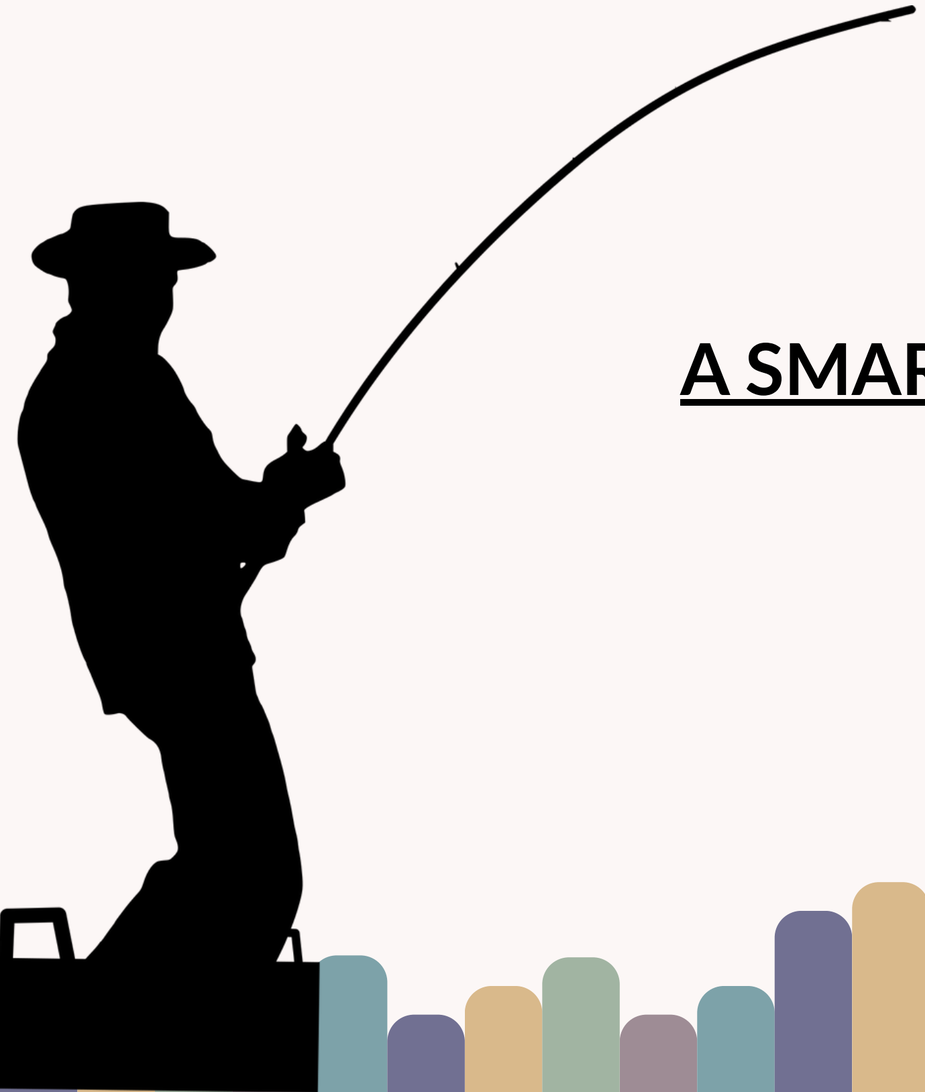


# PHISHNET

## A SMART NET THAT CATCHES PHISHING ACROSS CHANNELS (CASE - 6)


Akshat Srivastava - 1032222619

Pranjal Vishwakarma - 1032222756

Taha Kapadia - 1032222679

Tamanna Bansal - 1032222422

# IDEA TITLE

## PROBLEM STATEMENT

Fraudsters are increasingly exploiting advanced AI capabilities—such as **deepfake audio** to mimic a person's voice with high accuracy, **phishing emails** crafted to resemble legitimate communications, and coordinated **cross-channel attacks** that combine multiple mediums like calls, emails, and messaging apps—to convincingly impersonate individuals.Such impersonation attempts can compromise **sensitive information**, enable **unauthorized transactions**, and **erode public trust**, posing serious threats to financial security, personal privacy, and the overall resilience of digital ecosystems.

## NEED STATEMENT

There is an urgent need for a real-time, multi-modal fraud detection system to combat the rise of sophisticated impersonation attacks. Fraudsters now use deepfake audio, advanced phishing, and coordinated cross-channel tactics to bypass traditional defenses. Current security measures, often limited to single channels, fail to detect these complex threats in time. A next-generation solution must correlate voice biometrics, email metadata, linguistic cues, and behavioral patterns in real time to identify and stop fraud as it happens—**protecting privacy**, **preventing financial loss**, and **preserving trust in digital communications.**

## PROPOSED SOLUTION

- **Multi-Modal Detection**: Analyzes emails, calls, chats, and sites using NLP, voiceprint, OCR, and metadata, with watermark & context checks.
- **Explainable Decisions**: RL-tuned thresholds + LLM-generated regulator-ready explanations.
- **Proactive Prevention**: Alerts users, equips investigators, and deploys honeypot decoys.
- **Adaptive Hardening**: Red-team AI simulates attacks → Blue-team retrains → Continuous improvement.
- **Green & Compliant**: Carbon-aware compute, RL optimization, and immutable audit logs.

# PRIOR /EXISTING SOLUTIONS

## Audio / deepfake detection

Methods using spectral features (LFCC/MFCC), temporal-consistency checks, prosody & pause analysis, and ML / deep models (CNNs, ResNets, transformers) for classifying synthetic vs. real audio. Surveys and recent papers show steady progress but remaining generalization problems.

## Multi-modal / cross-modal systems

Recent research explores cross-modal feature fusion, correspondence learning, and unified models that jointly learn from audio, video, text and metadata to detect deception or manipulation. Benchmarks and new challenges are emerging for multimodal deception detection

## Email / phishing detection

Classic stacks combine header & metadata checks (SPF, DKIM, DMARC), feature-based ML classifiers on lexical/linguistic cues, URL and domain reputation, and graph-based link analysis to spot campaigns. Commercial and research tools layer NLP + metadata heuristics

## Adaptive & RL approaches in fraud

Proof-of-concept and academic work apply reinforcement learning for adaptive thresholding, dynamic feature weighting, and policy-based decisioning in fraud detection, showing promise in non-stationary settings.

# LIMITATIONS

CHANNEL SILOS

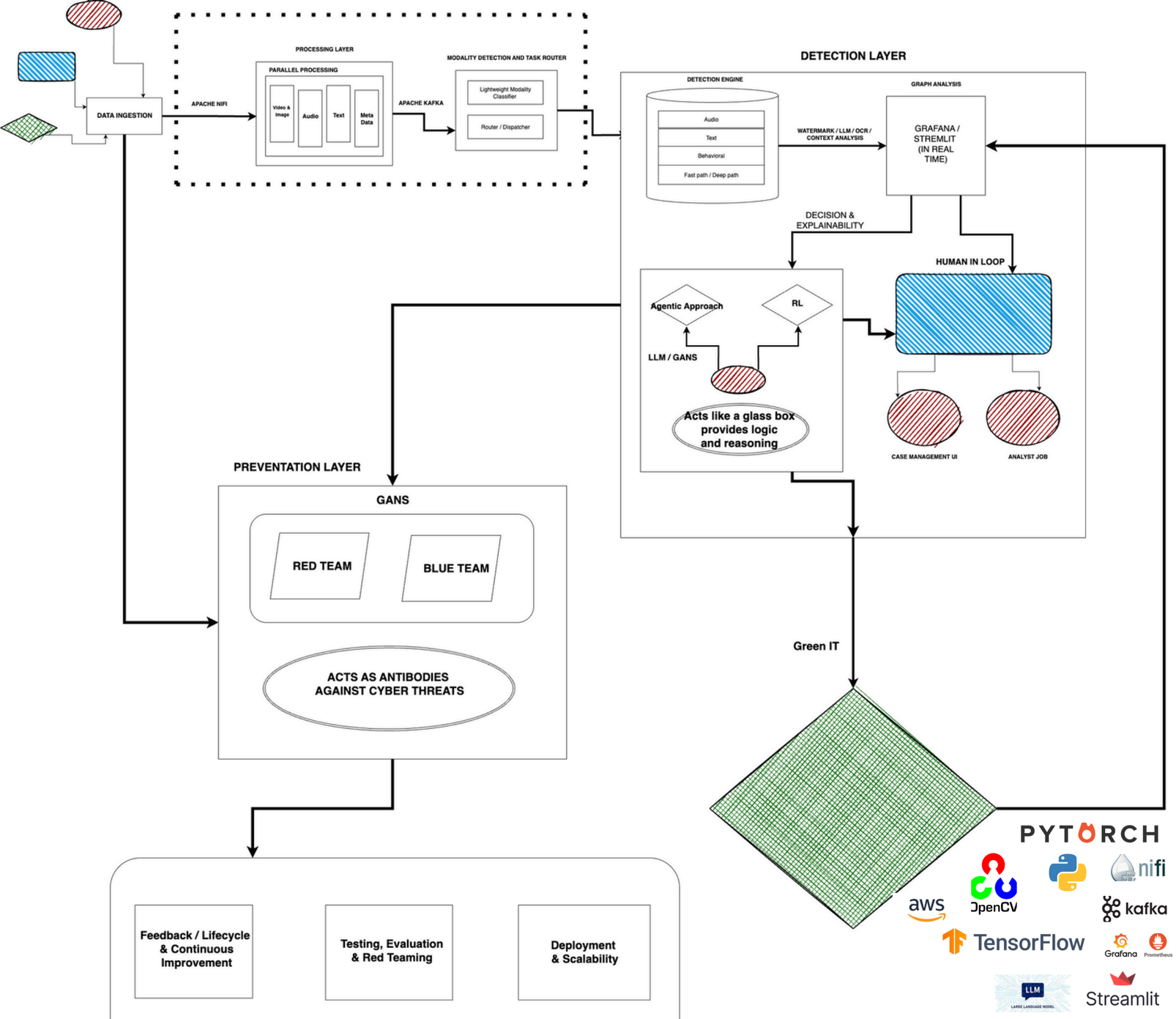
GENERALIZATION & ROBUSTNESS

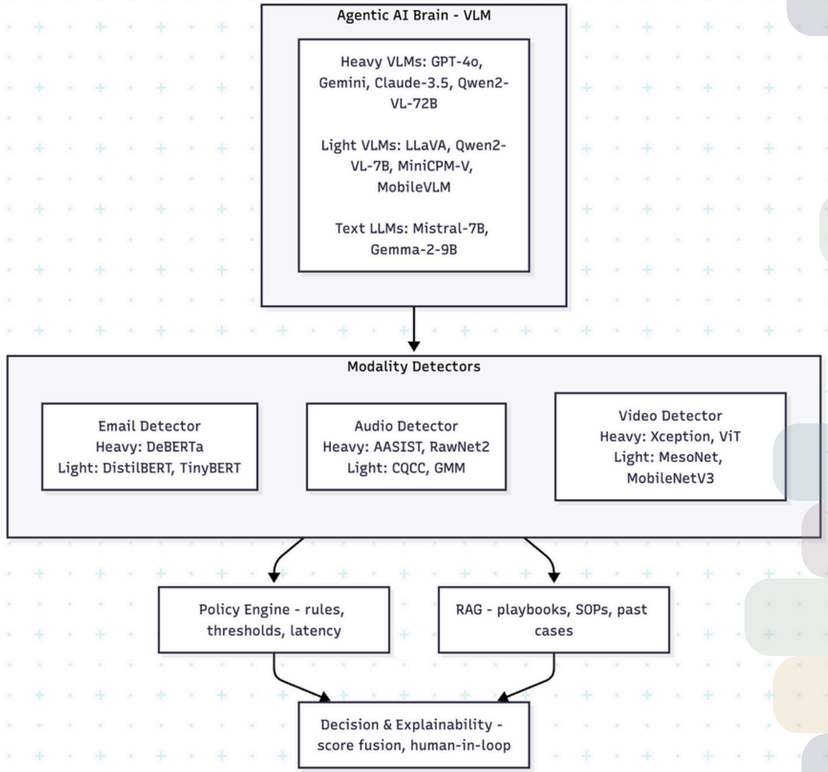
LATENCY VS. ACCURACY TRADEOFFS

LIMITED ADAPTIVE LEARNING

EXPLAINABILITY & COMPLIANCE

# TECHNICAL APPROACH




**Email phishing (text)**: A single LLM (e.g., Llama-3-Instruct, Mistral-7B) can classify reasonably well when fine-tuned, but a task model (DeBERTa/DistilBERT) is faster/cheaper at scale.

**Audio deepfake**: Needs acoustic artifacts (phase, spectro-temporal cues) that LLMs don't see. You need an audio anti-spoof model (AASIST/RawNet2). Transcribing to text and asking an LLM loses the spoof cues.

**Video deepfake**: Requires pixel/temporal artifacts (blending, lip-sync jitter) that a text LLM can't access. Use vision models (Xception, ViT/Swin, MesoNet).

# MVP WISE SCALING

## MVP-1

**Objective:**

- Build a working detection layer across text, audio, and video channels.
- Establish data → preprocessing → classification → explainability flow

.

**What we achieved:**

- **Datasets**: 3 raw datasets (text, audio, video), preprocessed & structured.
- **Classification Models:** Baseline ML/DL models trained per channel.
- **Agentic Reflection:** Channel-wise reasoning layer (explains predictions).
- **Labelling & Thresholds:** Applied custom thresholds for audio & video.
- **Developed UI:** Detection results integrated into UI dashboard.

## MVP-2

**Objective:**

- Move from offline detection to real-time multi-modal monitoring.
- Introduce proactive prevention mechanisms against phishing threats.

**Deliverables:**

- **Real-Time Integration:** Connect detection models with live data streams.
- **Visualization:** Deploy Grafana dashboards for real-time monitoring.
- **Prevention Layer:** GANs (Red Team vs Blue Team) simulating threats.
- **Mitigation Actions:** Quarantine, warning banners, escalation to analysts.

## MVP-3

**Objective:**

- Enable adaptive, explainable decision-making for prevention.
- Leverage AI agents and reinforcement learning for continuous improvement.

**Deliverables:**

- **Agentic AI Integration:** Add reasoning agents into prevention workflows.
- **Reinforcement Learning:** Adaptive policy learning from outcomes & feedback.
- **Human-in-Loop:** Analyst feedback strengthens prevention accuracy.
- **Explainability:** Glass-box reasoning for transparent and auditable decisions.
- **Outcome:** A self-learning, proactive cyber defense system that continuously evolves against new phishing threats while staying explainable and trustworthy.

# INDUSTRY-STANDARD SDLC

PRODUCTION

SAFETY

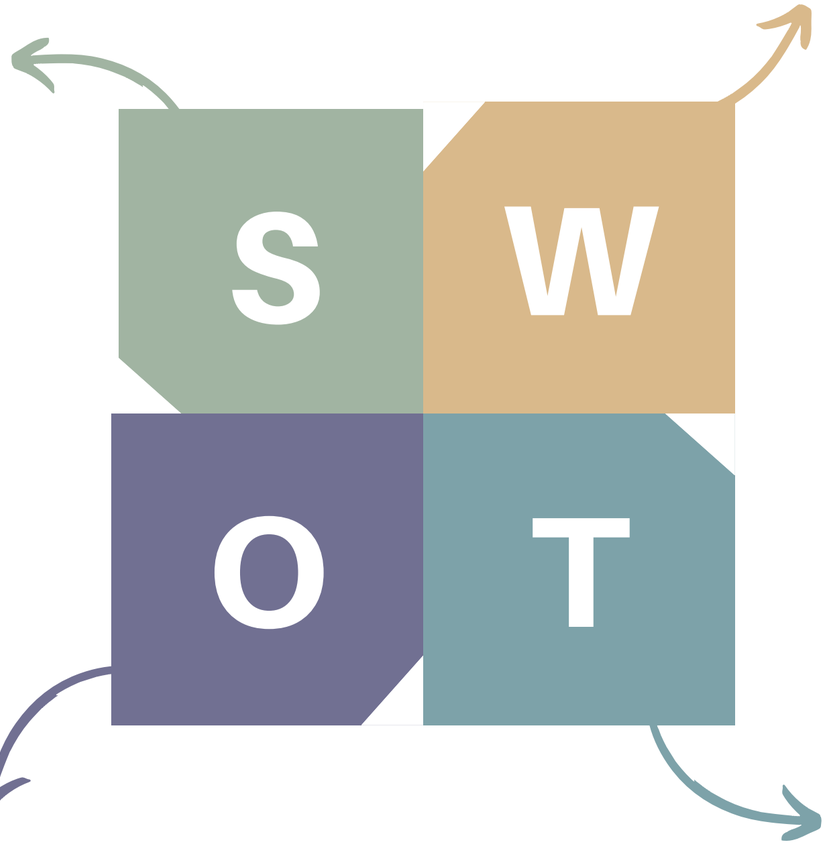
DEVELOPEMENT

DEPLOYMENT

# SWOT ANALYSIS

## STRENGTHS

- Cross-channel detection (email, call, chat, web).
- Multi-layered AI (detectors + RL + watermarking).
- Explainable AI for analyst trust.
- Privacy-first, no raw data storage.
- Green IT compliance (sustainable).




## WEAKNESSES

- Complex, harder to implement/maintain.
- Resource-intensive (infra + compute).
- Dependent on high-quality data.
- Integration challenges with legacy systems.

## OPPORTUNITIES

- Scale into banking, insurance, etc.
- Rising demand with deepfake & phishing growth.
- Add federated fraud graphs for cross-org defense.
- Expand into e-commerce, healthcare, telecom.

## THREATS

- Rapidly evolving GenAI threats.
- Risk of adversarial model attacks.
- Regulatory uncertainty.
- Organizational resistance to integration.

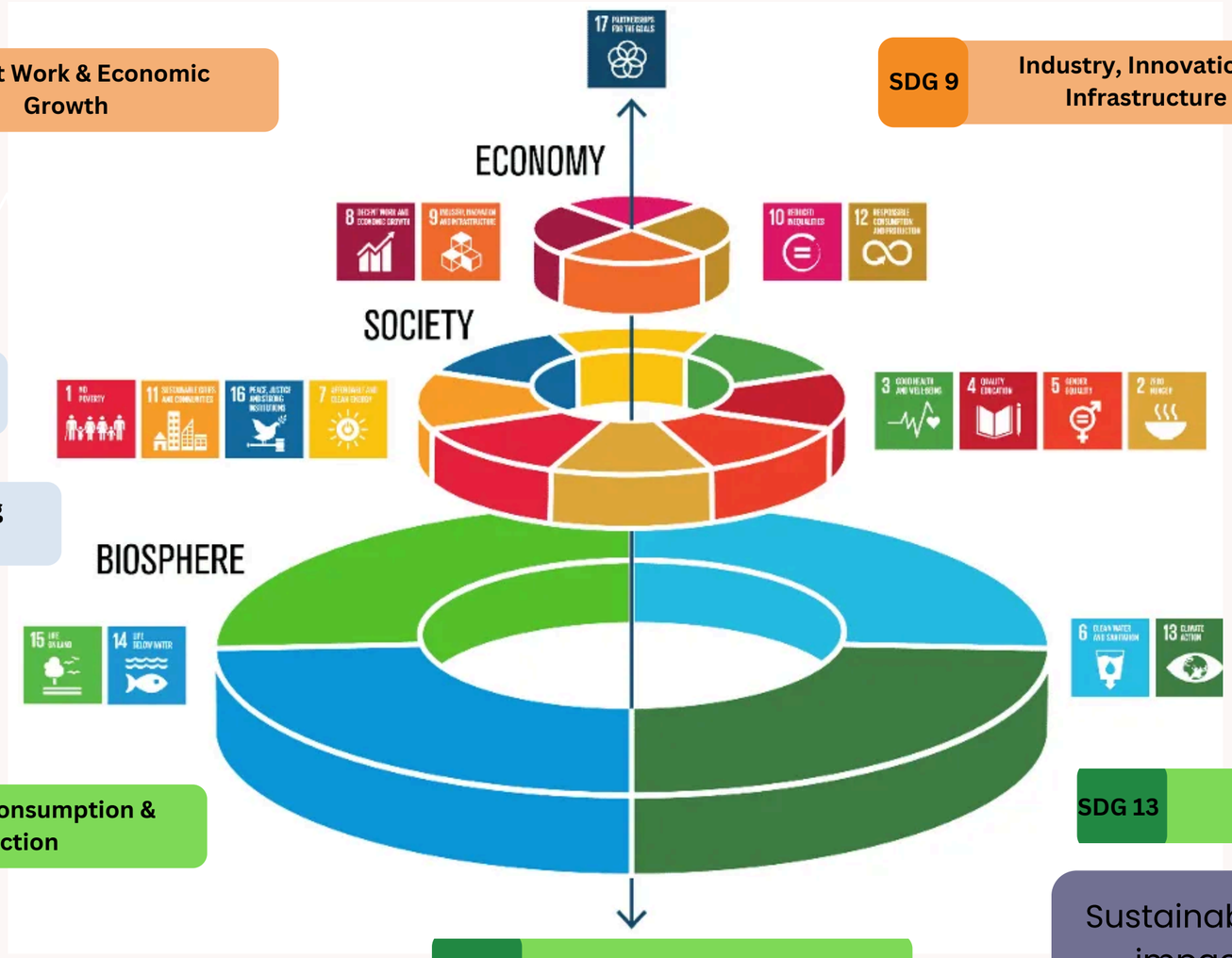
# IMPACTS AND BENEFITS

Operational Efficiency

**SDG 8** Decent Work & Economic Growth

**SDG 9** Industry, Innovation & Infrastructure

Multi Layered Security




**SDG 4** Quality Education

**SDG 16** Peace, Justice & Strong Institutions

**SDG 10** Reduced Inequalities

Reducing False Positives

Regulatory Compliance

**SDG 12** Responsible Consumption & Production

**SDG 13** Climate Action

Sustainability impact

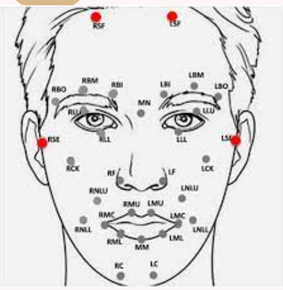
**SDG 15** Life on Land

# REFERENCES

## Detecting Audio-Visual Deepfakes with Fine-Grained Inconsistencies (2024)

Link

This paper proposes a detector that captures fine-grained audio-visual inconsistencies using a Conv1D audio extractor and a shallow ResNetConv3D visual extractor, enhanced by cross-attention to focus on critical facial regions. A temporally-local pseudo-fake augmentation introduces subtle mismatches for better generalization. The model achieves 97.7% AUC (DFDC) and 84.5% AUC (FakeAVCeleb), outperforming prior methods.
**Strengths**: fine-grained spatial/temporal modeling, cross-attention, pseudo-fake augmentation, strong generalization.
**Limitations**: limited datasets, shallow design, L2 reliance, slight in-dataset drop, no real-world testing.




## Provable Robust Watermarking for AI-Generated Text (2023)

Link

This paper proposes Unigram Watermarking, a method that biases token sampling in LLMs toward a secret "green list" of words to embed an invisible signal in generated text. It provides theoretical proofs of robustness against paraphrasing, editing, and adversarial attacks while maintaining fluency. Experiments across different models show that the watermark achieves high detection accuracy without degrading text quality.
**Strengths**: theoretically grounded; robust against text perturbations; low false positives; efficient detection; does not affect readability.
**Limitations:** only for LLM outputs, requires generation-time setup, vulnerable to advanced rephrasing.

## Adaptive Memory Networks With Self-Supervised Learning for Unsupervised Anomaly Detection (2023)

Link

This paper presents AMSL, a model that learns normal patterns and memory features to spot unusual behavior in data without needing examples of fraud. Tested on large time-series datasets (like the CAP dataset with 900M samples), it performed over 4% better than other leading methods and stayed reliable even with noisy input.
**Strengths**: strong accuracy, learns diverse patterns, handles large data, noise-tolerant.
**Limitations**: tested on limited domains, may struggle with very subtle anomalies, still centered on one type of approach.

## Watermarking for AI Content Detection: A Review on Text, Visual, and Audio Modalities (2025)

Link

This survey reviews watermarking approaches across text, image, and audio modalities, categorizing them into statistical, signal-based, and deep-learning-based methods. It analyzes robustness against transformations (compression, paraphrasing, cropping), stealthiness, and practicality for real-world deployment. The paper highlights challenges like adversarial removal, lack of cross-modal standards, and issues of privacy and governance.
**Strengths**: comprehensive taxonomy; cross-modal coverage; highlights practical deployment challenges; identifies gaps like lack of interoperability.
**Limitations**: survey-only (no new model), limited cross-dataset benchmarking, future multimodal risks not experimentally addressed.

# THANK YOU

## GROUP NAME - OVERFITTERS