible state immediately, as in Susceptible-Infected-Susceptible (SIS) dynamics concurrently explored in the general MAS safety community [21, 42, 57, 60, 80]. We adopt a two-population (user/item) homogeneous mean-field approximation since k is fixed by configuration and the catalog uniformly resampled, leaving agent degrees near-homogeneous by construction. Per-turn exposure follows the discrete-time contact form of [7].

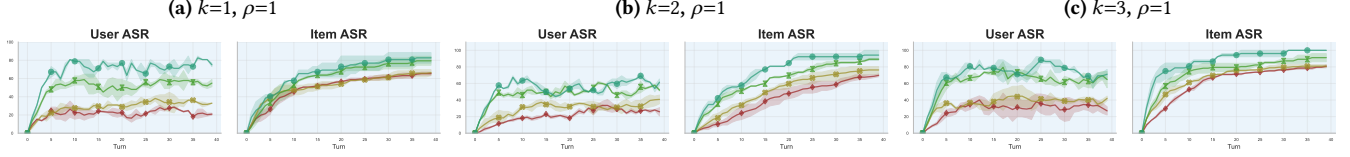


Figure 8: Attacker ratio ablation. Each panel fixes a topology config; curves show 10%/25%/50%/75% attackers.

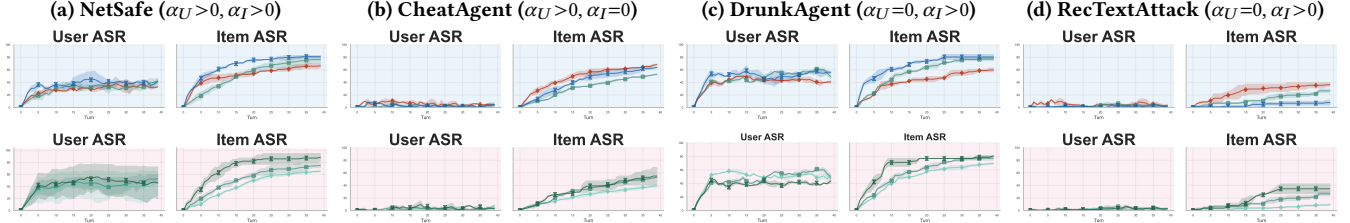


Figure 9: Adapted RecSys-native attacks that differ in styles and (α_U, α_I) , evaluated through the lens of connectivity k and ρ .

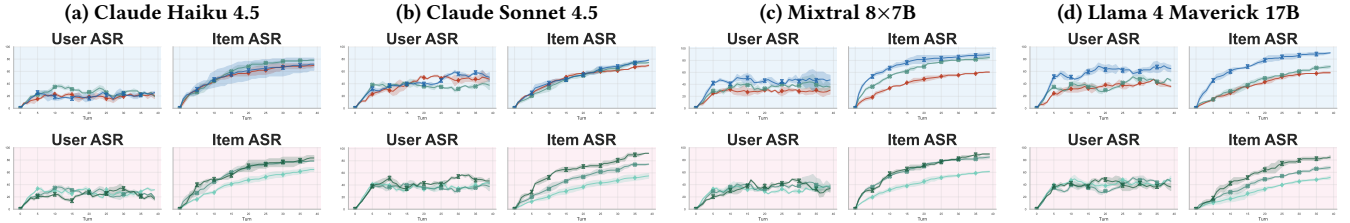


Figure 10: Cross-LLM consistency ablation (NetSafe, 25% attackers). Top: k ; bottom: ρ . Claude 4.5 variants exhibit better robustness, likely due to better world knowledge and stronger recovery from benign interaction.

Infection is driven by connectivity-driven exposure at partition-specific susceptibility scales β_U, β_I , offset by a contact-mediated, alignment-driven recovery rate γ and a system-level additive adjustment r :

$$\begin{aligned} \frac{d\rho_U}{dt} &\approx \beta_U(1-\rho_U)[1-(1-\rho_I)^k] - \rho_U[(1-\rho_I) \cdot k \cdot \gamma + r], \\ \frac{d\rho_I}{dt} &\approx \beta_I(1-\rho_I)[1-(1-\rho_U)^{\rho k}] - \rho_I[(1-\rho_U) \cdot \rho \cdot k \cdot \gamma + r] \end{aligned} \quad (1)$$

where an item agent receives $n_U k / n_I = \rho k$ user contacts per turn against a user agent's k . At steady state ($d\rho_U/dt = 0$), we freeze ρ_I at the seeded attacker prevalence α_I : a *first-order* (one-hop) approximation in which only seeded attackers transmit, disregarding secondary transmission from contaminated victims. Solving for the steady-state prevalence and reporting the contamination odds gives the **Recovery-Aware First-Order Connectivity** predictor, each partition's β absorbed into its own γ and r :

$$R_U = \frac{1 - (1 - \alpha_I)^k}{(1 - \alpha_I) \cdot k \cdot \gamma + r}, \quad R_I = \frac{1 - (1 - \alpha_U)^{\rho k}}{(1 - \alpha_U) \cdot \rho \cdot k \cdot \gamma + r} \quad (2)$$

by symmetric derivation on the item side. Correlation results are available in the linked codebase, where (γ, r) are empirically fitted per attack and partition to maximise mean Spearman correlation. We observe that predictor-outcome correlations are uneven across attackers' goals, roles, and temporal regimes, suggesting this first-order formulation is an over-simplification and that incorporating higher-order contact dynamics would be necessary for more accurate characterization.

8 Conclusions and Future Work

We systematically study how candidate count (k) and catalog concentration (ρ) shape vulnerability in multi-agent CF systems. Six attacks and four defenses are reproduced in the agentic CF setting and evaluated through the lens of connectivity. The two axes affect user and item partitions asymmetrically, and transient growth differently from steady-state contamination, yielding richer connectivity-vulnerability patterns than in general MAS; each is best tracked as a distinct risk. Practically, (1) more homogeneous user cohorts ($\rho \uparrow$) and higher user fan-out ($k \uparrow$) may amplify individual agent impact and warrant care when scaling or reconfiguring CF systems; (2) while LLM alignment drives recovery from contamination (§6.3), connectivity-aware guardrails remain necessary. We summarize eight key elements of reproducibility for CF in §5, which may help structure safety evaluation metric reporting in future work.

Limitations. (1) *Scale & Domain*: All experiments are conducted on a 100-user MovieLens cohort; generalisation to larger, sparser, or power-law catalog distributions remains open. (2) *Potential Confounders*: Adjusting k co-varies with prompt and ranking task structures, while resampling for ρ shifts item degree distributions. Observed effects are reported under these coupled changes. (3) *Utility Trade-off*: We focus on safety metrics (ASR, leakage); measuring recommendation quality impact (e.g., HR@k, NDCG) under active defenses could complement this view. (4) More systematic calibration of LLM judges, sensitivity to the judge's underlying model, statistical significance across repeated seeds, and adaptive threat models represent important avenues for future extension.

References

- [1] Mert Cemri, Melissa Z. Pan, Shuyi Yang, Lakshya A. Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. 2025. Why Do Multi-Agent LLM Systems Fail? arXiv:2503.13657 <https://arxiv.org/abs/2503.13657>
- [2] Yin Cheng, Liao Zhou, Xiyu Liang, Dihao Luo, Tewe Lee, Kailun Zheng, Weiwei Zhang, Mingchen Cai, Jian Dong, and Andy Zhang. 2026. Let the Agent Steer: Closed-Loop Ranking Optimization via Influence Exchange. arXiv:2603.27765 <https://arxiv.org/abs/2603.27765>
- [3] Yashar Deldjoo, Tommaso Di Noia, and Felice Antonio Merra. 2021. A Survey on Adversarial Recommender Systems: from Attack/Defense Strategies to Generative Adversarial Networks. doi:10.1145/3439729
- [4] Yang Feng and Xudong Pan. 2025. SentinelNet: Safeguarding Multi-Agent Collaboration Through Credit-Based Dynamic Threat Detection. arXiv:2510.16219 <https://arxiv.org/abs/2510.16219>
- [5] Adam Fournay, Gagan Bansal, Hussein Mozannar, Cheng Tan, Eduardo Salinas, Erkan Eric Zhu, Friederike Niedtner, Grace Proebsting, Griffin Bassman, Jack Gerrits, Jacob Alber, Peter Chang, Ricky Loynd, Robert West, Victor Dibia, Ahmed Awadallah, Ece Kamar, Rafah Hosn, and Saleema Amershi. 2024. Magentic-One: A Generalist Multi-Agent System for Solving Complex Tasks. arXiv:2411.04468 <https://arxiv.org/abs/2411.04468>
- [6] Mohammad Sajjad Ghaemi. 2025. A Survey of Collusion Risk in LLM-Powered Multi-Agent Systems. In *NeurIPS 2025 Workshop on Responsible Foundation Models*. <https://openreview.net/forum?id=Ylh8617Qyd>
- [7] S. Gómez, A. Arenas, J. Borge-Holthoefer, S. Meloni, and Y. Moreno. 2010. Discrete-time Markov chain approach to contact-based disease spreading in complex networks. doi:10.1209/0295-5075/89/38009
- [8] F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. 19 pages. doi:10.1145/2827872
- [9] Mengkang Hu, Yuhang Zhou, Wendong Fan, Yuzhou Nie, Bowei Xia, Tao Sun, Ziyu Ye, Zhaoxuan Jin, Yingru Li, Qiguang Chen, Zeyu Zhang, Yifeng Wang, Qianshuo Ye, Bernard Ghanem, Ping Luo, and Guohao Li. 2025. OWL: Optimized Workforce Learning for General Multi-Agent Assistance in Real-World Task Automation. arXiv:2505.23885 <https://arxiv.org/abs/2505.23885>
- [10] Chengkai Huang, Junda Wu, Yu Xia, Zixu Yu, Ruhan Wang, Tong Yu, Ruiyi Zhang, Ryan A. Rossi, Branislav Kveton, Dongruo Zhou, Julian McAuley, and Lina Yao. 2025. Towards Agentic Recommender Systems in the Era of Multimodal Large Language Models. arXiv:2503.16734 [cs.AI] <https://arxiv.org/abs/2503.16734>
- [11] Chengkai Huang, Junda Wu, Tong Yu, Julian McAuley, and Lina Yao. 2025. A Tutorial on Agentic LLM for Recommender Systems. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems (RecSys '25)*. Association for Computing Machinery, New York, NY, USA, 1417–1419. doi:10.1145/3705328.3748007
- [12] Xu Huang, Jianxun Lian, Yuxuan Lei, Jing Yao, Defu Lian, and Xing Xie. 2023. Recommender AI Agent: Integrating Large Language Models for Interactive Recommendations. arXiv:2308.16505 <https://arxiv.org/abs/2308.16505>
- [13] Yue Huang, Yu Jiang, Wenjie Wang, Haomin Zhuang, Xiaonan Luo, Yuchen Ma, Zhangchen Xu, Zichen Chen, Nuno Moniz, Zinan Lin, Pin-Yu Chen, Nitesh V Chawla, Nouha Dziri, Huan Sun, and Xiangliang Zhang. 2026. Emergent Social Intelligence Risks in Generative Multi-Agent Systems. arXiv:2603.27771 <https://arxiv.org/abs/2603.27771>
- [14] Zheng Hui, Xiaokai Wei, Yexi Jiang, Kevin Gao, Chen Wang, Frank Ong, Seun Yoon, Rachit Pareek, and Michelle Gong. 2025. Toward Safe and Human-Aligned Game Conversational Recommendation via Multi-Agent Decomposition. arXiv:2504.20094 <https://arxiv.org/abs/2504.20094>
- [15] Rishi Jha, Harold Friedman, Justin Wagle, and Vitaly Shmatikov. 2025. Breaking and Fixing Defenses Against Control-Flow Hijacking in Multi-Agent Systems. arXiv:2510.17276 <https://arxiv.org/abs/2510.17276>
- [16] Chumeng Jiang, Jiayin Wang, Weizhi Ma, Charles L. A. Clarke, Shuai Wang, Chuhan Wu, and Min Zhang. 2024. Beyond Utility: Evaluating LLM as Recommender. arXiv:2411.00331 <https://arxiv.org/abs/2411.00331>
- [17] Ishan Kavathekar, Hemang Jain, Ameya Rathod, Ponnurangam Kumaraguru, and Tanuja Ganu. 2025. TAMAS: Benchmarking Adversarial Risks in Multi-Agent LLM Systems. arXiv:2511.05269 <https://arxiv.org/abs/2511.05269>
- [18] Rana Muhammad Shahroz Khan, Zhen Tan, Sukwon Yun, Charles Fleming, and Tianlong Chen. 2025. *Agents Under Siege*: Breaking Pragmatic Multi-Agent LLM Systems with Optimized Prompt Attacks. arXiv:2504.00218 <https://arxiv.org/abs/2504.00218>
- [19] LangManus Contributors. 2025. *LangManus: A Community-Driven AI Automation Framework*. <https://github.com/Darwin-lfl/langmanus>
- [20] Donghyun Lee and Mo Tiwari. 2024. Prompt Infection: LLM-to-LLM Prompt Injection within Multi-Agent Systems. arXiv:2410.07283 <https://arxiv.org/abs/2410.07283>
- [21] Jun Leng, Yu Liu, Litian Zhang, Ruihan Hu, Zhuting Fang, and Xi Zhang. 2025. From static to adaptive: immune memory-based jailbreak detection for large language models. arXiv:2512.03356 <https://arxiv.org/abs/2512.03356>
- [22] Kaihong Li, Huichi Zhou, Bin Ma, and Fangjun Huang. 2025. Semantic-Shield: LLM-Powered Audits Expose Shilling Attacks in Recommender Systems. arXiv:2509.24961 <https://arxiv.org/abs/2509.24961>
- [23] Jianxun Lian, Yuxuan Lei, Xu Huang, Jing Yao, Wei Xu, and Xing Xie. 2024. RecAI: Leveraging Large Language Models for Next-Generation Recommender Systems. In *Companion Proceedings of the ACM Web Conference 2024* (Singapore, Singapore) (WWW '24). Association for Computing Machinery, New York, NY, USA, 1031–1034. doi:10.1145/3589335.3651242
- [24] Ruichao Liang, Le Yin, Jing Chen, Yebo Feng, Cong Wu, Xiaoyu Zhang, Huangpeng Gu, Zijian Zhang, and Yang Liu. 2025. Don't Trust Your Upstream: Exploiting LLM Multi-Agent System via Topology-Guided Adversarial Propagation. arXiv:2512.04129 <https://arxiv.org/abs/2512.04129>
- [25] Chen Lin, Si Chen, Meifang Zeng, Sheng Zhang, Min Gao, and Hui Li. 2024. Shilling Black-Box Recommender Systems by Learning to Generate Fake User Profiles. *IEEE Transactions on Neural Networks and Learning Systems* 35, 1, 1305–1319. doi:10.1109/tnnls.2022.3183210
- [26] Jinbo Liu, Defu Cao, Yifei Wei, Tianyao Su, Yuan Liang, Yushun Dong, Yan Liu, Yue Zhao, and Xiyang Hu. 2025. Topology Matters: Measuring Memory Leakage in Multi-Agent LLMs. arXiv:2512.04668 <https://arxiv.org/abs/2512.04668>
- [27] Jiahao Liu, Shengkang Gu, Dongsheng Li, Guangping Zhang, Mingzhe Han, Hansu Gu, Peng Zhang, Tun Lu, Li Shang, and Ning Gu. 2025. AgentCF++: Memory-enhanced LLM-based Agents for Popularity-aware Cross-domain Recommendations. arXiv:2502.13843 <https://arxiv.org/abs/2502.13843>
- [28] Reza Yousefi Maragheh and Yashar Deldjoo. 2025. The Future is Agentic: Definitions, Perspectives, and Open Challenges of Multi-Agent Recommender Systems. arXiv:2507.02097 <https://arxiv.org/abs/2507.02097>
- [29] Rui Miao, Yixin Liu, Yili Wang, Xu Shen, Yue Tan, Yiwei Dai, Shirui Pan, and Xin Wang. 2025. BlindGuard: Safeguarding LLM-based Multi-Agent Systems under Unknown Attacks. arXiv:2508.08127 <https://arxiv.org/abs/2508.08127>
- [30] Thanh Toan Nguyen, Quoc Viet Hung Nguyen, Thanh Tam Nguyen, Thanh Trung Huynh, Thanh Thi Nguyen, Matthias Weidlich, and Hongzhi Yin. 2024. Manipulating Recommender Systems: A Survey of Poisoning Attacks and Countermeasures. arXiv:2404.14942 <https://arxiv.org/abs/2404.14942>
- [31] Liangbo Ning, Wenqi Fan, and Qing Li. 2025. Exploring Backdoor Attack and Defense for LLM-empowered Recommendations. arXiv:2504.11182 <https://arxiv.org/abs/2504.11182>
- [32] Liang-bo Ning, Shijie Wang, Wenqi Fan, Qing Li, Xin Xu, Hao Chen, and Feiran Huang. 2024. CheatAgent: Attacking LLM-Empowered Recommender Systems via LLM Agent. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24)*. ACM, 2284–2295. doi:10.1145/3637528.3671837
- [33] Junjun Pan, Yixin Liu, Rui Miao, Kaize Ding, Yu Zheng, Quoc Viet Hung Nguyen, Alan Wee-Chung Liew, and Shirui Pan. 2025. Explainable and Fine-Grained Safeguarding of LLM Multi-Agent Systems via Bi-Level Graph Anomaly Detection. arXiv:2512.18733 <https://arxiv.org/abs/2512.18733>
- [34] Donguk Park, Dongwon Lee, and Yeon-Chang Lee. 2026. Echoes in the Loop: Diagnosing Risks in LLM-Powered Recommender Systems under Feedback Loops. arXiv:2602.07442 <https://arxiv.org/abs/2602.07442>
- [35] Romualdo Pastor-Satorras, Claudio Castellano, Piet Van Mieghem, and Alessandro Vespignani. 2015. Epidemic processes in complex networks. 925–979 pages. <https://link.aps.org/doi/10.1103/RevModPhys.87.925>
- [36] Vaidehi Patil, Elias Stengel-Eskin, and Mohit Bansal. 2025. The Sum Leaks More Than Its Parts: Compositional Privacy Risks and Mitigations in Multi-Agent Collaboration. arXiv:2509.14284 <https://arxiv.org/abs/2509.14284>
- [37] Qi Yao Peng, Hongtao Liu, Hua Huang, Qing Yang, and Minglai Shao. 2025. A Survey on LLM-powered Agents for Recommender Systems. arXiv:2502.10050 <https://arxiv.org/abs/2502.10050>
- [38] Markus Schedl, Vito Walter Anelli, and Elisabeth Lex. 2023. Trustworthy Recommender Systems: Technical, Ethical, Legal, and Regulatory Perspectives. In *Proceedings of the 17th ACM Conference on Recommender Systems* (Singapore, Singapore) (RecSys '23). Association for Computing Machinery, New York, NY, USA, 1288–1290. doi:10.1145/3604915.3609497
- [39] Yu Shang, Peijie Liu, Yuwei Yan, Zijong Wu, Leheng Sheng, Yuanqing Yu, Chumeng Jiang, An Zhang, Fengli Xu, Yu Wang, Min Zhang, and Yong Li. 2025. AgentRecBench: Benchmarking LLM Agent-based Personalized Recommender Systems. arXiv:2505.19623 <https://arxiv.org/abs/2505.19623>
- [40] Xu Shen, Yixin Liu, Yiwei Dai, Yili Wang, Rui Miao, Yue Tan, Shirui Pan, and Xin Wang. 2025. Understanding the Information Propagation Effects of Communication Topologies in LLM-based Multi-Agent Systems. arXiv:2505.23352 <https://arxiv.org/abs/2505.23352>
- [41] Wentao Shi, Xiangnan He, Yang Zhang, Chongming Gao, Xinyue Li, Jizhi Zhang, Qifan Wang, and Fuli Feng. 2024. 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 (SIGIR 2024)*. ACM, 1893–1903. doi:10.1145/3626772.3657683
- [42] Inderjeet Singh, Vikas Pahuja, Aishwariya Priya Rathina Sabapathy, Chiara Picardi, Amit Giloni, Roman Vainshtein, Andrés Murillo, Hisashi Kojima, Motoyoshi Sekiya, Yuki Unno, and Junichi Suga. 2026. Adversarial Intent is a Latent Variable: Stateful Trust Inference for Securing Multimodal Agentic RAG. arXiv:2602.21447 <https://arxiv.org/abs/2602.21447>

- [43] Param Thakkar and Anushka Yadav. 2024. Personalized Recommendation Systems using Multimodal, Autonomous, Multi Agent Systems. arXiv:2410.19855 <https://arxiv.org/abs/2410.19855>
- [44] Giovanni De Toni, Cristian Consonni, Erasmo Purificato, Emilia Gomez, and Bruno Lepri. 2026. With a Little Help From My Friends: Collective Manipulation in Risk-Controlling Recommender Systems. arXiv:2603.28476 <https://arxiv.org/abs/2603.28476>
- [45] Harold Triedman, Rishi Jha, and Vitaly Shmatikov. 2025. Multi-Agent Systems Execute Arbitrary Malicious Code. arXiv:2503.12188 <https://arxiv.org/abs/2503.12188>
- [46] Liwen Wang, Wenxuan Wang, Shuai Wang, Zongjie Li, Zhenlan Ji, Zongyi Lyu, Daoyuan Wu, and Shing-Chi Cheung. 2025. IP Leakage Attacks Targeting LLM-Based Multi-Agent Systems. arXiv:2505.12442 <https://arxiv.org/abs/2505.12442>
- [47] Lei Wang, Jingsen Zhang, Hao Yang, Zhiyuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Ruihua Song, Wayne Xin Zhao, Jun Xu, Zhicheng Dou, Jun Wang, and Ji-Rong Wen. 2024. User Behavior Simulation with Large Language Model based Agents. arXiv:2306.02552 [cs.IR] <https://arxiv.org/abs/2306.02552>
- [48] Shilong Wang, Guibin Zhang, Miao Yu, Guancheng Wan, Fanci Meng, Chongye Guo, Kun Wang, and Yang Wang. 2025. G-Safeguard: A Topology-Guided Security Lens and Treatment on LLM-based Multi-agent Systems. arXiv:2502.11127 <https://arxiv.org/abs/2502.11127>
- [49] Yuanrong Wang and Yingpeng Du. 2025. Shilling Recommender Systems by Generating Side-feature-aware Fake User Profiles. arXiv:2509.17918 <https://arxiv.org/abs/2509.17918>
- [50] Yancheng Wang, Ziyang Jiang, Zheng Chen, Fan Yang, Yingxue Zhou, Eunah Cho, Xing Fan, Xiaojiang Huang, Yanbin Lu, and Yingzhen Yang. 2023. RecMind: Large Language Model Powered Agent For Recommendation. arXiv:2308.14296 <https://arxiv.org/abs/2308.14296>
- [51] Yubo Wang, Min Tang, Nuo Shen, Shujie Cui, and Weiqing Wang. 2025. Privacy Risks of LLM-Empowered Recommender Systems: An Inversion Attack Perspective. arXiv:2508.03703 [cs.IR] <https://arxiv.org/abs/2508.03703>
- [52] Zongwei Wang, Min Gao, Junliang Yu, Xinyi Gao, Quoc Viet Hung Nguyen, Shazia Sadiq, and Hongzhi Yin. 2024. ID-Free Not Risk-Free: LLM-Powered Agents Unveil Risks in ID-Free Recommender Systems. arXiv:2409.11690 <https://arxiv.org/abs/2409.11690>
- [53] Zihao Wang, Tianhao Mao, XiaoFeng Wang, Di Tang, and Xiaozhong Liu. 2025. IndirectAD: Practical Data Poisoning Attacks against Recommender Systems for Item Promotion. arXiv:2511.05845 [cs.CR] <https://arxiv.org/abs/2511.05845>
- [54] Zongwei Wang, Junliang Yu, Tong Chen, Hongzhi Yin, Shazia Sadiq, and Min Gao. 2025. Towards Secure and Robust Recommender Systems: A Data-Centric Perspective. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining (Hannover, Germany) (WSDM '25)*. Association for Computing Machinery, New York, NY, USA, 1020–1023. doi:10.1145/3701551.3703484
- [55] Zongwei Wang, Junliang Yu, Min Gao, Wei Yuan, Guanhua Ye, Shazia Sadiq, and Hongzhi Yin. 2024. Poisoning Attacks and Defenses in Recommender Systems: A Survey. arXiv:2406.01022 [cs.CR] <https://arxiv.org/abs/2406.01022>
- [56] Zhefan Wang, Yuanqing Yu, Wendi Zheng, Weizhi Ma, and Min Zhang. 2024. MACRec: A Multi-Agent Collaboration Framework for Recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2760–2764. doi:10.1145/3626772.3657669
- [57] Qianshan Wei, Tengchao Yang, Yaochen Wang, Xinfeng Li, Lijun Li, Zhenfei Yin, Yi Zhan, Thorsten Holz, Zhiqiang Lin, and XiaoFeng Wang. 2025. A-MemGuard: A Proactive Defense Framework for LLM-Based Agent Memory. arXiv:2510.02373 <https://arxiv.org/abs/2510.02373>
- [58] Yuqian Wu, Yuhong Peng, Jiapeng Yu, and Raymond S. T. Lee. 2024. MAS4POI: a Multi-Agents Collaboration System for Next POI Recommendation. arXiv:2409.13700 [cs.IR] <https://arxiv.org/abs/2409.13700>
- [59] Yiqing Wu, Ruobing Xie, Zhao Zhang, Yongchun Zhu, FuZhen Zhuang, Jie Zhou, Yongjun Xu, and Qing He. 2023. Attacking Pre-trained Recommendation. arXiv:2305.03995 [cs.IR] <https://arxiv.org/abs/2305.03995>
- [60] Yutong Wu, Jie Zhang, Yiming Li, Chao Zhang, Qing Guo, Nils Lukas, and Tianwei Zhang. 2025. Cowpox: Towards the Immunity of VLM-based Multi-Agent Systems. arXiv:2508.09230 <https://arxiv.org/abs/2508.09230>
- [61] Yu Xia, Sungchul Kim, Tong Yu, Ryan A. Rossi, and Julian McAuley. 2025. Multi-Agent Collaborative Filtering: Orchestrating Users and Items for Agentic Recommendations. arXiv:2511.18413 <https://arxiv.org/abs/2511.18413>
- [62] Faouzi El Yagoubi, Godwin Badu-Marfo, and Ranwa Al Mallah. 2026. AgentLeak: A Full-Stack Benchmark for Privacy Leakage in Multi-Agent LLM Systems. arXiv:2602.11510 <https://arxiv.org/abs/2602.11510>
- [63] Bingyu Yan, Ziyi Zhou, Xiaoming Zhang, Chaozhao Li, Ruilin Zeng, Yirui Qi, Tianbo Wang, and Litian Zhang. 2025. Attack the Messages, Not the Agents: A Multi-round Adaptive Stealthy Tampering Framework for LLM-MAS. arXiv:2508.03125 <https://arxiv.org/abs/2508.03125>
- [64] Shiyi Yang, Zhibo Hu, Xinshu Li, Chen Wang, Tong Yu, Xiwei Xu, Liming Zhu, and Lina Yao. 2025. DrunkAgent: Stealthy Memory Corruption in LLM-Powered Recommender Agents. arXiv:2503.23804 <https://arxiv.org/abs/2503.23804>
- [65] Shiyi Yang, Xinshu Li, Guanglin Zhou, Chen Wang, Xiwei Xu, Liming Zhu, and Lina Yao. 2025. Retrieval-Augmented Review Generation for Poisoning Recommender Systems. arXiv:2508.15252 <https://arxiv.org/abs/2508.15252>
- [66] Haocheng Yu, Yaxiong Wu, Hao Wang, Wei Guo, Yong Liu, Yawen Li, Yuyang Ye, Junping Du, and Enhong Chen. 2025. Thought-Augmented Planning for LLM-Powered Interactive Recommender Agent. arXiv:2506.23485 <https://arxiv.org/abs/2506.23485>
- [67] Miao Yu, Shilong Wang, Guibin Zhang, Junyuan Mao, Chenlong Yin, Qijiong Liu, Qingsong Wen, Kun Wang, and Yang Wang. 2024. NetSafe: Exploring the Topological Safety of Multi-agent Networks. arXiv:2410.15686 <https://arxiv.org/abs/2410.15686>
- [68] Qiusi Zhan, Zhixiang Liang, Zifan Ying, and Daniel Kang. 2024. InjecAgent: Benchmarking Indirect Prompt Injections in Tool-Integrated Large Language Model Agents. arXiv:2403.02691 <https://arxiv.org/abs/2403.02691>
- [69] An Zhang, Yuxin Chen, Leheng Sheng, Xiang Wang, and Tat-Seng Chua. 2024. On Generative Agents in Recommendation. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*. <https://dl.acm.org/doi/abs/10.1145/3626772.3657844>
- [70] Hanrong Zhang, Jingyuan Huang, Kai Mei, Yifei Yao, Zhenting Wang, Chenlu Zhan, Hongwei Wang, and Yongfeng Zhang. 2025. Agent Security Bench (ASB): Formalizing and Benchmarking Attacks and Defenses in LLM-based Agents. In *International Conference on Learning Representations (ICLR)*. <https://arxiv.org/abs/2410.02644>
- [71] Jizhi Zhang, Keqin Bao, Wenjie Wang, Yang Zhang, Wentao Shi, Wanhong Xu, Fuli Feng, and Tat-Seng Chua. 2024. Prospect Personalized Recommendation on Large Language Model-based Agent Platform. arXiv:2402.18240 <https://arxiv.org/abs/2402.18240>
- [72] Junjie Zhang, Yupeng Hou, Ruobing Xie, Wenqi Sun, Julian McAuley, Wayne Xin Zhao, Leyu Lin, and Ji-Rong Wen. 2024. AgentCF: Collaborative Learning with Autonomous Language Agents for Recommender Systems. In *Proceedings of the ACM Web Conference 2024 (WWW)*. 3679–3689. doi:10.1145/3589334.3645537
- [73] Jinghao Zhang, Yuting Liu, Qiang Liu, Shu Wu, Guibing Guo, and Liang Wang. 2024. Stealthy Attack on Large Language Model based Recommendation. arXiv:2402.14836 <https://arxiv.org/abs/2402.14836>
- [74] Minxing Zhang, Zhaochun Ren, Zihan Wang, Pengjie Ren, Zhunmin Chen, Pengfei Hu, and Yang Zhang. 2021. Membership Inference Attacks Against Recommender Systems. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security (Virtual Event, Republic of Korea) (CCS '21)*. Association for Computing Machinery, New York, NY, USA, 864–879. doi:10.1145/3460120.3484770
- [75] Quan Zhang, Bingqi Zeng, Chijin Zhou, Gwhian Go, Heyuan Shi, and Yu Jiang. 2024. Human-Imperceptible Retrieval Poisoning Attacks in LLM-Powered Applications. arXiv:2404.17196 <https://arxiv.org/abs/2404.17196>
- [76] Tao Zhang, Kehui Yao, Luyi Ma, Jiao Chen, Reza Yousefi Maragheh, Kai Zhao, Jianpeng Xu, Evren Korpoglu, Sushant Kumar, and Kannan Achan. 2025. No-Human in the Loop: Agentic Evaluation at Scale for Recommendation. arXiv:2511.03051 <https://arxiv.org/abs/2511.03051>
- [77] Zaibin Zhang, Yongting Zhang, Lijun Li, Hongzhi Gao, Lijun Wang, Huchuan Lu, Feng Zhao, Yu Qiao, and Jing Shao. 2024. PsySafe: A Comprehensive Framework for Psychological-based Attack, Defense, and Evaluation of Multi-agent System Safety. arXiv:2401.11880 <https://arxiv.org/abs/2401.11880>
- [78] Yuyue Zhao, Jiancan Wu, Xiang Wang, Wei Tang, Dingxian Wang, and Maarten de Rijke. 2024. 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 (Washington DC, USA) (SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 1796–1806. doi:10.1145/3626772.3657828
- [79] Jialong Zhou, Lichao Wang, and Xiao Yang. 2025. GUARDIAN: Safeguarding LLM Multi-Agent Collaborations with Temporal Graph Modeling. arXiv:2505.19234 <https://arxiv.org/abs/2505.19234>
- [80] Yijin Zhou, Xiaoya Lu, Dongrui Liu, Junchi Yan, and Jing Shao. 2026. INFA-Guard: Mitigating Malicious Propagation via Infection-Aware Safeguarding in LLM-Based Multi-Agent Systems. arXiv:2601.14667 <https://arxiv.org/abs/2601.14667>
- [81] Zhenhong Zhou, Zherui Li, Jie Zhang, Yuanhe Zhang, Kun Wang, Yang Liu, and Qing Guo. 2025. CORBA: Contagious Recursive Blocking Attacks on Multi-Agent Systems Based on Large Language Models. arXiv:2502.14529 <https://arxiv.org/abs/2502.14529>
- [82] Lixi Zhu, Xiaowen Huang, and Jitao Sang. 2024. A LLM-based Controllable, Scalable, Human-Involved User Simulator Framework for Conversational Recommender Systems. arXiv:2405.08035 <https://arxiv.org/abs/2405.08035>
- [83] Yifan Zhu, Chao Zhang, Xin Shi, Xueqiao Zhang, Yi Yang, and Yawei Luo. 2025. MASTER: Multi-Agent Security Through Exploration of Roles and Topological Structures - A Comprehensive Framework. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (Eds.). Association for Computational Linguistics, Suzhou, China, 16895–16921. doi:10.18653/v1/2025.findings-emnlp.917