Sentinel

Perimeter defense for platform write paths

Sentinel

Perimeter defense that makes automated abuse uneconomical

Open Feed Network, Inc. · candortrustandsafety.com *Every claim below maps to a dated behavioral proof in the public evidence record.*

The problem

Bot farms don't announce themselves. They arrive as dozens of individually plausible accounts and only look like an attack when you can see them together. Open registrations invite them; closed ones kill communities. Sentinel is the layer that sees them together — and turns an attack into a cost the attacker bears, not your moderators.

What Sentinel does

Sentinel is inline perimeter defense for a platform's write paths. It scores activity on **behavior, never identity** — coordination, not newness — and when it detects a coordinated attack it escalates the cost onto the attacker rather than simply blocking and absorbing the load itself.

How it works, in principle

Every request is scored on coordination signals: how fast actions arrive, how much "different" traffic shares network space, and how much supposedly independent content follows one template. Individual users are never judged on who they are — only patterns of behavior trip the perimeter.

Rejection is not the only response. Automated attackers probing the edge can be routed into a **tarpit** — connections held open and answered slowly, burning the attacker's resources instead of your moderators' time — and sustained attacks are met with **escalating friction that shifts the cost of continuing onto the attacker**, with overflow capacity provisioned automatically only when a genuine multi-source attack is underway and released when it subsides. The perimeter absorbs; the platform never notices. *Exact signals, thresholds, and escalation parameters are not published.*

Verified behaviors

- **Coordinated attacks are absorbed automatically.** Under a real multi-source attack, the defensive escalation tier fires on its own, applies attacker-side friction, and provisions overflow capacity that is cleanly decommissioned afterward. *Verified July 2026.*
- **Legitimate users are untouched.** Ordinary human traffic never reaches the escalation thresholds and passes normally, before and after an attack.

Honest limitations

Detection is behavioral and probabilistic — thresholds are tuned over time, not perfect on day one. Sentinel defends the perimeter; it is not content moderation after admission.