# Elara URL Scan Platform – Final V2 Design

## Introduction

This document presents the **final V2 architecture** for the Elara cybersecurity platform. It is an incremental upgrade rather than a rewrite—v1 still exists for backward‑compatibility. The V2 engine introduces deterministic models with calibrated probabilities, unifies AI summarisation via **Gemini** on Vertex AI, and exposes a B2B **Central AI API** and **Model‑Training API**. All existing features (URL/message/file scanning, threat‑intel sync, deepfake detection, profile analysis, fact checking, recovery support, literacy coach) are preserved and integrated into the new consensus pathway.

### Why Vertex AI?

According to Google’s official materials, **Vertex AI** is a *unified machine‑learning platform* that streamlines data preparation, model training and deployment under one roof[[1]](https://cloudchipr.com/blog/vertex-ai#:~:text=Vertex%20AI%20is%20Google%20Cloud%E2%80%99s,using%20Google%20Cloud%27s%20robust%20infrastructure). It supports both AutoML and custom training, allowing teams to build models using no‑code workflows or frameworks like TensorFlow and PyTorch[[2]](https://cloudchipr.com/blog/vertex-ai#:~:text=Vertex%20AI%20lets%20you%20do,users%20of%20all%20skill%20levels). Vertex AI also provides a feature store, pipelines for automation, experiment tracking and a model registry[[3]](https://cloudchipr.com/blog/vertex-ai#:~:text=Beyond%20training%2C%20Vertex%20AI%20offers,MLOps%20features%20to%20operationalize%20AI). These capabilities make it the ideal platform for training and serving Elara’s models while maintaining reproducibility and observability.

### High‑Level Overview

V2 replaces the hand‑tuned scoring and multi‑LLM consensus of v1 with a **two‑stage model pipeline** and calibrated probabilities. It also consolidates the AI summarisation layer to Gemini. The system is divided into four domains: **Scanning & Analysis**, **Threat‑Intel & External Data**, **Admin & APIs**, and **Data Lake & Training Loop**.

## ASCII Architecture Diagram

External Users & Clients B2B Partners Admins  
 │ │ │  
 ┌───────────▼───────────┐ ┌───────────▼───────────┐ │  
 │ API Gateway & WAF │ │ Central AI API │ │  
 │ (JWT, OAuth2, rate │ │ (multi‑tenant) │ │  
 │ limits, routing) │ │ │ │  
 └───────────┬───────────┘ └───────────┬───────────┘ │  
 │ │ │  
 ┌──────────────────▼───────────┐ ┌────────────▼──────┐ ┌──▼──┐  
 │Scan & Analysis Request Router│ │Data & Training API│ │Admin│  
 │• v1/v2 dispatch │ │• Download features │ │UI │  
 │• Tenant config │ │• Upload training │ │ │  
 │• Dedup & caching │ │• Launch pipelines │ │ │  
 └──────────────┬──────────────┘ └───────────┬────────┘ └────┘  
 │ │ │  
 │ │ │  
 ┌────────────▼───────────┐ ┌──────────────────▼──────────────┐  
 │Threat‑Intel Gate │ │Data Lake (BigQuery) & Feature │  
 │• 18 sources │ │Store (Vertex) │  
 │• Weighted matching │ │• Stores scan & TI features │  
 │• Policy: tombstone/dual │ │• Ingests uploaded data │  
 │ tier‑1 hits │ │• Feeds training and inference │  
 └──────────────┬──────────┘ └───────┬─────────────────────────┘  
 │ │  
 Reachability Probe │  
 • ONLINE/OFFLINE/WAF │  
 │ │  
 ┌────────▼───────────┐ │  
 │Evidence Collector │ │  
 │• Headless browser │ │  
 │• WHOIS/DNS/TLS/ASN │ │  
 │• Screenshot + OCR │ │  
 └────────┬────────────┘ │  
 │ │  
 Feature Extraction │  
 • Lexical tokens │  
 • Tabular metrics │  
 • Causal signals (FOM, │  
 BID, RHD, auto‑download)│ │  
 │ │  
 Stage‑1 Models │  
 • URL encoders (XGBoost +│  
 URLBERT/PhishBERT) │  
 • Tabular risk model │  
 • Early exit on high │  
 confidence │  
 │ │  
 ┌───────────────High confidence?──────────────┐  
 │ │ │  
 │ Yes │  
 │ Return early verdict │  
 │ │ │  
 │ No │  
 ▼ │ ▼  
 Stage‑2 Models │ Stage‑2 Models (offline branch)  
 • Text (Gemma/Mixtral) │ • Only if applicable  
 • Image CNN (login layout) │  
 │ │  
 Combiner & Conformal Calib. │  
 • Fuse logits & causal signals │  
 • Produce probability + CI │  
 • Branch‑specific thresholds │  
 │ │  
 Policy Overrides │  
 • Tombstone │  
 • Dual tier‑1 hits │  
 • FOM+BID+young domain │  
 │  
 ▼  
 Verdict & Evidence → DB & Cache  
 │  
 Consensus Service  
 • Returns risk score, band, probability, CI  
 • Returns decision graph & recommendations  
 • Invoked by Central AI API & Ask Elara Chatbot  
 │  
 ┌─────────────┴─────────────────────┐  
 │ │  
 Gemini Router (Flash/Pro) Training Pipelines  
 • Summarise results & answer • Vertex AI pipelines train URL,  
 questions tabular, text, image & combiner models  
 • Use RAG for knowledge retrieval • Use Feature Store features  
 │ │  
 ▼ ▼  
 Ask Elara Chatbot Model Registry (Vertex)  
 • Answers user questions • Versioning & experiment tracking  
 • Cites TI sources

### Key Differences from v1

1. **Deterministic scoring with calibrated probabilities**: v1 uses static weights and LLM multipliers; v2 trains models on scan data and TI features, and calibrates the outputs with conformal prediction.
2. **Branch‑specific thresholds**: Each reachability state (ONLINE/OFFLINE/WAF/PARKED) has its own probability thresholds to convert the model output to risk bands, avoiding under‑scoring on offline scans.
3. **Central AI & Model‑Training APIs**: New endpoints expose verdicts, chat answers, and data access while enforcing tenant isolation and billing. These endpoints also accept uploaded training data for custom models.
4. **Gemini‑only summarisation**: All human‑readable explanations and chatbot responses go through Gemini via Vertex AI with smart routing and caching.

## Implementation Notes

1. **Data Lake & Feature Store**: Use BigQuery as the canonical store for scan features and threat‑intel hits. Standardise feature schemas and publish them to Vertex AI Feature Store so that both training and real‑time inference use the same definitions[[3]](https://cloudchipr.com/blog/vertex-ai#:~:text=Beyond%20training%2C%20Vertex%20AI%20offers,MLOps%20features%20to%20operationalize%20AI).
2. **Training Pipelines**: Build Vertex AI Pipelines that read from BigQuery, prepare data, and train the URL encoders, tabular risk models, text models (Gemma/Mixtral), image CNN and the combiner. Register the trained models in Vertex AI Model Registry. Employ the MLOps features (experiment tracking, model registry, and pipelines) to automate retraining and deployment[[3]](https://cloudchipr.com/blog/vertex-ai#:~:text=Beyond%20training%2C%20Vertex%20AI%20offers,MLOps%20features%20to%20operationalize%20AI).
3. **Model Deployment**: Deploy models to Vertex AI prediction endpoints. Use CPU nodes for small models (URL encoders, tabular) and GPU nodes for text and image models. Co‑host Stage‑1 models to save cost; Stage‑2 models run only on uncertain cases.
4. **Gemini Router**: Implement a router that uses Gemini 1.5 Flash for simple summaries and Gemini 1.5 Pro for complex reasoning. Use Vertex AI to unify billing and leverage your credits; maintain a fallback to the direct Gemini API in case Vertex is down. Cache responses keyed by (question, evidence) for 1–6 hours depending on volatility.
5. **Consensus Service**: Expose a service that receives a scan ID and optional modules (deepfake, profile, fact). It fetches the scan verdict, calls additional analyzers, and returns a combined result. This is consumed by the AI chatbot and the Central AI API.
6. **Admin & B2B UI**: Extend the admin dashboard to support selecting v1/v2 engines, uploading custom datasets, managing API keys and viewing usage. Provide API documentation and SDKs for partners to integrate the Central AI and Model‑Training APIs.

## Conclusion

The V2 architecture keeps all existing functionality while introducing a robust, scalable and cost‑efficient scanning pipeline. By leveraging Vertex AI’s unified ML platform[[1]](https://cloudchipr.com/blog/vertex-ai#:~:text=Vertex%20AI%20is%20Google%20Cloud%E2%80%99s,using%20Google%20Cloud%27s%20robust%20infrastructure) and automating training and deployment, the platform ensures continuous improvement. The introduction of the Central AI API and Model‑Training API unlocks new revenue streams without compromising security or privacy.

[[1]](https://cloudchipr.com/blog/vertex-ai#:~:text=Vertex%20AI%20is%20Google%20Cloud%E2%80%99s,using%20Google%20Cloud%27s%20robust%20infrastructure) [[2]](https://cloudchipr.com/blog/vertex-ai#:~:text=Vertex%20AI%20lets%20you%20do,users%20of%20all%20skill%20levels) [[3]](https://cloudchipr.com/blog/vertex-ai#:~:text=Beyond%20training%2C%20Vertex%20AI%20offers,MLOps%20features%20to%20operationalize%20AI) What Is Vertex AI? Streamlining ML Workflows on Google Cloud

<https://cloudchipr.com/blog/vertex-ai>