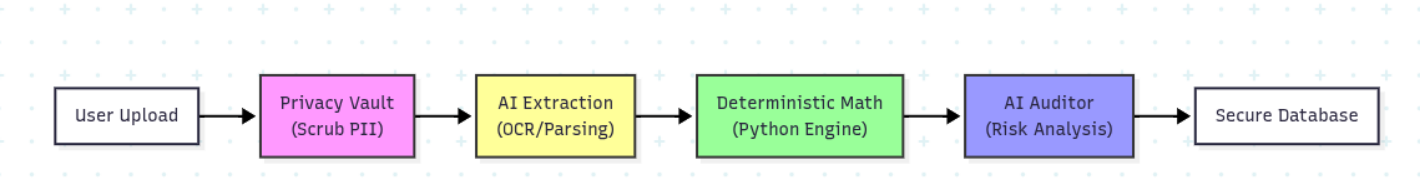
Sentinel Project Documentation

1. System Flow Diagram



2. System Explanation: Hybrid Architecture

Hybrid Architecture: Why We Separate AI and Math

In building Sentinel, I adopted a hybrid architecture that splits responsibilities between deterministic code and artificial intelligence based on the nature of the task. LLMs (Large Language Models) are powerful reasoning engines but can be unreliable calculators, prone to "hallucinations" when performing precise arithmetic operations.

To solve this, we strictly isolate the **payroll calculation core** from the **AI layer**. We use pure, deterministic Python functions for all monetary operations—gross pay, tax rates, and net pay calculations. This ensures 100% accuracy and auditability, eliminating the risk of an AI guessing a salary figure.

Conversely, we deploy AI agents where they excel: **Ambiguity and Reasoning**. The AI Auditor evaluates the *context* of a transaction (e.g., "Is 50 hours of overtime normal for a Junior Designer?"), a task that requires semantic understanding rather than strict math. This separation of concerns gives us the best of both worlds: the unyielding precision of code for the money, and the adaptive intelligence of agents for the oversight.

3. Security Note: The Privacy Vault Pattern

Security in Sentinel is enforced through the "Privacy Vault" pattern, a middleware layer that sits between the user input and the outside world. Before any data is passed to an AI model or stored in logs, it must pass through the Vault.

This layer performs two critical functions: 1. **PII Scrubbing**: Using rigorous regex patterns, it detects and redacts sensitive information like email addresses (*email*) and phone numbers (*phone*), replacing them with generic placeholders like `[REDACTED]`. 2. **Identity Anonymization**: It decouples user identities from their data by generating deterministic hashes (e.g., `User_C8DA`) for internal processing.

This ensures that even if an AI model were to leak data or a log file were exposed, no personally identifiable information would be compromised. The system operates entirely on anonymized contexts, preserving user privacy by design.