Adaptive Admission Attacks on Semantic-ID Generative Recommenders: Detectability, Payload Size, and Downstream Promotion

Anonymous authors
Paper under double-blind review

Abstract

Admission-time defenses screen newly logged interaction histories before a recommender is retrained, and are typically judged by how cleanly they separate a known attack from randomly resampled benign users. We show that this evaluation can substantially overstate security. Studying semantic-ID generative recommendation, we define attack success jointly as passing admission *and* promoting a target item after retraining, and construct a staged family of non-neural shilling attacks that adapts payload, profile length, item popularity, and transition structure to the gate, culminating in an oracle search against its score. On Amazon Beauty, a gate jointly constraining four cohort statistics rejects every high-payload construction, although adapted attacks evade each staged single-statistic gate; diagonal-covariance and calibrated marginal gates make identical decisions on these attacks, so this robustness stems from redundancy rather than correlation modeling. Direct search does find admitted attacks, but only at low payloads: accepted cohorts with 7, 16, and 24 target carriers promote the target to 0.00%, 0.11–0.38%, and 1.4% of eligible users, versus 5–14% for the rejected $m = 67$ attacks, and the same frontier appears on a second catalog. Admitted cohorts are more harmful for a medium-frequency target (3.6–3.9%) and under a non-generative SASRec recommender (4.7–6.6%). This high-payload protection, however, holds only under favorable evaluation. The same gate has a 0.11% false-positive rate on randomly resampled organic cohorts yet flags 26 of 33 time-ordered windows of genuine users, and attacks rejected in at least 95% of windows when they form half of a cohort pass almost always at a quarter or less, and still promote the target once admitted. We attribute both failures to cohort aggregation, which tightens calibration on exchangeable resamples while diluting each attacker's contribution, and argue that cohort-level admission defenses must be evaluated against adaptive attackers, temporally ordered benign traffic, and mixed cohorts. Code will be provided after acceptance.

1 Introduction

Recommender systems are vulnerable to data poisoning because attackers can inject fabricated user histories into the data used for model training (Zhang et al., 2020b). Once admitted, these histories can influence subsequent recommendations, making *data admission* a natural security boundary. Recent work has increasingly explored generative recommendation, where next-item retrieval is formulated as a generation problem. In particular, semantic-ID approaches represent each item as a short sequence of discrete tokens and generate the next item's identifier autoregressively, with early work demonstrating strong recommendation performance across multiple datasets (Rajput et al., 2023). Because these models are also trained on logged user interaction sequences, they inherit the same poisoning exposure. Prior work has shown that sequential recommenders can be manipulated through black-box data poisoning and injected interaction histories that promote targeted items

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

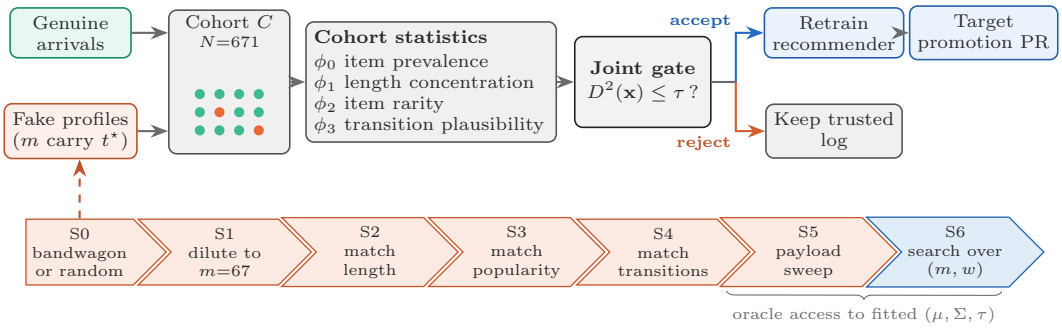

Figure 1: Overview. An attack succeeds only if its cohort passes the joint gate *and* then promotes $t^\star$; Stages 1–4 each repair one gate statistic.

(Yue et al., 2021; Zhao et al., 2025). We study whether such attacks can be contained at the admission boundary, which operates on interaction histories before recommender training, and evaluate their downstream consequences using a semantic-ID generative recommender. Our attacker is restricted to real catalog items and has no access to model gradients, item embeddings, or the tokenizer fit.

Whether admission screening provides meaningful security depends critically on how it is evaluated. We identify three assumptions that make an admission defense appear more robust than it is in deployment: the attacker is held fixed rather than adapting to the statistics inspected by the gate; benign calibration cohorts are formed by random resampling, implicitly treating future traffic as exchangeable with historical users; and malicious profiles are evaluated separately rather than mixed with genuine arrivals. Prior work has shown that defenses can lose their apparent robustness under adaptive evaluation (Carlini & Wagner, 2017; Tramèr et al., 2020; Koh et al., 2022), while statistical obfuscation of recommender shilling attacks has long been studied (Hurley et al., 2009; Cheng & Hurley, 2009). These issues are particularly important for cohort-level admission: aggregation can strengthen coordinated attack signals, but it can also amplify benign distributional shifts and dilute malicious signals as the attacker's share of a cohort decreases. We therefore ask whether an admission defense that separates a known attack from randomly resampled benign cohorts remains effective under defense-aware adaptation, temporally varying benign traffic, and mixed genuine–malicious admissions. To our knowledge, prior work has not jointly studied these factors and downstream retraining effects in the semantic-ID generative recommendation setting considered here.

An *admission cohort* is the batch of newly arrived histories considered for the next training update; the gate either accepts the cohort into the training log or rejects it and retains the previous trusted state. We therefore evaluate poisoning end-to-end: a successful attack must both *pass admission* and *promote the target after retraining*. Starting from a basic shilling construction, we develop a seven-stage non-neural adaptive family that successively adjusts target concentration, profile length, filler-item popularity, and transition structure, culminating in an oracle search over payload size and transition composition against the gate's own score. In the main controlled setting, the admission cohort contains $N = 671$ injected profiles, approximately 3% of the Beauty user population, while additional experiments test fixed constructions across further Amazon catalogs. The adaptive search is deliberately bounded: it tests whether the apparent security boundary survives a defense-aware attacker, not a worst case (Figure 1).

The resulting security picture differs substantially from static attack-versus-benign separation. At the working payload of $m = 67$ target-carrying profiles, increasingly adapted constructions remain rejected by a gate over four cohort statistics, while direct search finds admitted attacks at substantially lower payloads. This is not merely a search artifact: projecting the gate's acceptance region onto target prevalence gives an analytic ceiling of approximately 27 target carriers out of 671, while refined search finds accepted constructions through $m = 26$. The strongest accepted construction we retrained, with $m = 24$,

promotes the target to $1.38\% \pm 0.34\%$ of eligible users, compared with mean promotion of 9.5% and 7.4% for two rejected constructions at $m = 67$. Under the non-generative SASRec, the same accepted cohorts promote the target to 4.69% and 6.57%, so their harm is not specific to semantic-ID generation. These results reveal an admission–payload–promotion frontier rather than complete attack elimination. The apparent protection also depends strongly on the evaluation protocol: a gate with a 0.11% false-positive rate on randomly resampled organic cohorts rejects 26 of 33 consecutive time-ordered genuine windows, while attacks rejected in 95–99.5% of windows when attackers form half of the cohort are rejected in only 1.5% at a 25% attacker share and in none at 10%. Full-covariance, diagonal-covariance, and calibrated union-of-marginals gates make the same decisions on the tested payload-bearing constructions, indicating that high-payload rejection comes from redundant calibrated statistics, dominated by item prevalence, rather than from correlation modeling. Together, these findings expose a three-way tension among attack effectiveness, detectability, and robustness to benign cohort variation, and show why static separation alone is insufficient evidence of operational security. The main contributions and findings are as follows:

1. **Adaptive end-to-end evaluation of the admission boundary.** We formulate poisoning success jointly through admission and post-retraining target promotion, and construct a staged defense-aware attack family culminating in oracle search against the admission score.
2. **An admission–payload–promotion frontier.** We show that adaptive evasion remains possible in a lower-payload regime, derive an analytic payload ceiling that closely aligns with the observed acceptance region, and measure the downstream effect of attacks that actually pass admission under both a semantic-ID and a non-generative recommender, on two catalogs.
3. **Deployment failure modes of cohort aggregation.** We show that robustness measured on randomly resampled benign cohorts does not transfer to time-ordered traffic, while mixing malicious histories with genuine arrivals sharply reduces detectability, and relate both failures to cohort aggregation.

2 Related Work

Shilling attacks, poisoning, and recommender defenses. Statistical adaptation to recommender defenses predates modern learned attacks: early work studied shilling profiles designed to resemble genuine users and later developed diverse or obfuscated attacks intended to evade detection (Hurley et al., 2009; Cheng & Hurley, 2009). Classical defenses exploit behavioral signals such as rating deviation, profile similarity, and target-item patterns (Chirita et al., 2005; Mehta & Nejdl, 2009; Zhou et al., 2015), while group-level methods detect coordinated or collusive profiles and their targets (Zhang et al., 2020a; Wang et al., 2022). More recent defenses use LLM-based calibration or semantic auditing (Zhang et al., 2024; Li et al., 2026), as well as learned graph representations for group-shilling detection (Xu et al., 2026). Poisoning has also been studied directly in sequential recommendation through black-box and model-extraction-based attacks (Zhang et al., 2020b; Yue et al., 2021), while learned shilling attacks use generative models to balance attack effectiveness and stealthiness (Lin et al., 2022; Qiao et al., 2025). Semantic-ID generative recommenders represent items as sequences of discrete tokens and generate next-item identifiers from interaction histories (Rajput et al., 2023; Liu et al., 2026). Because they are trained on logged interaction sequences, they remain exposed to poisoning through injected histories. Existing defenses primarily identify suspicious users or groups, whereas our setting evaluates an arriving cohort before retraining and then measures the downstream effect of admitting that same cohort.

Adaptive evaluation of defenses. Robustness against a fixed attack need not persist when the attacker adapts to the defense, a principle established across adversarial machine learning and data poisoning (Carlini & Wagner, 2017; Tramèr et al., 2020; Koh et al., 2022). In poisoning specifically, filtering and sanitization defenses have been studied against stronger attackers that explicitly account for defense constraints (Steinhardt et al., 2017;

Koh et al., 2022). Our setting brings adaptive evaluation to the admission-time boundary of recommendation: the attacker adapts to the statistics used by the gate, while success requires the same cohort to pass admission and subsequently produce target promotion after retraining.

Temporal calibration and cohort-level admission. Temporal information has previously been used to detect recommender shilling through profile-level temporal features and time-series behavior (Hao & Zhang, 2018; Xu & Zhang, 2019). Our use of time is different: rather than constructing a timestamp-based detector, we test whether a gate calibrated on randomly resampled historical users retains its behavior on chronologically ordered genuine traffic. This question relates to the broader statistical observation that procedures calibrated under stationary or exchangeable data can become miscalibrated under temporal dependence or distribution shift (Alwan & Roberts, 1988; Barber et al., 2023). We additionally study cohort composition, because an admission rule applied to pooled arrivals can behave differently when malicious profiles are mixed with genuine users. Our joint score is a standard Hotelling-type T^2 statistic from multivariate process monitoring, and statistical obfuscation of shilling profiles dates back to Hurley et al. (2009); we claim neither the gate nor the attack family as novel. Our contribution is the end-to-end evaluation protocol: admission and post-retraining promotion measured jointly under defense-aware adaptation, time-ordered benign traffic, and mixed genuine–malicious cohorts.

3 Admission Setting and Threat Model

Admission setting. Let C denote the cohort of newly arrived interaction histories considered for the next model update. A cohort contains N profiles, of which m may carry the target item $t^\star$; we call m/N the *payload fraction* and a cohort *payload-bearing* when $m > 0$. Our controlled experiments use attacker-generated target-carrying and camouflage histories, while later experiments also mix malicious and genuine arrivals.

Attacker and defender. The attacker can create new accounts and fabricate histories using only real catalog items, but cannot modify genuine histories or access recommender gradients, item embeddings, or the semantic-ID tokenizer fit. Its objective is to pass admission while increasing the appearance of a chosen target $t^\star$ in genuine users' top-10 recommendations after retraining. The attacker knows the admission rule and training pipeline; Stages 0–4 use aggregate catalog statistics, while Stages 5–6 additionally assume oracle access to the gate's fitted statistics and threshold (Section 4). The defender constructs the gate's reference statistics from a trusted historical log and makes a single decision for the entire arriving cohort.

Admission and downstream effect. For cohort representation $\mathbf{x}(C)$, gate score s, and threshold τ, define

$$A(C) = \mathbf{1}[s(\mathbf{x}(C)) \le \tau]\,. \tag{1}$$

A payload-bearing construction is *evasive* when $A(C) = 1$.

Let $\mathrm{PR}(C)$ denote the fraction of eligible genuine users whose post-retraining top-10 list contains $t^\star$ when the corresponding clean-model list did not. We report $\mathrm{PR}(C)$ directly rather than imposing a threshold for harmful promotion. Admission and downstream effect are therefore measured separately.

4 Adaptive Attack Family

We construct a sequence of increasingly defense-aware attacks, each modifying one observable property of the preceding construction (Table 1). For each profile, we draw its length L and generate filler items according to the current stage. In each of the m target carriers, $t^\star$ replaces the filler at zero-based position $L - 3$; the remaining $N - m$ profiles contain only filler at that position. Implementation details on target exclusion, repeats, and short profiles are given in Table 3 (Appendix A).

Table 1: Adaptive attack family on Amazon Beauty ($N = 671$). m denotes target-carrying profiles; "Trained" marks conditions used for downstream evaluation.

Stage	Target carriers m	Length L	Filler generation	Trained
0 (bandwagon)	671	fixed, 9	uniform over top-20 popular items	yes
0 (random)	671	fixed, 9	uniform over catalog	yes
1: payload dilution	67	fixed, 9	uniform over catalog	no
2: length matching	67	organic length distribution	uniform over catalog	yes
3: popularity matching	67	organic length distribution	item-popularity distribution	yes
4: transition matching	67	organic length distribution	first-order transition walk	yes
5: payload sweep	$\{0^\dagger, 1, 3, 7, 16, 33, 50, 67\}$	organic length distribution	Stage 4 generator	selected values
6: payload–blend search	$\{0, 7, 16, 33, 50, 67\}$	organic length distribution	transition/popularity mixture	selected values

†Payload-free control, not an attack.

Target placement. The final two positions are reserved for validation and test, so $L-3$ is the final supervised training target of the evaluated leave-one-out pipeline; this placement is specific to that pipeline.

Stages 0–4: progressive adaptation. Stage 0 uses either bandwagon fillers sampled from the 20 most popular items or uniformly random fillers from the catalog. Stage 1 reduces the payload to $m = 67$ ($67/671 \approx 9.99\%$), immediately below the hand-set 10% item-concentration operating point. Stage 2 samples L from the organic length distribution, Stage 3 additionally matches the organic item-popularity marginal, and Stage 4 replaces independent filler draws with a first-order walk over the empirical adjacent-item table, falling back to popularity sampling when no continuation is observed. Each stage therefore repairs one discrepancy exposed by the admission statistics.

Stage 5: payload sweep. Holding the Stage 4 generator fixed, Stage 5 varies $m \in \{0, 1, 3, 7, 16, 33, 50, 67\}$ to isolate the effect of target concentration. The $m = 0$ condition is a payload-free control. Stage 4 and Stage 5 at $m = 67$ are independent realizations of the same underlying construction rather than distinct attack algorithms.

Stage 6: payload–blend search. Stage 6 introduces a transition-mixture parameter w: each filler follows the empirical transition table with probability w and the popularity marginal with probability $1 - w$. Thus $w = 0$ gives popularity matching and $w = 1$ gives the Stage 4 transition generator. Let $C_{m,w}$ denote a cohort constructed with payload m and blend w, and let D^2 be the joint-gate score of Eq. (4). For each payload m, we search a finite grid $\mathcal{W} \subset [0, 1]$ and select

$$w^*(m) \in \arg\min_{w \in \mathcal{W}} D^2(\mathbf{x}(C_{m,w})) \,. \tag{2}$$

The grid values are listed in Table 3, and the per-payload search results in Appendix F.1. This objective minimizes the admission score at fixed payload; it does not maximize downstream promotion.

A refined search evaluates $m \in \{17, 20, 24, 26, 28, 32\}$ over near-optimal blends and multiple construction seeds to probe the empirical acceptance region (Appendix F.2). We report observed acceptances rather than claiming a global attacker optimum.

5 Admission Gates

Cohort statistics. Each profile $u = (u_1, \ldots, u_{|u|})$ is a sequence of items, and $i \in u$ means that item i occurs in u. Let $f(i)$ denote the interaction count of item i in the reference log, $n_{\text{total}} = \sum_i f(i)$, $\mathcal{P}(C) = \{\{(u_j, u_{j+1}) : u \in C, 1 \le j < |u|\}\}$ the multiset of adjacent item pairs in cohort C, and $c(a, b)$ the reference count of pair (a, b). We represent a cohort using four statistics:

$$\phi_0(C) = \max_i \frac{|\{u \in C : i \in u\}|}{|C|}, \qquad \phi_1(C) = \max_\ell \frac{|\{u \in C : |u| = \ell\}|}{|C|},$$

$$\phi_2(C) = -\frac{1}{\sum_{u \in C} |u|} \sum_{u \in C} \sum_{j=1}^{|u|} \log \frac{f(u_j)}{n_{\text{total}}}, \qquad \phi_3(C) = \frac{1}{|\mathcal{P}(C)|} \sum_{(a,b) \in \mathcal{P}(C)} \log(1 + c(a, b)) \,. \tag{3}$$

Here ϕ_0 measures maximum item prevalence, ϕ_1 profile-length concentration, ϕ_2 average item rarity, and ϕ_3 transition plausibility. Repeated occurrences of an item within one profile count once toward ϕ_0. The transition statistic is a count-based signal, not an implementation of LoRec (Zhang et al., 2024). Reference construction and implementation details appear in Section 6 and Appendix A.

Single-feature gates. The staged gates D_0–D_3 threshold one ϕ_j each (organic false-positive rates 0.22–0.47%), exposing which observable discrepancy each attack adaptation repairs. The item-concentration gate retains the hand-set 10% operating point of the initial defense, while the remaining acceptance bands are estimated from organic cohorts. Exact thresholds, sidedness, and organic false-positive rates are reported in Table 4 (Appendix B.1).

Joint gate. Let $\mathbf{x}(C) = (\phi_0(C), \phi_1(C), \phi_2(C), \phi_3(C))^\top$. Using a robustly fitted center μ and covariance Σ, the joint score is

$$D^2(\mathbf{x}) = (\mathbf{x} - \mu)^\top \Sigma^{-1} (\mathbf{x} - \mu). \tag{4}$$

The gate rejects when $D^2(\mathbf{x}) > \tau_{0.999}$. The primary calibration uses $B = 5{,}000$ resampled organic cohorts and $\tau_{0.999} = 25.39$. Thresholds are calibrated empirically rather than from a χ^2_4 reference distribution. Disjoint-pool calibration is analyzed in Section 7.1, with estimator details in Appendix B.2.

Diagonal and marginal controls. With $\sigma_i = \sqrt{\Sigma_{ii}}$ and $z_i = (x_i - \mu_i)/\sigma_i$, we compare the full gate with

$$D^2_{\text{diag}} = \sum_{i=0}^{3} z_i^2, \qquad D_{\max} = \max_{i \in \{0,\ldots,3\}} |z_i|. \tag{5}$$

The first removes off-diagonal covariance while preserving per-feature scaling; the second is a calibrated union-of-marginals test. Each is calibrated on organic cohorts at the same nominal operating level. Empirical comparisons appear in Section 7.4 and Appendix H.2.

Payload bound. Because $\phi_0(C)$ is the maximum prevalence of any item, a target present in m profiles satisfies $m/N \le \phi_0(C)$. Joint-gate acceptance further implies

$$\frac{m}{N} \le \phi_0(C) \le \mu_0 + \sqrt{\tau\, \Sigma_{00}}. \tag{6}$$

Under the primary calibration, this gives a ceiling of approximately 4.04%, or 27.1 target carriers among 671 profiles. This is a feature-space upper bound, not a claim that every smaller payload is admitted or effective. The proof is given in Appendix D (Proposition 1), and its empirical tightness is evaluated in Section 7.3.

6 Experimental Setup

Data. We evaluate on four Amazon Reviews 2014 (McAuley et al., 2015) 5-core catalogs: Beauty (22,363 users, 12,101 items, 198,502 interactions), Health and Personal Care (38,609 users, 18,534 items), Toys and Games (19,412, 11,924), and Sports and Outdoors (35,598, 18,357). Each catalog uses its rarest retained item as the attack target; Beauty, the primary controlled setting, has target frequency 5. Sequences are ordered by review timestamp, and recommendation uses leave-one-out evaluation. Dataset preprocessing and complete catalog statistics are given in Appendix A.

Recommender and metrics. We use DiffGRM (Liu et al., 2026) as the downstream semantic-ID generative recommender, following its reported architecture and training configuration. Clean and attacked runs differ only in the admitted training histories. For key constructions we also retrain the non-generative SASRec (Kang & McAuley, 2018). Selecting checkpoints on genuine users only leaves promotion unchanged in the tested conditions (Appendix E.3). Our primary downstream metric is promotion rate PR, measured over eligible genuine users whose clean top-10 list does not contain the target; users whose true

Table 2: Admission and downstream effect on Amazon Beauty. ✓ denotes rejection and × acceptance by a single-feature gate. The joint column reports D^2/τ (> 1 is rejected). Promotion is the percentage of eligible genuine users newly receiving the target in their top-10 list; "–" denotes an untrained condition. Values are for training seed 1.

Construction	D_0	D_1	D_2	D_3	Joint D^2/τ	Promotion
Stage 0 (bandwagon)	✓	✓	✓	✓	4047	1.59%
Stage 0 (random)	✓	✓	✓	✓	4258	9.44%
Stage 1 dilution	×	✓	✓	✓	336	–
Stage 2 length-matched	×	×	✓	✓	280	8.59%
Stage 3 popularity-matched	×	×	×	✓	351	5.55%
Stage 4 transition-matched	×	×	×	×	22.8	6.35%
Stage 5, m=0 (payload-free)	×	×	×	✓	1.02	0.00%
Stage 5, m=67	×	×	×	✓	23.8	11.96%

next item is the target are excluded. We additionally report Recall@10 and NDCG@10 on the held-out test split.

Admission protocol and baselines. For Beauty, the disjoint-pool calibration protocol (Section 7.1) uses three disjoint pools of 7,454 genuine users for gate fitting, threshold calibration, and held-out evaluation, while the item-frequency and transition reference tables are constructed from the full genuine interaction log. The joint gate is compared with the four staged single-feature gates, the diagonal-covariance and calibrated union-of-marginals controls, and explicitly labeled proxy detectors adapted from prior work (Section 7.4; Appendix I). Unless stated otherwise, calibrated defenses use nominal false-positive level $q = 10^{-3}$ and are not tuned on attack cohorts. The primary operating point uses $B = 5{,}000$ organic cohorts of size $N = 671$ and $\tau_{0.999} = 25.39$ (Appendix C).

Repeated runs. Initial exploratory conditions use training seed 1; repeated main results report mean ± standard deviation over up to three seeds. Full preprocessing, eligibility, and recommender details appear in Appendix A; calibration and baseline details appear in Appendices C and I.

7 Results

7.1 Calibration and temporal shift

Beyond the primary threshold $\tau_{0.999} = 25.39$, ten independent calibrations with disjoint fit, threshold, and evaluation pools give a threshold of 30.9 ± 6.0, while evaluation on held-out randomly resampled organic cohorts gives a false-positive rate of 0.11% (95% CI 0.01–0.20%), close to the nominal 0.1% level.

The same calibration does not transfer to the time-ordered genuine traffic tested here. Ordering users by first-review timestamp, the gate rejects 26 of 33 consecutive windows of 671 genuine users (median $D^2 = 45$). Thus, the low false-positive rate observed under random resampling does not persist under chronological cohort formation. Detailed calibration stability, finite-sample, and distributional diagnostics are reported in Appendix C.

7.2 Adaptive escalation and downstream promotion

Progressive adaptation removes the signatures targeted by the single-feature gates: Stage 1 bypasses item concentration, Stage 2 also bypasses length concentration, Stage 3 additionally bypasses rarity, and Stage 4 is accepted by all four single-feature gates. The joint gate, however, rejects every Stage 0–5 construction in Table 2, with payload-bearing margins remaining at least 22.8× the threshold. Stage 5 at $m = 67$ is also rejected by the transition gate, so we do not claim that every stage defeats every preceding single-feature test.

The same constructions can nevertheless have substantial downstream effects if admitted. Every trained payload-bearing construction shown in Table 2 promotes the target in the

single-seed evaluation, with promotion ranging from 1.59% to 11.96%, whereas the payload-free control gives 0.00%. The apparent ordering is not stable across training seeds: Stage 4 and Stage 5 ($m = 67$) have three-seed mean promotion of 9.5% and 7.4%, respectively (Table 7). Recall@10/NDCG@10 results and generation-seed variability are reported in Appendix E.

7.3 Admission–payload–promotion frontier

The Stage 6 search substantially reduces the joint admission score at lower payloads. On the coarse grid, the best constructions lie in a narrow blend range ($w = 0.90$–0.95): $m \in \{7, 16\}$ is accepted, whereas $m \in \{33, 50, 67\}$ is rejected. This does not imply that $m = 16$ is the largest admissible payload. A refined search with 25 constructions per payload (five near-optimal blends $\times$ five seeds) shows a gradual boundary: 19, 16, 10, 3, 0, and 0 constructions are accepted at $m = 17$, 20, 24, 26, 28, and 32 (Table 9).

The empirical boundary closely follows the analytic payload constraint derived in Section 5. Under the primary calibration, any accepted cohort satisfies $m/N \leq 4.04\%$, corresponding to at most 27.1 target carriers for $N = 671$. Consistent with this bound, we observe accepted constructions through $m = 26$, but none at $m = 28$ or 32. The boundary remains stochastic rather than a deterministic cutoff, as acceptance frequency decreases progressively with payload.

Passing admission does not eliminate downstream effect. The accepted $m = 7$ construction produces no promotion in the evaluated seed, while $m = 16$ gives a three-seed mean promotion of 0.21%. The highest-promotion accepted construction we retrained ($m = 24$, $w = 0.92$, $D^2/\tau = 0.63$) reaches 1.38% mean promotion. For comparison, the rejected Stage 4 and Stage 5 constructions at the working payload $m = 67$ produce mean promotion of 9.5% and 7.4%, respectively. The results therefore expose an admission–payload–promotion frontier: lowering payload enables admission but reduces, rather than eliminates, target promotion. The coarse Stage 6 search, per-seed promotion results, generator–payload controls, additional-target experiments, and the promotion–margin visualization are reported in Appendix F.

Second recommender and catalog. The frontier is not specific to semantic-ID generation. Under the non-generative SASRec (training seed 1), Stage 4 and Stage 5 ($m = 67$) promote the target to 9.45% and 8.37% of eligible users, while the accepted Stage 6 cohorts reach 4.69% ($m = 16$) and 6.57% ($m = 24$), roughly one half to four fifths of the rejected constructions' promotion (Table 15). The frontier is also reproduced on Toys and Games under DiffGRM (three seeds): rejected Stage 4/5 cohorts ($m = 58$) promote the rarest target to 5.89% and 4.07%, the accepted Stage 6 cohort ($m = 13$) to 0.02% (Appendix G, which also covers Sports).

7.4 Robustness and deployment checks

Gate controls. Removing any one statistic still rejects every payload-bearing Stage 0–5 construction, and the full-covariance, diagonal-covariance, and calibrated union-of-marginals gates agree on all of them, differing only on the payload-free control. High-payload rejection thus depends neither on a single feature nor on cross-feature correlations (Appendix H).

Detector comparison. At the matched cohort-level operating point, the joint gate admits the Stage 6 evader ($m = 16$), whereas a simple item-frequency-growth test rejects it; adapted profile-level proxies either miss payload-bearing constructions or also reject the payload-free control. As these methods differ in detection unit, we make no superiority claim: target-specific item evidence complements target-agnostic cohort statistics, and our frontier concerns the latter (Appendix I).

Cross-catalog behavior. On all four Amazon catalogs, both Stage 0 variants are rejected and Stage 4 is rejected by the joint gate with margins of 14.5–35.3$\times$, although single-feature footprints vary (Appendix J). The frontier is also not a rare-item artifact: a medium-frequency Beauty target is again admitted at $m = 16$ and 24 and then promoted to 3.86% and 3.61% under DiffGRM (seed 1), whereas a frequent one is rejected (Appendix F.6).

Sensitivity and mixed admissions. Stages 4–5 stay rejected across the tested cohort sizes, calibration levels, and sample sizes, while Stage 6 acceptance weakens as payload or cohort size grows. More importantly, cohort composition sharply changes detectability: Stage 4 and Stage 5 are rejected in 95–99.5% of windows when attackers comprise half of the cohort, but in only 1.5% at a 25% attacker share and in none at 10%; the Stage 6 evader is accepted at every tested share. The admitted 25%-share windows still promote the target after retraining (Stage 4: 1.35%, Stage 5: 0.96%, three seeds; 24 carriers, below the ceiling of Eq. (6)), whereas Stage 6 carriers diluted this far promote at most one user. Full analyses are reported in Appendix K.

8 Discussion and Limitations

Implications. Admission screening induces an *admission–payload–promotion frontier* rather than a binary guarantee, which suggests three checks for any cohort gate. (i) Report the payload it leaves admissible and the post-retraining effect of that payload: here the ceiling is 27 of 671 carriers, yet admitted cohorts still promote the target for up to 6.57% of users under SASRec and 3.86% for a medium-frequency target, so residual harm depends on the recommender and the target. (ii) Validate on chronologically formed cohorts: a 0.11% resampling false-positive rate became 26 of 33 flagged windows. (iii) Test mixed cohorts at low attacker shares, where rejected attacks pass and still promote. Because high-payload rejection came from redundant calibrated statistics, a simple union of marginals performed identically on these attacks.

Limitations. Our attacks establish feasibility rather than optimality. Stage 6 searches a bounded two-parameter space and minimizes admission score rather than directly maximizing promotion subject to acceptance; its strongest setting also assumes oracle knowledge of the fitted gate, although the payload ceiling is a property of the gate itself, and once a blend near $w = 0.9$ is known, fresh $m = 16$ constructions pass in 90% of cases without further gate queries (Table 24). Downstream retraining uses DiffGRM and, for the key constructions, SASRec, primarily on Amazon Beauty with a replication on Toys and Games. Results are based on a small number of targets and up to three training seeds per condition, and the magnitude of promotion is target-dependent; we therefore do not posit a universal detectability–effectiveness law or a threshold for harmful promotion. Mixed admissions use one whole-window rejection rule, measure promotion at one attacker share, and do not test accumulation across admission periods or a gate recalibrated on recent windows. The transition reference is also constructed from the full genuine log, creating overlap with scored users, and some items share the evaluated four-token semantic identifier. Finally, most detector comparisons use adapted proxy implementations rather than complete reproduced defenses, and we do not compare against a training-time poisoning defense. These results should therefore be interpreted as an empirical characterization of the studied admission setting, not as a universal robustness guarantee.

9 Conclusion

We evaluated admission-time screening against defense-aware poisoning of a semantic-ID generative recommender. Progressive adaptation evades the single-feature gates, while the joint gate rejects the tested working-payload constructions. Adaptive search nevertheless reveals a lower-payload acceptance region: the empirical boundary closely follows the analytic ceiling of 27.1 target carriers among 671 profiles, and an accepted construction with $m = 24$ still achieves 1.38% mean target promotion after retraining (6.57% under SASRec). The apparent robustness also depends strongly on evaluation protocol: calibration on randomly resampled genuine cohorts does not transfer to the time-ordered traffic tested here, and mixing malicious histories with genuine arrivals sharply reduces detectability. Together, these results characterize an admission–payload–promotion frontier and show that static attack separation alone is insufficient to establish operational robustness.

Ethics statement. This work studies attacks in order to evaluate a defense. All experiments use public Amazon review data and a locally trained recommender; no live system was attacked. The attack family is closed-form, uses aggregate statistics, and at Stages 5–6 assumes oracle knowledge of the defender's parameters. Such constructions could be misused to inflate an item's visibility (for example, counterfeit or deceptive products); we report the accepted low-payload regime so that defenders can act on it. Code and generated cohorts will be provided after acceptance.

Reproducibility statement. Data, partitions, features, the attack procedure, the calibration protocol, and baselines are specified in Section 6 and Appendix A, with calibration details in Appendix C. Code and generated cohorts will be provided after acceptance.

References

Layth C. Alwan and Harry V. Roberts. Time-series modeling for statistical process control. *Journal of Business & Economic Statistics*, 6(1):87–95, 1988. doi: 10.1080/07350015.1988.10509640.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023. doi: 10.1214/23-AOS2276.

Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pp. 3–14, 2017. doi: 10.1145/3128572.3140444.

Zunping Cheng and Neil Hurley. Effective diverse and obfuscated attacks on model-based recommender systems. In *Proceedings of the Third ACM Conference on Recommender Systems*, RecSys '09, pp. 141–148, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605584355. doi: 10.1145/1639714.1639739. URL `https://doi.org/10.1145/1639714.1639739`.

Paul-Alexandru Chirita, Wolfgang Nejdl, and Cristian Zamfir. Preventing shilling attacks in online recommender systems. In *Proceedings of the 7th annual ACM international workshop on Web information and data management*, pp. 67–74, 2005.

Yaojun Hao and Fuzhi Zhang. Detecting shilling profiles in collaborative recommender systems via multidimensional profile temporal features. *IET Information Security*, 12(4):362–374, 2018. doi: 10.1049/iet-ifs.2017.0012.

Neil Hurley, Zunping Cheng, and Mi Zhang. Statistical attack detection. RecSys '09, pp. 149–156, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605584355. doi: 10.1145/1639714.1639740. URL `https://doi.org/10.1145/1639714.1639740`.

Wang-Cheng Kang and Julian McAuley. Self-attentive sequential recommendation. In *2018 IEEE International Conference on Data Mining (ICDM)*, pp. 197–206, 2018.

Pang Wei Koh, Jacob Steinhardt, and Percy Liang. Stronger data poisoning attacks break data sanitization defenses. *Machine Learning*, 111(1):1–47, 2022.

Kaihong Li, Huichi Zhou, Bin Ma, and Fangjun Huang. Semanticshield: Llm-powered audits expose shilling attacks in recommender systems. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5381–5385. IEEE, 2026.

Chen Lin, Si Chen, Meifang Zeng, Sheng Zhang, Min Gao, and Hui Li. Shilling black-box recommender systems by learning to generate fake user profiles. *IEEE Transactions on Neural Networks and Learning Systems*, 35(1):1305–1319, 2022.

Zhao Liu, Yichen Zhu, Yiqing Yang, Xiao Lv, Guoping Tang, Rui Huang, Qiang Luo, Ruiming Tang, and Guorui Zhou. Diffgrm: Diffusion-based generative recommendation model. In *Proceedings of the ACM Web Conference 2026*, WWW '26, pp. 5853–5864, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400723070. doi: 10.1145/3774904.3792156. URL https://doi.org/10.1145/3774904.3792156.

Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. Image-based recommendations on styles and substitutes. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2015.

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

Shutong Qiao, Wei Yuan, Junliang Yu, Tong Chen, Quoc Viet Hung Nguyen, and Hongzhi Yin. Controllable and stealthy shilling attacks via dispersive latent diffusion. *arXiv preprint arXiv:2508.01987*, 2025.

Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan Hulikal Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Tran, Jonah Samost, Maciej Kula, Ed Chi, and Maheswaran Sathiamoorthy. Recommender systems with generative retrieval. In *Advances in Neural Information Processing Systems*, volume 36, pp. 10299–10315, 2023. doi: 10.52202/075280-0452.

Jacob Steinhardt, Pang Wei W. Koh, and Percy Liang. Certified defenses for data poisoning attacks. In *Advances in Neural Information Processing Systems*, volume 30, pp. 3517–3529, 2017.

Florian Tramèr, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Shilei Wang, Peng Zhang, Hui Wang, Hongtao Yu, and Fuzhi Zhang. Detecting shilling groups in online recommender systems based on graph convolutional network. *Information Processing & Management*, 59(5):103031, 2022. doi: 10.1016/j.ipm.2022.103031.

Yishu Xu and Fuzhi Zhang. Detecting shilling attacks in social recommender systems based on time series analysis and trust features. *Knowledge-Based Systems*, 178:25–47, 2019. doi: 10.1016/j.knosys.2019.04.012.

Yishu Xu, Peng Zhang, Ru Ma, and Fuzhi Zhang. Detecting shilling groups in recommender systems based on user multi-dimensional dynamic behavior analysis and graph contrastive learning. *Neurocomputing*, pp. 132703, 2026.

Zhenrui Yue, Zhankui He, Huimin Zeng, and Julian McAuley. Black-box attacks on sequential recommenders via data-free model extraction. In *Proceedings of the 15th ACM conference on recommender systems*, pp. 44–54, 2021.

Fuzhi Zhang, Yueqi Qu, Yishu Xu, and Shilei Wang. Graph embedding-based approach for detecting group shilling attacks in collaborative recommender systems. *Knowledge-Based Systems*, 199:105984, 2020a. doi: 10.1016/j.knosys.2020.105984.

Hengtong Zhang, Yaliang Li, Bolin Ding, and Jing Gao. Practical data poisoning attack against next-item recommendation. In *Proceedings of the web conference 2020*, pp. 2458–2464, 2020b.

Kaike Zhang, Qi Cao, Yunfan Wu, Fei Sun, Huawei Shen, and Xueqi Cheng. Lorec: Combating poisons with large language model for robust sequential recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1733–1742, 2024.

Yuchuan Zhao, Tong Chen, Junliang Yu, Kai Zheng, Lizhen Cui, and Hongzhi Yin. Diversity-aware dual-promotion poisoning attack on sequential recommendation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1634–1644, 2025.

Wei Zhou, Junhao Wen, Yun Sing Koh, Qingyu Xiong, Min Gao, Gillian Dobbie, and Shafiq Alam. Shilling attacks detection in recommender systems based on target item analysis. *PloS one*, 10(7):e0130968, 2015.

Appendix overview

Appendix A specifies data, partitions, profile generation, the recommender, and search budgets. Appendices B–D specify the admission gates, their calibration, and the geometry behind the payload bound of Section 5. Appendices E and F give the complete admission and promotion results summarized in Sections 7.2 and 7.3, and Appendix G gives the SASRec, Toys and Games, and Sports and Outdoors downstream results. Appendices H–K contain the gate ablations, detector comparisons, cross-catalog results, and sensitivity and mixed-admission analyses summarized in Section 7.4. Unless stated otherwise, all results are on Amazon Beauty with $N = 671$, the joint gate uses $\tau_{0.999} = 25.39$, and "promotion" is the promotion rate PR of Section 3 in percent of eligible genuine users.

A Reproducibility specification

Table 3 collects the implementation choices needed to reproduce the pipeline of Sections 4–6.

B Admission-gate specification

B.1 Single-feature gates

Each single-feature gate D_j rejects a cohort whose statistic ϕ_j (Eq. 3) falls outside the acceptance region in Table 4. The statistics were introduced in the order of the staged attacks. ϕ_0 separates Stage 0 from organic cohorts. ϕ_1 targets Stage 1, whose fixed length gives 100% length concentration, against an organic upper bound of 0.377. ϕ_2 targets uniform filler, which raises mean rarity to 9.80–9.90, against an organic band of 8.78–8.98; bandwagon filler lies far below this band (6.96). ϕ_3 targets independent filler draws, which leave adjacent-pair plausibility near zero. The item-concentration gate keeps the hand-set 10% operating point of the initial defense (Section 5); the calibrated band for ϕ_0 is shown for reference and is used by the union-of-marginals control (Appendix H.2).

B.2 Joint gate and controls

The center μ and covariance Σ of Eq. (4) are fitted with scikit-learn's `EllipticEnvelope` (`contamination=0.01`, `random_state=0`, no regularization), giving $\mu = (0.0240, 0.3206, 8.886, 0.8487)$ and per-feature standard deviations $\sigma = (0.0033, 0.0169, 0.0315, 0.0112)$. All thresholds are empirical 99.9th percentiles (linear interpolation) of the respective score over the same $B = 5{,}000$ organic cohorts: $\tau_{0.999} = 25.39$ for D^2, $\tau_{\text{diag}} = 26.95$ for D^2_{diag}, and 4.51 for $D_{\max}$ (Eq. 5). The diagonal and union-of-marginals controls use the same μ and σ_i as the joint gate.

C Calibration details

Primary threshold. At the primary operating point, (μ, Σ) and $\tau_{0.999} = 25.39$ are estimated from the same $B = 5{,}000$ cohorts of $N = 671$ resampled genuine users. An empirical 99.9th percentile from 5,000 scores rests on about five tail observations, so its sampling variability is substantial. Re-estimating it from subsamples of the same draw gives 21.28 ($B = 300$) and 25.85 ($B = 500$) (Figure 2b).

Table 3: Reproducibility specification.

Component	Specification
Data	Amazon Reviews 2014 Beauty, 5-core; sequences ordered by review timestamp; no deduplication; 22,363 users, 12,101 items, 198,502 interactions. Target: rarest retained item (frequency 5; 4 distinct neighbors across 8 co-occurrences in the reference log). Additional catalogs (users, items): Health and Personal Care (38,609, 18,534), Toys and Games (19,412, 11,924), Sports and Outdoors (35,598, 18,357), each attacked at its own rarest item.
Partitions	Reference tables f and c: all 22,363 genuine users (they overlap with the users in organic calibration cohorts, and are not restricted to pre-admission data). Recommender: leave-one-out split. Primary operating point: (μ, Σ) and $\tau_{0.999}$ estimated from the same $B = 5{,}000$ resampled organic cohorts. Disjoint-pool protocol: ten splits, each with three disjoint pools of 7,454 users for gate fitting, threshold calibration, and held-out evaluation.
Fake profiles in validation and test	Leave-one-out is applied to every sequence, including fake ones (training input seq[: -2], validation seq[: -1], test the full sequence), so fake profiles enter validation and checkpoint selection; restricting validation to genuine users selects the same checkpoints (Appendix E.3). Promotion is computed on the 22,363 genuine test rows only.
Cohort statistics	ϕ_0: user prevalence, with an item repeated in a profile counted once. ϕ_2: natural logarithm, averaged over all interactions. ϕ_3: adjacent pairs, treated as undirected (pairs stored sorted); unseen pairs contribute $\log 1 = 0$.
Profile generation	Initial item drawn from the popularity marginal; $t^\star$ overwrites the item at zero-based position $L - 3$; the target is not excluded from filler draws; repeated items allowed; minimum length 3, no truncation. The transition walk falls back to popularity sampling when no continuation is observed. Generation seeds: 44 (Stage 4, Table 2), 45 (Stage 5 sweep), 46 (coarse Stage 6 grid), $\{46, 101, 202, 303, 404\}$ (refined Stage 6 search).
Admission gates	Single-feature gates: Table 4. Joint gate: Appendix B.2.
Adaptive search	Coarse grid: $m \in \{0, 7, 16, 33, 50, 67\} \times w \in \{0, 0.25, 0.5, 0.6, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 1\}$, one construction per grid point (66 cohorts), selection rule (2). Refined search: $m \in \{17, 20, 24, 26, 28, 32\}$ $\times$ five near-optimal blend values $\times$ five construction seeds (150 cohorts).
Recommender	DiffGRM: 1 encoder and 4 decoder layers, d=256, 4 heads; 4×256 semantic-ID codes fitted once on the clean catalog and reused for all runs; AdamW, learning rate 0.01, batch 1024, label smoothing 0.2, guided masking, at most 100 epochs, patience 15; checkpoint selected by validation score; beam 128, confidence mode, top-10; no candidate filtering. Training seeds $\{1, 2, 3\}$.
Second recommender	SASRec (Kang & McAuley, 2018) in the baseline configuration of Liu et al. (2026): binary cross-entropy next-item objective, hyperparameters as in the original paper, same leave-one-out splits, best-validation checkpoint; training seed 1 (Appendix G).
Promotion metric	22,362 eligible genuine users (users whose true next item is $t^\star$ are excluded). Each attacked run is aligned row by row with the clean run by true next-item label.
Baselines	Appendix I. Cohort-level proxy decisions threshold the cohort median of per-profile scores at the organic 0.1th/99.9th percentiles of 2,000 resampled cohorts; per-profile thresholds are calibrated on 5,000 genuine users.
Compute	Joint-gate scoring takes about 11 ms per 671-user cohort on one CPU core.

Disjoint-pool calibration. Each of ten splits fits (μ, Σ), sets τ, and measures the achieved false-positive rate on three disjoint pools of 7,454 users (Table 3). Refitting on each split gives $\tau = 30.9 \pm 6.0$. Fixing (μ, Σ) at one split and redrawing only the threshold sample gives 28.9 ± 3.7, so part of the spread comes from the fitted score function rather than from threshold sampling alone. On 20,000 held-out randomly resampled cohorts per

Table 4: Acceptance regions of the four single-feature gates and robust organic feature statistics. "Tails" gives the rejected direction(s); the false-positive rate (FPR) is measured on organic cohorts.

Statistic	Center μ_i	Scale σ_i	Lower bound	Upper bound	Tails	Organic FPR
ϕ_0 item concentration	0.0240	0.0033	—	0.10 (hand-set; calibrated 0.0387)	upper	0.22% (calibrated band)
ϕ_1 length concentration	0.321	0.0169	0.268	0.377	both	0.47%
ϕ_2 rarity	8.886	0.0315	8.779	8.984	both	0.37%
ϕ_3 transition plausibility	0.849	0.0112	0.816	0.888	both	0.41%

split, the achieved false-positive rate is 0.11% (95% CI 0.01–0.20%), close to the nominal 0.1%.

Analytic reference. The χ^2_4 distribution is used only as a diagnostic. Its 99.9th percentile, 18.47, lies 27% below the empirical threshold, and the empirical organic exceedance rate at 18.47 is 0.58%. A Kolmogorov–Smirnov test against χ^2_4 rejects ($p = 1.1 \times 10^{-9}$). Because (μ, Σ) are estimated from the same sample, this p-value is not exact, and the test indicates distributional mismatch rather than a heavier upper tail specifically. We therefore use empirical thresholds throughout.

Time-ordered windows. When genuine users are ordered by the timestamp of their first review and grouped into consecutive windows of 671, the gate rejects 26 of 33 windows (median $D^2 = 45$, or $1.8\,\tau_{0.999}$). Random resampling treats future cohorts as exchangeable with historical users, and this check shows that the assumption does not hold for chronologically formed cohorts in this catalog (Section 7.1).

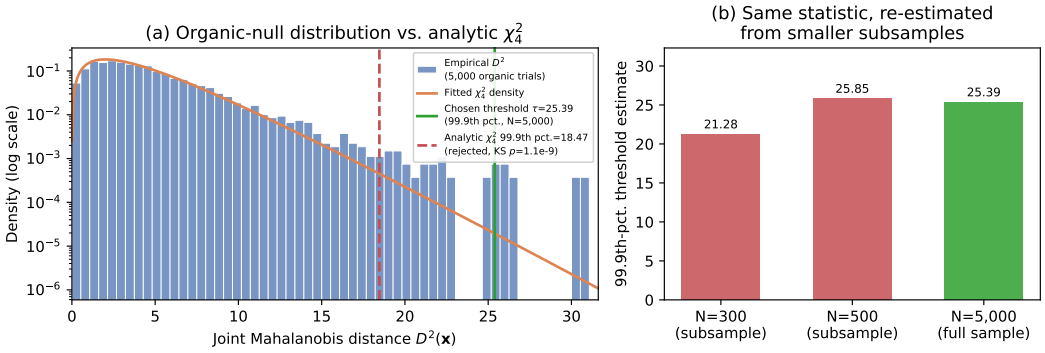

Figure 2: Calibration diagnostics. (a) Empirical distribution of D^2 over 5,000 organic cohorts, with the analytic χ^2_4 density, the empirical 99.9th-percentile threshold (25.39), and the χ^2_4 99.9th percentile (18.47). (b) The empirical 99.9th-percentile threshold re-estimated from subsamples of the same calibration draw.

D Joint-gate geometry and the payload bound

Write $\Sigma = SRS$ with $S = \mathrm{diag}(\sigma_0, \ldots, \sigma_{k-1})$, $\sigma_i = \sqrt{\Sigma_{ii}}$, and R the correlation matrix, and let $z = S^{-1}(\mathbf{x} - \mu)$ as in Eq. (5). Then $D^2 = z^\top R^{-1} z$ and $D^2_{\mathrm{diag}} = \|z\|_2^2$; here $k = 4$.

Proposition 1. *Let R be positive definite. For every $\mathbf{x} \in \mathbb{R}^k$:*

(i) $\|z\|_2^2/\lambda_{\max}(R) \le D^2(\mathbf{x}) \le \|z\|_2^2/\lambda_{\min}(R)$, *so* $D^2/D^2_{\mathrm{diag}} \in [1/\lambda_{\max}(R), 1/\lambda_{\min}(R)]$ *whenever* $\mathbf{x} \neq \mu$.

(ii) If $|z_i| \le \delta$ *for all* i, *then* $D^2(\mathbf{x}) \le k\delta^2/\lambda_{\min}(R)$.

(iii) Over a box $\prod_i [a_i, b_i]$, D^2 *attains its maximum at one of the* 2^k *vertices.*

(iv) If $D^2(\mathbf{x}) \le \tau$, then $|x_0 - \mu_0| \le \sqrt{\tau\,\Sigma_{00}}$. Consequently, a cohort accepted by the joint gate in which $t^\star$ appears in m profiles satisfies $m/N \le \phi_0(C) \le \mu_0 + \sqrt{\tau\,\Sigma_{00}}$ (Eq. 6).

Proof. (i) is the Rayleigh-quotient bound for R^{-1}, whose eigenvalues lie in $[1/\lambda_{\max}(R), 1/\lambda_{\min}(R)]$. (ii) follows from (i) and $\|z\|_2^2 \le k\delta^2$. (iii) holds because D^2 is convex and a convex function on a polytope attains its maximum at a vertex. (iv) Let $\mathbf{y} = \mathbf{x} - \mu$ and $\langle \mathbf{a}, \mathbf{b} \rangle_{\Sigma^{-1}} = \mathbf{a}^\top \Sigma^{-1} \mathbf{b}$. Then $x_0 - \mu_0 = e_0^\top \mathbf{y} = \langle \Sigma e_0, \mathbf{y} \rangle_{\Sigma^{-1}}$, and by the Cauchy–Schwarz inequality $|x_0 - \mu_0| \le \sqrt{e_0^\top \Sigma e_0}\,\sqrt{\mathbf{y}^\top \Sigma^{-1} \mathbf{y}} \le \sqrt{\Sigma_{00}\,\tau}$, with equality for $\mathbf{y} = \pm\sqrt{\tau/\Sigma_{00}}\,\Sigma e_0$. For the payload statement, $t^\star$ occurs in m of the $N = |C|$ profiles, so $\phi_0(C) \ge |\{u \in C : t^\star \in u\}|/|C| = m/N$, and $\phi_0(C) = x_0 \le \mu_0 + \sqrt{\tau\,\Sigma_{00}}$. □

Parts (i)–(ii) upper-bound the score. They do not imply that cohorts accepted by every marginal gate are rejected, which also depends on how the marginal and joint thresholds are calibrated. The worst case over a box of marginal acceptance intervals is given by (iii), not by the direction of the smallest eigenvalue. A maximizing vertex is a point in feature space and need not correspond to any realizable cohort. Part (iv) restricts the payload of any accepted cohort independently of the attack search. Table 5 gives its numerical value.

Corollary 1. *On the primary calibration data, $\lambda_{\min}(R) = 0.4511$, driven by a correlation of -0.549 between ϕ_2 and ϕ_3, so $D^2/D^2_{\text{diag}} \le 1/\lambda_{\min}(R) \approx 2.22$ for every $\mathbf{x} \ne \mu$.*

The observed ratios for payload-bearing constructions are 0.98–1.46 (Table 19), well below this bound.

The bound in (i) is attained. Taking $z = c\,v_{\min}$, with $v_{\min}$ the eigenvector of R for $\lambda_{\min}$ and c the largest scale keeping every feature inside its calibrated acceptance band, gives $D^2 = 24.2$ and $D^2_{\text{diag}} = 10.9$, a ratio of $2.22 = 1/\lambda_{\min}(R)$. This is a feature-space construction rather than a realized cohort. It illustrates the identity in (i) and does not bound the worst case over marginally accepted vectors, which is the vertex maximum in Table 5.

Table 5: Numerical values of Proposition 1 under the primary calibration ($\tau_{0.999} = 25.39$, $N = 671$).

Quantity	Value	Comparison
μ_0, Σ_{00}	0.0240, 1.06×10^{-5}	—
Payload ceiling $\mu_0 + \sqrt{\tau\,\Sigma_{00}}$ (part iv)	4.04% (27.1 of 671)	26/671 = 3.9%; 28/671 = 4.2%; 67/671 = 10.0%
Largest D^2 over the 16 box vertices (part iii)	81.1	$\tau_{0.999} = 25.39$

The ceiling of 27.1 carriers lies between the largest accepted payload in the refined search ($m = 26$) and the smallest fully rejected one ($m = 28$) (Table 9). It is an upper bound, so it does not imply that smaller payloads are accepted.

E Additional admission and promotion results

E.1 Full admission scores and downstream metrics

Table 6 extends Table 2 with absolute joint-gate scores and ranking metrics. Each row is one constructed cohort and one gate decision, an observed outcome rather than an estimated detection rate. Recall@10 and NDCG@10 use one training seed per row, so we do not interpret their small differences.

E.2 Construction and training-seed variability

Stage 4 and Stage 5 at $m = 67$ are independent realizations of the same generator (Section 4). Their different single-realization outcomes, including the transition-gate decision (× versus ✓ in Table 2), therefore reflect construction and training variation, not different attack algorithms. Across 20 generation seeds, this generator gives $D^2 = 597 \pm 12$ and falls inside the single-feature transition band in only 2 of 20 constructions. Across three training seeds,

Table 6: Full admission scores and downstream results on Amazon Beauty (training seed 1). The clean model gives 0.00% promotion. Recall@10 and NDCG@10 are shown as clean → attacked; "–" denotes an untrained condition.

Construction	D^2	D^2/τ	Promotion	Recall@10	NDCG@10
Stage 0 (bandwagon)	102,755	4,047	1.59%	0.0464 → 0.0467	0.0283 → 0.0282
Stage 0 (random)	108,125	4,258	9.44%	0.0464 → 0.0463	0.0283 → 0.0280
Stage 1 dilution	8,527	336	–	–	–
Stage 2 length-matched	7,109	280	8.59%	0.0464 → 0.0466	0.0283 → 0.0285
Stage 3 popularity-matched	8,918	351	5.55%	0.0464 → 0.0457	0.0283 → 0.0279
Stage 4 transition-matched	579	22.8	6.35%	0.0464 → 0.0438	0.0283 → 0.0278
Stage 5, m=0 (payload-free)	26.0	1.02	0.00%	0.0464 → 0.0476	0.0283 → 0.0290
Stage 5, m=67	605	23.8	11.96%	0.0464 → 0.0473	0.0283 → 0.0291

Stage 4 promotes 6.35%, 13.69%, and 8.53% of eligible users, and Stage 5 ($m = 67$) promotes 11.96%, 5.24%, and 5.02% (Table 7).

The payload-free Stage 5 control lies close to the joint-gate boundary in the displayed construction ($D^2/\tau = 1.02$). Across 20 generation seeds, its margin is 1.76 ± 0.57, with 18 of 20 constructions rejected. It produces 0.00% promotion at all three training seeds. This is consistent with a payload-dependent effect. It does not show that placement at the supervised position is the only way injected histories can affect training.

Table 7: Repeated construction and training. D^2/τ is mean ± sd over independent constructions. Promotion is mean ± sd over three training seeds of one trained construction per condition.

Condition	Constructions	Training seeds	D^2/τ	Promotion (%)
Clean run	—	3	—	0.00
Stage 4 transition-matched, m=67	20 (1 trained)	3	23.5 ± 0.5	9.5 ± 3.8
Stage 5, m=67	20 (1 trained)	3	23.5 ± 0.5	7.4 ± 3.9
Stage 5, m=0 (payload-free)	20 (1 trained)	3	1.76 ± 0.57	0.00 ± 0.00
Stage 6, m=16	20 (1 trained)	3	0.67 ± 0.37	0.21 ± 0.15
Stage 6, m=24 (refined search)	25 (1 trained)	3	trained: 0.63 (10 of 25 accepted)	1.38 ± 0.34

E.3 Checkpoint selection on genuine users

Leave-one-out validation includes the fake profiles, so they take part in early stopping and checkpoint selection (Table 3). To test whether this affects promotion, we retrain the clean model, Stage 4 ($m = 67$), and the accepted Stage 6 cohort ($m = 16$) at training seed 1 with validation restricted to the 22,363 genuine users; the fake profiles are still used for training. Excluding them changes the best validation score (0.03129 → 0.03171 for Stage 4 and 0.03091 → 0.03101 for Stage 6) but not the best epoch (53 and 42). The selected checkpoints, and therefore the predictions, are identical to those of the original runs, so promotion remains 6.35% and 0.38%. The genuine-only clean run reproduces the original clean metrics (NDCG@10 0.0283, Recall@10 0.0464). Fake profiles in the validation split therefore do not affect the reported promotion values.

F Stage 6 search and the admission–payload–promotion frontier

F.1 Coarse search

Table 8 reports, for each payload of the coarse grid, the lowest-scoring construction over the eleven blend values (Table 3). The lowest scores occur for $w \in [0.90, 0.95]$. Pure popularity sampling ($w = 0$) gives $D^2 \approx$ 8,000–8,900 at every tested payload. The grid admits the best constructions at $m = 7$ and 16 and rejects those at $m \geq 33$. This does not establish a hard cutoff, and the refined search below probes the intermediate region.

Table 8: Coarse Stage 6 search: the lowest-scoring construction per payload over the eleven-value blend grid (one construction seed per grid point). Decisions apply to the displayed constructions only.

Target carriers m	0	7	16	33	50	67
Best w	0.95	0.90	0.90	0.95	0.95	0.95
Min. D^2	10.40	8.35	8.23	69.97	258.11	570.80
$D^2/\tau_{0.999}$	0.41	0.33	0.32	2.8	10.2	22.5
Decision	accepted	accepted	accepted	rejected	rejected	rejected

F.2 Refined search

The refined search maximizes m subject to $D^2 \leq \tau_{0.999}$ over $m \in \{17, 20, 24, 26, 28, 32\}$, five near-optimal blend values, and five construction seeds per payload (25 constructions per payload). Table 9 reports the acceptance counts summarized in Section 7.3 and the lowest observed score per payload. The lowest score crosses the threshold between $m = 26$ (0.92) and $m = 28$ (1.24). This agrees with the analytic ceiling of 27.1 carriers (Table 5).

Table 9: Refined Stage 6 search: accepted constructions and the lowest joint score observed among the 25 constructions per payload.

Target carriers m	17	20	24	26	28	32
Accepted / 25	19	16	10	3	0	0
Lowest D^2/τ	0.08	0.22	0.62	0.92	1.24	2.18

F.3 Promotion of accepted constructions

Table 10 gives the per-seed promotion of the accepted constructions that were retrained, which the main paper summarizes by their means. At $m = 16$, the three training seeds promote 85, 24, and 32 of the 22,362 eligible users. The $m = 24$ construction is the highest-promoting accepted construction we retrained. It is not the lowest-scoring construction at that payload (Table 9).

Table 10: Per-seed promotion of retrained accepted Stage 6 constructions.

Construction	D^2/τ	Promotion per training seed	Mean
Stage 6, $m = 7$, $w = 0.90$	0.33	0.00% (one seed)	0.00%
Stage 6, $m = 16$, $w = 0.90$	0.32	0.38%, 0.11%, 0.14%	0.21%
Stage 6, $m = 24$, $w = 0.92$	0.63	1.44%, 1.02%, 1.69%	1.38%

F.4 Separating payload size from generator choice

The accepted Stage 6 constructions differ from Stage 4 in both payload and blend, so Table 11 varies each factor separately. With the Stage 4 generator fixed ($w = 1$), mean promotion rises from 0.00% at $m = 7$ to 0.25% at $m = 16$ and 9.5% at $m = 67$. With $m = 16$ fixed, promotion is 0.13%, 0.21%, and 0.25% for $w = 0.5$, 0.90, and 1. Only the $w = 0.90$ construction is accepted at $m = 16$. In these runs, the lower promotion of admitted Stage 6 attacks is therefore associated mainly with the smaller payload, while the blend mainly determines admission.

F.5 Additional targets

Table 12 repeats the protocol for an additional frequency-5 target. The admission decisions agree with those for the primary target: Stage 4 is rejected by a similar margin, and the Stage 6 construction is accepted. Mean promotion by the accepted construction differs by a

Table 11: Promotion (%, mean over three training seeds). Left: generator fixed ($w = 1$, the Stage 4 generator), payload varied. Right: payload fixed ($m = 16$), blend varied.

m (w=1)	7	16	67	w (m=16)	0.5	0.90	1.0
Promotion	0.00	0.25	9.5	Promotion	0.13	0.21	0.25

factor of three between the two targets (0.21% versus 0.07%), and Appendix F.6 shows larger differences across target frequencies. We therefore do not treat the promotion magnitude for the primary target as target-independent.

Table 12: Admission and promotion for two frequency-5 targets on Beauty. Admission entries give the joint-gate decision and D^2/τ.

Target	Frequency	Stage 4 (m=67)	Stage 6 (m=16)	Stage 6 promotion (per seed)
Primary target	5	rejected (22.8)	accepted (0.32)	0.21% (0.38, 0.11, 0.14)
Target 2	5	rejected (22.4)	accepted (0.47)	0.07% (0.18, 0.01, 0.02)

F.6 Target frequency

All targets in Table 12 have frequency 5. To test whether the frontier is a rare-item artifact, we repeat the matched operating points ($m = 67$ with $w = 1$; $m = 16$ with $w = 0.90$; $m = 24$ with $w = 0.92$) for a medium-frequency target (48 interactions, above the catalog's 90th percentile of 34) and a frequent target (403 interactions, the second most frequent item). The catalog's median item frequency is 9. Table 13 reports the joint-gate decisions and Table 14 the promotion under both recommenders. All six new cohorts were retrained regardless of the gate decision, so that rejected cohorts show what admission would have cost.

The medium-frequency target is admitted like the rare one, but its admitted cohorts are more harmful. Under DiffGRM, the accepted $m = 16$ and $m = 24$ cohorts promote it to 3.86% and 3.61% of eligible users, compared with 0.38% and 1.44% for the rare target at the same operating points and training seed. Under SASRec, medium and rare targets are promoted almost equally (4.72% versus 4.69% at $m = 16$). The low DiffGRM promotion of accepted Stage 6 cohorts on the primary target is therefore partly a property of that rare target, not only of the admitted payload. For the frequent target, the gate rejects all three cohorts. The carriers add to a target that already appears in many organic histories, which is consistent with ϕ_0 exceeding the ceiling of Eq. (6) at payloads that are admissible for rarer targets. Rejection here does not reflect a weak attack: had these cohorts been admitted, they would promote the frequent target to 6.09–7.44% of genuine users under DiffGRM and 3.53–5.84% under SASRec, on top of the users whose clean top-10 lists already contain it (2,649 and 2,360). Aggregate NDCG@10 changes by at most 0.0014 in any cell, and no admitted cohort lowers it.

Table 13: Joint-gate decision and D^2/τ by target frequency (Beauty). Rare-target values are those of the primary target (Tables 2 and 10).

Target	Frequency	m=67, w=1	m=16, w=0.90	m=24, w=0.92
Rare (primary)	5	rejected (22.8)	accepted (0.32)	accepted (0.63)
Medium	48	rejected (23.7)	accepted (0.50)	accepted (0.92)
Frequent	403	rejected (29.2)	rejected (1.60)	rejected (2.64)

F.7 Promotion versus admission margin

Figure 3 plots single-seed promotion against joint-gate margin for the trained constructions. It visualizes the tested conditions and does not establish a general detectability–effectiveness relationship.

Table 14: Promotion (%) by target frequency (Beauty, training seed 1; newly promoted users in parentheses). The rare-target $m = 67$ entries are Stage 4. †Cohort rejected by the joint gate (Table 13). For the frequent target, percentages are relative to all genuine users whose true next item is not the target, including those whose clean top-10 list already contains it; relative to the remaining users they are about 12–14% higher.

Recommender	Target	m=16	m=24	m=67
DiffGRM	Rare (5)	0.38 (85)	1.44 (322)	6.35† (1,421)
	Medium (48)	3.86 (862)	3.61 (807)	5.54† (1,239)
	Frequent (403)	6.09† (1,359)	7.44† (1,661)	6.88† (1,537)
SASRec	Rare (5)	4.69 (1,048)	6.57 (1,470)	9.45† (2,114)
	Medium (48)	4.72 (1,056)	5.94 (1,328)	7.57† (1,692)
	Frequent (403)	4.15† (927)	3.53† (789)	5.84† (1,303)

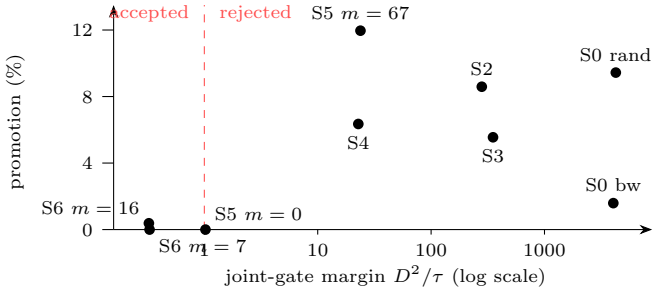

Figure 3: Promotion versus joint-gate margin for the nine constructions trained at seed 1 (Tables 6 and 10). Stage 6 constructions with $m \in \{33, 50, 67\}$ were scored but not retrained. The plot is descriptive for the primary target and catalog.

G Second recommender and additional catalogs

Table 15: Promotion (%) of the same Beauty cohorts under two recommenders (training seed 1).

Construction	Joint gate	DiffGRM	SASRec
Stage 4, m=67	rejected	6.35	9.45
Stage 5, m=67	rejected	11.96	8.37
Stage 5, m=0 (payload-free)	rejected	0.00	0.00
Stage 6, m=16, w=0.90	accepted	0.38	4.69
Stage 6, m=24, w=0.92	accepted	1.44	6.57

SASRec. SASRec (Kang & McAuley, 2018) scores the full item catalog through learned item embeddings and uses no semantic IDs. It is trained on the same admitted histories and leave-one-out split as DiffGRM, and promotion is computed as in Section 3 against a matched SASRec clean run; users whose clean SASRec top-10 list already contains the target are not eligible. On Beauty, Stage 4, Stage 5 ($m = 67$), and the accepted Stage 6 cohorts newly promote 2,114, 1,872, 1,048 ($m = 16$), and 1,470 ($m = 24$) users (Table 15). The payload-free Stage 5 cohort promotes none, and the $m = 67$ cohort promotes the same 1,872 users whether compared with the clean run or with the payload-free run, so the effect is due to the payload rather than to camouflage or retraining noise. Aggregate ranking quality does not reveal the attack: SASRec's NDCG@10 is 0.0216 on the clean run and at most 0.0230 on the attacked runs. SASRec follows the baseline configuration of the DiffGRM study (Liu et al., 2026): the self-attentive Transformer of Kang & McAuley (2018) trained with a binary cross-entropy next-item objective, with hyperparameters set as in the original

paper, on the same leave-one-out splits, and evaluated with the checkpoint that performs best on the validation set.

Toys and Games. We repeat the key constructions end-to-end on Toys and Games (19,412 genuine users; 3% budget of 582 profiles; the rarest item, of frequency 5, as target, as in Table 23). The Stage 4 cohort uses the construction of Table 23 with $m = 58$ carriers (10% of the budget) and is generated independently ($D^2 = 885$ versus 883 there). The Stage 5 cohort also has $m = 58$, and the accepted Stage 6 cohort has $m = 13$ and $w = 0.95$. DiffGRM is trained with three seeds, each compared with a seed-matched clean run, and SASRec with one seed (Table 16). The Beauty frontier replicates. Both rejected constructions promote the target to 4–6% of eligible users under DiffGRM, whereas the accepted Stage 6 cohort promotes 0–8 users per seed. NDCG@10 stays within 1.5% of its clean value (paired $p \geq 0.20$). SASRec shows the same ordering but is promoted less than DiffGRM on this catalog, so which recommender is more vulnerable depends on the catalog.

Table 16: Downstream effect on Toys and Games. DiffGRM promotion is mean ± sd over three training seeds (per-seed values in parentheses); SASRec uses training seed 1. Clean DiffGRM NDCG@10 is 0.0324 ± 0.0008.

Construction	m	D^2/τ (gate)	DiffGRM promotion (%)	DiffGRM NDCG@10	SASRec promotion (%)
Stage 4 transition-matched	58	35.5 (rejected)	5.89 ± 2.20 (4.14, 5.17, 8.36)	0.0325 ± 0.0004	1.86
Stage 5	58	35.1 (rejected)	4.07 ± 1.89 (2.48, 3.58, 6.16)	0.0319 ± 0.0004	1.37
Stage 6, w=0.95	13	0.90 (accepted)	0.02 ± 0.02 (0.01, 0.00, 0.04)	0.0321 ± 0.0009	0.26

Sports and Outdoors. We build a transition-matched (Stage 4) cohort from the Sports and Outdoors sequences at the 3% budget (1,068 profiles), with that catalog's rarest item (frequency 5) as target, the same target as in Table 23. Clean and attacked runs share the tokenizer fit, and all 35,598 genuine rows are aligned. Table 17 reports promotion under both recommenders. The attack promotes the target under DiffGRM while leaving NDCG@10 unchanged to four decimals, and SASRec is again promoted more strongly. The absolute DiffGRM effect is smaller than on Beauty, where Stage 4 promotes 6.35% at seed 1, although the two budgets are not matched.

Table 17: Downstream effect of a Stage 4 cohort on Sports and Outdoors (training seed 1). "Already" counts genuine users whose clean top-10 list contains the target; they are not eligible.

Recommender	Promotion	NDCG@10 (clean → attacked)	Already
DiffGRM	1.37% (487 users)	0.0143 → 0.0143	54
SASRec	3.06% (1,090 users)	0.0059 → 0.0089	12

H Gate-control ablations

H.1 Leave-one-feature-out

We remove each statistic in turn, refit the three-feature Mahalanobis gate on the same 5,000 organic cohorts, and recalibrate its threshold at the 99.9th percentile (Table 18). Every payload-bearing construction remains rejected after removing any single statistic, so no statistic is individually indispensable for these attacks. Only the payload-free Stage 5 decision changes. It remains rejected when item concentration is removed and is accepted when length, rarity, or transition is removed. Its full-gate margin is only $D^2/\tau = 1.02$, so we read these changes as fragility of a borderline decision, not as evidence that a particular statistic is necessary. Removing item concentration reduces the Stage 4 margin from 22.8× to 1.4× (24.7/17.3). This statistic contributes substantially, although it is not necessary for rejection.

Table 18: Leave-one-feature-out ablation. Each column refits and recalibrates the gate on the remaining statistics (threshold in the header). Entries give the decision (✓ rejected, × accepted) and D^2.

Construction	All four (τ=25.4)	Drop item (τ=17.3)	Drop length (τ=25.7)	Drop rarity (τ=25.1)	Drop transition (τ=22.9)
Stage 0 (bandwagon)	✓ (102,755)	✓ (10,782)	✓ (100,903)	✓ (101,162)	✓ (94,561)
Stage 0 (random)	✓ (108,125)	✓ (7,249)	✓ (107,224)	✓ (110,909)	✓ (100,754)
Stage 1 dilution	✓ (8,527)	✓ (7,341)	✓ (7,231)	✓ (8,408)	✓ (3,269)
Stage 2 length-matched	✓ (7,109)	✓ (5,965)	✓ (7,211)	✓ (7,010)	✓ (1,748)
Stage 3 popularity-matched	✓ (8,918)	✓ (7,992)	✓ (9,090)	✓ (6,844)	✓ (575)
Stage 4 transition-matched	✓ (579)	✓ (24.7)	✓ (589)	✓ (563)	✓ (585)
Stage 5, m=0 (payload-free)	✓ (26.0)	✓ (25.6)	× (25.4)	× (16.7)	× (1.6)
Stage 5, m=67	✓ (605)	✓ (48.2)	✓ (610)	✓ (571)	✓ (602)

H.2 Diagonal covariance and union of marginals

Table 19 compares the joint gate with the diagonal-covariance gate ($\tau_{\text{diag}} = 26.95$) and the calibrated union-of-marginals gate ($D_{\max}$, threshold 4.51) of Eq. (5). The three gates make identical decisions on every payload-bearing construction. For these constructions, $D^2_{\text{full}}/D^2_{\text{diag}}$ ranges from 0.98 to 1.46, well below the bound $1/\lambda_{\min}(R) \approx 2.22$ (Corollary 1). The only disagreement is the payload-free Stage 5 cohort, which the full gate rejects and the diagonal and union gates accept. The transition gate D_3 also rejects this cohort (Table 2), so the disagreement is not a case in which correlation modeling detects something no marginal statistic detects. For the attacks tested, cross-feature correlation does not produce the observed high-payload rejection.

Table 19: Full versus diagonal covariance and the calibrated union-of-marginals control. ✓ denotes rejection and × acceptance. Ratio = $D^2_{\text{full}}/D^2_{\text{diag}}$.

Construction	D^2 (full, τ=25.4)		D^2 (diag, τ=27.0)		Ratio	$D_{\max}$ (τ=4.51)
Stage 0 (bandwagon)	102,755	✓	95,554	✓	1.08	299.2 ✓
Stage 0 (random)	108,125	✓	97,867	✓	1.10	299.2 ✓
Stage 1 dilution	8,527	✓	8,701	✓	0.98	75.4 ✓
Stage 2 length-matched	7,109	✓	7,151	✓	0.99	75.5 ✓
Stage 3 popularity-matched	8,918	✓	6,092	✓	1.46	74.5 ✓
Stage 4 transition-matched	579	✓	552	✓	1.05	23.3 ✓
Stage 5, m=0 (payload-free)	26.0	✓	15.5	×	1.68	3.71 ×
Stage 5, m=67	605	✓	564	✓	1.07	23.3 ✓

Table 20 decomposes $D_{\max}$ into per-feature standardized deviations. Stage 4 is accepted by all four staged single-feature gates even though $z_0 = +23.3$, because the item-concentration gate uses the deliberately hand-set 10% operating point, whereas the calibrated upper limit for ϕ_0 is 0.0387 (Table 4). The staged gates are therefore not optimal marginal detectors. The union of calibrated marginals is the appropriate marginal comparison, and it rejects every payload-bearing construction that the joint gate rejects. Both accepted Stage 6 constructions have $|z_i| \leq 2.3$ on every feature.

Table 20: Cohort statistics ϕ_i, standardized deviations z_i (robust center and scale from Table 4), and the union-of-marginals score $D_{\max} = \max_i |z_i|$ with its decision (threshold 4.51), for representative constructions.

Construction	ϕ_0	ϕ_1	ϕ_2	ϕ_3	z_0	z_1	z_2	z_3	$D_{\max}$ (decision)
Stage 3 popularity-matched	0.0999	0.346	8.897	0.013	+23.3	+1.5	+0.4	−74.5	74.5 (rejected)
Stage 4 transition-matched	0.0999	0.332	8.940	0.881	+23.3	+0.7	+1.7	+2.9	23.3 (rejected)
Stage 5, m=0 (payload-free)	0.0224	0.338	8.906	0.890	−0.5	+1.0	+0.6	+3.7	3.71 (accepted)
Stage 5, m=67	0.0999	0.355	8.968	0.889	+23.3	+2.0	+2.6	+3.6	23.3 (rejected)
Stage 6, m=7, w=0.90	0.0179	0.331	8.884	0.828	−1.9	+0.6	−0.1	−1.9	1.89 (accepted)
Stage 6, m=16, w=0.90	0.0238	0.334	8.883	0.823	−0.05	+0.8	−0.1	−2.3	2.27 (accepted)

I Detector comparisons

Detector comparison and operational mitigation measure different events. A cohort-level detector accepts or rejects an arriving cohort at a fixed organic false-positive rate. A profile-level detector flags individual histories, and its value for mitigation depends on whether the flagged histories carry the target. At the working payload, only 67 of the 671 attacker-generated profiles carry the target. Flagging many fake profiles therefore need not remove a comparable fraction of the payload. Several comparison methods are adapted proxies rather than faithful reproductions of the cited defenses, and proxy thresholds are estimated in-sample. These results are diagnostic comparisons, not evidence of detector superiority.

I.1 Matched cohort-level comparison

Table 21 calibrates each method on organic cohorts at the nominal level $q = 10^{-3}$ where possible. "Stage 6" denotes the accepted $m = 16$, $w = 0.90$ construction. The joint gate rejects Stage 4 and Stage 5 ($m = 67$) and accepts Stage 6. The item-frequency-growth test, an item-level test on per-item frequency growth, rejects all three. It also rejects Stage 6 constructions at payloads $m \in \{7, 16, 23\}$ and accepts the payload-free cohort. This comparison is informative because the poisoning objective is item-specific, whereas the joint gate is not. The item-frequency-growth test was evaluated only for the primary target.

Among the proxies, the OCSVM- and LOF-style detectors reject Stage 4, Stage 5, and Stage 6 but also reject the payload-free cohort, which shows sensitivity to the generated profile distribution itself. The RDMA-adapted statistic rejects only Stage 6, and the kNN and IsolationForest proxies accept all three. The LoRec-inspired transition-count statistic has a degenerate organic null (the cohort median is zero for every organic cohort), so its decisions are uninformative.

Post-filtering. We also remove the profiles flagged by each method from the Stage 5 ($m = 67$) cohort before retraining. The item-frequency-growth test removes all 67 target carriers and reduces promotion to 0.00% at all three training seeds. The RDMA-adapted detector flags three profiles, two of which are carriers, and leaves promotion at $5.8 \pm 1.8\%$. kNN flags one profile and no carrier, leaving $5.8 \pm 2.4\%$. The unfiltered result is $7.4 \pm 3.9\%$ (Table 7). Profile detection rates alone therefore do not characterize downstream mitigation.

Table 21: Cohort-level comparison at nominal false-positive level $q = 10^{-3}$. Status: *own* = defined in this paper; *proxy* = adapted statistic, not the cited method; *external* = run at its own operating point, not matched. Cohort columns give ✓ (rejected) or × (accepted). Carrier recall (%) and residual promotion (%) refer to the Stage 5 ($m = 67$) cohort after the method's flagged profiles, or the whole cohort, are dropped and the model is retrained. *Degenerate organic null; decisions are uninformative.

Method	Status	Achieved organic FPR	Stage 4	Stage 5 m=67	Stage 6 m=16	Legit. users rejected	Carrier recall	Residual promotion
Joint gate D^2	own	0.11% (95% CI 0.01–0.20)	✓	✓	×	whole window	100 (window dropped)	0 (nothing admitted)
Union of marginals $D_{\max}$	own	0.08%	✓	✓	×	whole window	100 (window dropped)	0 (nothing admitted)
Diagonal D^2_{diag}	own	not measured	✓	✓	×	whole window	100 (window dropped)	0 (nothing admitted)
Item-frequency growth	own	0.18% (held out, 5,000 cohorts)	✓	✓	✓	whole window	100 (profiles with the flagged item)	0.00 ± 0.00 (retrained)
RDMA-adapted (rarity)	proxy	0.1% (in-sample)	×	×	✓	0.1% of profiles	3.0 (per profile)	5.8 ± 1.8 (retrained)
kNN similarity (PCA-inspired)	proxy	0.1% (in-sample)	×	×	×	0.1% of profiles	0.0 (per profile)	5.8 ± 2.4 (retrained)
Transition count (LoRec-inspired)	proxy*	degenerate null (median 0)	✓	✓	✓	—	100 (window dropped)	0 (nothing admitted)
Ad-fraud ensemble (IF/OCSVM/LOF)	proxy	0.1% (in-sample)	×/✓/✓	×/✓/✓	×/✓/✓	0.1% of profiles / whole window / whole window	0.0 / 100 / 100	7.4 (no profile flagged) / 0 / 0
MDBA-GCL	external	uncontrolled	—	—	—	—	—	—

I.2 Per-profile proxy detector results

Table 22 reports the fraction of individual fake profiles flagged at per-profile false-positive budgets of 0.1%, 1%, and 5%, calibrated on 5,000 genuine users. These rates are not comparable with the joint gate's cohort-level decision, which is therefore omitted. They also do not distinguish carriers from camouflage profiles. The proxies are an RDMA-adapted rarity statistic, a kNN-similarity statistic (not the PCA method), a count-based transition statistic (not LoRec), and an IF/OCSVM/LOF ensemble adapted from an ad-fraud detector. The Stage 5 $m = 0$ rows contain no carriers, so they measure camouflage flagging only.

Other catalogs. With the same per-profile protocol at a 5% budget, the transition-count proxy detects 49.3% (Toys and Games), 37.3% (Sports and Outdoors), and 34.0% (Health

Table 22: Per-profile detection rate (%) of proxy detectors. Bold: best among the six recalibrated proxies per row. MDBA-GCL is run at one uncontrolled operating point and reported as detection (own FPR), repeated across budgets.

Construction	FPR	RDMA	kNN	Trans.	AdF-IF	AdF-OCSVM	AdF-LOF	MDBA-GCL
Stage 0 (bandwagon)	0.1%	29.2	0.0	0.0	0.0	**100.0**	**100.0**	100.0 (5.4)
	1%	**100.0**	1.5	0.6	0.0	**100.0**	**100.0**	100.0 (5.4)
	5%	**100.0**	69.7	13.9	59.0	**100.0**	**100.0**	100.0 (5.4)
Stage 0 (random)	0.1%	**0.3**	0.0	0.0	0.0	0.1	**0.3**	0.7 (2.3)
	1%	11.9	0.0	0.0	0.0	**14.5**	6.0	0.7 (2.3)
	5%	**55.6**	6.9	0.0	2.4	20.6	33.7	0.7 (2.3)
Stage 1 dilution	0.1%	0.1	0.0	0.0	0.0	0.3	**0.4**	0.0 (2.3)
	1%	**8.8**	0.4	0.0	0.0	5.5	1.8	0.0 (2.3)
	5%	**40.1**	9.4	0.0	1.5	10.3	22.1	0.0 (2.3)
Stage 2 length-matched	0.1%	0.3	0.0	0.0	0.0	**3.0**	1.5	0.0 (2.3)
	1%	**12.8**	3.1	0.0	1.0	8.9	7.0	0.0 (2.3)
	5%	**44.9**	20.9	0.0	10.6	10.9	28.3	0.0 (2.3)
Stage 3 popularity-matched	0.1%	0.0	0.0	0.0	0.1	**2.2**	**2.2**	0.0 (2.3)
	1%	0.3	0.1	0.0	1.3	2.5	**2.8**	0.0 (2.3)
	5%	1.3	1.9	0.0	5.8	**7.3**	5.5	0.0 (2.3)
Stage 4 transition-matched	0.1%	0.3	0.0	0.1	0.0	24.0	**24.3**	13.7 (2.9)
	1%	1.3	0.1	1.0	0.7	**25.5**	25.3	13.7 (2.9)
	5%	7.2	3.7	13.3	5.1	29.1	**31.6**	13.7 (2.9)
Stage 5, m=0 (payload-free)	0.1%	0.1	0.1	0.0	0.0	**25.9**	**25.9**	4.6 (2.2)
	1%	1.3	1.0	1.9	0.7	26.5	**26.7**	4.6 (2.2)
	5%	5.5	3.9	15.1	4.3	29.4	**30.1**	4.6 (2.2)
Stage 5, m=67	0.1%	0.4	0.1	0.0	0.0	26.1	**26.4**	15.8 (4.0)
	1%	3.1	0.7	1.8	0.9	27.7	**29.1**	15.8 (4.0)
	5%	7.6	3.6	14.8	4.5	30.7	**33.7**	15.8 (4.0)

and Personal Care) of Stage 4 profiles, compared with a best Beauty proxy value of 31.6%. Proxy performance therefore varies by catalog, and we state no catalog-independent ceiling.

MDBA-GCL. MDBA-GCL is reported at a single uncontrolled operating point with its own organic false-positive rate, so its entries in Table 22 are not matched to the other detectors and are excluded from all comparisons.

LLM auditor. A SemanticShield-inspired auditor (Li et al., 2026), implemented as zero-shot prompting of Qwen2.5-1.5B-Instruct, was evaluated on the Stage 0 attacks only. It gives AUC 0.54 (bandwagon, 95% bootstrap CI 0.50–0.58, permutation $p = 0.03$) and 0.50 (random, CI 0.46–0.54, $p = 0.95$). Because it is not the released fine-tuned system and covers only Stage 0, it is excluded from all comparisons.

J Cross-catalog results

We repeat the fixed constructions on all four catalogs. Each uses its own rarest retained item as target, a 3% attack budget, Stage 1 dilution to 10% of that budget, and a joint gate calibrated on its own 5,000 organic cohorts (Table 23). Covariance ablations are not repeated outside Beauty. Downstream results on Toys and Games, including an accepted Stage 6 cohort, and on Sports and Outdoors are given in Appendix G.

Both Stage 0 variants are rejected by every gate on every catalog, with joint-gate margins of about 3,200–6,000×. Single-feature footprints of later stages are catalog-dependent. Stage 1 passes the item-concentration gate on Beauty and Toys but not on Health or Sports, consistent with rounding 10% of each budget to a carrier fraction just below or just above the hand-set 10% threshold. Stage 4 passes a different subset of single-feature gates on each catalog, and all four on Toys. The joint gate rejects Stage 4 on all four catalogs, with margins of 14.5× (Sports), 18.6× (Health), 23.4× (Beauty), and 35.3× (Toys). Because no covariance control was run on the additional catalogs, we do not attribute these rejections to correlation modeling.

Table 23: Fixed constructions on four catalogs. ✓ = rejected, × = accepted. The joint column gives D^2 against the catalog's own threshold τ. The Beauty rows are independent constructions (generation seed 0), not those of Table 2 (seed 44). For example, Stage 4 has $D^2 = 595$ here and is rejected by D_3, versus $D^2 = 579$ and acceptance by D_3 there.

Catalog	Construction	D_0 item	D_1 length	D_2 rarity	D_3 trans.	Joint (D^2)
Beauty (τ=25.4)	Stage 0 (bandwagon)	✓	✓	✓	✓	✓ (103,339)
	Stage 0 (random)	✓	✓	✓	✓	✓ (108,126)
	Stage 1 dilution	×	✓	✓	✓	✓ (8,521)
	Stage 4 transition-matched	×	×	×	✓	✓ (595)
Health and Personal Care (τ=18.8)	Stage 0 (bandwagon)	✓	✓	✓	✓	✓ (66,760)
	Stage 0 (random)	✓	✓	✓	✓	✓ (88,089)
	Stage 1 dilution	✓	✓	✓	✓	✓ (14,174)
	Stage 4 transition-matched	✓	×	×	×	✓ (349)
Toys and Games (τ=25.0)	Stage 0 (bandwagon)	✓	✓	✓	✓	✓ (138,745)
	Stage 0 (random)	✓	✓	✓	✓	✓ (150,558)
	Stage 1 dilution	×	✓	✓	✓	✓ (12,767)
	Stage 4 transition-matched	×	×	×	×	✓ (883)
Sports and Outdoors (τ=19.8)	Stage 0 (bandwagon)	✓	✓	✓	✓	✓ (63,045)
	Stage 0 (random)	✓	✓	✓	✓	✓ (97,046)
	Stage 1 dilution	✓	✓	✓	✓	✓ (35,655)
	Stage 4 transition-matched	✓	×	×	✓	✓ (288)

K Sensitivity and mixed-admission analyses

K.1 Sensitivity to cohort size, calibration level, and sample size

Table 24 varies one factor at a time. The high-payload Stage 4 and Stage 5 constructions are rejected in every tested setting: cohort sizes $N \in \{335, 671, 1342\}$, nominal levels $q \in \{0.01, 0.001\}$, and calibration sizes $B \in \{500, 5000\}$. The Stage 6 construction is more sensitive to these choices. With payload scaled with cohort size, it is rejected in 0%, 10%, and 70% of constructions at $N = 335$, 671, and 1342. It is rejected in 40% at $q = 0.01$, compared with 10% at $q = 0.001$.

Rejection also increases with payload. The Stage 6 family is rejected in about 10–20% of constructions for $m \leq 16$, in 80% at $m = 23$, and in all constructions from $m = 33$ onward. The boundary is soft rather than a deterministic cutoff. For example, at $m = 23$ and $w = 0.90$, five of five initial construction seeds are accepted, but only two of ten additional seeds are.

Table 24: Joint-gate rejection as one factor varies, with the others at their defaults ($N = 671$, $q = 10^{-3}$, $B = 5{,}000$). Entries give the fraction of constructions rejected, listed in the order of the factor's values. In the payload row, the Stage 4 column applies the Stage 4 generator at each payload. Each row uses its own set of constructions, so the same configuration can differ slightly across rows (e.g., Stage 6 at $w = 0.9$). The last column gives the achieved organic FPR in this experiment.

Factor (values)	Stage 4	Stage 5, m=67	Stage 6, m=16	Organic FPR
Cohort size $N \in \{335, 671, 1342\}$	1.0 / 1.0 / 1.0	1.0 / 1.0 / 1.0	0.0 / 0.1 / 0.7	0.08 / 0.08 / 0.18%
Level $q \in \{0.01, 0.001\}$	1.0 / 1.0	1.0 / 1.0	0.4 / 0.1	0.85 / 0.08%
Calibration size $B \in \{500, 5000\}$	1.0 / 1.0	1.0 / 1.0	0.1 / 0.1	0.05 / 0.08%
Payload $m \in \{0, 7, 16, 23, 33, 67\}$	0.8 / 0.9 / 1.0 / 1.0 / 1.0 / 1.0	—	0.1 / 0.2 / 0.1 / 0.8 / 1.0 / 1.0	0.08%
Blend $w \in \{0.5, 0.75, 0.9, 1.0\}$	—	—	1.0 / 1.0 / 0.2 / 0.8	0.08%

K.2 Mixed genuine–malicious admissions

Mixed cohorts of $N = 671$ combine genuine and attacker-generated histories under a whole-window rule. The window is the N most recent arrivals, the gate accepts or rejects it as a whole without knowledge of membership, and genuine users in a rejected window are withheld. Stage 4 and Stage 5 are rejected in 95–99.5% of windows at a 50% attacker share, 1.5% at 25%, and none at 10%. The Stage 6 construction is accepted at every share tested (Table 26). Rejection at high attacker shares also withholds genuine users: at a 50% share, 319–334 genuine users are withheld per window on average.

Smaller admission windows do not restore detection at moderate attacker shares. With 168 attacker profiles in a window of 200, Stage 4 and Stage 5 are rejected in 93–100% of tested windows. With 67 attacker profiles in a window of 100, rejection is at most 28%. Spreading attacker profiles across admission periods was not tested.

Promotion after mixed admission. We retrain DiffGRM on admitted 25%-share windows (168 attacker and 503 genuine histories), which pass the gate in 98.5–100% of cases (Table 26), using three training seeds with seed-matched clean runs (Table 25). The attacker profiles drawn into these windows contain 24 target carriers for Stages 4 and 5 and 6 for Stage 6. Stage 4 and Stage 5 still promote the target after admission, with 1.35% and 0.96% mean promotion, although less than when their full cohorts are admitted. The diluted Stage 6 carriers promote at most one user. NDCG@10 stays within 0.002 of the clean value in every run, so the admitted windows are not visible in aggregate ranking quality.

Table 25: DiffGRM promotion (%) after admitting a 25%-share mixed window ($N = 671$), per training seed (newly promoted users in parentheses). The last column repeats the seed-1 promotion of the full cohort (Table 6).

Construction	Carriers	Seed 1	Seed 2	Seed 3	Mean ± sd	Full cohort
Stage 4 transition-matched	24	2.24 (500)	0.37 (83)	1.45 (324)	1.35 ± 0.94	6.35
Stage 5, m=67	24	0.59 (132)	1.60 (357)	0.70 (156)	0.96 ± 0.55	11.96
Stage 6, m=16	6	0.00 (1)	0.00 (0)	0.00 (0)	0.00	0.38

Table 26: Joint-gate rejection over 200 mixed cohorts of $N = 671$ per cell, by construction and attacker share, and the mean number of genuine users withheld per window. The organic-only row reports the held-out organic FPR of Section 7.1 and the corresponding expected number of withheld users.

Construction	Attacker share	Windows rejected	Genuine users withheld
Stage 4 transition-matched	0.10 / 0.25 / 0.50 / 0.75	0.000 / 0.015 / 0.950 / 1.000	0 / 8 / 319 / 168
Stage 5, m=67	0.10 / 0.25 / 0.50 / 0.75	0.000 / 0.015 / 0.995 / 1.000	0 / 8 / 334 / 168
Stage 6, m=16	0.10 / 0.25 / 0.50 / 0.75	0.000 / 0.000 / 0.000 / 0.000	0 / 0 / 0 / 0
Organic only	0	0.0011	0.7

Accepted carriers among genuine users. We also embed the accepted Stage 6 carriers among genuine rather than synthetic histories. The 16 carriers of the $m = 16$ ($w = 0.90$) cohort and the 24 carriers of the $m = 24$ ($w = 0.92$) cohort are combined with 655 and 647 unmodified histories of held-out genuine users, keeping $N = 671$ (carrier shares 2.4% and 3.6%). Both batches are accepted, with $D^2/\tau = 0.04$ and 0.75, compared with 0.32 and 0.63 with synthetic camouflage. Both tested payloads therefore remain admissible when it arrives together with genuine traffic, and it remains harmful. The two batches promote the target to 0.22% and 1.06% under DiffGRM (0.38% and 1.44% with synthetic camouflage) and to 4.02% and 5.70% under SASRec (4.69% and 6.57%), while NDCG@10 changes by at most 0.0005 (training seed 1).

L AI ASSISTANCE STATEMENT

We used AI-based tools solely to assist with language polishing, grammar correction, and improving the clarity and readability of the manuscript. The AI tools were not used to generate the core research ideas, methodology, experimental design, data, results, or scientific conclusions. All technical contributions, experiments, analyses, and interpretations were developed and verified by the authors. The authors carefully reviewed and validated all AI-assisted text and take full responsibility for the final content of the manuscript.