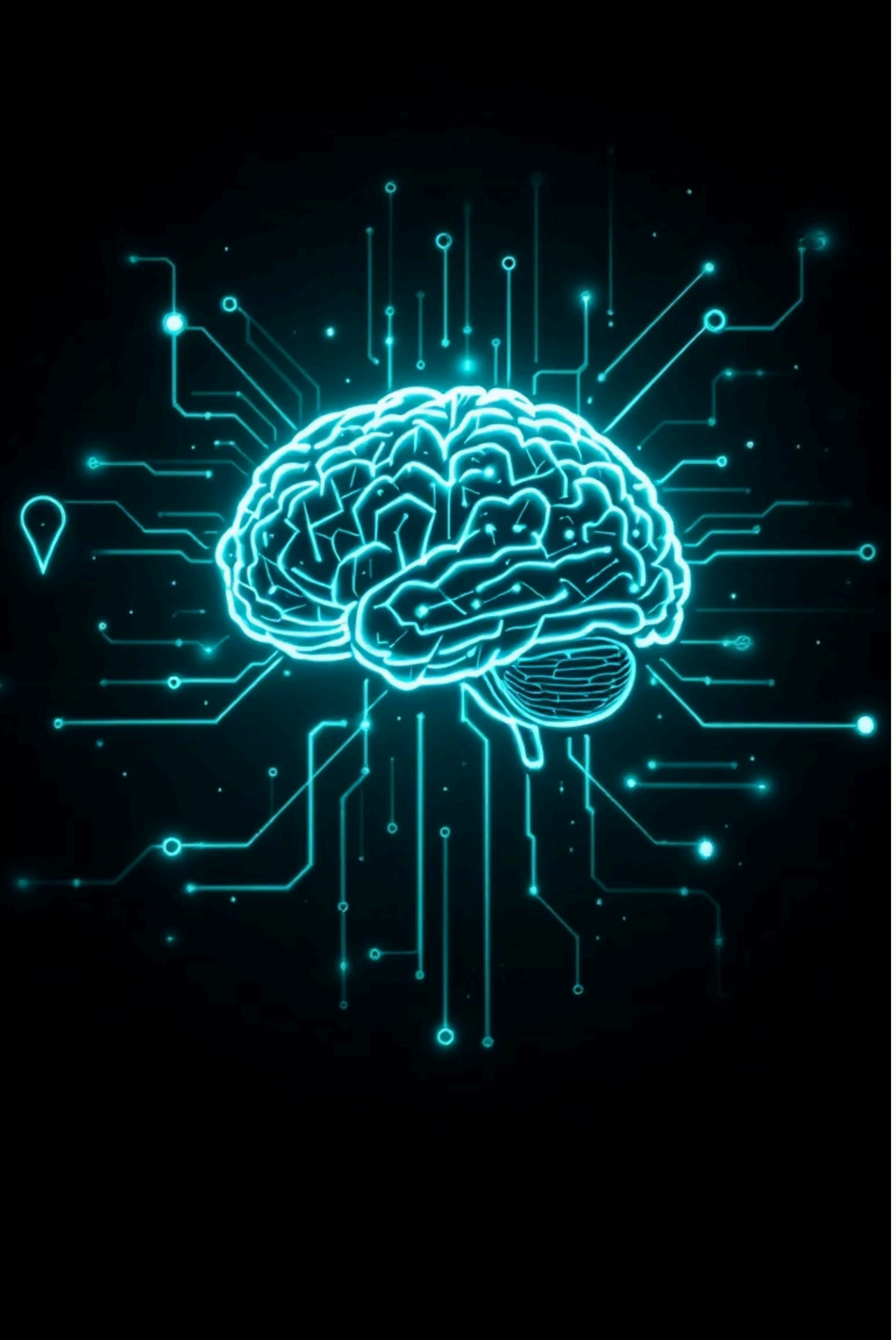


AI in Social Engineering and Phishing Campaigns

This presentation explores the application of Artificial Intelligence and large language models in crafting sophisticated phishing emails. We demonstrate how AI can simulate social engineering attacks, test their believability, and evaluate detection mechanisms. Our goal is to enlighten cybersecurity professionals on emerging threats using AI-driven techniques and to showcase defenses against them.

Collecting and Preparing Phishing Datasets

PhishTank Dataset

A community-driven repository of verified phishing URLs continuously updated and curated for accuracy.

Nazario Phishing Corpus

A comprehensive collection of email samples and associated metadata gathered from known phishing campaigns.

CIC Phishing URLs Dataset

Includes labeled URLs focusing on recent phishing trends, ideal for training detection algorithms.

These datasets enable the acquisition of real-world phishing data essential for realistic email generation and classification.

Generating Phishing Emails Using Large Language Models

1

Model Choice

Leverage GPT-based models like OpenAI GPT or open-source LLMs such as llama-cpp and GPT-J.

2

Contextual Prompts

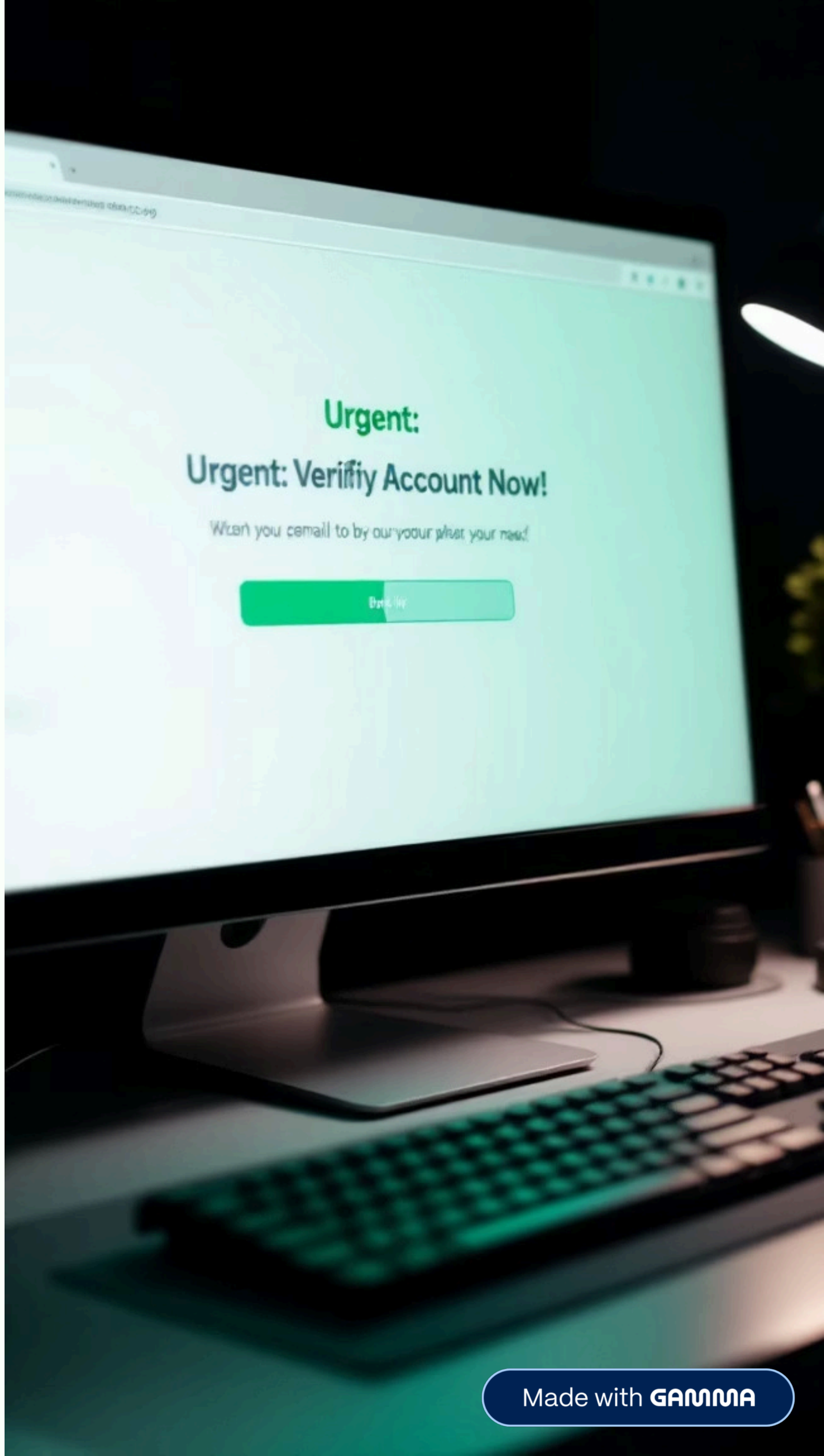
Use targeted prompts for themes like banking, password resets, and job offers to simulate real attacker tactics.

3

Example Prompt

"Generate a phishing email pretending to be from PayPal requesting the user to verify their account due to suspicious activity."

This approach replicates socially engineered content at scale with credible, context-aware language tailored to deceive recipients.



Training Automated Phishing Email Classifiers

Text Feature Extraction

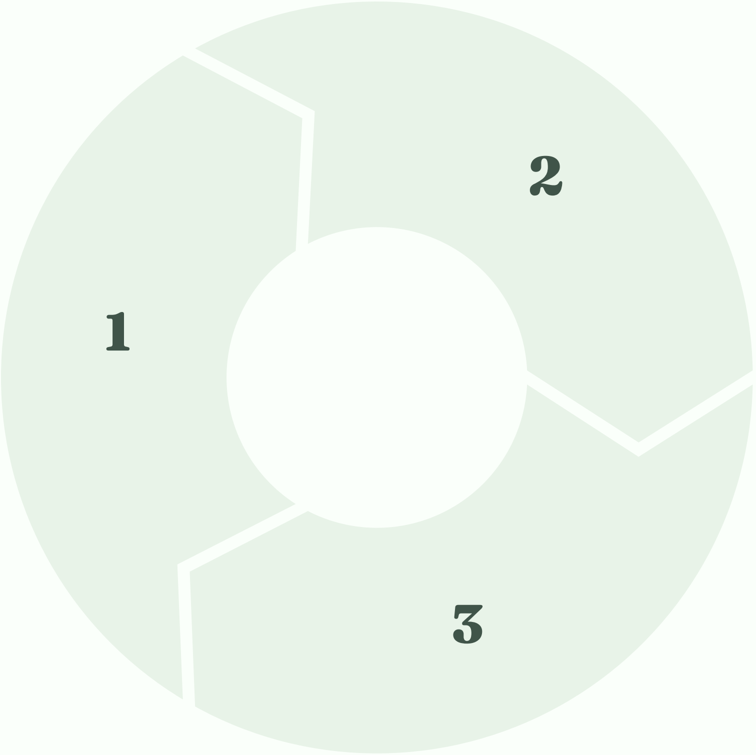
- TF-IDF to capture keyword importance
- BERT embeddings for semantic understanding

Classification Algorithms

- Logistic Regression: interpretable baseline
- Random Forest: robust ensemble method
- SVM: high-dimensional data separability

These techniques combine to create detection models evaluated rigorously on precision, recall, and F1 metrics to ensure reliable identification of phishing content.

Evaluating Human Believability of Phishing Emails



User Surveys

Ethically conduct controlled surveys to measure human detection and confusion rates.

LLM Believability Scoring

Deploy language models to rate email realism based on linguistic cues and context.

Combined Insights

Contrast model scores with human responses to identify vulnerability gaps in awareness.

These evaluation strategies help us understand phishing effectiveness and inform improved user training and system defenses.

Tools and Libraries for Development

Python

Primary programming language enabling versatile data handling and model integration.

Scikit-learn & NLTK

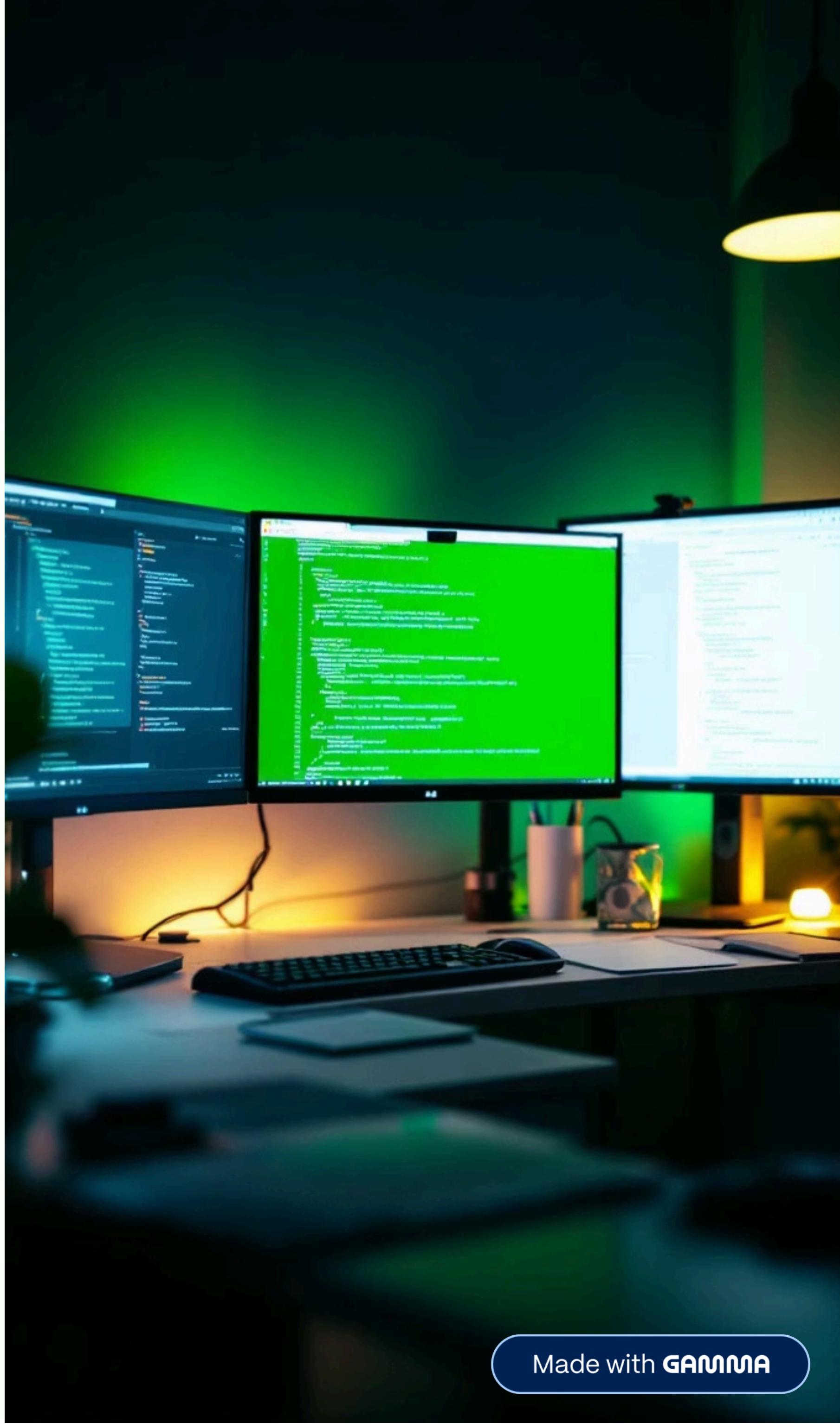
Libraries for ML models, feature processing, and natural language tasks.

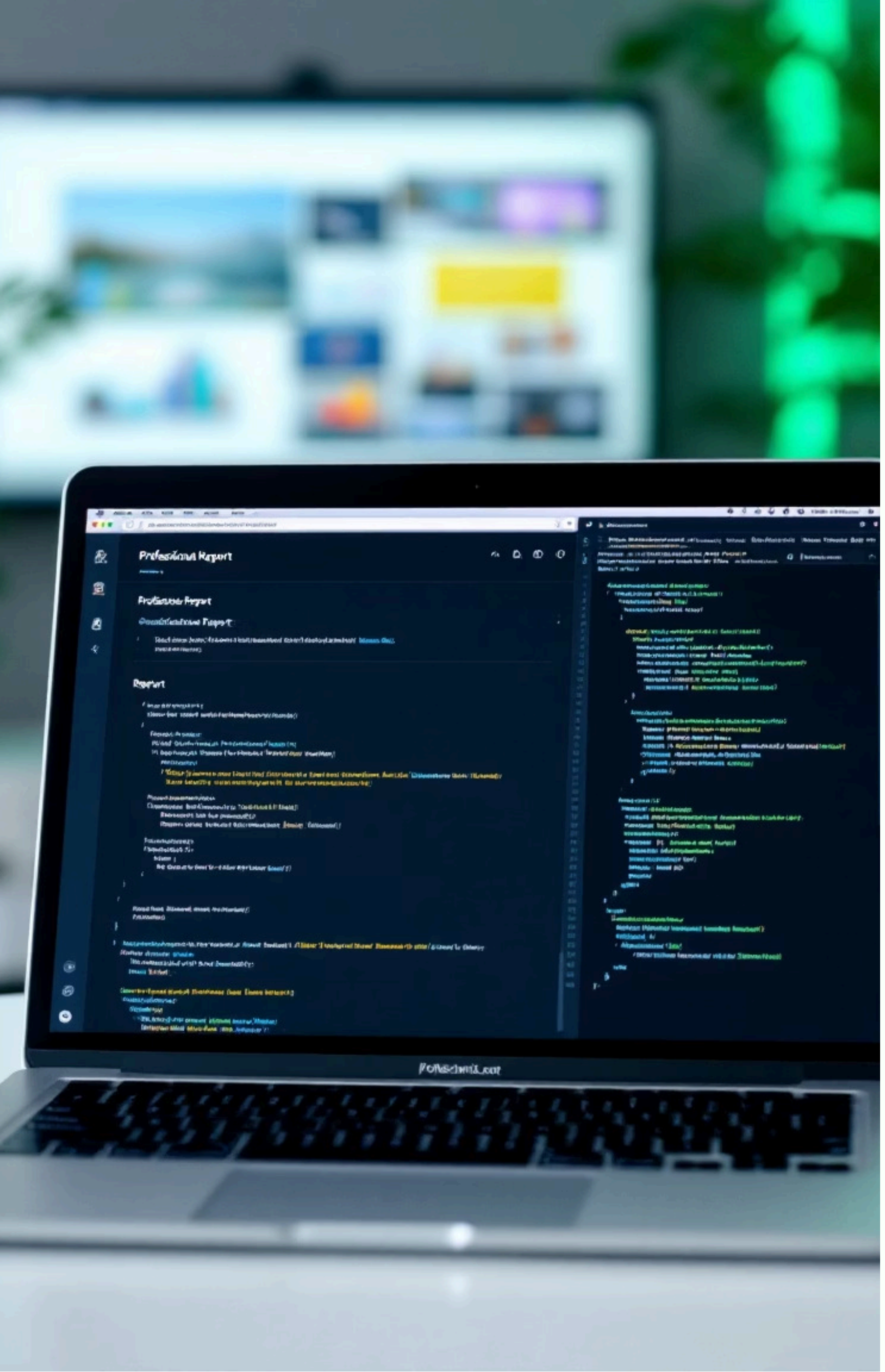
Transformers (Hugging Face)

State-of-the-art pretrained LLMs for text generation and embedding extraction.

Flask

Simple web framework to deploy demos or dashboards for interaction.





Deliverables and Outputs

Auto-generated Phishing Samples

Curated phishing emails showcasing AI-generated attack sophistication.

Trained Detection Model

Machine learning classifiers capable of distinguishing phishing and legitimate emails reliably.

Optional Dashboard/Demo

Interactive tool for testing inputs and visualizing model predictions in real time.

Comprehensive Report and Presentation

Detailed documentation summarizing methodology, results, and security implications.

Key Takeaways and Next Steps

1

AI Enables Realistic Phishing

Large language models amplify social engineering capabilities significantly.

2

Detection Requires Robust Models

Multi-technique classification improves phishing identification accuracy.

3

Human Awareness Remains Crucial

Education and evaluation bolster resistance to sophisticated email scams.

4

Future Work

Integrate adaptive AI detection and explore deeper cross-modal phishing signals.

This project highlights urgent cybersecurity challenges introduced by AI-generated phishing and outlines an analytical framework for defense and evaluation.