Securing AI Agents in FinTech: PII Redaction Pipelines

Practical patterns for deploying LLM agents in regulated finance - without leaking cardholder, banking or identity data.

Why this matters

Financial services teams in the US, UK, EU and the Gulf want to use LLMs for support triage, dispute analysis and KYC summarisation. But every input may contain cardholder data, IBAN numbers, national IDs, and other regulated PII. Send that to a hosted LLM and you have a compliance incident.

The PII redaction pipeline

1. Inbound text passes through a deterministic detector for IBAN, PAN, SSN, NI, Emirates ID, Iqama, etc.
2. A second-pass NER model catches names, addresses and free-form identifiers
3. Detected spans are replaced with stable, reversible tokens (e.g., [PAN_001])
4. The redacted prompt goes to the LLM
5. The response is re-hydrated only inside the trust boundary, never logged

Audit & evidence

- Append-only audit log of every redaction decision
- Hash-based reconciliation so auditors can prove no PII left the boundary
- Quarterly red-team exercises against the detector

> *Outcome: SOC 2 Type II and PCI DSS audits passed with zero AI-related findings.*

Where this fits

Banks in London, payment processors in Dubai, neobanks in Berlin, BNPL providers in Sydney, and KYC vendors in Riyadh have all deployed variants of this pipeline.