SMART INDIA HACKATHON 2025

- **Problem Statement ID –** 25161
- **Problem Statement Title-** Mitigating National Security Risks Posed by Large Language Models (LLMs) in AI-Driven Malign Information Operations
- **Theme-** Blockchain & Cybersecurity
- **PS Category-** Software
- **Team ID-** 74658
- **Team Name-** slayers

Case Study

Russia “CopyCop” LLM Disinformation Network (2023)

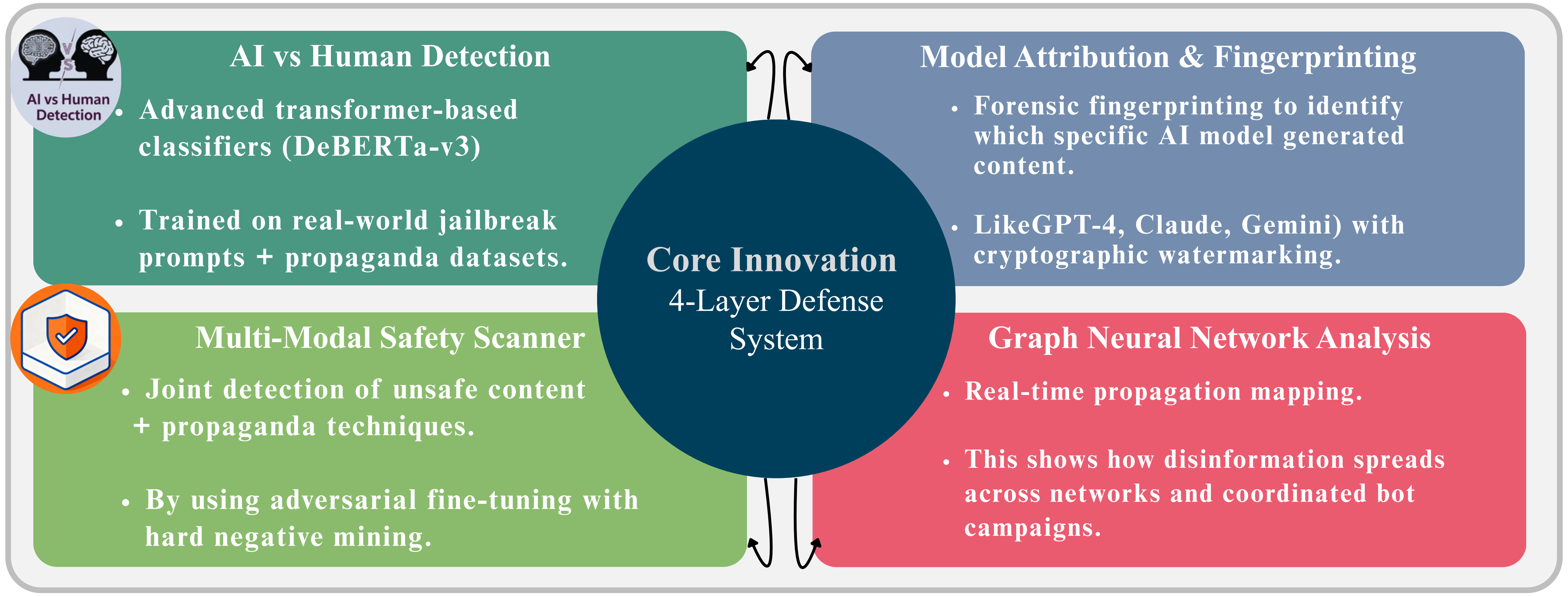
- CopyCop utilized large language models (LLMs), establishing over 200 fake news and fact-checking websites that imitated local media in the US, UK, France, Canada, and Norway, generating more than 19,000 AI-generated articles per month.
- Content tailored to polarize audiences on Ukraine, US politics, Israel-Gaza conflict, and European policy—aiming to erode support for Ukraine and discredit Western institutions.
- LLM prompt instructions (e.g., “take a cynical tone,” “target conservatives”) accidentally leaked in published articles, confirming AI manipulation
- Succeeded in seeding false narratives, risking trust in elections and media. Total loss is hard to quantify, but scale and strategic timing aim to destabilize democracies

CENTER FOR COUNTERING DISINFORMATION
FAKE NEWS
10101100110
01011011011
.COM
.ORG
.COM
.ORG
COPYCOP: RUSSIA'S LARGE-SCALE DISINFORMATION NETWORK TARGETING THE WEST

[Source of Information]

Our Solution: LLMaGen

India's First Comprehensive AI Content Detection & Forensics Platform



AI vs Human Detection

- Advanced transformer-based classifiers (DeBERTa-v3)
- Trained on real-world jailbreak prompts + propaganda datasets.

Model Attribution & Fingerprinting

- Forensic fingerprinting to identify which specific AI model generated content.
- LikeGPT-4, Claude, Gemini) with cryptographic watermarking.

Core Innovation

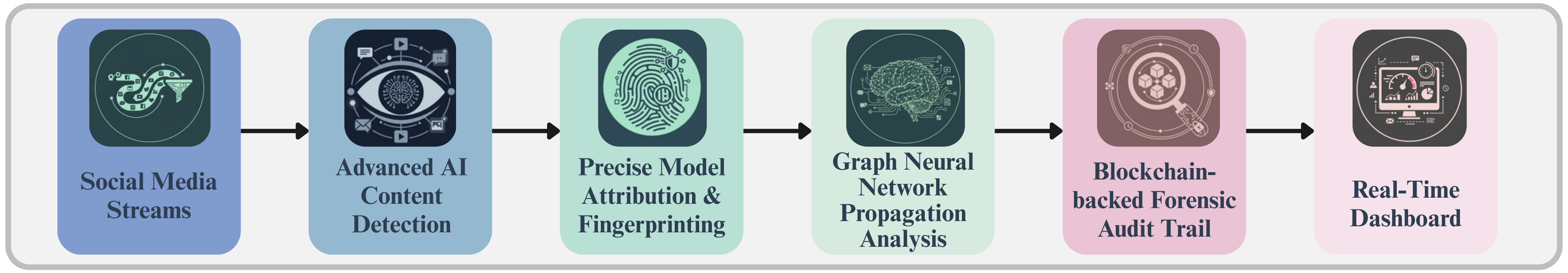
4-Layer Defense System

Multi-Modal Safety Scanner

- Joint detection of unsafe content + propaganda techniques.
- By using adversarial fine-tuning with hard negative mining.

Graph Neural Network Analysis

- Real-time propagation mapping.
- This shows how disinformation spreads across networks and coordinated bot campaigns.



Social Media Streams → Advanced AI Content Detection → Precise Model Attribution & Fingerprinting → Graph Neural Network Propagation Analysis → Blockchain-backed Forensic Audit Trail → Real-Time Dashboard

Key Technical Differentiators

Explainable AI (XAI)
Highlights exact words/sentences triggering detection for transparent decision-making

Federated Security
Privacy-preserving intelligence sharing with allied nations via blockchain audit trails

Live Graph Visualization
Interactive network analysis showing bot coordination and influence operations

Red-Team Hardened
Continuous adversarial testing ensuring robustness against evolving attack patterns

Continuous Adaptive Learning
Adversarial retraining and red-teaming ensure the model evolves against new AI jailbreaks and propaganda tactics.

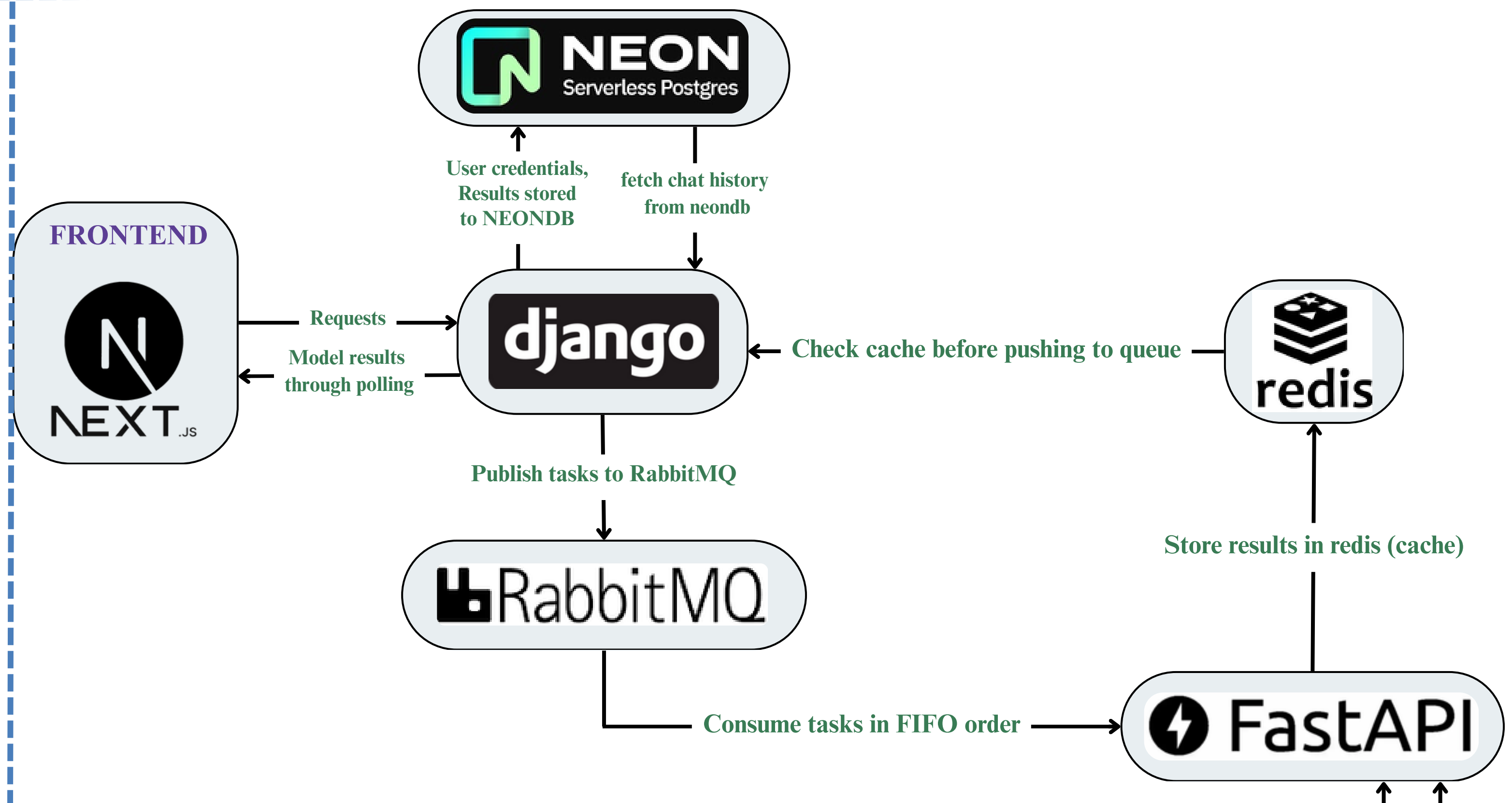
slayers

TECHNICAL APPROACH

SMART INDIA HACKATHON 2025

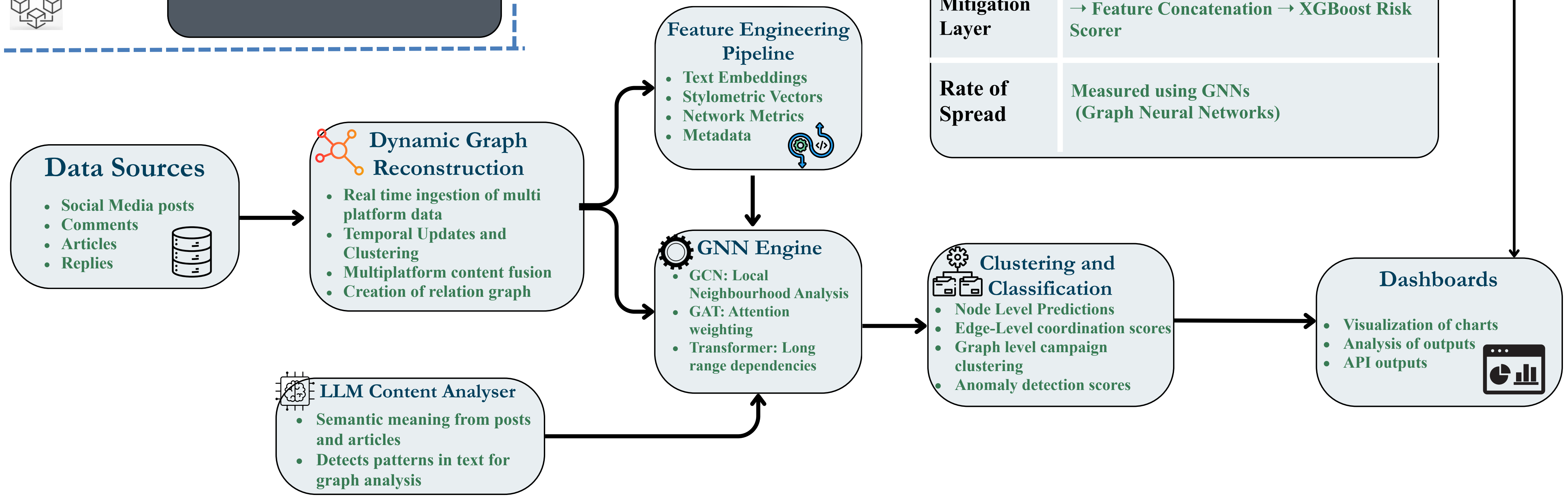
Tech Stack

- *Frontend*: NextJS
- *Backend*:Django Ninja, FastAPI/Flask (microservices)
- *TaskBroker*:RabbitMQ
- *DevOps*:Docker, GitHub Actions, GCP Cloud Run
- *DataBase*: NeonDB
- *Monitoring*: Prometheus+ Grafana
- *ML*:PyTorch, Hugging Face,scikit-learn, spaCy
- *GNN*:PyTorch Geometric / DGL
- *Blockchain*: Hashlib


NEON Serverless Postgres
User credentials, Results stored to NEONDB
fetch chat history from neondb
FRONTEND
NEXT.JS
Requests
Model results through polling
django
Check cache before pushing to queue
redis
Publish tasks to RabbitMQ
Store results in redis (cache)
RabbitMQ
Consume tasks in FIFO order
FastAPI

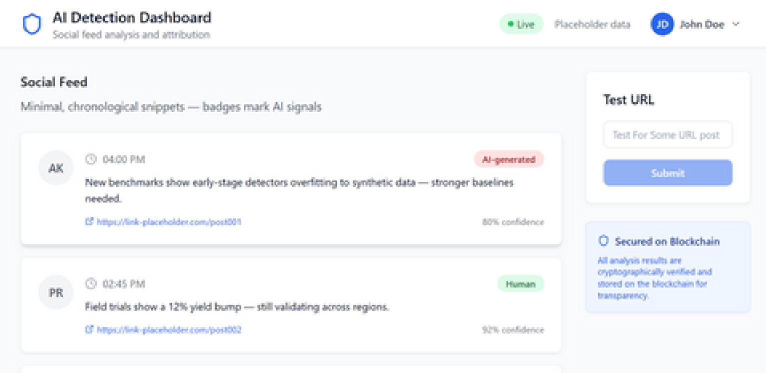
MODELS

AI VS REAL Text Detection	Raw Text→ Preprocessing → Transformer Embeddings(BERT)→ Feature Concatenation → Random Forest Classifier
Which AI Model (Attribution)	Raw Text→ Preprocessing →Embeddings → Feature Extraction → Multi-Class Random Forest Classifier
Risk Mitigation Layer	Raw Text→ Preprocessing →Embeddings → Feature Concatenation → XGBoost Risk Scorer
Rate of Spread	Measured using GNNs (Graph Neural Networks)


Data Sources
Social Media posts
Comments
Articles
Replies
Dynamic Graph Reconstruction
Real time ingestion of multi platform data
Temporal Updates and Clustering
Multiplatform content fusion
Creation of relation graph
Feature Engineering Pipeline
Text Embeddings
Stylometric Vectors
Network Metrics
Metadata
GNN Engine
GCN: Local Neighbourhood Analysis
GAT: Attention weighting
Transformer: Long range dependencies
Clustering and Classification
Node Level Predictions
Edge-Level coordination scores
Graph level campaign clustering
Anomaly detection scores
Dashboards
Visualization of charts
Analysis of outputs
API outputs
LLM Content Analyser
Semantic meaning from posts and articles
Detects patterns in text for graph analysis

Dashboards

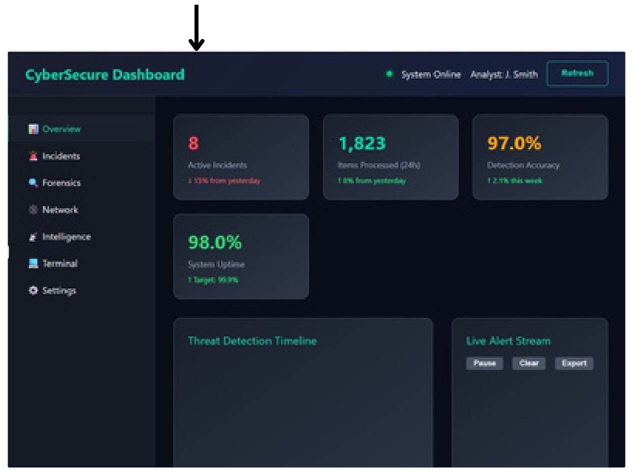
Dashboards
Consumer
Government



New benchmarks show early-stage detectors overfitting to synthetic data — stronger baselines needed.

Field trials show a 12% yield bump — still validating across regions.

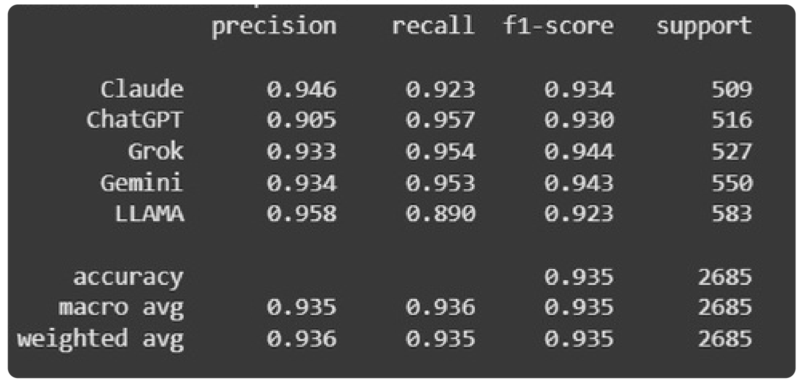
All analysis results are cryptographically verified and stored on the blockchain for transparency.



- Government Dashboard: Monitor flagged content, view blockchain-verified logs, and analyze moderation insights.
- Consumer Dashboard: User-friendly interface showing personal content status, appeal options, and transparency reports.
- Both dashboards provide clear visuals, analytics, and secure access, ensuring smooth interaction between authorities and users.

Model Attribution

- Tracks which model made each prediction or decision
- Logs model version, confidence score, and output details
- Helps identify performance differences between models
- Ensures accountability for every AI-generated result
- Supports explainability and continuous model improvement

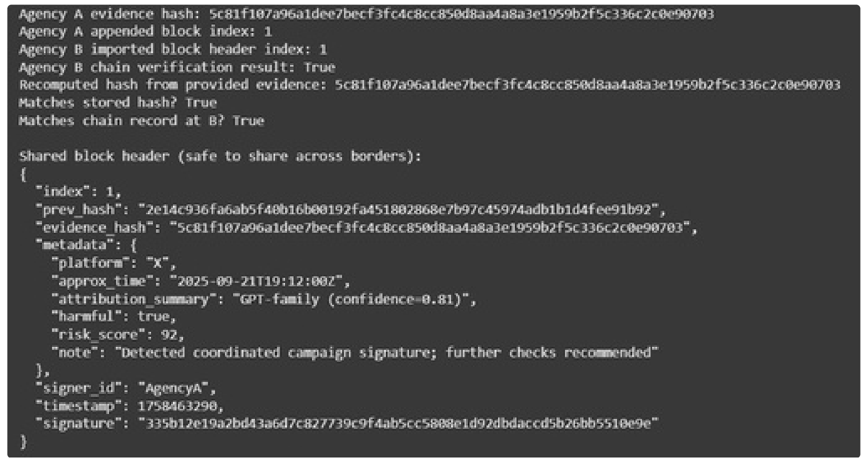


```
              precision    recall  f1-score   support

      Claude      0.946     0.923     0.934       509
     ChatGPT      0.905     0.957     0.930       516
        Grok      0.933     0.954     0.944       527
      Gemini      0.934     0.953     0.943       550
       LLAMA      0.958     0.890     0.923       583

    accuracy                          0.935      2685
   macro avg      0.935     0.936     0.935      2685
weighted avg      0.936     0.935     0.935      2685
```

Blockchain Audit Trail



```
Agency A evidence hash: 5c81f107a96a1dee7becf3fc4c8cc850d8aa4a8a3e1959b2f5c336c2c0e90703
Agency A appended block index: 1
Agency B imported block header index: 1
Agency B chain verification result: True
Recomputed hash from provided evidence: 5c81f107a96a1dee7becf3fc4c8cc850d8aa4a8a3e1959b2f5c336c2c0e90703
Matches stored hash? True
Matches chain record at B? True

Shared block header (safe to share across borders):
{
  "index": 1,
  "prev_hash": "2e14c936fa6ab5f40b16b00192fa451802868e7b97c45974adb1b1d4fee91b92",
  "evidence_hash": "5c81f107a96a1dee7becf3fc4c8cc850d8aa4a8a3e1959b2f5c336c2c0e90703",
  "metadata": {
    "platform": "X",
    "approx_time": "2025-09-21T19:12:00Z",
    "attribution_summary": "GPT-family (confidence=0.81)",
    "harmful": true,
    "risk_score": 92,
    "note": "Detected coordinated campaign signature; further checks recommended"
  },
  "signer_id": "AgencyA",
  "timestamp": 1758463290,
  "signature": "335b12e19a2bd43a6d7c827739c9f4ab5cc5808e1d92dbdaccd5b26bb5510e9e"
}
```

- Securely records all model actions and moderation results
- Stores data (outputs, timestamps, user actions) on an immutable ledger
- Prevents tampering or alteration of past records
- Ensures transparency, trust, and accountability
- Acts as a verifiable proof of moderation history
- Supports ethical and compliant AI operations

Feasibility & Viability

TECHNICAL READINESS

- **1+ months of development with working prototypes already built**
- **Real-world datasets**
- **Robust production stack:** NextJS, Django, GCP, Docker

INFRASTRUCTURE

- Cloud-native, containerized microservices for scale
- Ready integrations with Twitter/X, Reddit and Telegram APIs
- Blockchain-backed forensic trails for trust and traceability

PROOF OF CONCEPT

- **Live demo successfully tested and validated**
- **Achieved over 90% accuracy in early trials**

MARKET OPPORTUNITY

- Global SaaS moderation with $2B+ addressable market
- Twitter/X India: $50M+ annual content moderation budget
- $2B+ global market driven by election integrity needs (ASEAN, EU, UN) and cross-border intelligence sharing revenue.

SCALABILITY

- Built for multi-nation rollouts with secure cross-border data sharing
- Auto-scaling infrastructure along with continuous updates
- Edge-level integration for real-time speed and efficiency

SUSTAINABILITY

- Strategic partnerships with startups & research labs
- Ongoing R&D pipeline fueling long-term innovation

RISK MITIGATION

Challenge	Solution
Evolving Threats	→ Red-teaming
Scale Loads	→ Auto-scaling + edge deployment
Privacy	→ Only public data analyzed, zero personal data collected
False Positives	→ Explainable AI
Cross-Border Operations	→ Blockchain integration + Standard APIs

Slayers

MVP CLICKABLE LINKs:

Frontend : (demo)

1) Consumer dashboard
2) Government dashboard

Backend:

backend link

Target Audience

Public

- Checking AI generated content
- Attributing the content to specific models

Goverment Officials

- GNN powered analysis of post
- Dashboards representing threat data
- Blockchain based cross-border intelligence sharing

IMPACT

Social

- Safer Digital Ecosystem
- Stronger Public Trust and Transparency
- Protects Vulnerable Communities

Economic

- Preventing economic loss caused by Cyber Crimes
- Lowers down the Election & National security costs

Cultural

- Reduces Extremist influence sustaining the cultural inclusivity
- Promotes the use of Ethical AI

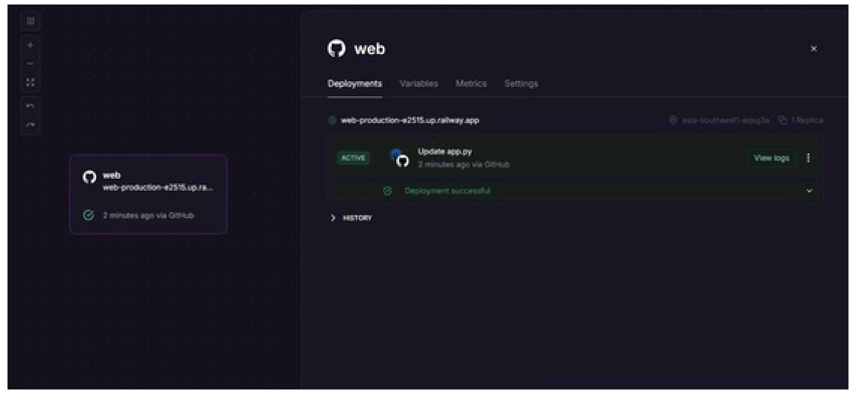
Quantifiable Impact			
Metric	**Before**	**After**	**Outcome**
Detection Time	2-4 hr	20 minutes	Atleast 50% Faster
Public Trust	Low	30 to 40% increment	More Credibility
Election Disinformation	High	20-30% reduction	Safer Elections
Phishing	Frequent	60% reduction	Stronger Protection
Fraud Loss	$1Billion +	$500 M - $750M saved	Enhanced Digital Security

BENEFITS

- Reduces election time misinformation incidents by 20-30 %
- Decrement in phishing incidents by 60%
- Faster cross border intelligence collaboration through Blockchain and AI

SMART INDIA HACKATHON 2025

AI vs HUMAN Model (Deployed Site)



web

Deployments Variables Metrics Settings

web-production-e2515.up.railway.app asia-southeast1-eqsg3a 1 Replica

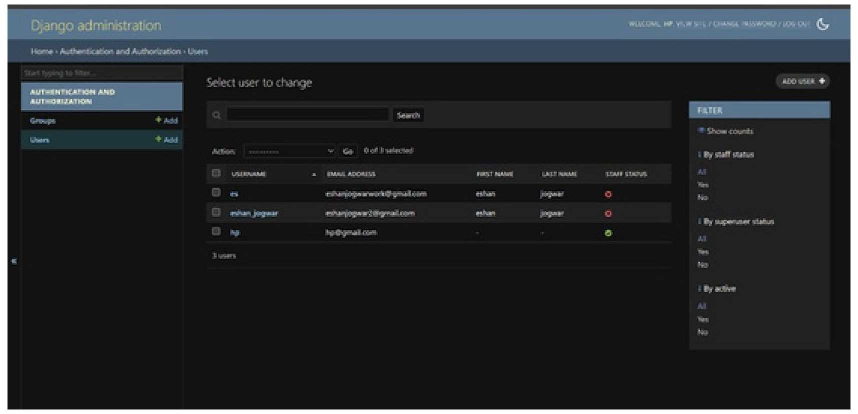
ACTIVE Update app.py 2 minutes ago via GitHub View logs

Deployment successful

HISTORY

web
web-production-e2515.up.ra...
2 minutes ago via GitHub

Backend Admin Panel



Django administration

WELCOME, HP. VIEW SITE / CHANGE PASSWORD / LOG OUT

Home › Authentication and Authorization › Users

AUTHENTICATION AND AUTHORIZATION
Groups + Add
Users + Add

Select user to change

ADD USER +

Search

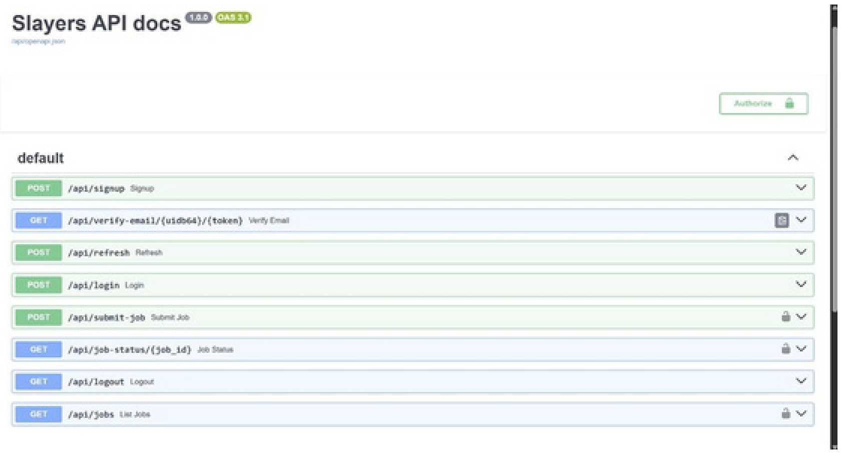
Action: --------- Go 0 of 3 selected

USERNAME	EMAIL ADDRESS	FIRST NAME	LAST NAME	STAFF STATUS
es	eshanjogwarwork@gmail.com	eshan	jogwar	✗
eshan_jogwar	eshanjogwar2@gmail.com	eshan	jogwar	✗
hp	hp@gmail.com	-	-	✓

3 users

FILTER
Show counts
By staff status: All, Yes, No
By superuser status: All, Yes, No
By active: All, Yes, No

APIs Documentation



Slayers API docs 1.0.0 OAS 3.1

/api/openapi.json

Authorize

default

```
POST /api/signup Signup
GET  /api/verify-email/{uidb64}/{token} Verify Email
POST /api/refresh Refresh
POST /api/login Login
POST /api/submit-job Submit Job
GET  /api/job-status/{job_id} Job Status
GET  /api/logout Logout
GET  /api/jobs List Jobs
```

slayers

SMART INDIA HACKATHON 2025

RESEARCH AND REFERENCES

- “CISA Cybersecurity Advisory 2024”: *LINK*
- “State of AI-Generated Media Detection” (Meta AI Research 2024): *LINK*
- Cao, Lele. "Watermarking for AI Content Detection: A Review on Text, Visual, and Audio Modalities." arXiv preprint arXiv:2504.03765 (2025). LINK
- Bhardwaj, Akashdeep, Salil Bharany, and SeongKi Kim. "Fake social media news and distorted campaign detection framework using sentiment analysis & machine learning." Heliyon 10.16 (2024). LINK

- “Detecting Machine-Generated Text” (ACL 2019): *LINK*
- “GLTR: Statistical Detection of Generated Text” (IEEE 2020): *LINK*
- “How Powerful Are Graph Neural Networks?” (ICLR 2019): *LINK*
- “GPT-who: AnInformation Density-based Machine-Generated Text Detector” : *LINK*
- Countering Disinformation Effectively: An Evidence-Based Policy Guide LINK

- MAGE Dataset (Multi-LLM AI Detection) : *AI vs. Human classifier*
- ArguGPT Propaganda Dataset: *Safety Scanner*
- MMFakeBench Multimodal Misinformation Benchmark : *Multimodal content*
- DetectRL Adversarial Robustness Framework: *Red-teaming*
- Kaggle Harmful Speech Dataset: *Text detection*

Solution Summary Table

Requirement No.	Technical Requirements as per Problem Statement	How the requirement is fulfilled	Impact	✗ ✓
1	Real-Time AI-Generated Detection	Transformer models + multimodal analysis	Early, accurate detection	✓
2	Attribution and Forensics	Watermarking + stylometric tracing	Tracks AI model sources	✓
3	Graph-Based Threat Intelligence	GNN mapping disinformation clusters	Visualizes coordinated misinformation	✓
4	Cross-Border Intelligence Sharing	Federated protocols + Blockchain Audit	Secure international data sharing	✓
5	Automated Risk Assessment	Dashboards with real-time risk scoring	Actionable insights for response	✓
6	Vendor Collaboration and Red-Teaming	Partnerships + adversarial testing	Enhances robustness and security	✓
7	Privacy and Compliance	Federated learning + explainable AI	Ensures privacy and trust	✓

Drive Link for More MVPs and Results: LINK

OR

LLMaGEN