Building an ML Automoderation System for LLM-Augmented Online Platforms

Safety, scale, and decisioning for user-to-user and LLM-generated content

Andrei Shcherbinin
ML Engineer / ML Team Lead
Social Discovery Group

SOCIAL DISCOVERY GROUP

CONF42

Why Moderation Changes in LLM-Augmented Products

SOCIAL DISCOVERY GROUP

LLMs do not just increase content volume — they change the nature of abuse.

CONF42

Dating Platforms: Safety Is Product Quality

In dating, moderation is part of the core product experience, not a back-office function.

The Core Problem: Rare Violations, High Cost of Mistakes

Moderation is not a normal classification problem.

	Predicted normal	Predicted violation
Actually normal	OK	False positive: user friction PRODUCT FRICTION COST
Actually violation	False negative: safety risk SAFETY EXPOSURE COST	OK

FP = product friction

FN = safety exposure

> Under extreme imbalance, accuracy and ROC AUC can look good while missing the real question: how many dangerous messages we miss, and how many normal users we wrongly affect.

Accuracy is not the goal. Controlled risk is the goal.

Policy Is Part of the Model Specification

SOCIAL DISCOVERY GROUP

In moderation, policy is not documentation around the model — it is part of the model itself.

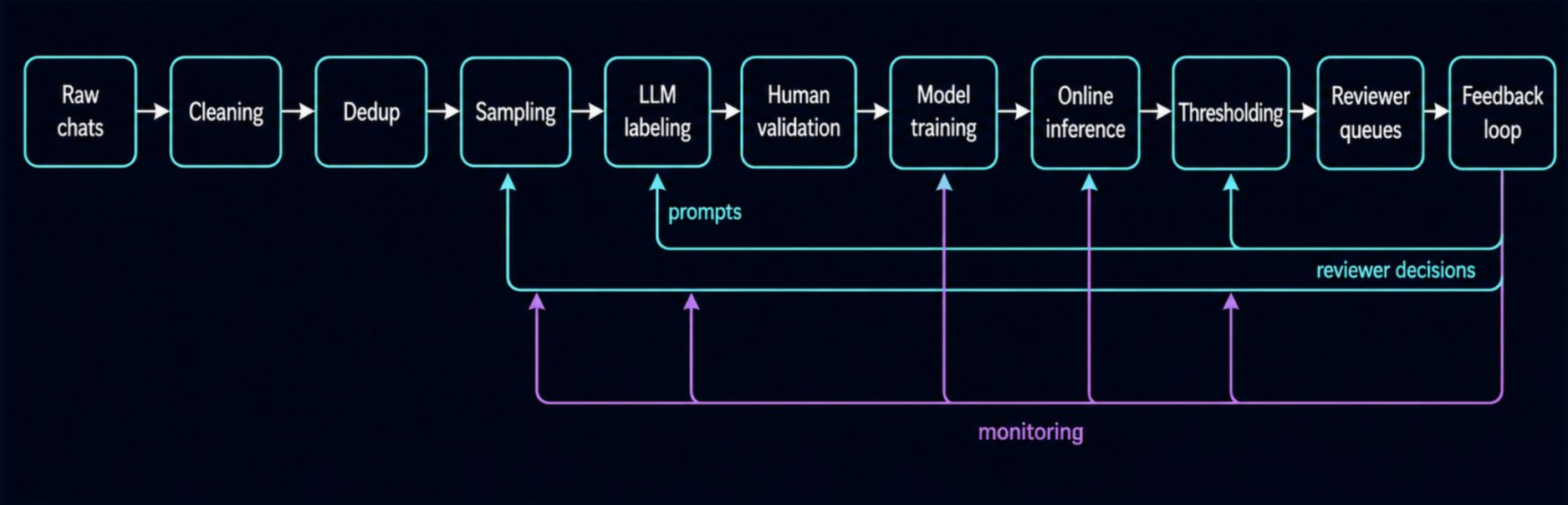
Policy object	ML object	Production object
Violation type	category label	reviewer queue
Severity	risk score	priority
Confidence	probability	threshold
Context	features / prompt	escalation
Reviewer decision	corrected label	retraining data

In strong moderation systems, taxonomy design, labeling instructions, active learning, and policy routing are core system components — not just data preparation.

End-to-End Moderation Blueprint

SOCIAL DISCOVERY GROUP

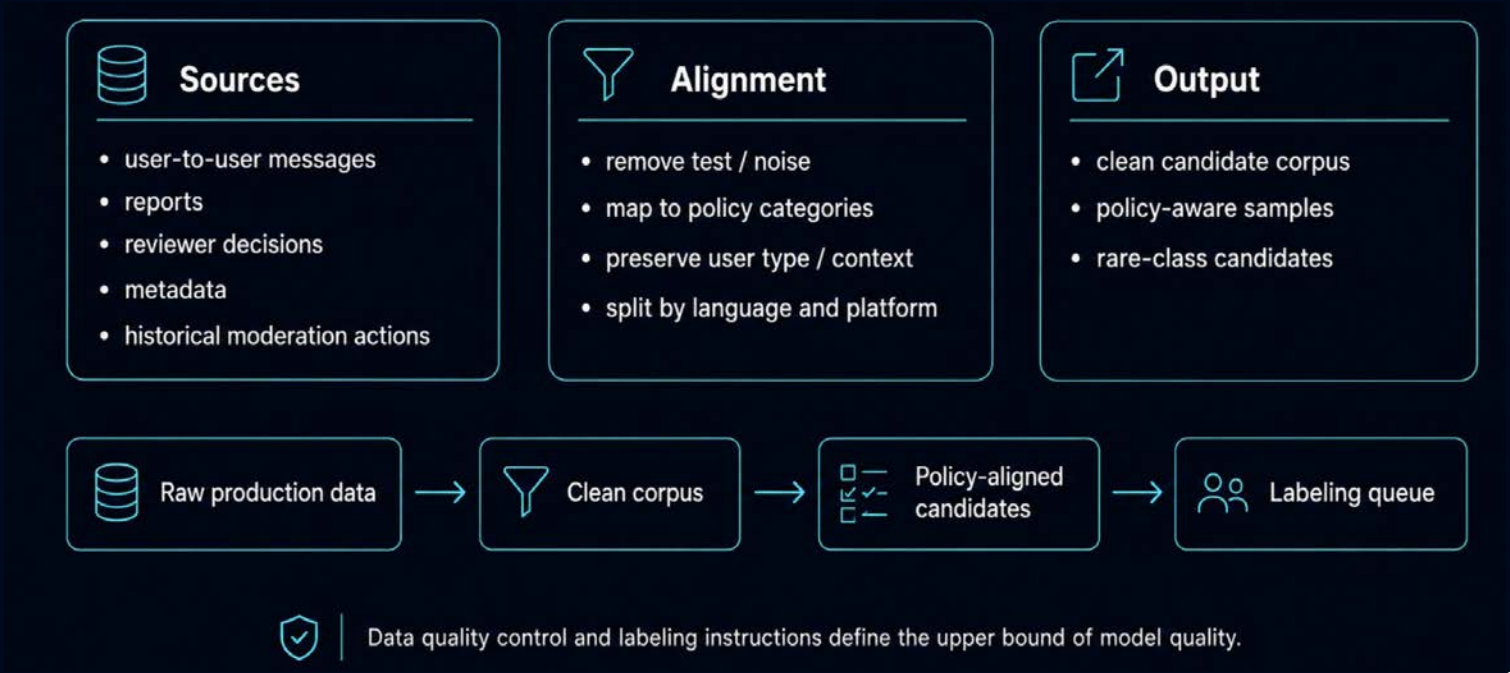
Automoderation is an operating system for safety decisions, not one classifier.



CONF42

Data Collection & Policy Alignment

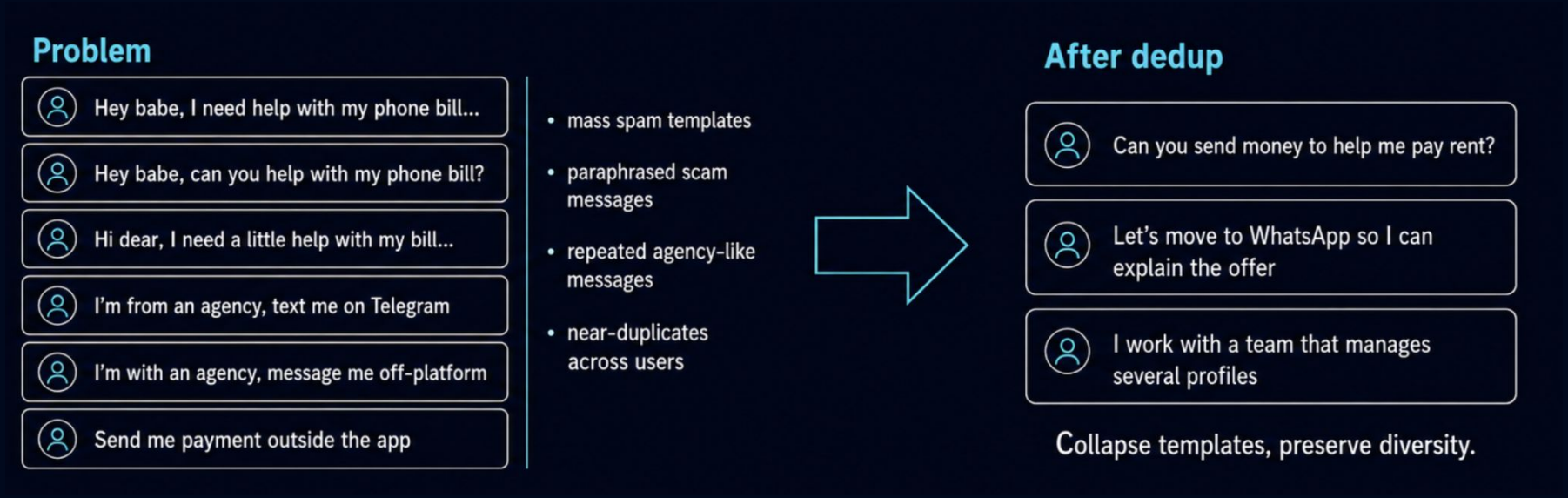
Raw messages are not training data until they are aligned with policy.



Cleaning & Dedup: Don’t Train on Templates

SOCIAL DISCOVERY GROUP

Without deduplication, the model learns templates instead of abuse semantics.

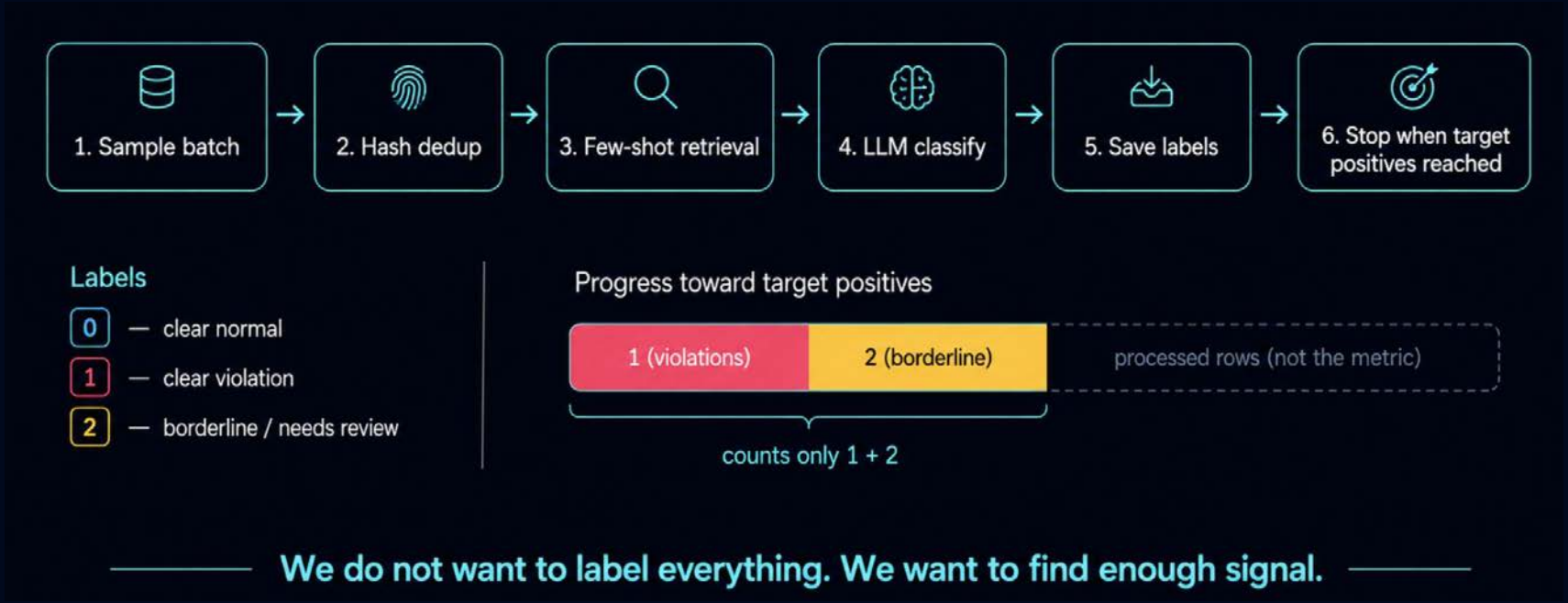


Deduplication is not just compression.
It is evaluation hygiene.

CONF42

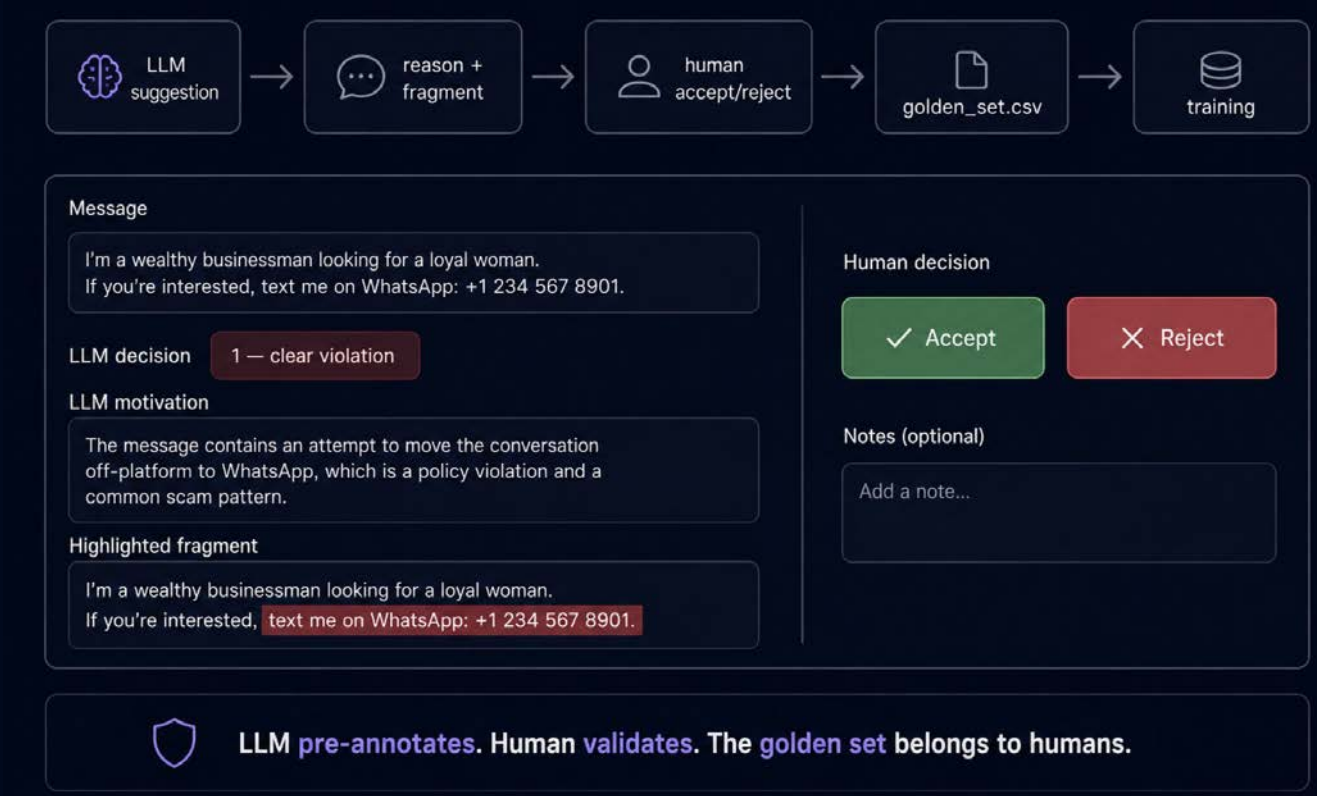
LLM Sampler-Labeler for Rare Classes

For rare violations, progress should be measured by positives found, not rows processed.



Human-in-the-Loop: LLM Labels Are Not Ground Truth

SOCIAL DISCOVERY GROUP



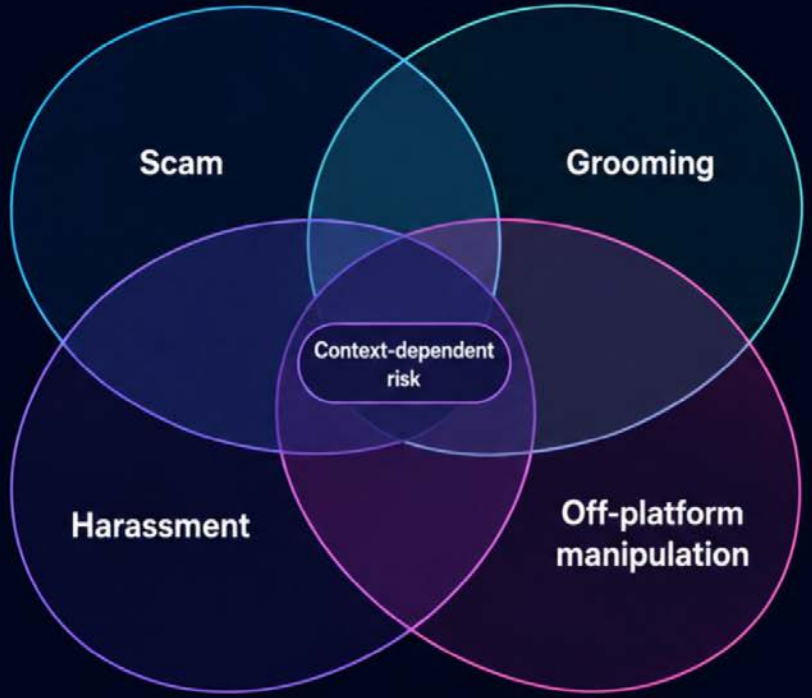
CONF42

Gray Zones: Where Generic Moderation Fails

The hardest violations often do not look like violations in one isolated message.

Examples

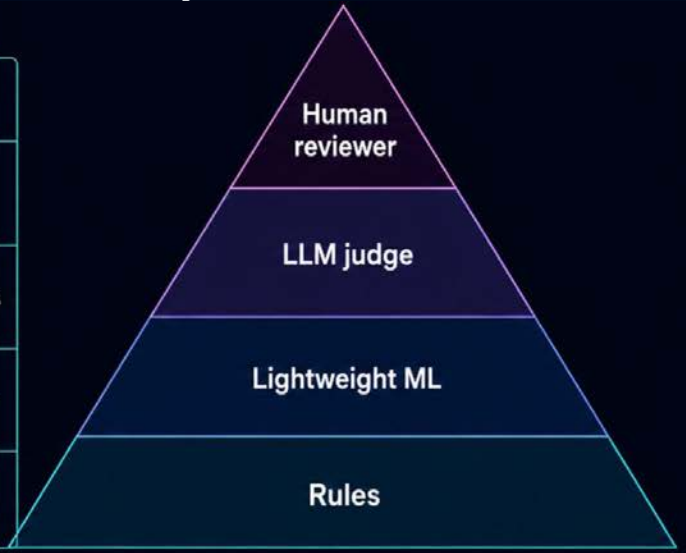
- grooming-like behavior
- harassment with indirect wording
- off-platform manipulation
- scam and romance fraud
- coded language
- culturally specific toxicity



Hybrid Model Stack

A hybrid stack wins because harms differ structurally.

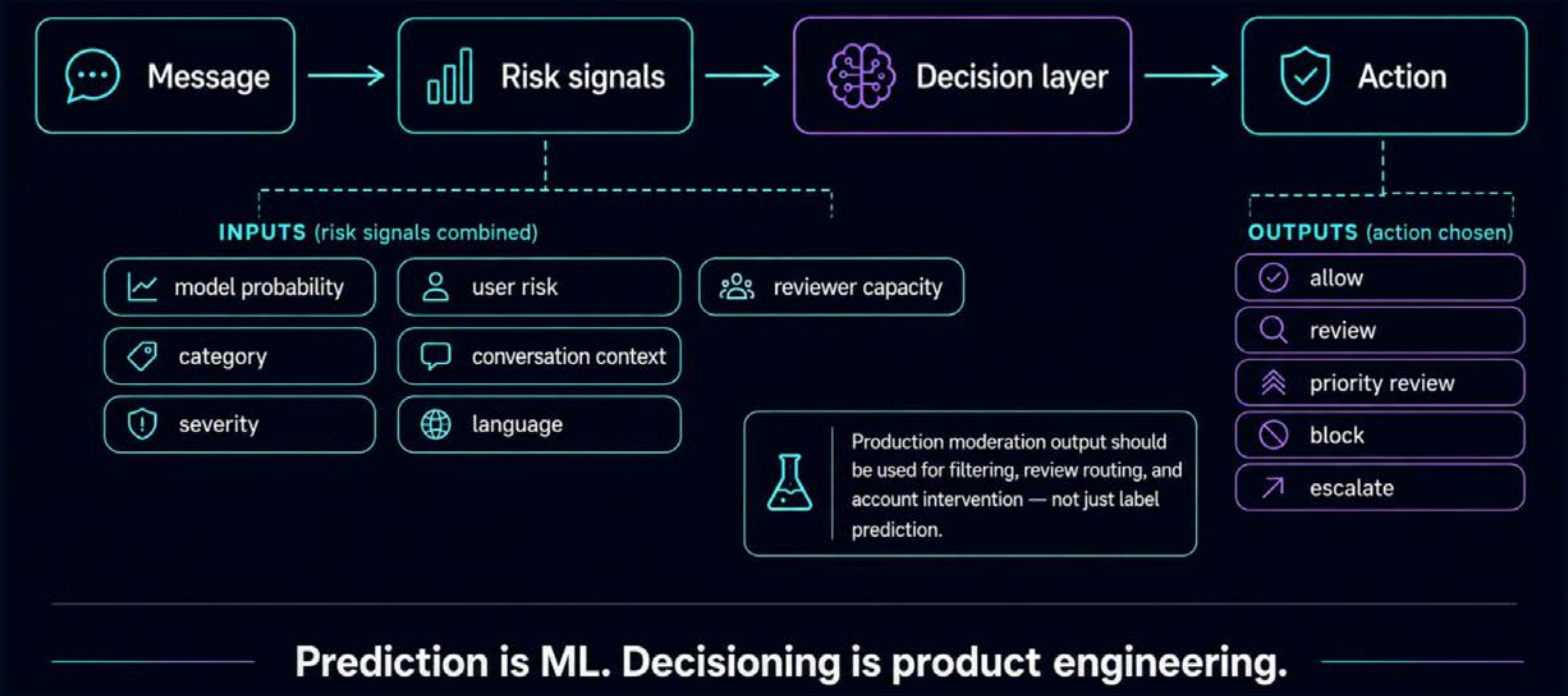
Layer	Best for	Why it works	Limitation
Rules / keywords	links, obvious abuse, obfuscation	fast, cheap, explainable	misses context
Lightweight ML	high-volume categories	scalable online inference	needs strong labels
LLM / MLLM	gray zones, reasoning, context	understands nuance	expensive, slower
Human review	final authority	policy judgment	limited capacity



Not all moderation tasks require expensive LLM inference. But nuanced harms require LLM / MLLM triage and human escalation.

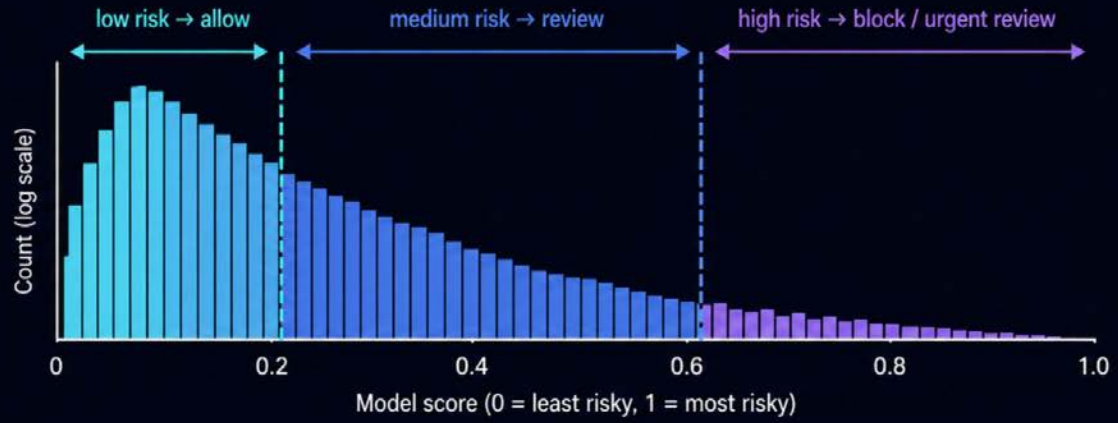
Real-Time Decision Layer

The model score is not the final moderation decision.



Thresholding Under Extreme Imbalance

Thresholds are operational controls, not fixed ML constants.



Thresholds need recalibration when policy changes, abuse patterns change, or the underlying model changes.

Category	Threshold style	Reason
Spam	higher automation	low ambiguity
Scam	high precision + escalation	business and user risk
Harassment	review-heavy	context-sensitive
Grooming	lower threshold for review	high safety cost

Evaluation & Error Analysis Loop

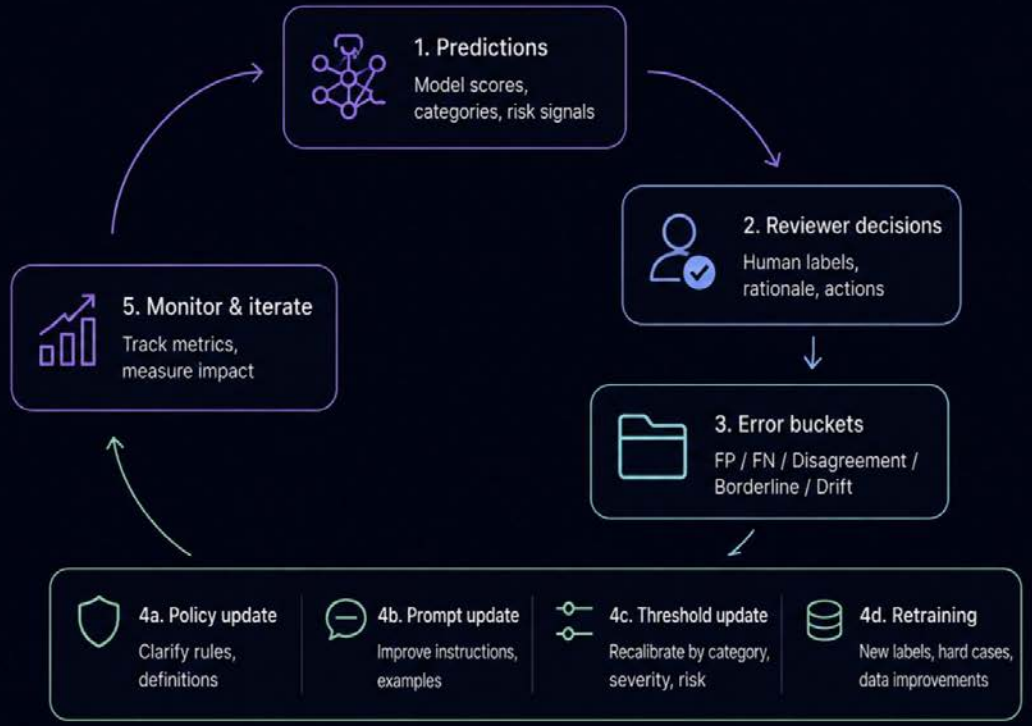
SOCIAL DISCOVERY GROUP

Metrics

- precision at low recall
- recall on high-severity classes
- false positive rate by category
- queue precision
- reviewer agreement
- time-to-review
- appeal / reversal rate
- drift by language, country, platform

Research hook

In subjective and imbalanced tasks, disagreement is signal.
Borderline cases should not be hidden — they should drive policy and model improvement.



Remember: Offline metrics are not enough for moderation quality.

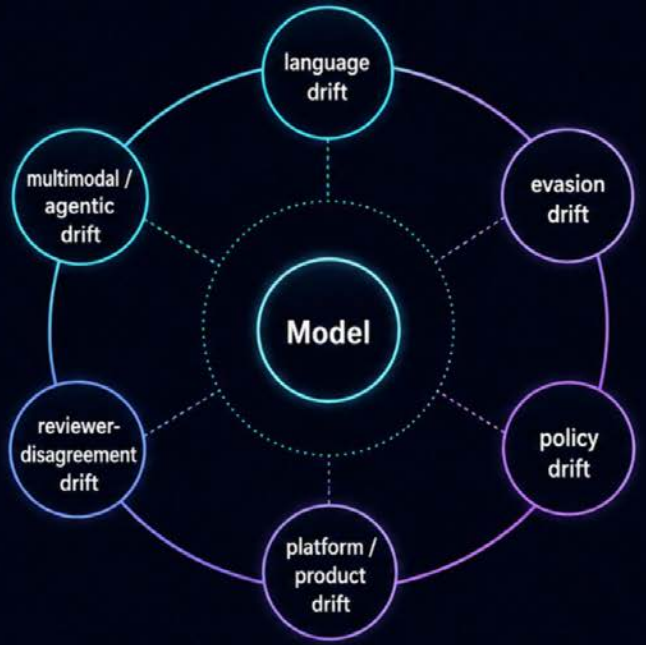
CONF42

Production Rollout & Monitoring

Abuse Evolves: Drift Is Not Only Statistical

In moderation, drift means users are adapting against the system.

Scammers actively change wording, use coded language, and move to indirect behavior.
Static evaluation becomes stale quickly.



Scam evolves faster than static benchmarks.

Impact

SOCIAL DISCOVERY GROUP

The system reduces manual load while preserving safety coverage.

Millions
of messages / day

-60%
reviewer load

90%+
violations caught

Important: Interpret with operational definitions.

- which policy classes
- which time window
- what "caught" means
- how borderline cases are counted
- whether reviewer denominator changed

Impact metrics in moderation are easy to misread without operational definitions because class imbalance, disagreement, and queueing effects change the denominator.

Key Takeaways

SOCIAL DISCOVERY GROUP

LLM-powered moderation works when it is designed as a system.

1. Policy first

Policy taxonomy defines what the model can learn.

2. Hybrid architecture wins

Rules, ML, LLMs, and humans solve different parts of the problem.

3. Rare-class thinking matters

Sampling, metrics, and thresholds must handle imbalance.

4. Queue design matters

Reviewer attention is the scarcest resource.

5. Feedback loop is the product

Abuse evolves, so moderation must continuously adapt.

SOCIAL DISCOVERY GROUP

Thank you!

LinkedIn QR

Andrei Shcherbinin
ML Engineer / ML Team Lead
Social Discovery Group

CONF42