Project Milestone 1: Statement of Work

Project: Rescam — Email Phishing Detection with Intent & Generative Context Understanding

Title & Authors

- **Team Name:** Rescam
- **Authors (Name • Email):**
  - Amit Berger • **aberger@mba2026.hbs.edu**
  - Harry Price • hprice@mba2026.hbs.edu

- **Submission Date:** 2025-09-25

Background & Motivation

Email phishing—especially business email compromise (BEC), credential harvesting, and invoice fraud—continues to drive breaches and losses. Modern phishing campaigns use brand impersonation, context from prior threads, and clean infrastructure that bypasses naive filters. We aim to build a practical, MLOps-grade **email-only** detection system that (a) identifies phishing, (b) classifies the attacker's **intent**, and (c) explains decisions using **generative AI** to reason over message text, headers, and linked content.

Problem Statement (short)

Build and evaluate a production-oriented **email phishing** detector that infers attacker **intent** (e.g., credential harvest, payment fraud, malware delivery, account takeover) via generative context understanding, while leveraging headers and URL intelligence to deliver accurate, explainable classifications suitable for real-world deployment.

Step 3: Data Sources

Source of Data (open source)

- **Enron-Spam (AUEB)** — labeled ham/spam emails with headers/bodies
- **SpamAssassin Public Mail Corpus** — ham/spam with full headers
- **CSDMC2010 SPAM** — competition corpus with labeled training emails
- **Phishing Email Curated Sets (Nazario, Nigerian Fraud, etc.)** — mbox archives of verified phishing emails (curated on Zenodo/Figshare)
- **(Optional) Intent-labeled community set** — small HF community dataset with intent/technique fields (to be validated)
- **URL intelligence (for enrichment only):** PhishTank/OpenPhish community feeds

Description of Datasets

- **Enron-Spam:** ~33k emails (ham+spam) across multiple user mailboxes; realistic business ham.
- **SpamAssassin:** ham/spam folders with rich headers; good for header parsing and baseline models.
- **CSDMC2010:** labeled training set for email spam; useful for additional variety.
- **Curated phishing (Nazario/Nigerian Fraud, etc.):** thousands of verified phishing emails (mbox) spanning many brands and years.

Key Attributes

- **Text content:** subject and body (HTML/plain), quoted thread context
- **Email headers:** From/Reply-To/Return-Path, Received chain; auth results (SPF/DKIM/DMARC when present)
- **URL features:** domain age/TLD, homoglyphs/Levenshtein to brand, entropy/length, path/query tokens, landing-page title

Relevance to the Project

These corpora enable training and evaluation of content+header models; curated phishing sets supply **true phishing** exemplars. URL feeds support enrichment and cross-checking of links referenced by emails. Together they support **intent detection** and **LLM reasoning** over realistic signals.

Data Quality Concerns & Mitigations

- **Age & concept drift.** *Mitigation:* evaluate over time-split folds; augment with recent curated phishing emails and URL feeds.
- **Label noise/imbalance.** *Mitigation:* dedupe, normalize, stratify; class weights/focal loss; calibrated thresholds.
- **Header sparsity/inconsistency.** *Mitigation:* robust parsers; degrade gracefully to content-only.
- **PII handling.** *Mitigation:* redact emails/phones/usernames at training time; document governance; limit retention.

Important: Proposals must include concrete, available datasets and the above details to be accepted.

Step 4: Scope & Preliminary Design

Objectives

- Build an **email** detector
- Predict **binary phishing** and **intent class** (credential harvest, payment/billing fraud, malware delivery, account takeover/BEC, other)
- Provide **natural-language rationales** (LLM) and highlight risky spans/headers/URLs
- Expose a scalable inference service and simple UI for testing and demo

Minimum Components Mapping

- **Large/Heterogeneous Data:** multiple email corpora + curated phishing mbox; header + text + URL enrichment
- **Scalability:** stateless API (FastAPI) with async URL expansion; containerized; horizontal autoscaling
- **Complex Models:** LLM-based intent scorer
- **Computationally Expensive Inference:** LLM reasoning + URL analysis; offer quantized/distilled LLM and caching for latency SLAs

Preliminary Architecture (high level)

- **Ingestion & Versioning:** pipelines to fetch email corpora and curated phishing sets; data versioning & lineage
- **Featurization:** robust parsing; HTML text extraction; header anomaly features; URL expansion + domain features
- **Modeling:**
  - **Intent taxonomy & labels:** seed with curated phishing + weak labeling rules; validate with human spot-checks
  - **LLM intent scorer:** few-shot prompts to classify intent + generate explanation
  - **Fusion classifier:** combine LLM scores with header/URL/tabular features classifier for final verdict
- **Serving:** REST API; feature store for URL/domain features; streaming feed updates
- **Monitoring:** drift, performance & cost telemetry; canary deployments; red-team prompts for robustness
- **Track and Notify on Misclassified Emails** To ensure continuous improvement and user trust, the system will include a **feedback loop** for emails incorrectly classified as spam or phishing.

Access & Ingestion Strategy

- **Primary (OAuth-connected dashboard):** Users sign in and grant read-only access to their mailbox; we process **new emails only** via provider notifications or safe polling and surface alerts inside the dashboard. Minimal data retention (configurable); store derived features and verdicts where possible.
- **Fallback (Forward-to-Analyze):** Users forward suspect emails to a designated address we control. Ingestion decodes MIME, parses headers, and runs the same pipeline. Optional auto-reply with verdict & explanation.
- **Security & Privacy:** least-privilege scopes; encrypted storage and transport; PII redaction where feasible; transparent consent & data policy; audit logging.

Application Mock (textual wireframe)

- **Screen 0 — Connect mailbox:** "Connect Gmail/Outlook"; OAuth consent; show connected status and scope info
- **Screen 1 — Alerts feed:** timeline of flagged emails (sender, subject, intent, severity, confidence); filters and search
- **Screen 2 — Email details:** verdict + **intent label**; LLM rationale; highlighted risky spans/URLs; header anomalies (SPF/DKIM/DMARC)
- **Screen 3 — Settings:** choose ingest mode (OAuth vs Forward-to-Analyze), retention window, auto-reply toggle, webhook/notification preferences

Learning Emphasis

Focus on MLOps fundamentals and **LLM-augmented** detection: reproducible data/experiments, evaluation under drift, scalable serving, monitoring, and safe prompting.

Limitations & Risks

- **Domain shift vs. modern kits** → mitigate with curated phishing updates and evaluation under drift
- **Label scarcity for intent** → mitigate with weak supervision + targeted human review
- **Ethical/PII constraints** → mitigate with redaction, access controls, and clear governance

Milestones & Tentative Deadlines

- **M1 (today, 09/25):** Submit SOW; register team; finalize **intent taxonomy**, dataset list, and architecture; decide ingest modes; scaffold OAuth app registration; create forwarding mailbox
- **By 09/29:** Incorporate staff feedback; configure OAuth consent + tenant/test mailbox; implement connector skeleton (OAuth flow, token storage); draft privacy/data policy
- **10/03:** Connector MVP (list new messages; safe polling); MIME parse + header/URL feature extraction; baseline classifier; first LLM prompt for intent
- **10/10:** adapter experiments for intent; fusion model v1; alerts pipeline; **Dashboard MVP** (connect mailbox + alerts feed)
- **10/17:** Real-time ingest path (notifications/queue); explanations UI; RBAC & audit logs; performance & drift monitors
- **10/24:** Hardening (rate limits/backoff), latency/cost optimizations, privacy review, red-team tests, final demo prep

Selected References (for R&D)

- Metsis, Androutsopoulos, Paliouras (CEAS 2006) — Enron-Spam dataset
- SpamAssassin Public Mail Corpus
- CSDMC2010 SPAM competition corpus
- Nazario / Nigerian Fraud phishing email corpora (curated)
- Recent work on LLM-based explainable phishing detection (2023–2025)