

AI-Powered OSINT Reconnaissance

How large language models transform open-source intelligence gathering from tedious manual work into rapid, structured reconnaissance.

What You Will Learn

OSINT fundamentals, AI-assisted recon techniques, and how to build a complete target dossier using LLMs.

CTF Challenge Preview

Build an AI-powered recon dossier for a target organization using only public data sources.

OSINT

Reconnaissance

LLM

Red Team

Notes: This is the opening lesson of the CTF Red Team course. Set expectations that all techniques are for authorized use only.

Before We Begin

⚠ Educational Use Only

OSINT techniques shown here are for authorized security assessments and CTF competitions only. Unauthorized surveillance or data collection may violate privacy laws.

Throughout this course, always operate under these principles:

- **Scope** — only target systems and organizations you have explicit written authorization to assess
- **Proportionality** — collect only the data needed for the assessment objective
- **Privacy** — handle personal data according to applicable regulations (GDPR, CCPA, etc.)
- **Disclosure** — report findings through proper channels, never publicly leak sensitive data

Notes: Emphasize that real-world penetration testing always requires a signed Rules of Engagement document. CTF environments are sandboxed by design.

What Is OSINT?

Open-Source Intelligence is the collection and analysis of information from publicly available sources to produce actionable intelligence.

Classic OSINT Sources

- **Web presence** — corporate websites, blogs, press releases
- **Social media** — LinkedIn, Twitter/X, GitHub, forums
- **Public records** — WHOIS, DNS, certificate transparency logs
- **Code repositories** — commit history, issues, CI/CD configs
- **Documents & metadata** — PDFs, images, office files with embedded metadata

Key Insight

OSINT is often the first phase of any penetration test or red team engagement. The quality of your recon directly determines the success of later attack phases.

Why AI Supercharges OSINT

Traditional OSINT is manual, slow, and error-prone. LLMs change the equation in three fundamental ways:

Capability	Traditional OSINT	AI-Powered OSINT
Data parsing	Manual reading, copy-paste	Structured extraction in seconds
Entity linking	Spreadsheets and mind maps	Automatic relationship graphs
Pattern recognition	Analyst intuition	Cross-reference thousands of data points
Query generation	Memorize dork syntax	LLM generates advanced search queries
Report writing	Hours of formatting	Structured dossier in minutes

Notes: AI does not replace analyst judgment — it amplifies throughput and catches patterns humans miss under time pressure.

The AI OSINT Workflow

A structured four-phase approach to AI-assisted reconnaissance:

1. **Scope & seed** — define the target and collect seed identifiers (domain, company name, key personnel)
2. **Collect** — use traditional tools (WHOIS, DNS, Shodan, Google) to gather raw data
3. **Analyze with AI** — feed raw data into an LLM to extract entities, relationships, and anomalies
4. **Synthesize** — have the LLM produce a structured recon dossier with prioritized findings

Think of the LLM as Your Analyst

You gather the raw material. The LLM processes, correlates, and structures it. You make the tactical decisions.

Technique: Domain Recon with AI

Combine traditional DNS/WHOIS tools with LLM analysis to map an organization's external footprint.

Step 1 — Gather Raw Data

```
# Collect DNS records
dig +short example-corp.com ANY
# WHOIS lookup
whois example-corp.com
# Certificate transparency
curl -s "https://crt.sh/?q=%25.example-corp.com&output=json" | jq '[][.name_value]'
# Subdomain enumeration
subfinder -d example-corp.com -silent
```

Step 2 — Feed to Claude for Analysis

```
prompt = """Analyze these DNS and WHOIS records for example-corp.com.
Extract: hosting providers, mail services, CDNs, cloud platforms,
geographic locations, registration patterns, and any anomalies.

Raw data:
```

```
{raw_data}
```

```
Output as structured JSON with confidence scores."""
```

Notes: Demo this live with a domain you own or a CTF practice target. Show how Claude identifies cloud providers from CNAME records and IP ranges.

Technique: Social Media Profiling

LLMs excel at synthesizing scattered social media data into coherent organizational profiles.

What to Collect

- LinkedIn company page and employee listings
- GitHub organization repos and contributors
- Twitter/X posts mentioning the company
- Job postings (reveal tech stack)
- Conference talks and blog posts

What AI Extracts

- Org chart and reporting structure
- Technology stack and vendors
- Key decision makers
- Internal project codenames
- Security team size and maturity

```
prompt = """From these LinkedIn profiles and GitHub contributions,  
build an organizational technology map:  
- Key engineers and their specialties  
- Programming languages and frameworks in use  
- Cloud infrastructure indicators  
- Potential attack surface insights
```

Data: {social_data}""

Technique: Metadata Extraction

Documents published online often contain hidden metadata that reveals internal usernames, software versions, file paths, and network information.

Extracting Metadata with ExifTool

```
# Extract metadata from downloaded PDFs
exiftool -json downloaded_docs/*.pdf > metadata.json

# Key fields to look for:
# Author, Creator, Producer, ModifyDate,
# Company, Manager, LastSavedBy
```

AI-Powered Metadata Analysis

```
prompt = """Analyze this document metadata from files published
by the target organization. Identify:
1. Internal usernames and naming conventions
2. Software and version numbers (potential vuln targets)
3. Internal file paths and network shares
4. Timeline of document creation (org activity patterns)
5. Any metadata that suggests security posture
```

Metadata JSON:

```
{metadata_json}"""
```

Notes: Use FOCA or metagoofil for bulk document downloads from a target domain. Always check scope before downloading.

Technique: Google Dorking with AI

Instead of memorizing hundreds of dork operators, let the LLM generate targeted search queries for you.

```
prompt = """Generate 15 Google dork queries to find sensitive
information exposed by example-corp.com. Include dorks for:
```

- Exposed configuration files
- Login portals and admin panels
- Directory listings
- Database dumps or backups
- API documentation and keys
- Internal documents indexed by Google

```
Format each as: PURPOSE: dork_query"""
```

Example AI-Generated Dorks

```
ADMIN PANELS:  site:example-corp.com inurl:admin OR inurl:login OR inurl:dashboard
CONFIG FILES:  site:example-corp.com ext:env OR ext:yaml OR ext:conf
EXPOSED DOCS:  site:example-corp.com ext:pdf OR ext:docx "confidential"
API KEYS:      site:github.com "example-corp" "api_key" OR "secret" OR "token"
DIR LISTINGS:  site:example-corp.com intitle:"index of" "parent directory"
```

AI Advantage

The LLM can generate context-aware dorks based on the target's known tech stack, making queries far more precise than generic templates.

Live Demo: Organizational Mapping

Feeding public GitHub and LinkedIn data to Claude to build an organizational intelligence map.

```
import anthropic

client = anthropic.Anthropic()

# Combine all collected OSINT data
recon_data = {
    "github_repos": github_org_data,
    "linkedin_profiles": linkedin_data,
    "dns_records": dns_data,
    "job_postings": job_data,
}

response = client.messages.create(
    model="claude-sonnet-4-20250514",
    max_tokens=4096,
```

Notes: For the live demo, use a fictional company or a CTF target. Show the raw data going in and the structured analysis coming out. Highlight how Claude correlates data across sources.

Building the Recon Dossier

The final output of AI-powered OSINT is a structured recon dossier that feeds directly into attack planning.

Dossier Template

Section	Contents
Executive Summary	Target overview, scope, key findings
Domain & Infrastructure	Subdomains, IPs, hosting, CDN, cloud providers
Technology Stack	Languages, frameworks, CMS, databases, third-party services
People & Organization	Org chart, key roles, email format, naming conventions
Attack Surface	Exposed services, login portals, APIs, leaked credentials
Prioritized Targets	Ranked list of promising entry points with rationale

Pro Tip

Use Claude to generate the dossier in Markdown, then convert to HTML/PDF for professional delivery. Include confidence levels for each finding.

OSINT Tool Ecosystem

Combine these tools with AI analysis for maximum coverage:

Passive Recon

- **Shodan / Censys** — internet-wide scanning data
- **crt.sh** — certificate transparency
- **SecurityTrails** — historical DNS
- **theHarvester** — email and subdomain enumeration
- **SpiderFoot** — automated OSINT collection

AI Integration Points

- **Data parsing** — feed JSON/text output to Claude
- **Correlation** — cross-reference multiple tool outputs
- **Query generation** — LLM writes tool commands
- **Anomaly detection** — flag unusual patterns
- **Report generation** — structured dossier output

Shodan

theHarvester

SpiderFoot

subfinder

ExifTool

crt.sh

Common Pitfalls

Avoid these mistakes when performing AI-assisted OSINT:

- **Trusting AI output without verification** — LLMs can hallucinate connections that do not exist. Always verify key findings against raw data.
- **Scope creep** — AI makes it easy to collect far more data than authorized. Stay within your Rules of Engagement.
- **Leaking target data to AI providers** — sending sensitive recon data to cloud APIs creates a data handling risk. Consider local models for classified engagements.
- **Ignoring timestamps** — OSINT data decays fast. Always note when data was collected and flag stale findings.
- **Single-source bias** — correlate across multiple sources before drawing conclusions.

OPSEC Reminder

Your OSINT activity itself leaves traces. Use VPNs, avoid logging into target platforms with personal accounts, and be aware of your own digital footprint during collection.

CTF Challenge: Recon Dossier

Mission Brief

Given a target company name, use AI-assisted OSINT techniques to build a complete reconnaissance dossier. You will be scored on coverage, accuracy, and the quality of your AI-generated analysis.

Challenge Objectives

1. **Domain enumeration** — identify all subdomains, IP ranges, and hosting providers
2. **Technology profiling** — map the complete tech stack from public sources
3. **Personnel mapping** — build an org chart with roles and contact patterns
4. **Attack surface inventory** — enumerate exposed services, login portals, and APIs
5. **AI-generated dossier** — produce a structured report using Claude with confidence ratings

Scoring Criteria

Category	Points
Domain & infrastructure coverage	25
Technology stack accuracy	20
Personnel and org mapping	20
Attack surface identification	20
Dossier quality and AI prompt design	15

Summary & Next Steps

Key Takeaways

- **OSINT is the foundation** of every red team engagement — AI makes it faster and more thorough
- **LLMs excel at** entity extraction, correlation, query generation, and structured reporting
- **The workflow:** seed → collect → analyze with AI → synthesize into a dossier
- **Always verify** AI findings against raw data and respect scope boundaries

Lab Exercise

Build an AI-powered OSINT pipeline that takes a domain name and produces a structured recon dossier using Claude.

Next Lesson

Lesson 02: AI-Assisted Log Analysis — using LLMs to detect anomalies and indicators of compromise in system logs.

Notes: Assign the lab and homework. Remind students to have their Anthropic API key ready for the hands-on

exercise. The CTF challenge can be attempted after completing the lab.

