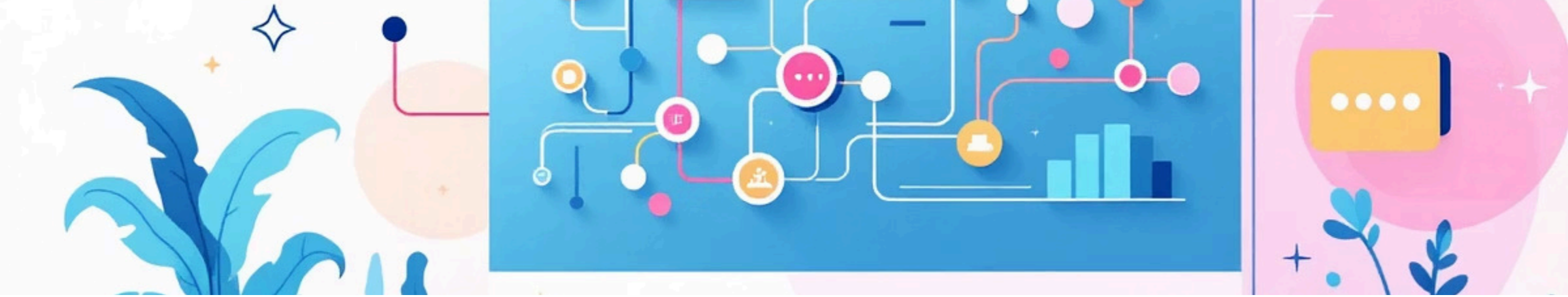
HarmLens

Content Moderation Co-Pilot



Not Just Another AI Chatbot

HarmLens is an API-first content moderation system designed for production environments. Unlike ChatGPT for ad-hoc analysis, HarmLens processes millions of posts automatically, routes content to moderation queues, and triggers actual platform actions.

ChatGPT vs HarmLens

Platform Use Case

ChatGPT: Ad-hoc chat analysis

HarmLens: Production moderation infrastructure

Speed

ChatGPT: ~5-10 seconds

HarmLens: <500ms

Scale

ChatGPT: Manual, one at a time

HarmLens: 100,000+ posts/hour automated

Integration

ChatGPT: Copy-paste responses

HarmLens: REST API, webhooks, batch

Actions

ChatGPT: Just analysis

HarmLens: Actually routes/removes content

Cost at Scale

ChatGPT: $0.50/post → $5,000/day

HarmLens: $0.001/post (self-hosted)

ChatGPT is like hiring a consultant. HarmLens is like buying infrastructure.

Key Innovation: Blockchain Audit Trail

Immutable Records

Every moderation decision recorded on blockchain

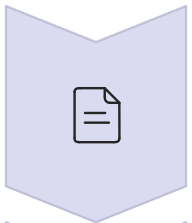
IPFS Storage

Full data stored on decentralized storage

Verifiable Compliance

Cryptographically verifiable for compliance

How HarmLens Works



Text Input

Post content extracted

Signal Extraction

5 core signals detected

Weighted Score

Risk calculation applied

Explainability

Reasons generated

Action Trigger

Platform routing

Five Core Signals

1

Emotion Intensity (30%)

Detects fear, anger, urgency patterns. Increases impulsive sharing likelihood.

2

Call-to-Action (25%)

Rule-based detection of share, forward, boycott keywords. Measures mobilization potential.

3

Toxicity/Targeting (20%)

Detects harassment, dehumanizing language, group targeting using BERT models.

4

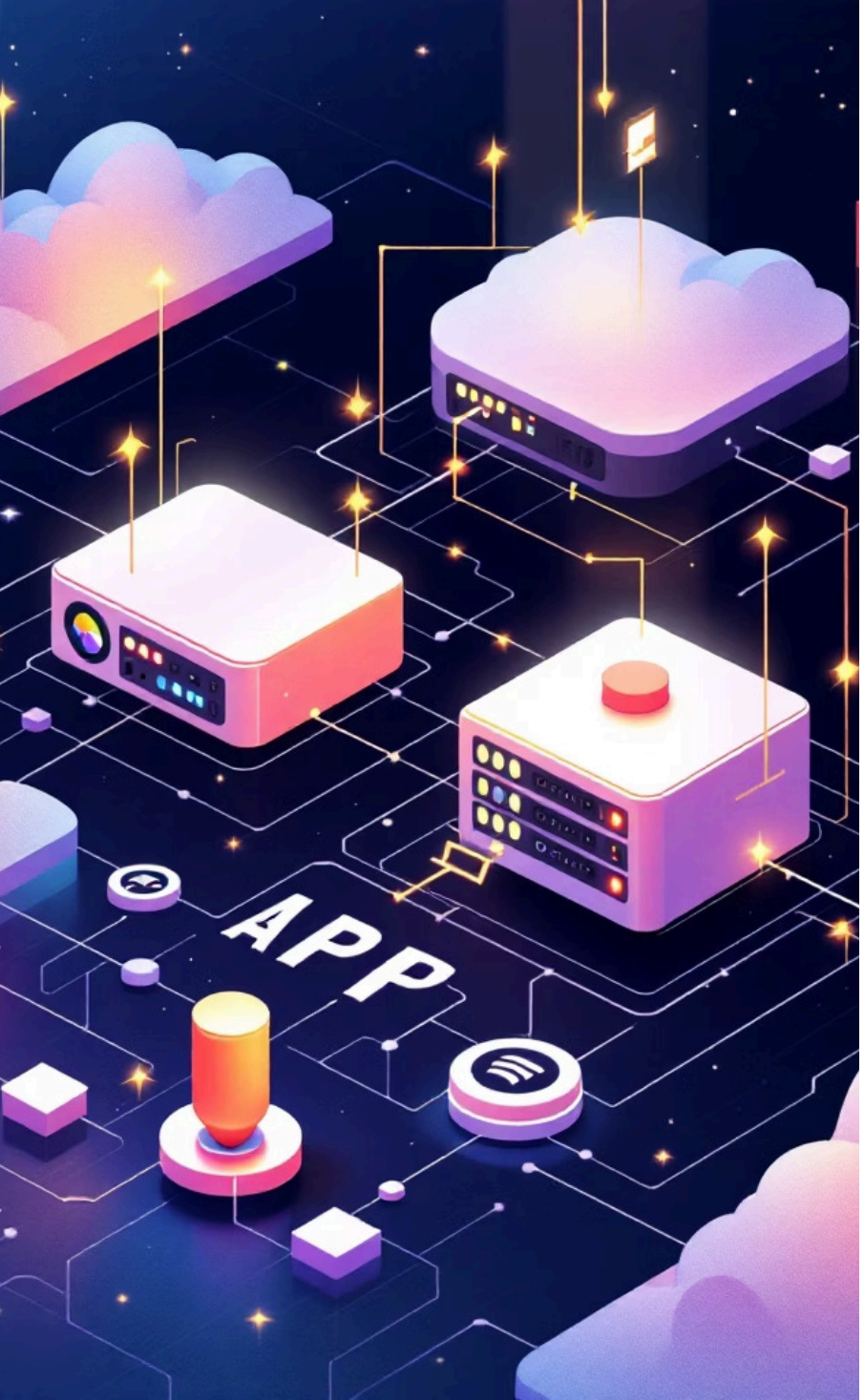
Context Sensitivity (15%)

Health, elections, communal tension, disasters. High-stakes contexts where misinformation escalates harm.

5

Child Safety (10%)

Rule-based guardrail detecting minor mentions with risky framing. Auto-escalates to specialized review.



Production-Ready Infrastructure

API-First Architecture

Platforms call endpoint, receive action recommendations. REST API with webhooks and batch processing.

Automated Workflows

Not just scores—actually performs actions: remove posts, route to queues, send alerts.

Built for Scale

Batch process 1M posts overnight. Self-hosted deployment eliminates per-request costs.

Compliance Ready

Full audit logs, explainable AI, GDPR-compliant, human-in-the-loop workflows, blockchain verification.

Quick Start Options

01

Demo UI

Clone project, install dependencies, run Streamlit demo at localhost:8501

02

API Server

Install FastAPI dependencies, run server at localhost:8000 with interactive Swagger docs

03

With Blockchain

Start local blockchain, deploy smart contract, run API server with immutable audit trail

Platform Integration Examples

Reddit Bot

Auto-moderate r/all in real-time. Streams submissions, analyzes content, removes high-risk posts automatically.

Webhook Alerts

Get instant notifications for high-risk content. HarmLens POSTs to your endpoint when triggered.

Batch Processing

Scan 100k posts overnight. Wake up to flagged content list for morning review.

Business Model & Pricing

Target Customers

Startups: Discord servers, Reddit alternatives, new social apps

Mid-size: Regional social networks, dating apps, gaming platforms

Enterprise: Large social media companies needing compliance

Revenue Model

- SaaS subscriptions (Professional tier)
- Usage-based pricing (Enterprise overages)
- White-label licensing (custom deployments)
- Professional services (custom fine-tuning)

99.8%

Cost Savings

vs ChatGPT per-post pricing

10-20x

Faster

vs LLM chat response time

$0.001

Per Post

Self-hosted infrastructure cost