AI-Driven Local Phishing Detection Agent

Updated Design Summary, Pipeline Review, and Roadmap (v1–v3)

1. Updated Executive Summary

This system is a privacy-first, local-only AI agent designed to monitor a user's inbox for phishing activity in near real-time. It avoids centralized brand databases and cloud-based reputation feeds, instead relying on structural signals, email authentication metadata, lightweight sandbox analysis, semantic reasoning, and explainable risk scoring. All sensitive data is handled ephemerally and deleted after analysis.

2. Core Design Principles (Reaffirmed)

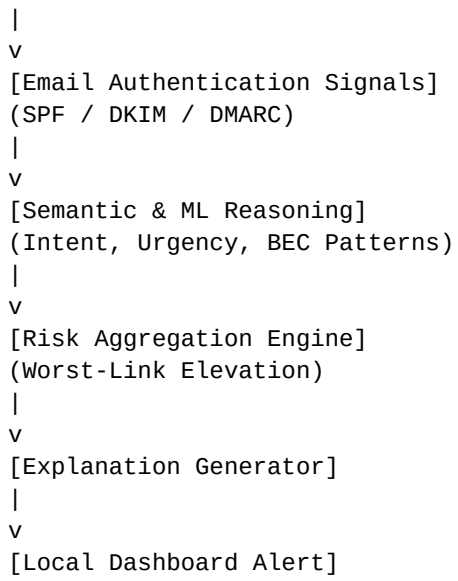
1. Local-only execution with zero third-party data sharing
2. Explainable risk inference instead of binary legitimacy claims
3. AI-assisted reasoning to avoid static brand or URL databases
4. Least-privilege email access (Mail.Read / Mail.ReadBasic)
5. Encrypted ephemeral storage with guaranteed deletion
6. Worst-link elevation: the most suspicious artifact defines risk

3. Updated Pipeline Architecture

The pipeline has been expanded to incorporate simple but high-impact security signals that are often overlooked in early phishing detection systems, including temporal anomalies, reply-to mismatches, display-name deception, and protocol abuse detection.

4. Updated Surface-Level Flowchart

```
[New Email Event]
|
v
[OAuth Ingestion]
|
v
[Header Parsing]
(From, Reply-To, Auth Results, Time)
|
v
[Body & HTML Parsing]
|
v
[URL & Artifact Extraction]
(Links, Shorteners, Protocols)
|
v
[Basic Sandbox Analysis]
(DNS, TLS, Redirect Chains)
```


|
v
[Email Authentication Signals]
(SPF / DKIM / DMARC)
|
v
[Semantic & ML Reasoning]
(Intent, Urgency, BEC Patterns)
|
v
[Risk Aggregation Engine]
(Worst-Link Elevation)
|
v
[Explanation Generator]
|
v
[Local Dashboard Alert]

5. Version 1 (v1) – Hardened Foundation

1. Local Gmail & Outlook monitoring with least-privilege OAuth
2. Header analysis including Reply-To mismatch and temporal anomalies
3. Robust URL extraction with protocol abuse and shortener detection
4. Basic sandbox analysis: DNS, HTTPS, redirect depth, domain age
5. Email authentication checks (SPF, DKIM, DMARC)
6. Semantic intent analysis and ML phishing probability scoring
7. Explainable risk score with controlled vocabulary reasons
8. Localhost dashboard with alert-focused UX

6. Version 2 (v2) – Intelligence & Deception Awareness

1. HTML deception heuristics (hidden links, anchor mismatch)
2. Advanced redirect chain scoring and TLD change detection
3. Negative evidence detection (missing personalization, generic phrasing)
4. User-controlled trust actions with guardrails
5. Caching, performance optimization, and timeout budgets

7. Version 3 (v3) – Advanced Analysis & Deployment

1. Optional headless browser sandbox execution
2. Attachment and image-based phishing analysis
3. Pluggable, opt-in brand intelligence modules
4. Feedback-informed scoring refinement
5. Dockerized deployment and portability

8. Design Opinion: URL Analysis Strategy

URL analysis is treated as structural risk inference rather than reputation lookup. The system does not attempt to label URLs as safe or malicious, but instead evaluates consistency, intent alignment, and technical plausibility. Redirect chains, protocol misuse, newly registered domains, and certificate timing inconsistencies are weighted more heavily than surface-level HTTPS presence. This approach is resilient against novel phishing infrastructure and minimizes dependence on external databases.

9. Final Assessment

With the inclusion of overlooked but simple signals, the pipeline now reflects how a human security analyst reasons about phishing. The system remains lightweight, privacy-preserving, and extensible, while delivering high trust and explainability to the end user.