

t-defence

Retrieval Augmented Generation (RAG) per la consultazione locale e sicura di documenti:

standard di sicurezza per la software supply chain

#TDefenceBusiness

AI4Cyber

Summary

Abstract 04

Introduzione 05

Un caso d’uso dello Scoring Module 06

Funzionamento 07

Estrazione delle restrizioni 08

Associazione alle tecniche MITRE ATT&CK 09

Modelli utilizzati 10

Sperimentazione 11

Validazione 11

Risultati 12

Sviluppi futuri 13

Bibliografia 14

AI4Cyber

AI4Cyber studio è il nuovo spazio di ricerca applicata di T-Defence, dedicato alla ricerca e applicazione dell'**Intelligenza Artificiale in ambito cybersecurity**.

Al suo interno, diversi specialisti con competenze ed esperienze eterogenee collaborano per affrontare le sfide poste dal panorama delle minacce informatiche, accompagnare i clienti nell'adozione dell'AI nei propri processi aziendali e condividere le attività di ricerca con la comunità tecnico-scientifica.

Il team cura l'**intero ciclo di sviluppo dei sistemi basati su AI**: dalla raccolta e dal preprocessing dei dati, all'addestramento e validazione dei modelli, fino al deployment in produzione.

Le soluzioni proposte si fondano su tecnologie avanzate di **Machine Learning**, **Deep Learning** e **Large Language Models**, e possono essere personalizzate in base alle specifiche esigenze del cliente.

L'AI Team è costantemente impegnato in attività di ricerca e sperimentazione, con un'attenzione particolare agli **aspetti etici**, alla **trasparenza** e alla **tutela della privacy**.

L'obiettivo è proporre soluzioni innovative che siano in linea con i principi di **equità**, **inclusività** e **rispetto dei diritti fondamentali**.

Abstract

La sicurezza della software supply chain rappresenta oggi una delle principali sfide per le organizzazioni che adottano componenti software di terze parti all'interno dei propri processi. In questo contesto, gli standard di sicurezza internazionali e nazionali, come le norme della famiglia ISO/IEC 27000, le pubblicazioni del NIST o le linee guida AgID, costituiscono il riferimento principale per definire i requisiti che un software deve rispettare. Tuttavia, questi documenti sono redatti in linguaggio naturale e risultano difficilmente traducibili in controlli verificabili in modo automatico.

Il presente studio propone un nuovo caso d'uso dello Scoring Module del sistema già descritto in [3] e denominato "Security Standard". Il modulo implementa un agente AI in grado di analizzare un documento normativo, estrarne le restrizioni di sicurezza applicabili al software e associare ciascuna di esse alle tattiche, tecniche e procedure (TTP) della matrice MITRE ATT&CK Enterprise [5]. L'agente produce in output un file CSV che può essere utilizzato da un tool esterno il quale, prendendo in input un software e il CSV relativo alle TTP del documento, calcola uno score riguardo la sicurezza del software in analisi.

Il modulo si integra con il sistema descritto in [1] [2] [4], operando sui documenti che esso è già in grado di elaborare. L'intera operazione si svolge in due fasi: una di inferenza, affidata a un LLM eseguito in locale e guidato da un prompt altamente tecnico, e una di matching, che combina classificazione, riformulazione dal punto di vista dell'attaccante e scoring di similarità semantica con soglie calibrate sperimentalmente.

La validazione, condotta su tre standard di dimensioni crescenti (ISO/IEC 27001:2022, ISO/IEC 27002:2022 e NIST SP 800-53 Rev. 5), ha mostrato un'accuratezza compresa tra l'85% e il 97%, confermando la validità dell'approccio anche su documenti di quasi 500 pagine.

Autori:

- Mattia Marino: AI Security Engineer
- Simona Sorgente: AI Officer

Introduzione

Negli ultimi anni gli attacchi rivolti alla software supply chain sono cresciuti sia in numero che in sofisticazione. Compromettere un componente software, una libreria o un aggiornamento consente infatti a un attaccante di raggiungere contemporaneamente un elevato numero di organizzazioni. Per questo motivo, valutare il comportamento di un software prima di introdurlo nel proprio perimetro è diventata un'attività centrale nei processi di gestione del rischio.

Gli standard di sicurezza già esistenti forniscono il quadro di riferimento per questa valutazione, ma tali documenti sono pensati per essere letti da esseri umani e non sono direttamente machine-readable. I requisiti software sono espressi in forma discorsiva, spesso mescolati a obblighi organizzativi, e risultano difficili da confrontare con il comportamento osservato di un eseguibile.

Il problema assume quindi due dimensioni. Da un lato è necessario estrarre dai documenti normativi le sole restrizioni che riguardano il software, riformulandole in modo atomico, preciso e privo di ambiguità. Dall'altro è necessario collegare ciascuna restrizione a una rappresentazione condivisa delle azioni malevole, così che la violazione di un requisito possa essere riconosciuta in modo oggettivo.

Svolgere manualmente queste attività su standard composti da centinaia di pagine richiede settimane di lavoro specialistico e deve essere ripetuto a ogni nuova versione del documento. Il caso d'uso "Security Standard" nasce per automatizzare questo processo, affiancando l'analista con un agente AI che opera interamente in locale.

Il risultato prodotto dal modulo è un file CSV contenente le restrizioni estratte e le TTP associate. Tale file può poi essere caricato in una piattaforma terza dedicata all'analisi della sicurezza della software supply chain al fine di analizzare se il software analizzato soddisfa le restrizioni o meno.

Un caso d'uso dello Scoring Module

Il caso "Security Standard" non introduce un nuovo paradigma all'interno del sistema, ma rappresenta un'ulteriore applicazione dello Scoring Module presentato in [3]. Come nel caso d'uso dedicato alla selezione del personale, anche qui il sistema si basa sull'impiego degli LLM come strumenti di strutturazione semantica, trasformando un documento eterogeneo in un insieme di elementi formalmente omogenei, e su un meccanismo di scoring che valuta la corrispondenza tra questi elementi e un riferimento esterno.

La differenza risiede nella natura degli oggetti confrontati. Nel caso precedentemente menzionato riguardante la selezione del personale, il documento viene confrontato con una richiesta formulata dall'utente in linguaggio naturale. Nel caso d'uso "Security Standard", invece, la richiesta consiste nell' estrazione delle restrizioni dal documento in analisi seguita dal matching con le TTP della matrice MITRE ATT&CK, mentre le TTP associate alle restrizioni costituiscono la base con cui calcolare uno scoring di sicurezza di un dato software confrontandole con le vulnerabilità estratte da esso. Il calcolo effettivo dello score viene poi effettuato da una piattaforma terza che utilizza il CSV come input.

Queste scelte progettuali hanno permesso di riutilizzare gran parte dei componenti già consolidati, come l'accesso al database centralizzato e la generazione degli embedding, concentrando lo sviluppo sulle sole parti specifiche del dominio, ovvero la formulazione del prompt di estrazione e le strategie di filtraggio dei risultati. Il prompt di estrazione è strutturato in modo tale che il modello agisca come un analista di sicurezza, imponendo delle regole da seguire pedissequamente e fornendo esempi di riferimento per le restrizioni da formulare. Le strategie di filtraggio, invece, consistono nel classificare le restrizioni estratte o come tecniche o di governance, in modo da considerare solo quelle riguardanti gli aspetti tecnici del software. Successivamente, le restrizioni vengono trasformate dal punto di vista dell'attaccante, in modo tale da avvicinarsi ulteriormente allo stile MITRE ATT&CK. Infine, si procede con il matching effettivo tra le restrizioni estratte e le TTP MITRE ATT&CK, completando così la pipeline.

Funzionamento

Il modulo “Security Standard” è stato realizzato come un servizio autonomo, affiancato a quelli già presenti nel sistema. Il servizio non replica la logica esistente del sistema, ma la riutilizza secondo due modalità complementari:

- opera sui documenti già caricati ed elaborati;
- condivide i moduli applicativi per l’accesso al database centralizzato, la lettura delle sezioni dei documenti, la generazione degli embedding e il salvataggio dei file prodotti.

In questo modo il nuovo modulo opera sugli stessi dati e con gli stessi meccanismi impiegati dal resto del sistema, garantendo coerenza nella gestione delle informazioni e semplificando la manutenzione. I file CSV generati vengono salvati nel database centralizzato, all’interno dello spazio dedicato all’utente. Volendo poi approfondire il funzionamento del modulo, esso si articola in due fasi in cascata, con la seconda fase che riceve in input il risultato della prima:

- **Inferenza**, in cui ogni sezione del documento viene sottoposta a un LLM incaricato di estrarre le restrizioni di sicurezza applicabili al software;
- **Matching**, in cui le restrizioni estratte vengono classificate, riformulate dal punto di vista dell’attaccante e associate alle tecniche MITRE ATT&CK tramite similarità semantica.

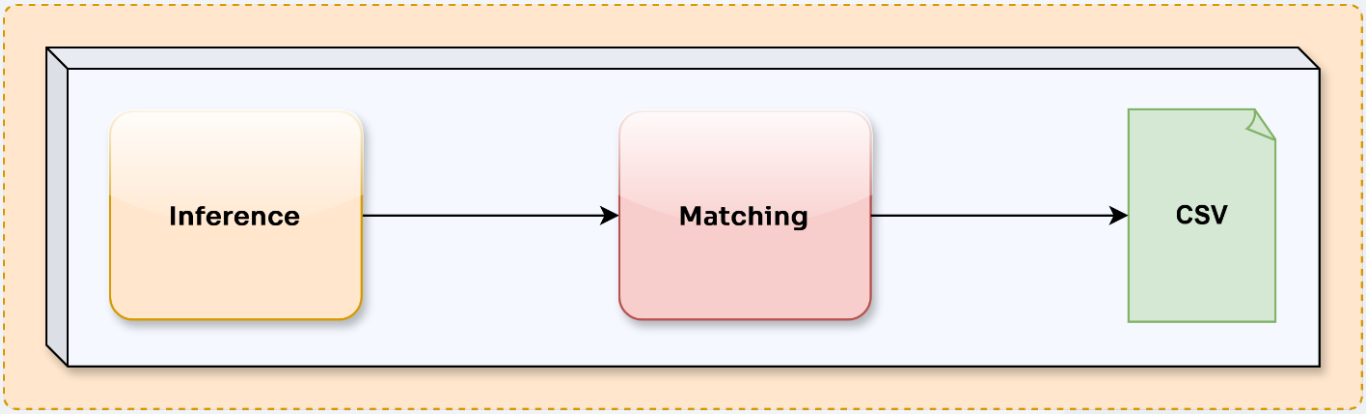


Figura 2: Le fasi della pipeline del modulo “Security Standard”.

Al termine del processo viene generato un file CSV composto dalle colonne ID, Name e List of techniques, che contengono rispettivamente l’identificativo progressivo della restrizione, il suo testo e l’elenco delle tecniche associate. Nel file in output vengono riportate solo le restrizioni a cui è stata associata almeno una tecnica.

Nella tabella sottostante viene riportato un esempio parziale dell'output del modulo.

ID	Name	List of techniques
1	The software must not use untrustworthy hardware, firmware, or software development and acquisition processes.	T1495, T1554, T1195.001, T1505, T1542.002
2	The software must not bypass security and privacy engineering principles during component integration.	T1505, T1671
3	The software must not operate without continuous monitoring of its security and privacy controls.	T1553, T1176, T1653, T1518.001
4	The software must not function without documented security and privacy practices integrated into operational processes.	T1505, T1176, T1027.002, T1671, T1072
5	The software must not allow unauthorized access to information or information systems.	T1187, T1001.003, T1553, T1562.001, T1027.002
6	The software must not allow unauthorized use of information or information systems.	T1001.003, T1553, T1187, T1027.002, T1014
7	The software must not allow unauthorized disclosure of information or information systems.	T1027.002, T1001.003, T1027, T1553, T1027.010
8	The software must not allow unauthorized disruption of information or information systems.	T1490, T1553, T1562.001, T1001.003, T1562.012
9	The software must not allow unauthorized modification of information or information systems.	T1562.001, T1001.003, T1553, T1562.012, T1027.002

Estrazione delle restrizioni

Nella fase di inferenza il sistema recupera l'elenco delle sezioni del documento e il relativo testo. Ogni sezione viene sottoposta singolarmente al modello linguistico, accompagnata dal suo titolo e dal nome del documento, così da fornire il contesto necessario all'interpretazione dell'estratto per permettere al modello di lavorare su un contesto coerente e circoscritto.

Il modello è istruito ad agire come un analista di sicurezza, con il compito di individuare tutte le restrizioni applicabili al software contenute nell'estratto e di riformularle come divieti brevi, atomici e privi di ambiguità, richiamando lo stile dei titoli delle tecniche MITRE ATT&CK. Ogni restrizione esprime un solo divieto, in modo concreto e tecnico, escludendo le clausole di natura puramente organizzativa. La configurazione del modello inoltre privilegia il determinismo, così da rendere i risultati riproducibili tra esecuzioni successive.

Le risposte del modello vengono infine sottoposte a una fase di post-elaborazione, che considera solo le restrizioni conformi al formato atteso, scarta quelle non sufficientemente informative e rimuove i duplicati raccolti dalle diverse sezioni, mantenendo l'ordine di comparsa nel documento.

Associazione alle tecniche MITRE ATT&CK

La fase di matching ha lo scopo di associare ciascuna restrizione alle tecniche della matrice MITRE ATT&CK Enterprise [5]. Il processo di matching di ogni restrizione si articola nelle tre sotto-fasi di seguito riportate.

1. Classificazione. Non tutte le restrizioni estratte possono essere associate a una tecnica avversaria. Requisiti relativi alla completezza della documentazione, alla revisione periodica delle policy o alla separazione dei ruoli descrivono infatti obblighi di governance, che non trovano corrispondenza in un comportamento tecnico osservabile. Per questo motivo ogni restrizione viene classificata dal modello o come tecnica o di governance, e queste ultime vengono escluse dalle fasi successive. In caso di esito incerto, la restrizione viene considerata tecnica al fine di limitare la perdita di informazione.

2. Riformulazione dal punto di vista dell'attaccante. Le restrizioni sono scritte dal punto di vista del difensore, mentre le descrizioni MITRE ATT&CK descrivono le azioni dell'attaccante. Dato che confrontare i due testi porterebbe a valori di similarità poco significativi, ogni restrizione tecnica viene riformulata dal modello come descrizione della tecnica avversaria adottando la terminologia propria di MITRE ATT&CK.

3. Calcolo dello score di similarità. Il testo riformulato viene trasformato in embedding e confrontato con i corrispondenti embedding di tutte le tecniche tramite cosine similarity. Le tecniche candidate vengono quindi selezionate applicando in cascata più criteri:

- Confronto dello score rispetto a una soglia assoluta, al di sotto della quale una tecnica viene considerata rumore;
- Valutazione dello score rispetto a una soglia relativa, per cui vengono mantenute solo le tecniche il cui score si discosta di poco da quello della tecnica migliore;

- Imposizione di un limite massimo al numero di tecniche associabili a ciascuna restrizione.

I valori delle soglie sono stati fissati a seguito di un'approfondita valutazione sperimentale, con l'obiettivo di associare a ogni restrizione soltanto le tecniche realmente pertinenti.

Durante la sperimentazione è inoltre emerso che alcune tecniche tendono ad associarsi a un numero eccessivo di restrizioni, a causa della natura generica della loro descrizione. Per contrastare questo fenomeno sono stati introdotti due filtri aggiuntivi: il primo, statico, applica un criterio di selezione più severo a un insieme ristretto di tecniche individuate empiricamente; il secondo, dinamico, opera sull'intero documento e, quando una tecnica compare in una quota eccessiva delle restrizioni tecniche, la rimuove da tutte quelle in cui non risulta tra le associazioni più rilevanti.

Modelli utilizzati

Coerentemente con l'impostazione dell'intero sistema, tutti i modelli impiegati vengono eseguiti in locale, senza che alcun dato lasci l'infrastruttura dell'utente. Nella configurazione adottata per questo studio sono stati utilizzati i seguenti modelli:

- Gemma 4 26B [6]: modello linguistico eseguito tramite OLLAMA [7], utilizzato sia per l'estrazione delle restrizioni sia per la classificazione e la riformulazione nella fase di matching, scelto per la capacità di gestire contesti estesi e di rispettare istruzioni di formato rigorose;
- mxbai-embed-large [8]: modello di embedding, lo stesso impiegato dal resto del sistema, utilizzato per il confronto semantico tra le restrizioni riformulate e le tecniche MITRE ATT&CK.

Sperimentazione

Al fine di validare il caso d'uso "Security Standard" è stata condotta un'approfondita sperimentazione, utilizzando standard di sicurezza differenti per struttura, finalità e dimensione:

- ISO/IEC 27001:2022 [9], standard internazionale per i sistemi di gestione della sicurezza delle informazioni;
- ISO/IEC 27002:2022 [10], raccolta di controlli di sicurezza con relative linee guida di implementazione;
- NIST SP 800-53 Rev. 5 [11], catalogo di controlli di sicurezza e privacy per sistemi informativi e organizzazioni.

La scelta di questi documenti consente di valutare il comportamento del sistema su testi che spaziano da poche decine a circa cinquecento pagine, e che alternano in misura diversa requisiti tecnici e obblighi di natura organizzativa.

La sperimentazione è stata eseguita su un VPS dotato di GPU NVIDIA RTX A6000 con 48 GiB di VRAM, processore Intel Xeon Icelake e 32 GB di RAM.

Validazione

In una prima fase si è tentato di utilizzare un modello di frontiera come oracolo, impiegando Claude Opus 4.7 [12] per verificare la correttezza delle restrizioni estratte e delle tecniche associate. Sono state fornite le stesse istruzioni e lo stesso documento sia al sistema descritto che a Claude Opus 4.7, chiedendo in seguito di associare le restrizioni generate alla matrice MITRE ATT&CK allegata.

Sebbene i primi risultati prodotti dall'oracolo siano apparsi positivi, proseguendo la sperimentazione essi si sono dimostrati poco affidabili, sia per l'impossibilità di controllare la variabilità delle risposte, che rendeva i risultati non riproducibili, sia per l'imprecisione dell'associazione con le TTP. Si è quindi scelto di affidare la validazione a un analista umano, che ha esaminato le restrizioni prodotte dal sistema e ha calcolato le metriche di interesse.

Risultati

Per ciascun documento sono state effettuate più esecuzioni, che hanno fornito risultati coerenti tra loro e in linea con quanto atteso dal sistema. La tabella seguente riporta, per ciascun documento, il numero di pagine, il numero di restrizioni estratte, l'accuratezza rilevata dall'analista e il tempo complessivo di elaborazione.

Documento	Pagine	N° restrizioni	Accuratezza	Tempo di esecuzione
ISO/IEC 27001:2022	26	51	85%	6 min.
ISO/IEC 27002:2022	164	838	93%	56 min.
NIST SP 800-53 Rev. 5	492	1389	97%	2 h.

I risultati mostrano un'accuratezza elevata su tutti i documenti analizzati, con valori che crescono all'aumentare della natura tecnica del testo. Il valore più basso, registrato su ISO/IEC 27001:2022, è attribuibile alla diversa struttura del documento; si tratta infatti di uno standard incentrato sui requisiti del sistema di gestione, in cui gli obblighi organizzativi prevalgono su quelli tecnici e l'estrazione di restrizioni applicabili al software risulta più difficile. Al contrario, NIST SP 800-53 descrive in modo puntuale controlli tecnici ben delimitati, che si prestano naturalmente alla riformulazione in forma di divieto.

Sviluppi futuri

La sperimentazione condotta ha mostrato come il caso d'uso "Security Standard" rappresenti una soluzione efficace per migliorare la sicurezza della software supply chain, offrendo agli utenti uno strumento in grado di trasformare documenti normativi complessi in standard di sicurezza personalizzati e verificabili. L'implementazione e l'utilizzo del sistema basato su AI si è rivelato determinante per ottenere risultati accurati in tempi contenuti.

Un primo sviluppo futuro riguarda sicuramente la generalizzazione del caso d'uso, così da consentire all'utente di estrarre requisiti da qualsiasi tipologia di documento, superando il vincolo degli standard di sicurezza formalmente strutturati.

In questa direzione si colloca anche l'estensione del sistema a linee guida, best practice e fonti analoghe, oltre agli standard ISO, NIST e AgID, che spesso contengono indicazioni operative di grande valore ma espresse in forma meno rigorosa.

Un ulteriore filone riguarda la fase di matching, nella quale si intende sperimentare metodi alternativi alla sola similarità semantica, ad esempio l'impiego di uno Small Language Model (SLM) addestrato o istruito specificamente per l'associazione tra restrizioni e tecniche avversarie.

Si prevede inoltre di estendere il sistema all'estrazione di restrizioni di sicurezza a partire da descrizioni dettagliate di attacchi reali, come report di threat intelligence o analisi di incidenti, invertendo di fatto il flusso attuale e consentendo di derivare nuovi requisiti direttamente dalle minacce osservate.

Infine, l'introduzione di meccanismi di feedback umano (human-in-the-loop) permetterebbe di affinare progressivamente le istruzioni di estrazione e i criteri di matching sulla base delle correzioni dell'analista, rendendo il sistema adattivo.

Bibliografia

[1] Retrieval-Augmented-Generation (RAG) per la consultazione locale e sicura di documenti: https://t-defence.it/wp-content/uploads/2025/07/Report_AI4Cyber.pdf

[2] Retrieval-Augmented-Generation (RAG) per la consultazione locale e sicura di documenti: miglioramenti e nuove funzionalità: https://t-defence.it/wp-content/uploads/2025/08/Report_AI4Cyber.pdf

[3] Retrieval-Augmented-Generation (RAG) per la consultazione locale e sicura di documenti: una pipeline per la standardizzazione dei dati e lo scoring semantico: https://t-defence.it/wp-content/uploads/2026/01/260129-AI4Cyber-Retrieval-Augmented-Generation.pdf

[4] Retrieval-Augmented-Generation (RAG) per la consultazione locale e sicura di documenti: indicizzazione a grafo e retrieval ibrido per la modalità FAT RAG: https://t-defence.it/wp-content/uploads/2026/09/Report_Retrieval_Augmented_Generation.pdf

[5] MITRE ATT&CK Enterprise Matrix: https://attack.mitre.org/matrices/enterprise/

[6] Gemma 4 26B: https://ollama.com/library/gemma4:26b

[7] OLLAMA: https://ollama.com/

[8] mxbai-embed-large: https://ollama.com/library/mxbai-embed-large

[9] ISO/IEC 27001:2022: https://www.iso.org/standard/27001

[10] ISO/IEC 27002:2022: https://www.iso.org/standard/75652.html

[11] NIST SP 800-53 Rev. 5: https://csrc.nist.gov/pubs/sp/800/53/r5/upd1/final

[12] Claude: https://www.anthropic.com/claude



t-defence

Next | Donexit | Foramil | Innodesi

Via Giacomo Peroni, 452 - 00131 Roma
tel. 06.45752720 - info@defencetech.it
www.t-defence.it

#TDefenceBusiness