INCIDENT POST-MORTEM / INTERNAL DOCUMENT

Root Cause Analysis

By Mandy Hathaway AI Ethics Specialist & Technical Writer mandyhathaway.com

Content Moderation Model Drift: Elevated False Positive Rate Across Two-Week Production Window

INCIDENT SUMMARY

Incident ID	INC-2024-0314
Date Opened	March 14, 2024
Date Resolved	March 27, 2024
Duration	13 days
Severity	SEV-2 — Significant user impact, no data loss
System Affected	Meridian Content Moderation API v2.4.1
Author	Mandy Hathaway, AI Systems & Documentation
Reviewers	ML Engineering, Trust & Safety, Product Operations

IMPACT	Over a 13-day window, the content moderation model flagged approximately 18,400 user posts for policy violations that human reviewers subsequently cleared. False positive rate rose from a baseline of 3.2% to 21.7%. Approximately 6,100 accounts received erroneous automated warnings. No content was permanently removed without human review, but user-facing warnings caused measurable increases in support volume and account deactivations.

INCIDENT TIMELINE

Mar 1	**Scheduled retrain**	Model retrained on expanded dataset including Q4 2023 and Q1 2024 user-generated content. Standard validation metrics passed. Model promoted to production.
Mar 7	**First signal**	Trust & Safety team notes a 12% increase in appeal volume over the prior week. Attributed to normal variation. No action taken.
Mar 11	**Second signal**	Support tickets referencing erroneous content warnings reach 3x baseline. ML on-call reviews error logs but finds no anomalies in system health metrics.
Mar 14	**Incident declared**	Product Operations flags an 18-point jump in false positive rate on weekly review dashboard. SEV-2 declared. Incident response initiated.
Mar 15	**Model rollback**	Meridian v2.4.1 rolled back to v2.3.8. False positive rate begins declining within 4 hours.
Mar 17	**Root cause identified**	Analysis of training data diff reveals significant over-representation of satire and irony-heavy content from newly added sources. Model learned to associate ironic phrasing patterns with policy violations.
Mar 22	**Remediated retrain**	Model retrained with corrected dataset. New validation suite added. Staged rollout begins.
Mar 27	**Incident closed**	Full production rollout complete. False positive rate at 3.1%, within baseline. Affected accounts receive automated clarification and warning retraction.

ROOT CAUSE ANALYSIS

The primary cause of the incident was training data composition drift introduced during the March 1 retrain. Two newly added content sources — a political satire aggregator and a comedy community forum — contributed approximately 340,000 posts to the training corpus, representing 14% of the total dataset. Both sources contained high concentrations of ironic and sarcastic language that superficially resembled policy-violating content.

How the model learned the wrong pattern

The model was trained to detect hate speech and harassment using a broad set of linguistic features including capitalization patterns, specific phrase structures, and sentiment intensity. Ironic content — particularly political satire — frequently employs these same features deliberately. A satirical post mocking racist language, for example, contains many of the same surface features as an earnest racist post.

Prior training datasets had not included satire sources at significant scale, so the model had not been exposed to this pattern in sufficient volume to learn the distinction. The expanded dataset tipped the balance: the model learned that ironic phrasing correlated with violations because, in the new training data, it often did — just not for the reasons the model needed to understand.

FINDING

The model was not malfunctioning. It was doing exactly what it was trained to do. The problem was in what it was trained on.

Why validation did not catch it

The standard validation suite tested precision and recall against a held-out sample drawn from the same distribution as the training data. Because the held-out sample included the same satire sources, the model performed well on validation — it had learned the patterns in those sources correctly.

What the validation suite did not test was performance on out-of-distribution content: real user posts that did not resemble the training distribution. The held-out test set is only as good as its coverage of the real input space, and in this case that coverage had a significant gap.

Contributing factors

- No data source audit was performed before adding new sources to the training corpus. Source composition was reviewed for volume and recency but not for linguistic or topical characteristics.
- The appeal volume spike on March 7 was correctly identified as a signal but incorrectly dismissed. A formal threshold for appeal volume escalation was not in place.

- System health metrics monitored latency, error rates, and throughput — not model behavior. A model performing incorrectly at high speed looks identical to one performing correctly.
- The validation suite had not been updated since v2.1 and did not include irony or satire test cases, despite these being known edge cases in content moderation literature.

WHAT WE RULED OUT

Root cause analysis involves ruling out plausible causes as much as identifying the actual one. The following were investigated and excluded.

Infrastructure changes

No infrastructure changes were deployed between February 28 and March 14. The incident is not attributable to serving infrastructure, latency changes, or API version mismatches.

Label quality in existing training data

A sample audit of 2,000 training examples from pre-existing sources found a labeling error rate of 2.1%, consistent with historical baselines. Label quality in the existing corpus was not a contributing factor.

Threshold misconfiguration

The confidence threshold for automated flagging was reviewed and found to be unchanged from v2.3.8. The issue was in the model's probability estimates, not in how those estimates were converted to decisions.

Adversarial input

Early in the investigation, adversarial manipulation of the model was considered as a possible cause. Log analysis found no coordinated patterns in the flagged content suggesting deliberate exploitation. The distribution of flagged posts was consistent with normal platform content.

CORRECTIVE ACTIONS

P1	ML Eng	Add data source characterization step to all training pipeline runs. New sources must pass linguistic diversity	**Complete**

		and topic distribution review before inclusion.	
P1	ML Eng	Expand validation suite to include irony, sarcasm, and satire test cases. Minimum 500 examples per category sourced from real platform content.	**Complete**
P1	Trust & Safety	Define formal escalation threshold for appeal volume. Increase of >20% week-over-week triggers automatic ML review.	**Complete**
P2	ML Eng	Add behavioral monitoring to production metrics dashboard. Track false positive rate on a 24-hour rolling basis using sample human review.	**In Progress**
P2	Product Ops	Design and implement user-facing communication template for erroneous warning retractions.	**Complete**
P2	ML Eng	Implement staged rollout protocol for all future model versions. No model to reach full production without 48-hour canary period at 5% traffic.	**In Progress**
P3	Documentation	Write internal guide on training data composition best practices. Distribute to all teams involved in dataset curation.	**In Progress**
P3	ML Eng	Retrospective review of v2.3.x validation suite for other potential coverage	**Scheduled**

		gaps. Report due 30 days from incident close.	

LESSONS LEARNED

On training data

Data volume is not a proxy for data quality. Adding more data makes a model more confident in whatever patterns it finds — including the wrong ones. The question to ask before expanding a training corpus is not just 'is this data representative of what we want to detect' but 'does this data introduce patterns that could confuse the model's existing decision boundaries.'

On monitoring

System health metrics and model behavior metrics are not the same thing. A model can be fast, available, and entirely wrong. Behavioral monitoring — tracking what the model actually decides, not just whether it decides quickly — needs to be a first-class part of any production ML system.

On signals

The appeal volume spike on March 7 was a real signal. It was seven days before the incident was formally declared. In hindsight the threshold for acting on that signal was too high. Early signals in ML systems are often ambiguous, but the cost of investigating a false alarm is much lower than the cost of a delayed response to a real one.

On validation

A validation suite reflects the assumptions of whoever built it. Ours assumed that the distribution of test data would remain similar to the distribution of training data. That assumption held for three model versions and then didn't. Validation suites need to be actively maintained and stress-tested against known edge cases in the domain — not just updated when something breaks.