Spot the Fake: AI-Powered Detection of Fraudulent Websites, Apps & Digital Content

AI-Powered Detection of Fraudulent Websites & Digital Content

Cipher Cop 2025

Team Lead: Harsh Jain
Team Members: Nikhil Singh, Rishiraj Gupta, Sumit Kothari

Proactive Protection Through Multi-Modal AI Analysis

The Digital Minefield

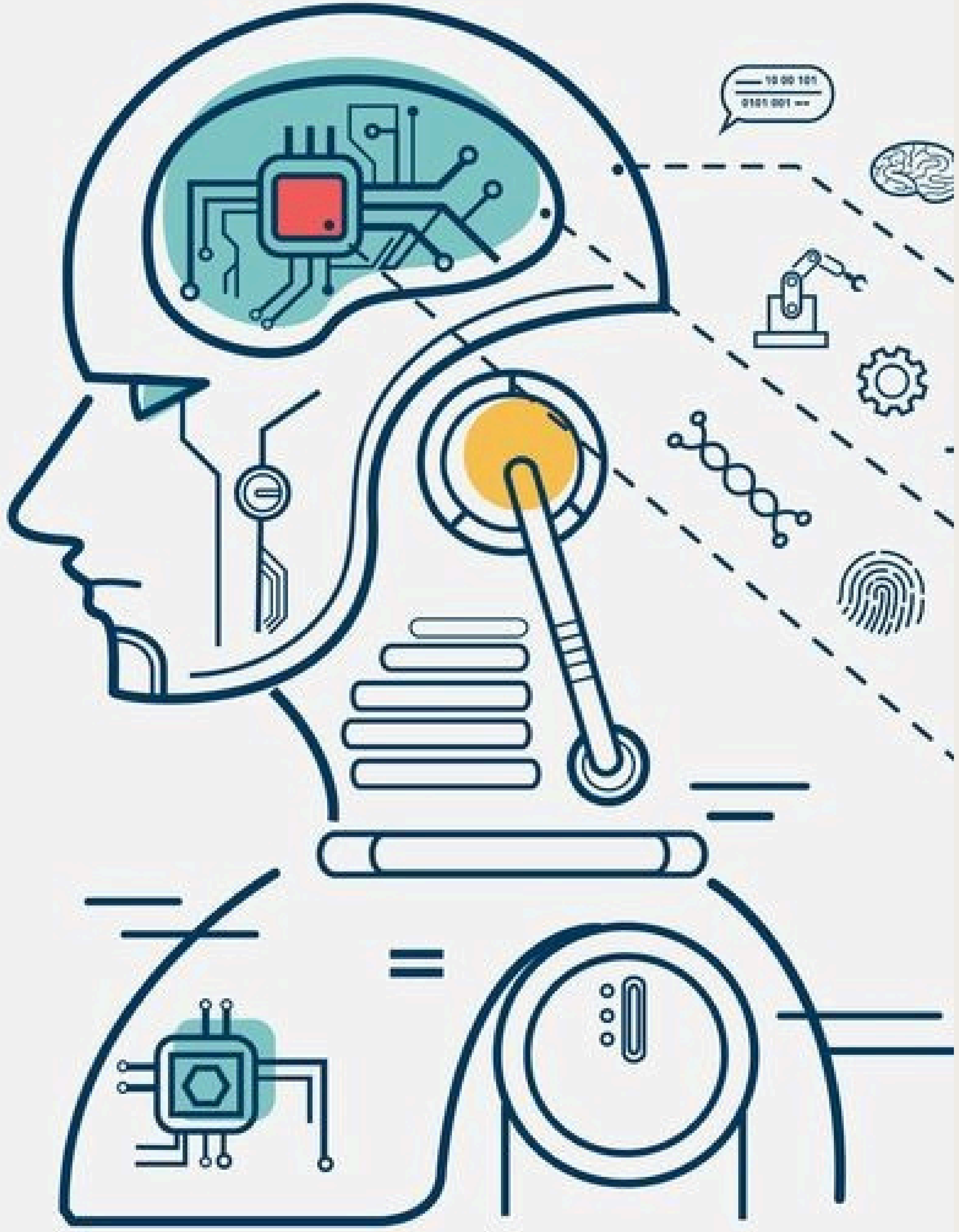
The Digital Epidemic:

- Explosion of sophisticated fake websites and apps cloning trusted brands.
- Used for phishing, data theft, financial scams, and malware distribution.

Why Current Solutions Fall Short:

- **Reactive, not Proactive:** Reliance on user reports and manual takedowns.
- **Evolving Threats:** Scammers constantly adapt, making simple rule-based systems ineffective.

Our Mission: To build a **proactive, intelligent system** that identifies fraud *before* users become victims.



Our Solution - A Multi-Layered AI Shield

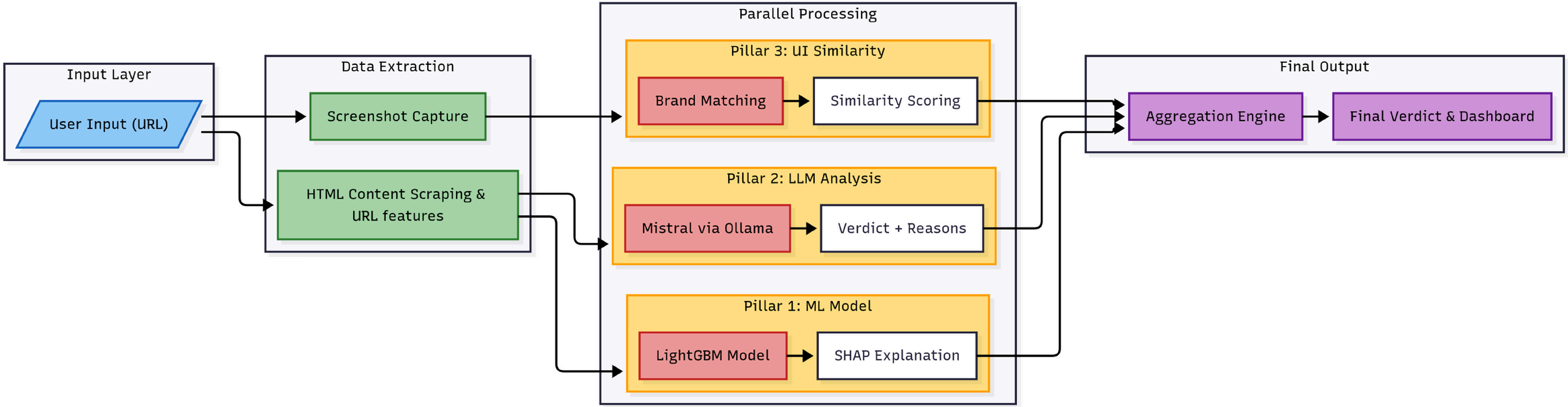
We developed a holistic defense system that analyzes multiple dimensions of digital content.

Three Pillars of Detection:

1. **Machine Learning (LGBM):** Analyzes URL structure and features for classic phishing hallmarks.
1. **Large Language Model (Mistral):** Understands the *content* and context of the webpage for semantic analysis.
1. **UI Similarity Engine:** Uses Computer Vision to detect visual mimicry of known brands.

Key Innovation: Combining these three approaches into a single, weighted scoring system for superior accuracy and explainability.

Slide 4: Technical Architecture Overview



Tech Stack: Python, Streamlit, LightGBM, Ollama (Mistral), Selenium, OpenCV, pytesseract, SHAP, BeautifulSoup.

Deep Dive 1 - The ML Engine (LightGBM)

What it does:

Analyzes over 30 heuristic features extracted directly from the URL.

Features Include:

- Length of URL/hostname, count of special characters (@, -, ~, %)
- Presence of IP addresses, punycode, or known shortening services.
- "Phishy" keywords (e.g., **login**, **verify**, **account**).
- Suspicious TLDs.

Output:

A probability score and a **SHAP explanation** showing *which features* most influenced the decision (e.g., "**nb_dots=10** was a strong phishing indicator").

Strength:

Extremely fast and great at catching obvious phishing patterns.

Deep Dive 2 - The LLM Analyst (Mistral)

What it does:

Acts as a cybersecurity expert reading the page's content.

Process:

1. Extracts and cleans main text from the HTML.
1. Sends the content to a locally-run Mistral model via Ollama.
1. Forces a structured JSON response with a verdict, risk level, and, crucially, a list of **evidence snippets**.

Sample LLM Output:

{"verdict": "phishing", "risk_level": "suspicious", "evidence_snippets": \["'Your account will be suspended' urgency trigger", "Mismatch between domain 'secure-paypal-update.com' and branded content"\]}

Strength:

Understands nuance, social engineering tactics, and contextual clues that URL analysis misses.

Deep Dive 3 - The UI Similarity Detective

What it does:

Answers the question "Does this *look* like a real PayPal (or other brand) site?"

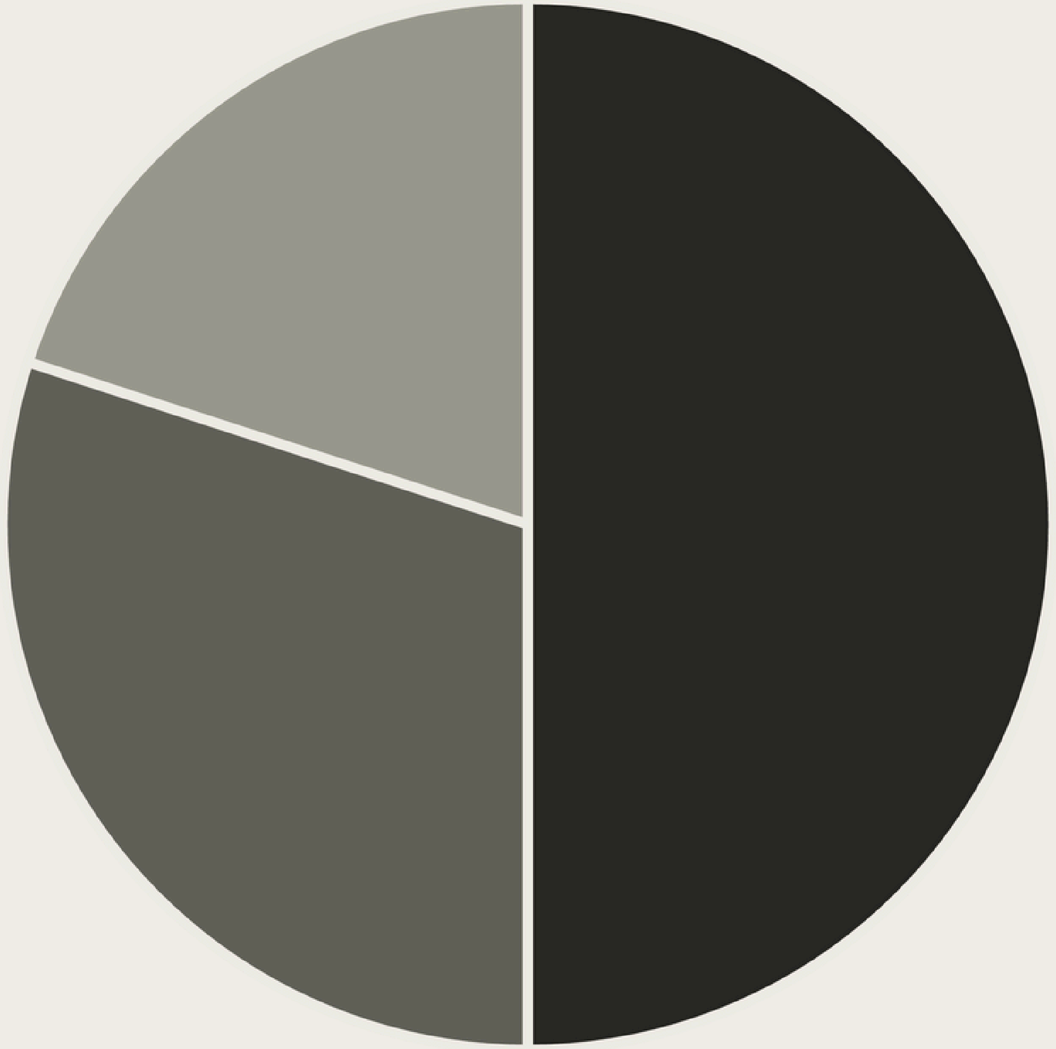
Process:

1. Takes a screenshot of the target site.
1. Extracts the domain and performs fuzzy matching against a library of brand reference images (**brand_ref.png**).
1. Computes a composite similarity score based on: ***Perceptual Hash (pHash):*** *Layout and structural similarity*. **Color Histogram:** Color scheme and palette matching. * **OCR + Text Similarity:** Text content and its stylistic presentation.

Strength:

Catchs sophisticated visual clones that might bypass other checks.

The Ensemble Decision - Smarter Together



Our Weighted Scoring Algorithm:

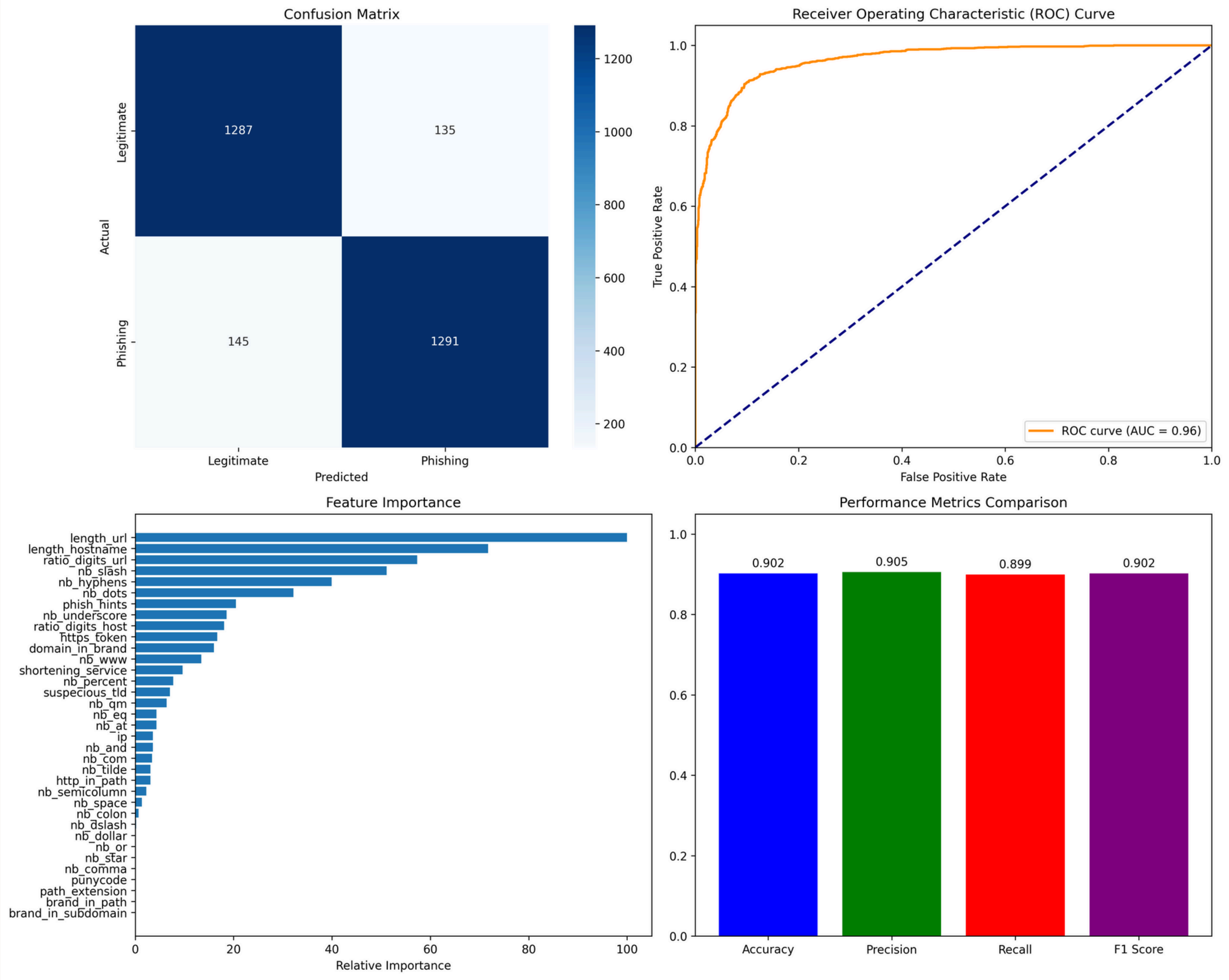
- **LGBM Score (50% Weight):** The foundational, quantitative check.
- **LLM Score (30% Weight):** The qualitative, contextual check.
- **UI Similarity Score (20% Weight):** The visual authenticity check.

The system aggregates these into a **Final Legitimacy Score**.

Why Ensemble?

No single method is perfect. Combining them reduces false positives/negatives.

LightGBM Performance Metrics



Results & Impact

Successfully Created:

A proactive, multi-modal detection system that meets the hackathon's objectives.

Key Achievements:

- **Explainable AI:** Every decision is supported by clear evidence from each module.
- **Holistic Approach:** Combines structural, semantic, and visual analysis for comprehensive coverage.
- **Prototype Ready:** A user-friendly web app that could be extended into a browser extension.

Potential Impact:

Protects users from financial and data loss, reduces the effectiveness of phishing campaigns, and increases trust in digital platforms.