Google Cloud H2S

Gen AI Exchange Hackathon

Team Name : Code Fellas222

Team Leader Name : Haritha

Problem Statement : AI-Powered Tool for Combating Misinformation

Project Name: FactForge

FactForge is a proactive, multimodal misinformation & scam-defence platform. It is designed to:

- Hunt scam/rumour artifacts across the open web and social channels.
- Index them into a retrieval-augmented knowledge base (RAG index).
- Provide fast, evidence-anchored verdicts and micro-lessons to users and community moderators (via web and mobile interfaces).

Problem Solved:

The rapid viral spread of scams and misleading content (including text, images, and links) results in financial loss, public-health risks, and misinformation cascades. Existing solutions typically fall short because they are:

- Reactive (waiting for user queries rather than proactively searching).
- Single-modal (failing to combine text, image, and metadata forensics).
- Lacking operational tools for communities and moderators to act quickly and safely.

What makes us different?

Differentiation Factor	FactForge Approach	Existing Solutions
Discovery	Proactive hunting via a hybrid crawler (Scrapy + Playwright) discovers emergent scams.	Only react to user queries or utilize static monitoring.
Evidence	Multimodal evidence collection: text + image forensics + metadata (WHOIS, payment IDs).	Single-modal analysis (often just text).
Verdicts	Explainability-first RAG returns structured JSON (verdict, trust-score, evidence list, micro-lessons).	Slow, traditional fact-checking with minimal educational feedback.
Flexibility	Dynamic Keyword & Pattern System allows real-time updates to capture new trends.	Cannot adapt due to static keywords list

Mechanics of FactForge

The solution is a multi-step pipeline that ensures rapid identification and effective containment:

1. The Hybrid Crawler finds new scams and artifacts.

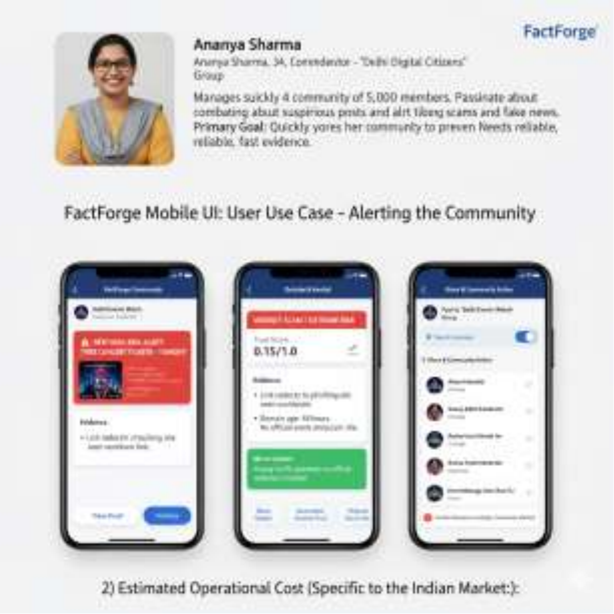
2. Enrichment workers extract provenance and payment signals.

3. The Classifier scores the risk and applies thresholds.

4. High-confidence artifacts are indexed into the Vector DB.

5. RAG retrieval surfaces these artifacts to the LLM explainer when similar claims appear.

6. The Social layer converts these verified detections into immediate actionable alerts for local communities and moderators.


FactForge

Ananya Sharma

Ananya Sharma, 34, Comendavtor - "Delhi Digital Citizens" Group

Manages suickly 4 community of 5,000 members. Passinate about combating abut susperious posts and alrt tibeg scams and fake news. **Primary Goal:** Quickly yores her community to preven Needs reliable, reliable, fast evidence.

FactForge Mobile UI: User Use Case - Alerting the Community

2) Estimated Operational Cost (Specific to the Indian Market:):

Core Features

1. Proactive Crawler & Indexer:
    - Scrapy + Playwright pipeline capturing HTML, screenshots, and redirect chains.
    - Includes Dynamic keyword CRUD operations via REST endpoints.

2. Enrichment & Forensics:
    - OCR (Tesseract), perceptual image hashing, and reverse-image signals.
    - Extraction of metadata (WHOIS/domain-age) and payment-IDs (UPI/phone/acc).
    - Utilizes a Pattern library (50+ indicators) for robust scam recognition.

3. Automated Screening & Scoring:
    - Heuristics + transformer classifier (scam_prob).
    - Weighted Scoring System factoring source reputation (0.3), content quality (0.25), factual consistency (0.2), emotional manipulation (0.15), and technical red flags (0.1).

Core Feature:

4. Vector Index & RAG:
   - Sentence-transformer embeddings indexed in Milvus/FAISS.
   - Retrieval-Augmented Generation (RAG) surfaces artifacts to the LLM explainer.
5. Explainability & Education Platform:
   - TrustMeter, evidence bullets, micro-lessons, and one-line verification tips.
   - Integration of Adaptive quizzes, gamification, and personalized learning paths.
6. Social Platform & Communities:
   - Public, global feeds, and post actions (upvote/comment/share).
   - Support for public, restricted, and private invite-only groups with moderator queues.
7. Security & Compliance:
   - HMAC-signed audit logs and PII redaction.
   - Tamper-evident signed logs to reduce false positives.

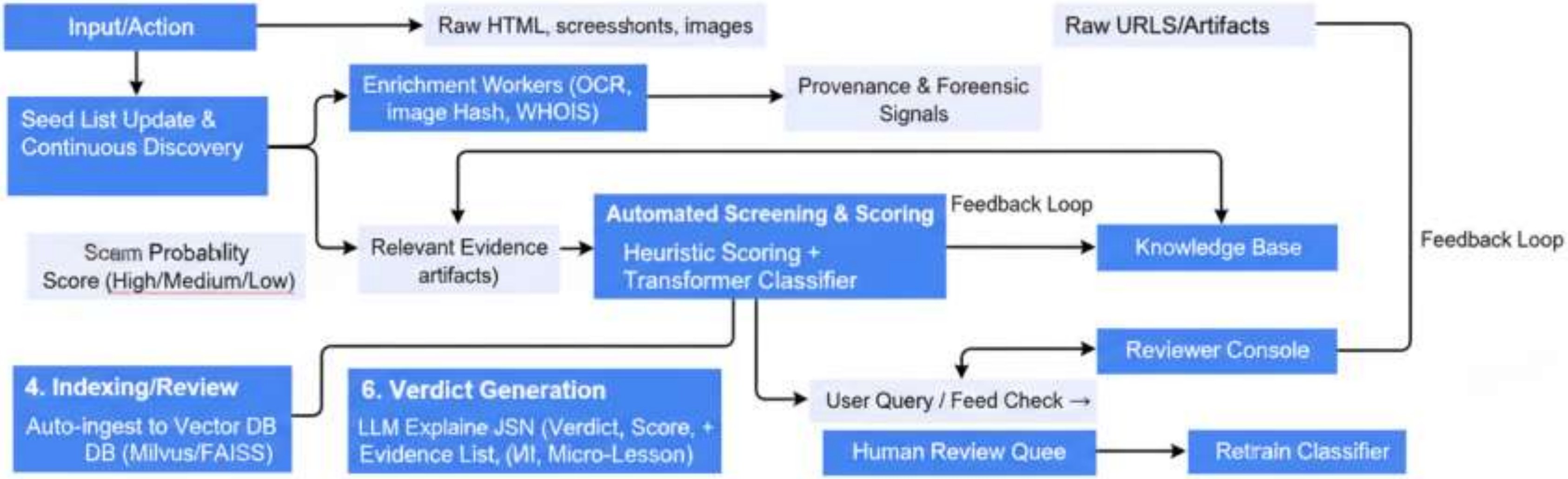
Process Flow

Step	Component / Action	Output / Destination
1. Discovery	Seed list update & continuous discovery	Raw URLs/Artifacts
2. Crawling	Crawler (Scrapy / Playwright)	Raw HTML, screenshots, images
3. Enrichment	Enrichment Worker (OCR, image hash, WHOIS)	Provenance and Forensic Signals
4. Screening	Heuristic scoring + Transformer Classifier	Scam Probability Score (High, Medium, Low)
5. Indexing/Review	**High-Score:** Auto-ingest to Vector DB (Milvus/FAISS) **Medium-Score:** Human Review Queue	Knowledge Base / Reviewer Console

Process Flow

Step	Component / Action	Output / Destination
6. Retrieval	User Query or Feed Check → RAG Retriever	Relevant Evidence (Vector DB artifacts)
7. Verdict Generation	LLM Explainer (uses RAG + classifier signals)	Structured JSON (Verdict, Score, Evidence List, Micro-Lesson)
8. User Action	Frontend (Check UI / Feed / Communities)	User receives verdict, shares to group, or posts to public feed
9. Feedback Loop	Reviewer Action on artifacts	Retrain Classifier

FactForge High-Level Process Flow



Input/Action

Raw HTML, screesshonts, images

Raw URLS/Artifacts

Seed List Update & Continuous Discovery

Enrichment Workers (OCR, image Hash, WHOIS)

Provenance & Forensic Signals

Feedback Loop

Scam Probablity Score (High/Medium/Low)

Relevant Evidence artifacts)

Automated Screening & Scoring

Heuristic Scoring + Transformer Classifier

Knowledge Base

Feedback Loop

Reviewer Console

4. Indexing/Review

Auto-ingest to Vector DB DB (Milvus/FAISS)

6. Verdict Generation

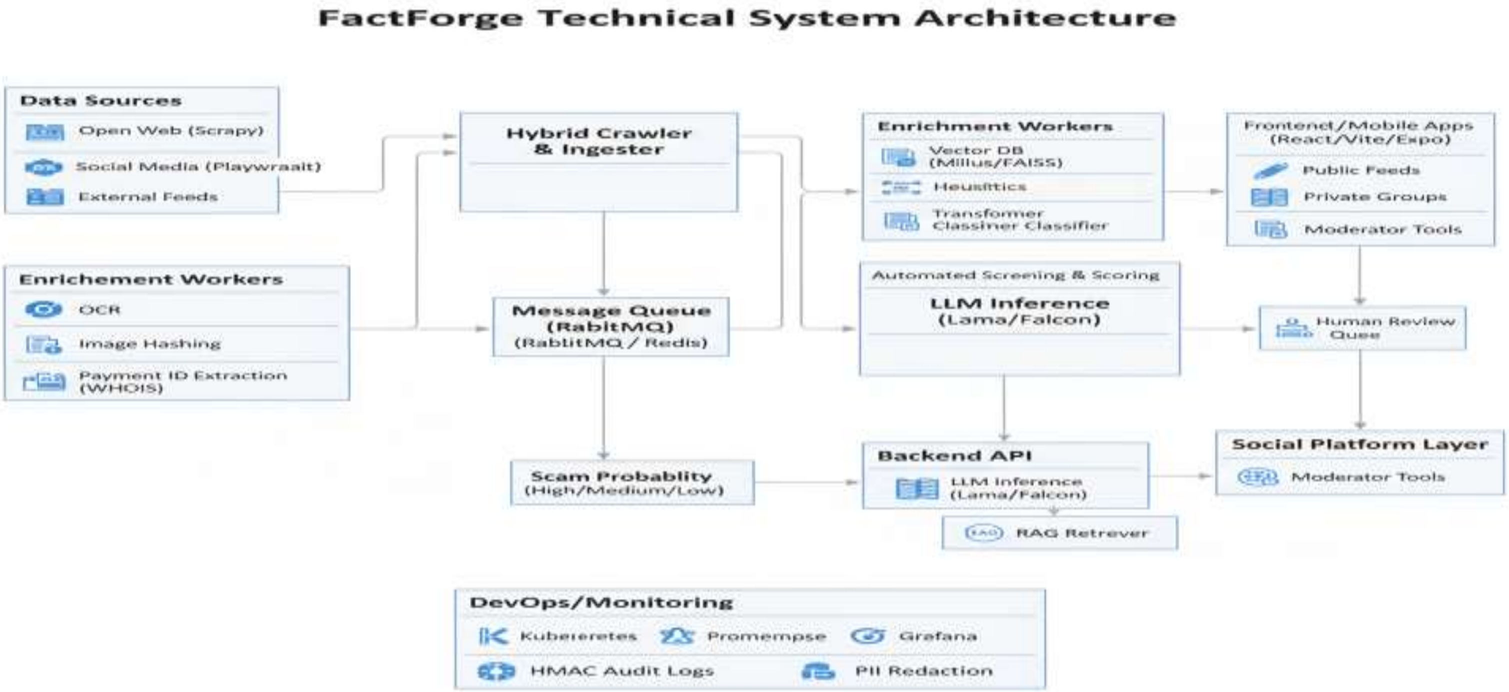
LLM Explaine JSN (Verdict, Score, + Evidence List, (MI, Micro-Lesson)

User Query / Feed Check →

Human Review Quee

Retrain Classifier

Architecture diagram of the proposed solution



FactForge Technical System Architecture

Data Sources
Open Web (Scrapy)
Social Media (Playwraait)
External Feeds

Enrichement Workers
OCR
Image Hashing
Payment ID Extraction (WHOIS)

Hybrid Crawler & Ingester

Message Queue (RabitMQ)
(RablitMQ / Redis)

Scam Probablity
(High/Medium/Low)

Enrichment Workers
Vector DB (Millus/FAISS)
Heusittics
Transformer Classiner Classifier

Automated Screning & Scoring
LLM Inference
(Lama/Falcon)

Backend API
LLM Inference (Lama/Falcon)

RAG Retrever

Frontenet/Mobile Apps
(React/Vite/Expo)
Public Feeds
Private Groups
Moderator Tools

Human Review Quee

Social Platform Layer
Moderator Tools

DevOps/Monitoring
Kubereretes
Promempse
Grafana
HMAC Audit Logs
PII Redaction

Technologies used:

Area	Key Technology
Frontend/Mobile	**React + Vite; Expo** (React Native via Bolt)
Crawler	**Scrapy + Playwright**
Backend/APIs	**Python FastAPI** (async)
Queue/Messaging	**RabbitMQ** or Redis Streams
Machine Learning	**Transformers** (Hugging Face)
Vector Processing	**sentence-transformers**

Area	**Key Technology**
Vector DB	**Milvus** (self-hosted) or FAISS
LLM Inference	**LLaMA/Falcon** (self-hosted via TGI/vLLM
Data Storage	**PostgreSQL + S3/Cloud Storage**
DevOps	**Kubernetes** (Prod) / Docker Compose (Dev)
Monitoring	**Prometheus + Grafana**

Why FactForge?

1. Tackling a High-Impact Problem
Misinformation & scams cause major harm (financial loss, public health risk).
Traditional tools can’t keep up with viral spread. FactForge is built for speed & impact.

2. Hybrid Approach for Better Results
Proactive Discovery: Hybrid crawler actively hunts emerging scams (not just keywords).
Explainability-First RAG: Every verdict comes with evidence + micro-lesson to build user resilience.

3. Actionable Containment
Detection alone isn’t enough.
Built-in social layer & private groups allow verified alerts to spread quickly and locally, reducing harm.

4. Scalable, Robust Architecture
Built on FastAPI + Kubernetes + Vector DB.
Solves persistence & scalability issues of earlier systems, enabling real-time, high-volume monitoring.

Google Cloud H2S

Gen AI Exchange Hackathon

Thank you