# Private AI e Digital Forensics

Negli ultimi anni l'Intelligenza Artificiale, e in particolare i Large Language Models (LLM), si è affermata come uno strumento di supporto trasversale in un numero crescente di discipline. Anche un ambito storicamente prudente e metodologicamente rigoroso come la Digital Forensics sta iniziando a confrontarsi con queste tecnologie.

La capacità degli LLM di sintetizzare grandi volumi di dati, individuare pattern latenti e assistere l'analista umano nella correlazione delle evidenze rende l'AI estremamente attraente nel contesto investigativo.

L'adozione indiscriminata di soluzioni AI in cloud introduce però una serie di criticità che, in ambito forense e investigativo, non possono essere ignorate.

Questo articolo analizza la possible adozione della *Private AI* come alternativa concreta e sostenibile all'uso di AI gestite da terze parti in cloud.

Il focus è sull'applicazione alla computer e mobile forensics: verranno esaminate le differenze architetturali tra AI cloud e AI locale, le motivazioni pratiche e normative che spingono verso soluzioni on-premise, i requisiti hardware minimi, i principali framework disponibili e, infine, un esempio operativo di utilizzo di una Private AI basata su RAG (Retrieval-Augmented Generation) per "dialogare" con le evidenze digitali.

## AI in cloud e AI locale: differenze

# sostanziali

## AI in cloud gestita da terze parti

Le soluzioni di Intelligenza Artificiale in cloud, come i servizi LLM esposti tramite API, presentano vantaggi immediatamente percepibili: non richiedono investimenti iniziali in hardware, rendono disponibili modelli di grandi dimensioni costantemente aggiornati e consentono un'integrazione rapida nei flussi di lavoro esistenti.

In ambito forense, però, questi benefici si accompagnano a criticità strutturali difficilmente eludibili; infatti, l'invio di dati a servizi esterni implica una perdita di controllo sulle informazioni, poiché i contenuti trasmessi alle API possono essere loggati, conservati o analizzati dal fornitore.

A ciò si aggiungono problematiche di compliance normativa che coinvolgono il GDPR, il segreto istruttorio, le policy aziendali e i requisiti di sovranità del dato.

Nel complesso, l'AI in cloud introduce un *trust boundary* esterno che risulta spesso incompatibile con le esigenze di molte indagini digitali.

# AI locale (Private AI)

Con il termine *Private AI* si intende l'esecuzione di modelli di Intelligenza Artificiale – LLM, modelli di embedding e sistemi RAG – all'interno di ambienti completamente controllati, come infrastrutture on-premise, ambienti air-gapped o cloud privati sotto il pieno controllo dell'organizzazione.

Questo approccio consente di mantenere un controllo completo sul dato, migliora l'auditabilità e la riproducibilità delle analisi, favorisce la conformità normativa e permette una

personalizzazione profonda dei modelli e dei workflow operativi.

Nel contesto della Digital Forensics, la Private AI non rappresenta una scelta ideologica, ma un vero e proprio requisito operativo: è spesso l'unico modo per avere la certezza che le evidenze digitali non escano mai dal perimetro investigativo.

# RAG: Retrieval-Augmented Generation

Il *Retrieval-Augmented Generation* (RAG) è un framework che potenzia i modelli linguistici collegandoli a fonti di dati esterne e aggiornate, in questo modo l'LLM può recuperare informazioni specifiche e utilizzarle per generare risposte più accurate, contestuali e affidabili, senza la necessità di un nuovo addestramento del modello, quindi in termini pratici, il RAG agisce come un vero e proprio "bibliotecario" per l'AI.

Il funzionamento del RAG si articola in tre fasi concettuali, in primo luogo avviene **il *retrieval***, durante il quale il sistema ricerca le informazioni più pertinenti all'interno di una base di conoscenza esterna composta da documenti, database o altre fonti strutturate.

Questa base di conoscenza è costituita da *embedding vectors*, ovvero rappresentazioni numeriche multidimensionali dei testi.

**Segue la fase di *augmentation***, in cui le informazioni recuperate vengono integrate nel prompt originale, arricchendo il contesto fornito al modello.

**Infine, nella fase di *generation***, l'LLM utilizza il contesto aumentato per produrre una risposta più precisa e coerente rispetto alla propria conoscenza interna statica.

Questo approccio permette di accedere a informazioni aggiornate, riduce il rischio di allucinazioni grazie all'ancoraggio a fonti verificabili e consente di impiegare

modelli generici in domini altamente specifici senza ricorrere a costosi processi di fine-tuning.

In sistemi come **NBMultiRag** (https://github.com/nannib/nbmultirag), anche i documenti non testuali vengono trasformati in testo; I file audio vengono trascritti tramite modelli come Whisper; immagini e video sono descritti mediante librerie di computer vision e OCR e qualsiasi altro formato, attraverso tecniche dedicate, viene convertito in testo e successivamente in embedding, entrando così a far parte della base di conoscenza interrogabile.

# Fine-tuning: definizione e implicazioni

Il *fine-tuning* è il processo di addestramento specialistico di un modello di Intelligenza Artificiale già pre-addestrato su un insieme di dati più piccolo e specifico.

Si parte da un modello dotato di una conoscenza generale, che viene affinato esponendolo ad esempi mirati per un compito particolare, con l'obiettivo di adattarne le capacità a un dominio specifico senza doverlo riaddestrare da zero.

Il fine-tuning consente una specializzazione molto spinta del modello, permettendogli di apprendere terminologie tecniche, formati specifici e stili di comunicazione difficilmente ottenibili con il solo prompt engineering, infine una volta addestrato, il modello richiede prompt più brevi, con un miglioramento dell'efficienza in fase di inferenza, e beneficia del *transfer learning* anche quando il dataset disponibile è relativamente limitato.

Accanto a questi benefici emergono limiti rilevanti, come i costi di addestramento che sono elevati, sia in termini di risorse computazionali sia di tempo umano necessario alla preparazione dei dati, inoltre la conoscenza del modello

rimane statica al momento del fine-tuning e può diventare rapidamente obsoleta in contesti dinamici.

Per questi motivi, esiste inoltre il rischio di *oblio catastrofico*, con perdita di capacità generali, e una complessità tecnica che aumenta la probabilità di overfitting.

# Perché adottare una Private AI in Digital Forensics

Le motivazioni che possono spingere investigatori, laboratori forensi e SOC ad adottare una Private AI sono molteplici.

Le evidenze digitali – immagini forensi, dump di memoria, log, e-mail, cronologie di navigazione – contengono spesso dati personali, segreti industriali e informazioni altamente sensibili; quindi, l'uso di una Private AI garantisce che tali dati non escano mai dal perimetro investigativo.

Dal punto di vista legale e reputazionale, l'uso improprio di servizi cloud per l'analisi di dati forensi può esporre professionisti e organizzazioni a sanzioni e contestazioni.

Anche quando i servizi cloud offrono garanzie di riservatezza, rimane valido il principio secondo cui il cloud è sempre "il computer di qualcun altro". Ridurre al minimo la circolazione di dati sensibili resta una pratica prudenziale.

È fondamentale chiarire che la Private AI deve essere intesa come strumento di supporto cognitivo e non come sostituto dell'analista; infatti, le risposte fornite dal sistema non devono essere accettate acriticamente, ma sempre verificate manualmente o tramite strumenti forensi tradizionali, insomma l'ultima parola deve rimanere umana, per evitare inesattezze o allucinazioni.

# Requisiti hardware minimi

Contrariamente a una percezione diffusa, una Private AI non richiede infrastrutture proibitive.

- **Configurazione minima di laboratorio**

- **CPU**: 8–16 core moderni;
- **RAM**: 32–64 GB;
- **Storage**: SSD NVMe da almeno 1–2 TB;
- **GPU** (opzionale ma consigliata):
    - 8–12 GB di VRAM per modelli 7B quantizzati;
    - 16–24 GB di VRAM per modelli 13B.

- **Configurazioni avanzate**

Per laboratori strutturati:

- GPU multiple (A4000/A5000, RTX 4090 o equivalenti);
- RAM superiore a 128 GB;
- separazione tra nodo di indicizzazione e nodo di inferenza.

L’uso della quantizzazione (4-bit o 8-bit) rende praticabile l’impiego di LLM locali anche su workstation forensic, infatti riducendo la precisione numerica dei pesi si abbassano i requisiti di memoria e i tempi di inferenza, con una possibile perdita di accuratezza accettabile nella maggioranza dei casi.

- **Impatto della quantizzazione sulla memoria su un modello da 20 miliardi di parametri**

| Formato | Bit per parametro | Modello 20B parametri | RAM richiesta (≈) |
|---|---|---|---|
| FP32 | 32 bit (4 byte) | 20B × 4 byte | 80 GB |
| FP16 / BF16 | 16 bit (2 byte) | 20B × 2 byte | 40 GB |
| INT8 | 8 bit (1 byte) | 20B × 1 byte | 20 GB |

| | | | |
|---|---|---|---|
| INT4 | 4 bit (0,5 byte) | 20B × 0,5 byte | 10 GB |

- **Parametri di configurazione rilevanti**

Nell'uso di framework LLM locali con RAG, alcuni parametri di campionamento sono particolarmente rilevanti:

- **Temperatura**: valori bassi (0.1–0.2) rendono il modello più deterministico e conservativo; valori elevati aumentano la varietà e creatività delle risposte ma anche il rischio di errori.
- **Top-K**: limita la scelta alle K parole più probabili, eliminando opzioni a bassa probabilità spesso responsabili di risposte incoerenti.

Per ridurre le allucinazioni si adottano temperature basse e valori di Top-K contenuti, vincolando il modello a risposte statisticamente più solide.

Infine, è importante anche scegliere anche un modello di "incorporamento" (embedding model) che serve a generare incorporamenti vettoriali: lunghi vettori di numeri che rappresentano il significato semantico di una data sequenza di testo.

# Schemi architetturali di riferimento

- **Architettura A – Stand-alone su workstation forense**

- storage locale delle evidenze;
- pre-processing e normalizzazione del dataset;
- motore di embedding locale (embedding model);
- LLM locale;
- interfaccia CLI o web isolata.

Ideale per consulenti individuali, indagini ad altissima sensibilità e ambienti air-gapped, offrendo massimo controllo,

con scalabilità limitata.

- **Architettura B – Modulare on-premise**

- repository centrale delle evidenze;
- nodo di indicizzazione RAG;
- nodo di inferenza LLM con GPU;
- sistema di logging e audit;
- interfacce multi-utente.

Adatta a laboratori forensi strutturati, CERT e SOC, questa garantisce scalabilità e tracciabilità, a fronte di maggiore complessità.

- **Architettura C – Cloud privato o sovrano**

- infrastruttura virtualizzata dedicata;
- data residency garantita;
- servizi di indicizzazione e inferenza isolati;
- accesso autenticato e auditabile.

Combina scalabilità e sovranità del dato, richiedendo governance e competenze cloud mature.

# Framework e strumenti

- **LLM locali**

- LLaMA e derivati (Mistral, Mixtral);
- Qwen;
- Phi;
- GPT-OSS.

Gestibili tramite Ollama, llama.cpp e vLLM.

- **Alcuni framework RAG**
- LangChain;
- LlamaIndex;
- Haystack;
- GPT4All;

- AnythingLLM;
- Cheshire Cat AI;
- Nuvolaris.
- **Il mio personale contributo**
- **nbmultirag**: framework leggero e flessibile per pipeline RAG multi-sorgente in contesti riguardanti fonti di dati di diversa tipologia.

| | |
|---|---|
| FILE PER IL RAG | |

# Parlare con le evidenze: esempio operativo

Scenario: un computer aziendale è sospettato di esfiltrazione dat e dall'analisi forense vengono estratti:

- file EML di posta elettronica;
- cronologia web;
- log di sistema;
- documenti di lavoro.

La pipeline Private AI + RAG prevede:

- normalizzazione delle evidenze in formato testuale;
- indicizzazione tramite embedding locali;
- interrogazione con LLM locale;
- logging completo di query e risposte.

Esempi di query:

- "Esistono correlazioni temporali tra accesso a servizi cloud e invio di e-mail esterne?"
- "Quali documenti sono stati aperti nei 30 minuti precedenti a un upload web?"
- "Riassumi le comunicazioni e-mail con allegati sospetti."

Il risultato non costituisce una prova automatica, ma uno strumento di triage cognitivo.

# LLM + RAG, fine-tuning o prompt engineering?

- **LLM + RAG**: approccio consigliato, flessibile e senza modifica del modello.
- **Fine-tuning + RAG**: utile solo in contesti altamente strutturati, costoso e complesso.
- **System prompt + RAG**: spesso sufficiente come compromesso iniziale.



# Limiti, rischi e anti-pattern

L’adozione della Private AI non elimina i rischi intrinseci associati all’uso dell’Intelligenza Artificiale, specialmente in ambito forense.

Anche in presenza di sistemi RAG, un LLM può produrre risposte non pienamente supportate dalle evidenze o interpretare in modo ambiguo artefatti complessi, inoltre, i modelli locali, possono presentare limiti di precisione rispetto a soluzioni cloud di grandi dimensioni.

Dal punto di vista operativo, l’assenza di logging strutturato, il mancato versioning di modelli e dataset, la contaminazione involontaria tra casi investigativi e configurazioni non documentate rappresentano rischi concreti per la validità e la difendibilità delle analisi.

Tra gli anti-pattern più ricorrenti rientrano l’utilizzo delle risposte dell’AI come evidenza probatoria diretta, l’automazione delle conclusioni investigative, l’introduzione di servizi cloud esterni per comodità durante fasi critiche e

l’applicazione di fine-tuning su dati sensibili senza adeguate misure di controllo.

Il principio guida rimane invariato: la Private AI deve essere trattata come uno strumento di supporto cognitivo e non come un sostituto del metodo forense.

# Checklist di adozione della Private AI

- **Governance e metodo**

- definizione chiara dello scopo;
- integrazione nel metodo forense esistente;
- dichiarazione dei limiti dell’AI nei report.

- **Dati ed evidenze**

- evidenze conservate in ambienti controllati;
- separazione rigorosa dei dataset per caso;
- documentazione delle trasformazioni;
- assenza di trasmissione verso servizi esterni.

- **Modelli e configurazione**

- scelta di LLM locali adeguati;
- versioning di modelli ed embedding;
- preferenza per RAG rispetto al fine-tuning;
- documentazione dei system prompt.

- **Infrastruttura**

- hardware dedicato;
- isolamento dell’ambiente di inferenza;
- logging e conservazione dei log;
- backup a fini probatori.

- **Aspetti legali e revisione umana**
- conformità normativa (GDPR, DPIA);
- revisione critica delle risposte dell’AI;

- verifica manuale delle correlazioni;
- riproducibilità e difendibilità in sede giudiziaria.

# Conclusioni

La Private AI rappresenta un’evoluzione naturale degli strumenti di supporto alla Digital Forensics e non sostituisce il metodo forense, ma lo potenzia, permettendo di interagire in modo semantico con grandi volumi di evidenze senza compromettere controllo, riservatezza e validità legale.

In un contesto normativo e tecnologico sempre più complesso, la scelta tra AI in cloud e AI locale non è solo tecnica, ma anche etica e professionale.

Per la Digital Forensics la direzione è sempre più chiara: l’AI deve stare dove stanno le evidenze.

Profilo Autore

Dott. Giovanni Bassetti, laureato in Scienze dell’Informazione, iscritto all’Albo dei Consulenti Tecnici presso il Tribunale Civile di Bari al n. 17, Project Manager della Live Distro internazionale CAINE per le indagini informatiche forensi, nonché fondatore della comunità italiana online dedicata alle investigazioni digitali (Computer Forensics Italy), autore di numerosi articoli pubblicati su riviste del settore nazionali ed internazionali. È stato consulente per alcuni procedimenti penali di rilevanza nazionale, è autore di un testo sulla computer forensics ed è segretario e membro fondatore di ONIF (Osservatorio Nazionale Informatica Forense).
Inoltre, ha partecipato come relatore e docente a numerosi convegni e corsi nazionali ed internazionali dedicati all’informatica forense ed intelligenza artificiale.

Condividi sui Social Network:
https://www.ictsecuritymagazine.com/articoli/private-ai-digita

l-forensics/