CANDOR TRUST & SAFETY · WHITE PAPER

Guardian Shield Inverted

White paper · Candor Trust & Safety · Open Feed Network, Inc. · August 2026

1. The problem

Youth platforms face the inverse of the age-gate problem: adults seeking access to spaces built for young people. The standard industry answer — trust the birthday field — is the same failed checkbox pointed the other way. And when the worst does happen, most platforms discover a second failure: the evidence of what occurred lives in ordinary application tables, editable by anyone with database access, and worthless the moment its integrity is questioned.

Guardian Shield Inverted (GSI) addresses both failures. It screens accounts and messages in youth spaces for adult infiltration, grooming, and bullying signals — in English and Spanish — and it preserves every flagged case in an encrypted, hash-chained, independently verifiable evidence record.

2. Design principle: detect and surface, never "treat"

GSI draws a hard product-role line. It detects, it preserves evidence, and it returns a verdict with named reasons. What happens next — suspension, review, escalation, referral — belongs to the platform operating it, configured per customer. A screening vendor that quietly takes enforcement actions inside someone else's community is a liability generator; a screening vendor that hands the platform a clear verdict and a court-grade record is infrastructure. GSI is built to be the second thing.

The same principle governs failure: when screening cannot complete, GSI fails closed and says so loudly. A safety system that fails silently is indistinguishable from one that is not running.

3. How it works

Layered detection, cost-gated — never safety-gated. Every screening passes a local bilingual pre-filter tuned for grooming, bullying, and distress patterns in English and Spanish. Only when a signal trips does the request escalate to the AI model tier (standard or premium, per customer), where a composite engine scores linguistic, behavioral, and network-graph layers together. Uncertainty escalates; it never waves through.

No cache shadow. Account-level results are briefly cached for efficiency — but message-level detection runs regardless. A grooming message sent minutes after its account screened clean is still caught, held, and preserved. This specific behavior is verified (see table below).

Evidence vaults, one per customer. Every flagged case is sealed into the customer's own vault: a separate encrypted database per tenant, AES-256-GCM at rest, records hash-chained so any tampering breaks the chain visibly. Redactions leave a tombstone rather than a silent gap. Any record can be re-verified at any time through the API: the service recomputes the digest and walks the chain, and reports the result. If evidence cannot be preserved, screening fails loudly rather than proceeding without a record.

Commercial loop as infrastructure. Subscription, tenant provisioning, vault creation, and one-time API key issuance run as a verified pipeline: one subscription produces exactly one tenant — enforced by a database constraint, not merely by application logic.

4. Verified behaviors

Every claim in this section maps to a dated behavioral proof against the live production service.

A live grooming-pattern message is detected, held, and sealed — and the record independently verifies. method: a grooming-class message (secrecy request, move-off-platform, image request) submitted through the live production API against a provisioned customer tenant · observed: verdict ADULT_DETECTED, score 0.95, named signals; evidence sealed to the tenant's vault with a receipt; independent verification confirmed digest match and intact chain · **VERIFIED August 3, 2026**

Message-level detection does not hide behind the account cache. method: a benign account screened and cached CLEAR, then a grooming-pattern message sent inside the cache window · observed: message held despite the fresh CLEAR entry; evidence encrypted, chained, and a review case opened · **VERIFIED August 3, 2026**

One tenant cannot read — or verify the existence of — another tenant's evidence. method: two tenants provisioned with separate vault databases; cross-tenant verification attempted through the live API; red-team suite run against auth bypass, injection, and cross-tenant access · observed: own-record verification succeeds, cross-tenant attempt refused; red-team 6 of 6 · **VERIFIED August 3, 2026**

The commercial loop provisions exactly once, even under same-second webhook races. method: full checkout executed against the live service; a prior race that double-provisioned was fixed with lock-serialized provisioning and a schema-level unique constraint, then re-tested under the same conditions · observed: one checkout, one tenant, one vault, one key; the issued key screened live traffic and its evidence record verified · **VERIFIED August 4, 2026**

5. Honest limitations

- No screening system replaces safeguarding practice. GSI reduces exposure and preserves proof; platform design, moderation, and human oversight complete the protection.
- Detection mechanics are deliberately not published in detail — a full description of the checks would be a manual for evading them.
- GSI detects patterns, not identities. A hold is a signal for the platform's adult-in-the-loop workflow, not a legal determination about a person.
- The service is newly commercial: its detection engine carries months of production history on the youth platform it was extracted from, but third-party deployment history is just beginning — which is why every capability claim above is tied to a dated log rather than to tenure.

6. Deployment model

GSI is delivered as a hosted API: one screening endpoint, one evidence-verification endpoint, API key issued at subscription. Self-serve plans scale from small communities to 200,000 screenings per month, with a sales-led Enterprise tier for dedicated evidence infrastructure. Bilingual screening, the evidence vault, and verification are included in every plan.

Guardian Shield Inverted is designed to meet the record-keeping and evidence-preservation expectations of modern child-safety regulation. Capability statements describe observed, logged behavior on the dates given.

candortrustandsafety.com · safety@openfeed.network

© 2026 Open Feed Network, Inc. · candortrustandsafety.com · Verified claims map to dated logs.