

Data Leakage Prevention with LLMs

Ensure your LLM applications are truly secure against data exposure. Taught by industry expert Jonathan Barrios, this hands-on course guides you through tracing sensitive data across your entire AI stack, from initial user prompts and RAG systems to agent tool calls and system logs. You'll step into real-world scenarios using Google Colab notebooks to implement robust data protection controls layer by layer. Through practical exercises, you'll master prompt scanning, automated PII redaction, granular retrieval permissions, prompt injection mitigation, output filtering, and secure logging. By the end of this course, you will have built a complete, working data protection pipeline built on synthetic data that you can adapt to your own applications or systems.

[CBT Nuggets course material](#) →



WEEK 1

Explain How Data Leaks Through AI Apps

Course Setup and What This Skill Covers
Software Evolution and Data Movement, Exposure, and Leakage
Tracing a Secret Through Every Surface
Challenge ☒

Detect Sensitive Data in Prompts

Introduction
A Checkpoint Before the Model
Match, Finding, Policy, and the First Detector
Regular Expressions and the Span Contract
Regular Expressions and the Span Contract Continued
Challenge

Classify Data by Sensitivity

Introduction
Detection Versus Sensitivity and the Four Levels
Default Classification Rules as Policy
Classifying Every Finding, Context, and Testing
Challenge

Redact Sensitive Data Before Sending

Introduction
Detection Is Not Protection
Mapping Sensitivity Levels to Policy Actions
Redaction Placeholders and Finding Spans
Right-to-Left Redaction and the Pre-Model Boundary
Challenge

Protect Data in File Uploads

Introduction

Files as a Separate Input Boundary

Metadata Validation Versus Content Inspection

Extracting, Scanning, and Protecting File Content

Challenge

Protect Data in RAG Systems

Introduction

Documents Need Security Metadata

Preserving Metadata Through Chunking

Deterministic Retrieval and the Retrieval Packet

Challenge

Apply Permissions to AI Retrieval

Introduction

Identity Must Travel With the Request

Authorization Checks and Filtering Before Ranking

Permission-Aware Retrieval and the Context Packet

Challenge

Protect Multi-Turn AI Conversations

Introduction

Conversation State Is the Real Prompt

Scanning the Whole Conversation State

Replay Policy and the Protected Multi-Turn Request

Challenge

Test Prompt Injection for Data Leakage

Introduction

Prompt Injection Versus Data Leakage

Building a Deterministic Leakage Oracle

Attack Success Rate and Output Protection

Challenge

Protect Data in AI Agents

Introduction

Safe Output Is Not Safe Execution

Protecting Tool Arguments and Tool Results

Projection, Bounded Steps, and the Protected Agent Runner

Challenge

Filter Sensitive AI Outputs

Introduction

The Output Boundary Reuses the Scanner

Output Detection Versus Output Policy

Presidio, Redaction, and the Output Audit Record

Challenge

Protect Sensitive Data in AI Logs

Introduction

Logs Are Another Data Destination

Schema Allow Lists and the Log Sanitizer

The Safe Log Writer, Traces, Metrics, and Export

Challenge

Combine LLM Data Protection Controls

Introduction

Findings, Classifications, Actions, and the Shared Policy

Tool, Log, and Authorization Adapters

The End-to-End Protected Request

Challenge

Respond to an AI Data Leak

Introduction

Detection, Triage, and Containment

Containment by Boundary and Safe Evidence

Evidence Manifest, Provenance, and Recovery

Challenge

Design an LLM Data Protection System

Introduction

The Architecture Control Matrix

Governing Policy Without Rebuilding

Failure Paths, Invariants, and the Release Decision

Challenge