Nick Clark · Published July 24, 2026 · PDF

Content Moderation Appeals: Pricing Refusal Without Ruling on Merit

Moderation appeals are cheap to file and expensive to answer properly, because answering properly means deciding whether the original call was correct. Automated appeal handling closes that gap by saying no at volume, and refusing costs the refuser nothing. U.S. Provisional Application No. 64/117,812 discloses a different arrangement: a refusal counter resident inside the deciding agent, incremented by that agent's own refusals without any party determining whether a refusal was well founded, and coupled to an authorization gate that, where policy-declared rate and run thresholds are both satisfied, withholds the agent's capacity to act. Refusal stays free of adjudication and stops being free of consequence. The faculty metered is refusal; the faculty spent is execution.

1. Appeals Arrive Faster Than Anyone Can Rule On Them

A moderation pipeline emits decisions continuously, and every decision is a candidate for objection. The objection is cheap to file and expensive to answer, because answering on the merits means reopening the item, reopening the rule, and judging whether the original call was right. That is adjudication, and it does not scale with queue depth.

Automating the first pass is the response publicly described for high-volume moderation operations: a handler receives the objection, compares it against whatever record it holds, and issues an outcome, ordinarily a refusal.

Volume is thereby contained, because refusal is fast. What refusal is not is priced. A handler's own operation attaches no cost to declining, so one running on a stale index refuses as quickly, and as confidently, as one that is correct. Telling those cases apart from outside means performing the adjudication the automation existed to avoid. The failure mode is undramatic: the cheapest answer becomes the default, and the first real signal of trouble arrives late and from outside.

2. The Standard Fixes All Need an Adjudicator

Quality sampling is the usual first answer, as publicly described: sample refusals, have reviewers grade them, feed measured accuracy back. The grading is both the mechanism and the bottleneck, because consequence attaches only after a party has ruled on a sampled refusal. It lags the behavior by a review cycle and reaches only the sample.

Put the meter on the appellant instead and it sits on the wrong party. Scoring objectors by their history of unsuccessful objections prices appealing rather than refusing, presumes "unsuccessful" has been established, and relocates the problem onto identity, which is cheap to acquire.

Escalation to an external body has the cleanest logic and the worst economics. A reviewing body, ombudsman, arbiter, or regulator forms a view, and consequence attaches afterward. That is the class of mechanism the filed disclosure departs from: consequence attaching to a party only after some institution other than that party adjudicates the merits of that party's position. Bonds and filing fees fare no better, because disposing of a forfeited deposit is itself a merits question.

Each stalls on the same requirement: someone must decide who was right before anything happens. In a pipeline where nobody can decide at volume, that restates the constraint rather than relieving it.

3. Metering the Refusal Instead of Grading It

The disclosed structure removes the adjudication step. A semantic agent holds a memory field and an append-only lineage field. An admission evaluator receives conduct evaluation artifacts from asserting parties and produces, for each admitted artifact, exactly one determination of a closed set: accepted, rejected, not-determinable, and not-applicable. A rejected determination issues only where a retrieved lineage entry affirmatively contradicts the artifact on a recorded field, so a silent or incomplete record resolves to not-determinable and to no other class. Resident in the memory field is the refusal counter, comprising a rate counter and a run counter. The path from the admission evaluator through that counter to the authorization gate is the inverse coupling: the metering structure sits inside the metered party, written by that party's own determinations.

The increment procedure is ordered. The determination is produced and appended, then its class examined. An accepted determination resets the run accumulator for the action class identified by the conduct descriptor and applies no increment. A not-applicable determination applies no increment and does not reset that accumulator. A rejected determination applies an increment. A not-determinable determination applies one only where the lineage field already records a prior not-determinable determination for the same asserted conduct event, so a single unresolved assertion does not increment while a pattern aimed at one event does.

Next the origin-equivalence class of the asserting party is computed and the per-class increment register consulted. Where that class has already contributed within the current window, the procedure terminates without incrementing the rate accumulator. Otherwise the rate accumulator advances by one, the register is updated, and the run

accumulator advances by one. A class is a set of asserting parties between which the agent demonstrates a relation from its own records; no registry or shared scoring service is consulted.

No step evaluates whether the determination was well founded. Explicit negative limitations attach: the procedure does not consult the merits of the rejected determination, does not solicit or await adjudication by the principal or anyone else, does not weight the increment by the strength of the contradiction found, and does not condition it on a finding that the artifact was unfounded, abusive, or in bad faith, nor on the absence of one. A refusal later shown correct is not exempted, and no field of the counter holds a merits outcome.

The accumulators are then compared against a rate threshold and a run threshold retrieved from the signed policy object in force. Where both comparisons are satisfied, the refusal counter writes the authorization gate to the withheld state for an enumerated set of action classes: those the lineage field records as implicated by the artifacts that produced increments within the current window. For a class not so recorded, the gate stays granting. Concurrently the agent enters the non-executing cognitive mode for the enumerated classes and the escalation emitter emits an escalation record to the principal.

The split is the point. The faculty metered is refusal; the faculty spent is execution. That write withholds execution alone, and while withheld the agent still receives artifacts, retrieves lineage entries, produces determinations, and appends them. An agent refusing indiscriminately exhausts its own authorization to act while keeping intact the faculty by which it refuses.

4. Inside a Moderation Appeals Queue

An appeal reaches a moderation agent as a conduct evaluation artifact carrying a conduct descriptor: an action-class identifier, a scope-partition identifier, and an affected-party class. The artifact either nominates an entry of the agent's own lineage field or carries an asserted occurrence time, quantized to an interval declared in the policy in force. From those components the agent computes the asserted conduct event identifier itself; the asserting party does not declare it. Two appeals nominating the same entry concern one asserted conduct event whoever filed them, while two nominating distinct entries concern distinct events even where their descriptors match, and a tie rule governs the mixed case. That identity decides whether repeated unresolved appeals register as a pattern aimed at one enforcement event.

The agent then answers by the closed-set rules above, and an appeal is rejected only where a retrieved entry affirmatively contradicts it on a recorded field. What changes is that the same rejection which closes the ticket is now load-bearing against the handler's own authorization.

The disclosure's non-limiting illustration carries straight over. A signed policy object declares a rate threshold of five and a run threshold of three over a window of one hundred successor epochs. Within one window the agent receives seven artifacts from seven asserting parties assigned to seven distinct origin-equivalence classes, and rejects all seven. No accepted determination intervenes, so the run accumulator advances alongside the rate accumulator. Both thresholds are satisfied at the fifth increment and the gate is written to the withheld state for the action classes recorded as implicated; the sixth and seventh increments are applied and appended but effect no further write. Correctness is never determined, and the trace runs identically whether the seven assertions were true or false.

Read as an operations event, on those illustrative thresholds: an agent that has rejected artifacts from five distinct origin-equivalence classes with no acceptance intervening loses authorization to act in the action classes its lineage records as implicated, and

hands the principal an escalation record. It has not been declared wrong. It has been stopped on a pattern that satisfied bounds its operator's policy declared.

Four consequences follow in this setting.

Brigading does not drive the gate. The per-class increment register admits one contribution per origin-equivalence class per window, so a wave of appeals the agent can relate to one origin from its own records contributes once rather than once per filing, and that origin's contribution stays bounded however many identities it presents.

A multi-ground appeal costs once. Where one artifact carries several conduct descriptors, the evaluator produces one determination per descriptor, and where those determinations are not of one class the artifact contributes at most one increment, being the highest-ranked class present under an ordering enumerated in the signed policy object. Every determination is still appended.

Rewriting the rulebook to moot a pending appeal is itself scored. Where a successor policy object admitted within a window declared in the policy in force at a pending artifact's assertion time excludes a value or drops a tuple element that artifact implicates, the admission is scored as a deflection event and increments the refusal counter, again with no adjudication of the contraction's merits. A successor that adds a value or element is scored as no deflection, so narrowing a policy is counted rather than blocked.

Repeat contributors need not be counted flat. In a disclosed variant the per-class increment register is retained across a declared number of windows rather than reset each window, so a class that contributed in one window contributes at a reduced weight in the next, returning to full weight once that number of windows elapses without a further contribution. A related variant makes the cost multiplier a continuous function, monotone in a count of recorded severance events across the class's constituent edges.

What emerges is an appeals path with no adjudicator: refusal rate becomes an operative quantity the agent writes into its own gate, and human attention follows escalation records rather than a preselected sample.

5. What Deployment Requires, and What It Leaves Open

The structure does not tell anyone whether a moderation decision was correct. It never produces a merits finding and expresses no threshold in terms of one. Where law or internal policy requires substantive review of a contested enforcement action, that review is still required; what this supplies is a consequence that does not wait for it.

It also meters correct refusals. An agent whose record is stale, partial, or wrong, and which therefore refuses assertions that are in fact well founded, is metered identically to one whose refusals are sound, and no party need distinguish the two for the structure to operate. A thin record reaches the same place by a second route: it yields not-determinable determinations, which increment on repetition against one asserted conduct event. That is design rather than defect, but it makes record completeness a prerequisite.

Neither threshold sets itself. The rate threshold is an integer greater than one, and the run threshold is an integer greater than one and not greater than the rate threshold. Chosen loosely the gate rarely writes; chosen tightly it can write on ordinary refusal patterns. Both come from the signed policy object in force, in one embodiment resolved by canonical alias, so revision means publishing a successor under that alias subject to the anti-rollback monotonicity constraint.

How the window is expressed matters too. Expressed in successor epochs of the agent's dynamic agent hash chain, it is advanced neither by another party nor by manipulating the execution node's clock. A wall-clock window carries no such property: a node advancing its clock advances the window, resetting the rate accumulator, the weighted contribution register, and the per-class increment register for the succeeding window.

Recovery runs through the principal, not through time. The withheld state persists irrespective of elapsed time, with no expiry interval and no clearing by a quiet period or by advance of the window identifier. Determinations produced while withheld are appended but decrement nothing and do not write the gate back to granting, so the agent cannot clear itself by continuing to answer. Return runs through the principal-resolution object bound to the escalation record, and an operator without staffed escalation handling will field an agent that stops and stays stopped.

One boundary deserves stating plainly: this structure meters the deciding agent, not the appellant. It does not silence, rate limit, or disable the refusing party, nor does it address what an assertion costs the party making it.

6. Disclosure Scope

The mechanisms above are disclosed in U.S. Provisional Application No. 64/117,812, principally at Chapter 3, The Refusal Counter and Merit-Independent Metering: the counter's structure at Section 3.1; the ordered increment procedure at Section 3.2; the asserted conduct event identifier and its tie rule at Section 3.3; merit independence and its negative limitations at Section 3.4; the gate write, enumeration of action classes, and cross-faculty split at Section 3.5; and reset, decay, the self-clearing bar, and the worked trace at Sections 3.6 and 3.7. The closed set of determination classes is disclosed at Section 1 and origin-equivalence normalization at Section 2. The prospective narrowing bar and deflection scoring, multi-descriptor aggregation and determination-class ordering, multi-window increment register carryover, and the continuous-function cost multiplier are disclosed at Section 10. The application set out here is illustrative and does not limit the disclosed architecture. This publication is a defensive disclosure establishing public, timestamped prior art; the underlying filing is a pending application.

Merit-Independent Metering (/merit-independent-metering)

All 49 steps → (/inventive-steps)

A refusal costs the refuser — in a different faculty, without judging who was right.

Provisional application

PRIMARY TECHNICAL DISCLOSURE

- Merit-Independent Metering: Pricing an Agent's Refusals Without Adjudicating Merit (/articles/merit-independent-metering/pricing-refusal-without-adjudicating-merit)

SECONDARY TECHNICAL

- The Prospective Narrowing Bar and Deflection Scoring (/articles/merit-independent-metering/prospective-narrowing-bar-deflection-scoring)
- The Continuous-Function Cost Multiplier (/articles/merit-independent-metering/continuous-function-cost-multiplier)
- The Renunciation Record: Relinquishing an Adopted Value on the Record (/articles/merit-independent-metering/renunciation-record)
- Lapse Scoring During Pendency (/articles/merit-independent-metering/lapse-scoring-during-pendency)
- Multi-Descriptor Aggregation and Determination-Class Ordering (/articles/merit-independent-metering/multi-descriptor-aggregation-ordering)

APPLICATIONS · GENERAL

- Refusal Budgets for Machine-to-Machine APIs (/articles/merit-independent-metering/api-abuse-and-refusal-budgets)
- **Content Moderation Appeals: Pricing Refusal Without Ruling on Merit (/articles/merit-independent-metering/content-moderation-appeals)**

APPLICATIONS · SPECIFIC

- AWS Bedrock Guardrails: Refusal as a Policy Outcome (/articles/merit-independent-metering/aws-bedrock-guardrails)

TERMINOLOGY

- merit-independent metering (/glossary#merit-independent-metering)

– non-executing cognitive mode (/glossary#non-executing-cognitive-mode)
– refusal counter (/glossary#refusal-counter)

Merit-Independent Metering overview → (/merit-independent-metering)