Whatnot Refund Policy
Refund Abuse: Diagnosis, Policy, and 30/60/90 Rollout

To: Whatnot Commerce S&O Leadership **From:** Kevin Astuhuaman **Date:** April 21, 2026

Recommendation in one sentence. Approve ~$22.4K of refund spend concentrated in a top-quintile cohort of refund-requesting buyers (53.8% of approved $ from 24 buyers who generate 47.8% of all refund requests); ship a five-tier evidence-led policy that targets this cohort without penalizing cold-start buyers, aiming to reduce baseline-adjusted suspected excess by 25–40% by Day 60.

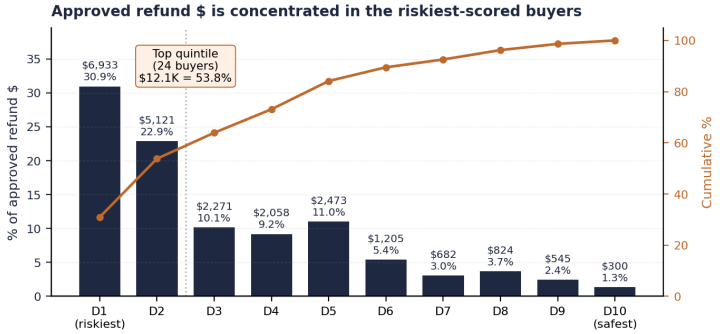
Assumptions (3). (1) "Gameable reason" (99% of refunds) is the eligibility universe, not an abuse label; the analytical lens is **buyer concentration** within it. (2) Tenure proxy = `buyer_total_orders` (no signup date in schema). (3) Descriptive analysis uses 12-month aggregates; **production policy computes features at refund-request time**.

1. Diagnosis — scale, drivers, exposure

Scale. 12 months, 2,697 orders, 297 refund requests, 228 approved. Approved refund spend = **$22,413** (8.1% of $278K observed order value). Requested = $30,459; denied (row-based) = $8,046.

Exposure is concentrated, not diffuse. Ranked by a 2-feature ex-ante risk-likelihood score (prior gameable-requested-$ ratio minus prior seller baseline, plus prior gameable-request-count percentile):

Cohort	Buyers	Approved $	% of approved $	Requests	% of requests
Top decile	12	$6,933	**30.9%**	92	31.0%
Top quintile	24	$12,055	**53.8%**	142	**47.8%**



(See attached Pareto chart — "Approved refund $ is concentrated in the riskiest-scored buyers." Risk deciles use only prior_excess_ratio and prior_gameable_request_count; approved $ is outcome, not part of the score. current_requested_refund_amount is never a tier-assignment input — only evidence burden and queue priority.)

Drivers. *Primary:* buyer concentration (top-quintile = 2.76× lift vs. random). *Secondary:* within the top-24 cohort, the top 3 hard-to-verify reason codes — "Marked delivered, not received" (37%), "Missing item" (22%), "Wrong item" (18%) — are **76% of requests vs. 31% outside** (+45pp); precisely the reasons that can be filed without proof. *Background:* high-value orders (Q4 by `order_value_usd`) = 59% of approved $ — worth mentioning, not a standalone tier driver.

Abuse vs. legitimate exposure (bounded, not a point claim): upper-bound $12,055 (top-quintile approved $) – seller-baseline expected refund-value proxy $358 = **baseline-adjusted suspected excess $11,697**; likely legitimate refunds ~$10,716 (outside-cohort + cohort baseline). The baseline-adjusted figure is what policy can plausibly address; the residual is not proof of abuse.

2. Policy — 5-tier action ladder

Ordered decision tree. **Overrides fire first** (positive buyer evidence approves; contradictions route to review). **Denial never on first touch** — requires formal Tier-3 or Tier-4 evidence process with reminder grace.

Tier	Specific logic (ex-ante features, prior/current only)	Action	Path back
1. Trusted	prior_excess_ratio ≤ P50 **AND** prior_gameable_request_count ≤ 1	Auto-approve	—
2. Standard	P50 < excess ≤ P80 **OR** count = 2 (non-hard-to-verify)	Auto-approve + education on 2nd gameable reason	—
3. Evidence required	P80 < excess < P90 **OR** `is_gameable_reason=1` paired with any prior gameable signal	Evidence request; **24h + reminder + 48h grace** before any escalation; graduated evidence by $ band	Evidence verified → approve
4. Manual review	excess ≥ P90 **AND** (count ≥ 3 **OR** prior_gameable_requested_$ ≥ P80) **AND** buyer_total_orders ≥ 5	Human reviewer approves or denies	Reviewer clears → approve
5. Deny + appeal	Tier-4 eligibility **AND** (evidence missing after reminder+48h **OR** submitted evidence directly contradicted **OR** ≥2 prior confirmed contradictions)	Deny with appeal loop; reviewer has factual deny authority on direct contradictions at any tier	Successful appeal or 90d clean → demote

Support floors (why Tier 4/5 can't trigger on ratio alone): needs `buyer_total_orders ≥ 5` (P25) **and** either 3+ prior gameable requests **or** ≥$397 (P80) prior gameable-requested $. Protects low-sample buyers from noisy ratios.

Evidence standards — top 3 hard-to-verify reasons (graduated by $):

Reason	Low-$ (≤P50)	Med-$ (P50–P90)	High-$ (≥P90) or repeat-risk
Marked delivered, not received	Buyer attestation + address confirm	Carrier claim/investigation or delivery-photo mismatch	Carrier investigation; police report only for high-$ stolen-package claims
Missing item	Unboxing photo	Unboxing photo/video vs. listing	Same + packaging weight/dimensions
Wrong item	Photo of received item	Photo vs. listing photo	Same + return of wrong item to seller

Damaged/misrepresented, buyer-seller dispute, and misbid/RTS use the same framework — full table in appendix.

3. Tradeoffs & guardrails

Tradeoff A — exposure reduction vs. refund-request friction. Top-quintile cohort = **47.8% of refund requests**, so Tier 3+ rules touch a material share of the queue (actual Tier 3+ volume reported post-simulation). **89% of observed orders (2,400/2,697) never file a refund request — their experience is unaffected.**

Tradeoff B — rule transparency vs. gaming adaptivity. Explainable rules + appeal loop chosen over opaque ML scorer. ML is a v2 evolution.

Guardrails (each with threshold, owner, pause action):

Guardrail	Threshold	Owner	Pause action
Appeal-overturn rate	>20% for 2 weeks	Trust/risk	Pause Tier 4/5 expansion; recalibrate
Manual-QA overturn on sampled Tier 3–5	>10%	QA lead	Halt new rollout; retrain reviewers
Repeat-purchase delta vs. matched controls	< −10pp at 60d	Data/analytics	Review false-decline cases; soften tier
Post-refund CSAT / complaint rate	CSAT drop >5pp or complaints 2×	CX ops	Pause; review evidence standards
New-buyer (`buyer_total_orders` ≤ 5) denial concentration	>15% of denials	Product	Tighten cold-start protection
Ops manual-review SLA	p90 turnaround >72h	Operations	Add capacity or tighten Tier 4 entry

Pre-launch legit-customer-risk proxy: of 24 top-quintile buyers eligible for 90d lookforward, **7 (29%)** kept ordering without repeat gameable requests — cohort-level false-positive signal to monitor in shadow mode before enforcement.

4. Rollout — 30/60/90

Owner functions: trust/risk, operations, data/analytics, product, CX. **Sequencing:** W1–2 shadow mode (score, don't act); W3–4 Tier 1–2 automation live; W5–8 Tier 3 evidence flow live; W9–12 full ladder + appeal loop instrumented.

Day	Success metric	Owner
30	Calibrated QA panel agreement ≥ 85% on 50-case blind double-review; sampled false-positive audit (20 Tier 4/5 cases, or all if <20) false-positive rate ≤ 10%; Pareto validated on fresh month	Trust/risk + Data/analytics
60	Pilot ambition: baseline-adjusted suspected excess in top-quintile cohort reduced **25–40%** (~$2.9K–$4.7K); appeal-overturn <20%; manual-review volume within capacity	Operations + Product
90	Sustained reduction; repeat-purchase and CSAT within ±2pp of baseline; manual-review SLA <48h; seller-side NPS unchanged	Trust/risk + Operations

Day-1 operational change: CX agents see tier + top-2 contributing signals + reason-code evidence standard on every refund request, instead of raw order history.

Appendix — interactive analysis: refund-policy.kevinastuhuaman.com · live policy simulator (calibrate tier thresholds, re-route the 297 historical requests), SQL query replay (DuckDB-WASM in-browser — 12 editable queries), 30/60/90 projection with pilot aggressiveness toggle, **every cited number verifiable on hover** (click any SQL badge → query runs locally, matches displayed value). **Also available:** **normalized workbook (xlsx)** with 5 clean tables (dim_buyer / dim_seller / fact_order / fact_refund_request / dim_buyer_risk_snapshot), data dictionary, 4 pre-built pivot tables + formula-driven policy simulator. Parquet bundle for DuckDB / Snowflake / Redshift replication at `/data/*.parquet`.

Note on data: the provided dataset is incomplete (no confirmed-abuse labels or account-age fields); thresholds are descriptive pilot values and would be recalibrated on production-scale prior-at-request features before Tier 4/5 go live. Evidence sources are Whatnot dataset only.