Project Proposal:

AI-Driven Phishing Detection and Mitigation Assistant

Executive Summary

Based on the growing threat of phishing attacks, I propose the development of a prototype AI-Driven Phishing Detection and Mitigation Assistant, to help user to prevent, and distinguish the phishing to protect their sensitive personal information, such as bank account, password, credit card numbers, or ID card numbers. This project aims to leverage NVIDIA's cutting-edge AI and cybersecurity frameworks to create a robust, intelligent, and adaptive system. By integrating NVIDIA NeMo Agent Toolkit, NeMo Microservices, and NVIDIA Morpheus, we will build a multi-agent system that not only detects sophisticated phishing attempts with high accuracy but also learns and evolves from new threats in real time.

Table of Contents

1.Introduction

1.1 Background

Phishing, a type of cyber-attack in Emails, Text messages, phone calls, social media messages, fake websites. With the advancements in AI technology today, if someone were to photograph your face or record your voice, your personal information could be used for illegal activities. However, precisely because AI technology is mature, we can also attempt to protect our sensitive data. Various regions are devising ways to deal with phishing such as Hong Kong, the government figured out a new way to identify the phishing in SMS. Only SMS messages sent by verified companies will contain the # symbol. This effectively and immediately distinguishes between genuine and fake messages. But, how about email?

It remains the leading attack vector for social engineering and credential theft. However, attackers' techniques are constantly evolving, and some sophisticated spoofed emails can still bypass checks. Although there are email phishing identification mechanisms and signs that will automatically detect it is phishing or not, such as domain verification, unnormal labels, and then put the emails into spam email.

Like I mentioned, every day there are new techniques involved, like new domain, new sentences, new ideas, or maybe some coincide such as you just bought a package, and then a phishing website sent you an email saying: Your package failed to be delivered (reason: insufficient shipping fee), then you will believe it and pay.

Machine learning is far behind the new ideas of attackers. How can we maximally avoid those phishing? “Unlike many other types of cyberattacks, phishing targets the human layer. Emails, text messages, phone calls, and even QR codes can serve as delivery mechanisms. These “lures” are crafted to look authentic, mimicking brands, colleagues, or executives, and they typically use urgency or fear to apply pressure.” (Dulce.2025)

We could solve the phishing from a psychological angle, we could let the system tell users up front that a message is fake or likely fake especially the URL, then show concrete evidence and reassurance, because the attackers are taking advantage of the victims' panic. By telling them why it's fake, this will calm them down from the beginning and allow them to analyze whether it's true or false.

2.Objectives

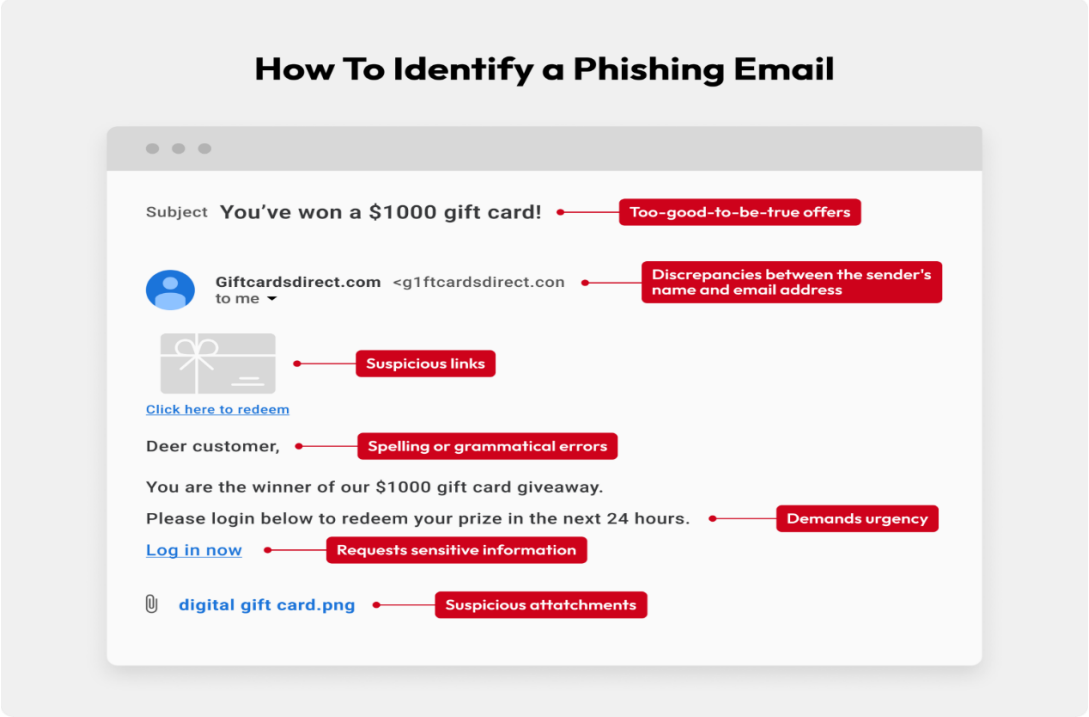
The core objective is to design and implement a prototype system that ingests incoming emails, analyzes them for phishing indicators using AI models, and takes automated, risk-based mitigation actions.

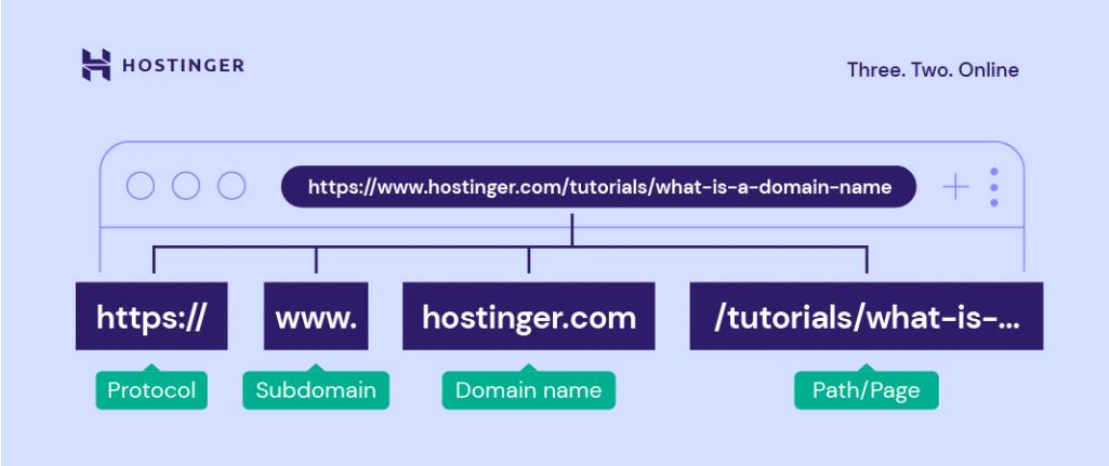
1. Psychology-Aware Feature Extraction
    - Parse email headers, body, attachments, and especially URLs into structured features capturing both technical and psychological cues.
2. Risk Scoring with Evidence
    - Use deep learning to compute a phishing risk score and extract concrete evidence (for example, fake-looking URLs, impersonation language, anomalous sender).
3. Agent-Based Decision and User Warning
    - Let specialized agents decide to Allow, Flag, or Quarantine, and for risky emails, clearly tell users the message is likely fake and show the key evidence.
4. Feedback-Driven Continuous Learning
    - Learn from new attacks, analyst review, and user reactions to improve both detection and user-facing messages over time.
5. End-to-End Prototype with NVIDIA
    - Deliver a working prototype showing the full flow—from detection to psychologically informed user warnings—powered by NVIDIA technologies.

3.Nvidia Components choosing

3.1 NVIDIA Morpheus

- Ingests incoming emails and parses headers, body, URLs, and attachments into structured features suitable for Machine Learning analysis.
- High-throughput processing of millions of emails per second
- Built-in cybersecurity ML models (NLP, GNN, digital fingerprinting)
- Anomaly detection using statistical methods
- Calculate the risk SCORE





Morpheus transforms raw email strings into numerical features and vectors for AI models analysis.

1. Email Parsing Stage

Raw email: "From: support@yourbank.com Subject: Urgent: Account Locked!"

↓

Structured data: headers dict + body text + URL list + attachment metadata

2. Feature Extraction Stages

"support@yourbank.com" → numerical features:

- domain_age = 3 days [new = risky]

- SPF_check = failed [0.0 score]

- sender_reputation = -0.7

↓

"yourbank.com" → embedding vector [0.23, -0.15, 0.89, ...] (BERT/transformer)

↓

"Urgent: Account Locked!" → NLP features:

- urgency_score = 0.92

- credential_request = True (1.0)

- impersonation_score = 0.87

3. Matrix Output for ML

Email ID | header_features | content_vector | url_risk | FINAL_SCORE

123 | [0.1, -0.3,..] | [0.23,...512] | 0.91 | 0.94 ← phishing!

4. Triton Inference

The numerical matrix goes to GPU-accelerated models (BERT, custom classifiers) running on NVIDIA Triton, producing your phishing risk score.

6. Runs the models

Compute a phishing risk score for each email and output whether it is likely phishing or benign.

7. Train the models

By feeding them common phishing email, we can feed the real-world cases such that it can learn, analyst the new threats.

- Subtle Misspellings: Attackers use slight variations, such as replacing ‘o’ with ‘0’ or changing ‘.com’ to ‘.next’
- Deceptive Subdomains: The domain looks real, but is actually a subdomain of a malicious site (e.g., apple.com.security-update.com).
- IP Address Instead of Domain: URLs that are just a string of numbers (e.g., http://192.168.1.1) are rarely legitimate for commercial sites.
- URL Shorteners: Links using bit.ly or tinyurl are used to hide the true, malicious destination.

Example:

- Subtle misspellings: y0urbank.com, paypa1.com
- Deceptive subdomains: apple.com.security-update.com
- IP addresses: http://192.168.1.1
- URL shorteners: bit.ly/abc123

NVIDIA Morpheus has continuous learning capabilities, it can learn by itself

- Generative AI Simulation: Learns from new attack within hours, by feeding it real world sample or simulated sample from NVIDIA NeMo

- Multi-Dimensional Detection
  - Content Analysis
  - Behavioral Analytics
  - Relationship Mapping
  - Temporal Pattern Recognition

3.2 NVIDIA NeMo Agent Toolkit for Multi-Agent Systems

- Allow developers to build multiple cooperating AI agents (for example, URL agent, header agent, policy agent, explanation agent) that consume Morpheus scores and features.

- These agents decide whether to allow, flag, or quarantine the email, and can generate user-facing messages that tell users an email/URL is likely fake and explain why.

- Generate Threat Variants with NeMo for Morpheus

- Make the decision

NeMo Agent Toolkit uses APIs to talk to Morpheus, and it excels at connecting to any tools/ data sources and even human-in-the-loop chat.

NeMo agent Toolkit is more like a decision maker, for example:

```
if risk_score > 0.9: quarantine()

elif risk_score > 0.8: flag()

else: allow()
```

The Morpheuse can tell you the percentage, but it will not make a decision. Instead, we can use NeMo agent Toolkit to make a decision, if the risk score bigger than 97% then, quarantine, elseif the risk bigger than 80%, will be flagged and receive the email, else it will pass and receive the email. After that, it will reason about context and intent.

Example 1: Urgent Invoice Payment

Morpheus sees: "0.99, quarantine."

NeMo Agent sees: "This vendor is fake, because their email was not the vendor's email, the payment linked to unknow webpage, quarantimed!"

Example 2: Password Reset Request

Morpheus sees: "0.96, quarantine."

NeMo Agent sees: "The URL for the password change page is incorrect; it redirects to a fake password change page, quarantimed!"

Example 3: Charity Donation Request

Morpheus sees: "0.81, flag."

NeMo Agent sees: "The donation is genuine, but the URL of the payment page is not from the charity, flagged!"

Example 4: HR Benefits Update

Morpheus sees: "0.23, allow."

NeMo Agent sees: "Sender and links verified to legitimate internal HR portal, no anomalies found, allowed."

3.3 How Morpheus work with NeMo Agent Toolkit

Morpheus makes the decision first. After that, it will tell NeMo Agent Toolkit via APIs. Next, NeMo Agent Toolkit will create agents that generate human explanations.

Together, Morpheus is “what”, NeMo Agent Toolkit is “why”.

Morpheus forcuses on high-speed detection, and the NeMo Agents provide contextual, human-friendly explanations.

Example:

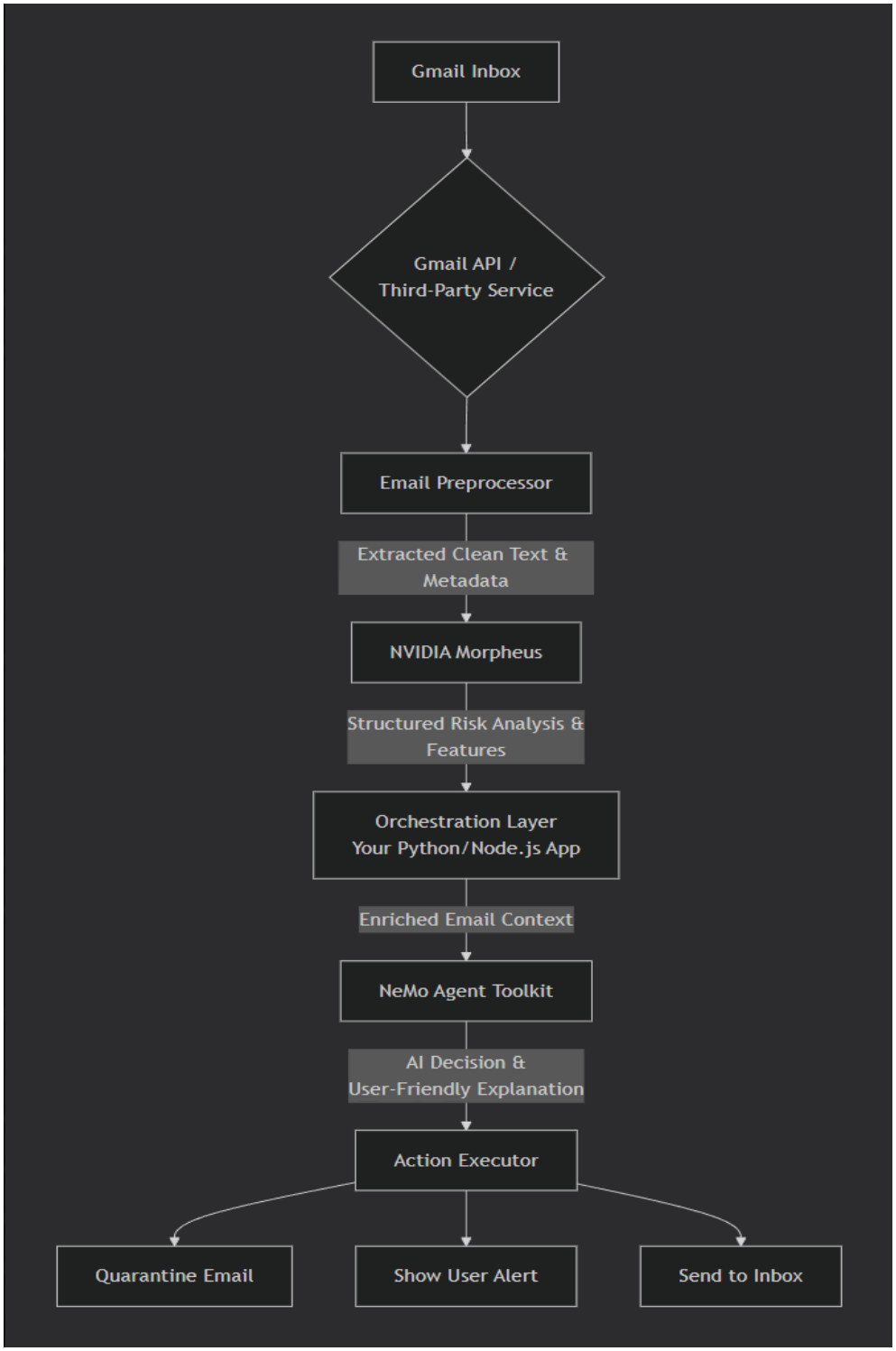
1. Morpheus → Detects threat → "0.99, quarantine"

2. NeMo Agent Toolkit → Receives Morpheus output

3. NeMo Agent → Analyzes WHY → Generates explanation

4. User sees: "Quarantined: Vendor email fake, payment link suspicious"

Overall summary



1. **Step 1: Fetch from Gmail**

   A separate service (like a Python script or cloud function) uses the **official Gmail API** to fetch new emails. This service acts as the "bridge."

2. **Step 2: Preprocess & Structure**

   The raw email (HTML, headers, attachments) is cleaned and parsed into a structured JSON format that your AI pipeline can understand.

3. **Step 3: Analyze with Morpheus**

   This structured data is sent through your **NVIDIA Morpheus** pipeline. Morpheus runs its ML models and outputs a **risk score** and a list of detected features (e.g., "suspicious_url": true).

4. **Step 4: Pass to NeMo Agent**

   Your main application code (the **Orchestration Layer**) takes the email data *plus* Morpheus's analysis and passes it to the **NeMo Agent Toolkit** as context.

5. **Step 5: Agent Reasons & Acts**

   Your NeMo Agent (e.g., the DecisionAgent) uses this complete context to make its final judgment and generate a user-friendly explanation.

6. **Step 6: Execute Action**

   Based on the agent's decision, your orchestration layer calls the **Gmail API again** to apply the label, move the email, or send a notification back to the user's inbox.

[Gmail] → [Fetch Service] → [Morpheus Analysis] → [NeMo Agent Decision] → [Action]

3.4 NeMo Microservices for agent continuous learning

The decision not to use NeMo Microservices for continuous learning here is a practical, scope-driven choice for a prototype, not a dismissal of their value. Their role is central to a full production system.

For the prototype, manually simulating the learning step keeps you agile. The NeMo Microservices as the intended pathway for scaling the system and enabling true continuous learning after the core concept is validated. So in this project, NeMo Agent Toolkit for Multi-Agent Systems and NVIDIA Morpheus would be enough to the prototype.

In future, we could use NeMo Microservices for agent continuous learning work with NVIDIA Morpheus and NeMo Agent Toolkit for Multi-Agent Systems.

- Performance
  - Scalability: Individual microservices scale independently
  - GPU pooling: Shared GPU resources across all services
- Flexibility
  - Mix models: Different LLMs for specialized tasks
  - A/B testing: Test new models without pipeline disruption
  - Multi-language: Easy integration of language-specific models
- Maintenance
  - Independent updates: Update NLP models without touching security models
  - Fault isolation: One service failing doesn't break entire pipeline
  - Monitoring: Granular metrics per service

Example

1. Email arrives → API Gateway

2. Orchestrator calls 3 services in parallel:

   - NeMo Microservice 1: Extract entities (people, orgs, dates)

   - NeMo Microservice 2: Generate text embeddings

   - Morpheus: Analyze threats and calculate risk SCORE

3. NeMo Agent combines all outputs get the risk SCORE from Morpheus via APIs

4. NeMo Microservice 3: Generate user-friendly explanation

5. Action Service: Apply label/move email in Gmail

4.Conclusion

This project delivers an AI-driven phishing assistant defined by **intelligent synergy** and **adaptive defense**:

1. **Architecture for Intelligence:** We combine **NVIDIA Morpheus** for high-speed detection with the **NeMo Agent Toolkit** for contextual reasoning. Morpheus finds the threat; the NeMo Agent explains *why* it's a threat in plain language for the user.

2. **Disrupting the Attack Psychology:** The system's key innovation is its **pre-emptive warning**. By immediately and clearly telling a user an email is suspicious and showing concrete evidence (like a mismatched URL), it disrupts the psychological triggers—urgency, curiosity, fear—that phishing relies on. This effectively **inoculates users** against manipulation at the point of attack.

3. **Foundation for the Future:** The prototype is built with a clear path to becoming **self-learning**. By integrating **NeMo Microservices** for continuous model evaluation and fine-tuning, the system can evolve autonomously to counter new threats, ensuring long-term resilience.

In summary, this solution is more than a filter; it's an **adaptive security layer** that reduces risk by combining scalable AI detection with human-centric communication to strengthen the first line of defense: the user.

Appendix

Dulce, S. (2025, August 26). *What is phishing? Everything you need to know*. https://zeronetworks.com/blog/what-is-phishing-everything-you-need-to-know

Trevino, A., & Trevino, A. (2025, April 10). *How to spot a phishing email*. Keeper Security Blog - Cybersecurity News & Product Updates. https://www.keepersecurity.com/blog/2023/09/22/how-to-spot-phishing-emails/

Dwiastuti, L. (2025, December 19). *What is a URL? Uniform Resource Locator explained*. Hostinger Tutorials. https://www.hostinger.com/au/tutorials/what-is-a-url

NVIDIA NEMO Agent Toolkit Overview — NVIDIA NEMO Agent Toolkit (1.3). (n.d.). https://docs.nvidia.com/nemo/agent-toolkit/latest/index.html

Overview of NEMO Microservices — NVIDIA NEMO Microservices documentation. (n.d.). https://docs.nvidia.com/nemo/microservices/latest/about/index.html

Getting Started with Morpheus — Morpheus (25.06). (n.d.). https://docs.nvidia.com/morpheus/getting_started.html

Luk Ho Lung

HKSTP Talent Foundry bootcamp

January 26th 2026