

Truth Shield

Credibility analysis with the reasoning shown

Truth Shield

Credibility analysis with the reasoning shown — never a black-box score

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

The problem

Disinformation moves faster than fact-checking, and opaque “trust scores” make things worse — a number nobody can interrogate is just another authority to distrust.

What Truth Shield does

Truth Shield analyzes public posts for credibility and produces an assessment **with its reasoning displayed** — what it weighed, what it found, why it concluded what it concluded. A reader can disagree, because they can actually see it. Content is labeled, not silently removed; the core is open source, free for the communities that need it most.

How it works, in principle

Each post is assessed for credibility signals and source quality, and returns a verdict, a confidence level, and a plain-language rationale shown alongside the content. The system is built to **protect legitimate speech**: opinions, satire, and ordinary low-stakes claims are labeled as such and never suppressed, and only provably false disinformation is acted on. It operates bilingually in English and Spanish, recognizing that disinformation campaigns frequently target bilingual communities. *The analysis prompts and scoring bands are part of the open-source core; this paper describes behavior, not tuning.*

Verified behaviors

- **A fabricated conspiracy claim is scored critical with its reasoning shown**, in both English and Spanish. *Verified 2026.*

- **Ordinary, low-stakes content is not penalized merely for lacking a citation**, and opinion is labeled opinion and never suppressed. *Verified July 28, 2026.*
- **A genuinely verified claim displays as verified in the feed**, not silently mislabeled. *Verified July 28, 2026.*

Honest limitations

Truth Shield surfaces signals and reasoning — it is not an arbiter of truth, and contested, evolving stories are exactly where human judgment matters most.