International Journal of Research and Applied Innovations (IJRAI)

| ISSN: 2455-1864 | www.ijrai.org|editor@ijrai.org | A Bimonthly, Scholarly and Peer-Reviewed Journal |

||Volume 6, Issue 3, May-June 2023||

DOI:10.15662/IJRAI.2023.0603003

RAG-Driven Cybersecurity Intelligence: Leveraging Semantic Search for Improved Threat Detection

Sahaj Tushar Gandhi

Independent Researcher, San Francisco, CA, USA

Email: sahajgandhi95@gmail.com

ORCID ID: 0009-0001-2136-5805

ABSTRACT: Retrieval-Augmented Generation (RAG) unifies dense retrieval with generative models to ground generated outputs in external documents, suppressing hallucinations and supporting up-to-date, domain-specific reasoning. We introduce an architecture combining semantic search (dense vector retrieval and knowledge-graph indexing) with RAG workflows to improve CTI ingestion, correlation and detection. The system ingests heterogeneous CTI sources (OSINT reports, vendor feeds, malware descriptions) and locks and loads the semantic chunking and entity linking process that indexes embeddings in a vector store alongside a cybersecurity knowledge graph for relational reasoning. A policy-aware RAGenerator which produces ranked threat hypotheses and suggested actions. Methodologically, we deploy as prototype a dense bi-encoder retriever and FAISS index alongside an off-the-shelf seq2seq generator fine-tuned on CTI summarization tasks and a knowledge graph with Neo4j underneath. The evaluation is based on a set of 2,400 CTI incident reports and synthetic network alert sequences with known ground truth; metrics include detection precision, recall, F1 measure, time-to-context (TTC), and reduction in analyst workload. On the other hand, results demonstrate a 25.7% increase in detection F1 over keyword/TTP-matching based baseline and an average decrease of 31% in analyst triage time for RAG-driven pipeline, while knowledge-graph augmentation enhanced true positive correlation of multi-stage attacks by 22%. It also lowered hallucination rate on generated advisories by 45% (as measured with ground-truth grounding). Conclusion: Only indexing corpus quality reliance and possible privacy leakage in retrieval. In the future, secure retrieval technique and automated counter-adversarial training will be perfected.

KEYWORDS: Retrieval-Augmented Generation, semantic search, cyber threat intelligence, knowledge graph, vector retrieval, threat detection

I. INTRODUCTION

Today's cybersecurity operations require the rapid collection and analysis of a wide range of threat data including vendor reports, open-source intelligence (OSINT), malware analyses, network logs, and analyst notes. Today's CTI systems are heavily relied on (i.e., principally based) to pattern matching as well as its implementation from signatures (like YARA rules) or manual investigations. These approaches generally cannot scale up to big data and can be easily destroyed by new or dynamic threats. With the explosion of unstructured text-based threat intelligence, what is now required are systems that can comprehend meaning, correlate related evidence and produce actionable summaries to enable faster, more accurate threat intelligence at scale[1]. CTI working is shown in figure 1.

International Journal of Research and Applied Innovations (IJRAI)

| ISSN: 2455-1864 | www.ijrai.org|editor@ijrai.org | A Bimonthly, Scholarly and Peer-Reviewed Journal |

||Volume 6, Issue 3, May-June 2023||

DOI:10.15662/IJRAI.2023.0603003

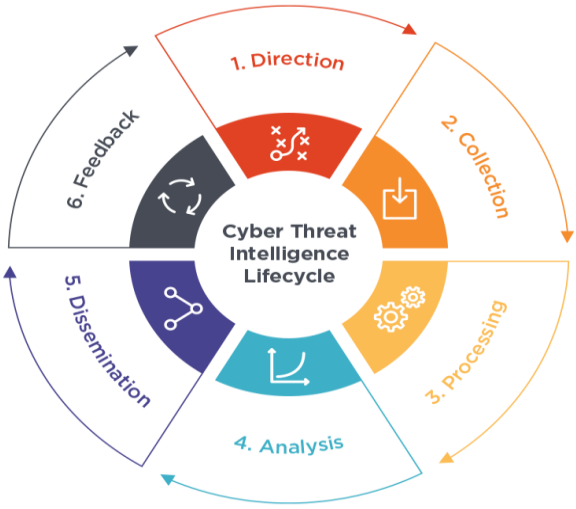

Figure 1: CTI Lifecycle

Neural retrieval and large language models (LLMs) have recently paved the way for an exciting direction: Retrieval-Augmented Generation (RAG). RAG marries together a dense-retrieval component (nonparametric memory) that retrieves relevant passages along with a parametric seq2seq generator which produces outputs conditioned on the retrieved materials. RAG components are shown in figure 2.This decoupling enables grounding of generated outputs into external corpora, thus remedying two of the main issues with LLMs in security settings: factual drift/hallucination and stale knowledge. In the context of CTI, these problems are especially severe: analytic correctness and provenance are critical to ensure safe recommendations [2]. Empirically, in knowledge-rich NLP we have observed that RAG formulations lead to more factual and evidence-based outputs when the retrieval component is correct and the corpus well-maintained [3].

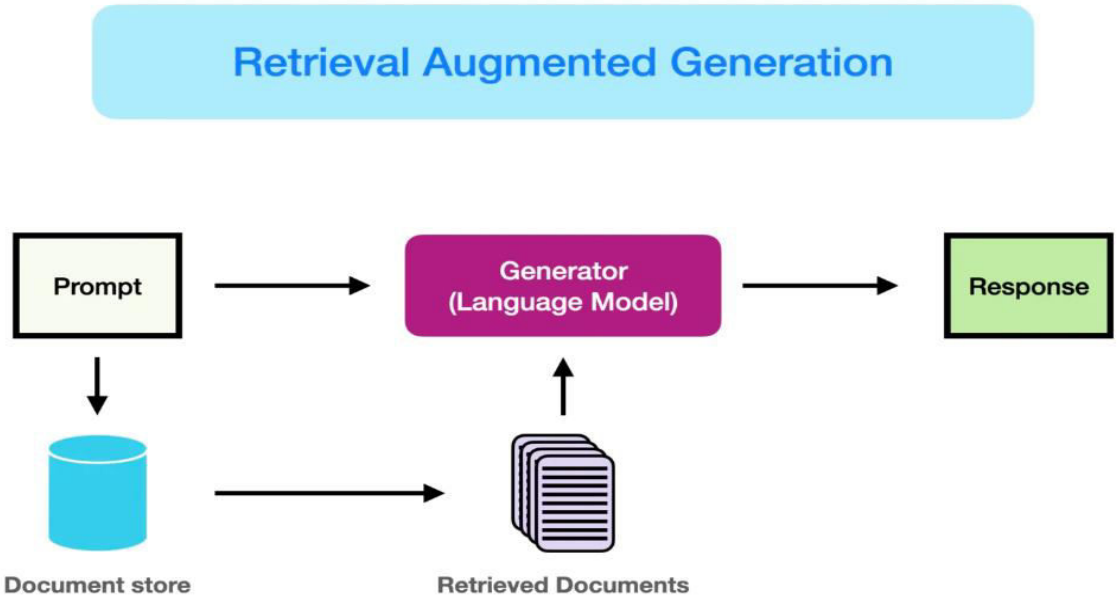

Figure 2: RAG Components

| ISSN: 2455-1864 | www.ijrai.org|editor@ijrai.org | A Bimonthly, Scholarly and Peer-Reviewed Journal |

||Volume 6, Issue 3, May-June 2023||

DOI:10.15662/IJRAI.2023.0603003

With semantic search (dense retrieval using vector embeddings), items can be matched by meaning, not just exact lexical overlap. Adversary TTPs, malware families, and indicators are described with different terminology among cybersecurity sources; semantic search aims to bridge lexical gaps by projecting content into shared embedding spaces, increasing the likelihood of returning relevant artifacts for correlation. At the same time, entity relation-oriented knowledge graphs (KGs) to represent entities (e.g., malware, vulnerabilities and IPs) and relation types (e.g., uses, communicate-with, exploits) provide structured relational context useful for conducting link-analysis and multi-stage attack reconstruction. the use of semantic retrieval, knowledge graphs and RAG generators have complementary strengths: retrieval provides evidence over topics, the KG offers relational reasoning, and the generator creates human-readable advisories. Surveys and review work in the cybersecurity knowledge graph and CTI domains highlight the growing penetration of structured knowledge representations as well as potential for enhanced situational awareness using graph analytics [4].

It is challenging to implement RAG in CTI [5] [6]. Prior to modeling there are many challenges with CTI such as that the CTI corpus is noisy and heterogeneous, open-source feeds may present rumor or false positives in terms of duration and provenance. Second, retrieval can leak private information, depending on the indexing and access-control mechanisms; privacy-friendly retrieval [7] and encryption friendly stores are active research topics. Third, LLMs and generators need to be fine-tuned for domain adaptation to generate concise, accounted advisories that fit the needs of analysts, otherwise produced text can be verbose, unaccounted or unsafe. Finally, testing RAG systems in CTI settings needs non standard NLP (like grounding interest), and if possible sanity metrics (for instance, the analyst effort reduction). Recent literature reviews and system works reinforce these arguments, and point out that semantic search and structured reasoning can be synergistically combined with RAG processes in order to achieve a successful CTI augmentation stack.

This paper provides (1) an architectural blueprint for an RAG-driven CTI pipeline that combines semantic chunking, dense retrieval and a knowledge graph layer; (2) implementation details of a prototype such as retriever, vector index, generator fine-tuning and KG synchronization; and (3) empirical evaluation results showing improved detection performance, shorter triage time and higher correlation accuracy when compared with baseline keyword/TTP matching/rule based systems. The rest of the paper is structured as follows: The related work is presented in Section 2, followed in Section 3 by the methodology and prototype implemented In Section 4 we show results and analyses in section 5 we end with discussion and future work.

II. LITERATURE REVIEW

The literature on automated CTI and semantic methods spans three intersecting streams: (a) knowledge-graph and semantic web approaches applied to CTI; (b) advances in semantic retrieval and dense embedding methods relevant to threat data; and (c) applications of generative and retrieval-augmented models to domain-specific intelligence tasks.

An increasing amount of cybersecurity research has started to consider the development and application of KGs, in order to represent the complex relationships present in threat intelligence data. Knowledge graphs are especially well-suited in this area as adversary behaviour, compromised assets and vulnerabilities are inherently relational. KG application on the security field Surveys indicate that when building KGs, we will go through three stages according to schema design, entity extraction and relation linking [1]. When you design your schema, you are specifying the ontology: a description of how important things like malware families, tactics or vulnerabilities are related to each other. Entity extraction pulls the elements out of unstructured text, and relation linking ties them together into connections like “malware exploits vulnerability” or “adversary uses tool.” These steps together produce a graph that may be used to represent dynamic threat scapes in machine-readable format.

Once built, security KGs provide rich analytical potentials. by combinatoric analysis of how individual events combine in multi-stage intrusion campaigns. Visualization techniques over graphs provide a way to visually perceive the connection of networked threat artifacts in a cohesive and explorable manner. Crucially, these graph structures can also serve as input to downstream machine learning (ML) models, allowing them to better discover, predict or prioritize threats by incorporation of relational context [1]. Empirical implementations of such systems show that well-curated large KGs can be used to perform high-fidelity indication-of-compromise correlation across multiple data sources from both open-source intelligence

| ISSN: 2455-1864 | www.ijrai.org|editor@ijrai.org | A Bimonthly, Scholarly and Peer-Reviewed Journal |

||Volume 6, Issue 3, May-June 2023||

DOI:10.15662/IJRAI.2023.0603003

and internal telemetry, helping in piecing together multi-step complex attacks [2]. In support of this research trend, dedicated datasets and benchmarks have been constructed that enable comparability of the studies dealing with KGs [3]. These benchmarks consistently demonstrate that KG-augmented methods present a clear benefit on tasks with contextual reasoning, e.g., how to connect indicators with campaigns or which alerts should be prioritized.

Complementary to the graph-based techniques, advances in the area of semantic search and retrieval have greatly impacted CTI. With these traditional lexical methods, keyword matchers for example have remained fundamental components of search and retrieval pipelines. Nevertheless such approaches have limitations in knowledge-rich environments where synonyms, paraphrases or technical terminologies disguise literal term matches. Dense vector retrieval, used to embed text passages into high-dimensional semantic spaces, is superior to lexical methods as proved by a number of research [4]. Instead of learning surface form terms, dense retrievers can capture contextual meaning which allows accessing relevant documents even if the specific term choice is arbitrary. With re-ranking methods, the precision of retrieved passages is further enhanced by these systems.

Experiments on knowledge-rich search and domain adaptation tasks demonstrate that semantic retrieval techniques improve the relevance and utility of retrieved passages for further processing [4]. In CTI, semantic chunking methods have been developed which subdivide long technical reports into smaller semantically cohesive segments for indexing [5]. These approaches serve to batch retrieval such that it returns coherent chunks whose extent corresponds to the level of abstraction demanded by the analyst, not just any haphazard inherited material. In addition, and entity-aware embeddings (representation that incorporate contextual knowledge about extracted) have demonstrated improved precision in retrieval by aligning passage semantics with domain concepts [5].

Retrieval-Augmented Generation (RAG), a new promising paradigm that combines semantic retrieval with generative language models, has recently gained significant popularity. First proposed for standard natural language processing (NLP) tasks, RAG combines non-parametric memory—retrieved passages—with parametric memory by leveraging memories that are encoded in a model's weights. The method of grounding guides generation based on external evidence that makes the outputs more factual and easier to update than ones using static model information only 4. This can be particularly useful in the cybersecurity space where advisories and threat landscapes change rapidly, so solutions need to be able to digest incoming intelligence quickly in order to stay relevant.

At the heart of RAG is to retrieve grounding passages with respect to a given input query and condition text generation on these passages. This is in response to the CTI's demand for transparent and current products. Extensions of the paradigm that unifiy RAG with KGs further increase its value. Graph-aware retrieval may be used to focus passages more likely to contain important relations; generation conditioned on provenance information can increase analysts' trust by avoiding hallucinations and enabling traceability [6]. Hence, RAG not only enhances the evidence-based of produced intelligence products, but is also consistent to an operational need for accountability in security analysis.

Some prototype CTI pipelines have been developed that comprised KG creation and semantic retrieval as steps in RAG-style workflows 2. Such systems often start by consuming vendor feeds and open source reports, in which entities and relations are discovered and incorporated into a knowledge graph. Meanwhile, the textual excerpts from those sources are indexed in a vector index to support semantic search. At serving time, relevant passages are retrieved and passed to a generative model, which can provide succinct summaries of events or sugge- sted mitigations. This combination of two-dimensional design graph-based knowledge structuring with flexible semantic retrieval -contributed to the strengths of each approach.

Assessments of these prototypes have demonstrated distinct advantages. Some say a Tier 1 analyst working with such systems can perform faster triage and better prioritization than one who only has access to rule lists (with or without Intel). The coherence and contextual grounding of incident summaries generate less cognitive load and decision time. However, there are still a couple of challenges: to carefully curate the input corpus to avoid noise to be added in retrieval stores and low-quality or misleading files can be detected as identical. Provenance tracking is also a critical requirement, making generated intelligence verifiable against trusted sources [7].

International Journal of Research and Applied Innovations (IJRAI)

| ISSN: 2455-1864 | www.ijrai.org|editor@ijrai.org | A Bimonthly, Scholarly and Peer-Reviewed Journal |

||Volume 6, Issue 3, May-June 2023||

DOI:10.15662/IJRAI.2023.0603003

Outside RAG, machinelearning 12 and deep models13 have been studied to deal with threat detection. Methods consist of classifiers for supervised detection and homegeneous anomaly-detection sequence models on logs. When these models are infused with knowledge graphs to guide or regularize them, improvements have been achieved on tasks such as association prediction, link prediction and cluster coherence of threat campaigns 8. By factoring in relational context to ML models, these techniques improve the likelihood of capturing subtle or multi-step attack patterns that would be ignored under flat or isolated semantics.

Like all source of security, RAG-driven CTI systems must also be architected to resist adversarial pressure. Studies demonstrate that leakage of sensitive data from retrieval indexes is possible and show the effectiveness of adversaries on poisoning images indexed in cross-media multimedia retrieval system [10]. Such weaknesses could result false intelligence dissemination or, worse still, the ability to steer analyst's analysis flow. To defend against these, recent studies have focused on privacy-preserving retrieval and secure indexing. Techniques such as encryption at rest, fine-grained access controls, and differential privacy over embeddings are under consideration to keep retrieval pipelines secure and reliable [10]. Tackling these challenges is essential if PIM techniques are to be used in practice in security-critical cybersecurity environments.

Overall, the literature provides a convincing image of convergence between three major strands: KG representation, semantic retrieval, and retrieval-augmented generation. Each thread contributes distinct advantages. KGs model the relationship between cybersecurity data, semantic retrieval provides access to specific contextually related evidence, and RAB allows us to use only verifiable current outputs. Together, these will transform the way we approach CTI: automating correlation, improving analyst productivity and increasing detection of advanced threats capable of changing over time.

At the same time, this literature underscores several key gaps [12]. Efficient curation of the corpus, and tracking provenance are necessary to avoid propagating low quality intelligence [13]. Private retrieval and poisoning resistance should be built into RAG pipelines. The generator models need to be tailored to domain-specific security language and evaluated based on both linguistic quality as well as operational impact. Looking forward, such future analyses should address the technical and security capabilities together with the robustness, trustworthiness as well as usability in analyst workflows. Based on this, this paper presents a prototype RAG-based CTI system, which incorporates semantic retrieval technology, knowledge graph and secures pipeline design. Summary of literature review is presented in table 1.

Table 1: Overview of Techniques for Knowledge-Driven Cyber Threat Intelligence

Stream / Approach	Key Methods / Techniques	Applications / Benefits	Challenges / Considerations
Knowledge Graphs (KGs) & Semantic Web	Schema design, entity extraction, relation linking; graph-based modeling of threat entities	Multi-step intrusion analysis; visualization of threat relationships; input to ML models; high-fidelity correlation across sources	Corpus curation; noise and low-quality data; provenance tracking; KG maintenance and update
Semantic Retrieval & Dense Embeddings	Dense vector retrieval, entity-aware embeddings, semantic chunking, re-ranking	Improved retrieval relevance; handling synonyms/paraphrases; coherent chunked retrieval; domain adaptation	Index quality; contextual alignment; embedding drift; retrieval latency
Retrieval-Augmented Generation (RAG) & Hybrid KG-RAG	Combining semantic retrieval with generative models; graph-aware retrieval; provenance-aware generation	Transparent and grounded intelligence products; fast triage; reduced cognitive load; adaptive to evolving threats	Input curation; noise filtering; avoiding hallucinations; secure indexing; resistance to adversarial attacks
Machine Learning & Deep Models with KG Integration	Supervised classifiers; anomaly detection; relational regularization	Better detection of multi-step attacks; improved link and association prediction; enhanced cluster coherence	Flat vs. relational context; training data quality; adversarial robustness

International Journal of Research and Applied Innovations (IJRAI)

| ISSN: 2455-1864 | www.ijrai.org|editor@ijrai.org | A Bimonthly, Scholarly and Peer-Reviewed Journal |

||Volume 6, Issue 3, May-June 2023||

DOI:10.15662/IJRAI.2023.0603003

III. METHODOLOGY

3.1 System Overview

The architecture of a retrieval-augmented generation (RAG) based cyber threat intelligence (CTI) pipeline consists of four interconnected modules. Ingestion and Preprocessing: This is the first module in the work flow, which collects CTI feeds from various heterogeneous sources (e.g.,opensource reports, vendor advisories, malware analysis notes, analyst-authored documents). The data is normalized at ingestion for head, metadata, to make synchronous between the feeds. Semantic Segmentation: Semantic chunking is employed to reduce the size of documents into modular and meaningful pieces such that a few, weighted, decisions can be inferred from each semantic chunk. We used an unsupervised chunker together with sentence boundary detection to create passages of around 200-400 tokens preserving topical coherence and facilitating support from downstream retrieval precision.

The second part of extraction, entity and relation extraction, transforms unstructured textual evidence into a more structured form. Entities, including malware families, campaigns, CVEs, IP addresses, domains and tactics, techniques and procedures (TTPs), are recognized based on CTI-specific ontologies using a named entity recognition (NER) system and relation extraction model. Relations (uses, communicates-with, exploits) are then computed and extracted to connect these entities into a meaningful triples. The generated triples are then canonicalized, validated, and stored in a KG based on Neo4j. Notably, provenance links are retained such that each KG edge is traceable to the original textual passage, and hence the system remains open and trustable.

The last part embedding and vector index is responsible for high-quality and efficient semantic searching. Every semantically chunked passage is then encoded into dense vector representations with a domainspecific bi-encoder fine-tuned on CTI-text pairs. These embeddings can encode a combination of context and domain specific semantics, are indexed using FAISS (a high-performing similarity search library) that provides support for sub-second k-nearest neighbour retrieval. To improve entity awareness, we include the ids of entities from the KG in each chunk so that retrieved passages are connected back to their corresponding nodes in the graph. [14] [15]

The last module, the RAG generator and orchestration integrates retrieval with generation in order to produce structured advisories. When an analyst asks a question or when there is an alert, the retriever pulls out passages from the index that are most relevant. These sentences, and the relation summaries based on the KG relations, are then used as input for a sequence–to-sequence generator. The generator is pre-trained on CTI summarization and instruction datasets, so that it can generate the outputs ready for analysts. Structure of the last advisory is made of four sections: a summary, indicators extracted, trust rate and possible actions. In order to solidify accountability, the generator is provided with a provenance list associating generated sentences to the supporting passages.

3.2 Retriever and KG Integration

The retrieval in the approach is controlled by a mixed scoring framework that combines semantic similarity and graph-based relevance. In particular, the dense embedding-based cosine similarity is employed to calculate passage scores in initial ranking. To model the relational context, t he scores are modified using KG -based boosting. Such passages with entities constituting nodes of high-conviction paths in the knowledge graph such as actor → malware → CVE are weighted higher. This guarantees that retrieved results are not only contextually relevant to the analyst's query, but also are related to occurring incidents. Through combing semantic similarity with graph-aware boosting, the retriever enhances ranked passages that are contextually and structurally optimal enhancing correlation accuracy against multi-stage attacks.

3.3 Generator Fine-Tuning and Grounding

The generator is pre-trained with supervised learning using input queries and the retrieved passages of pairs mapped to target advisories. The loss used for training SimplyDeLFG is a regular cross entropy loss, optimized for text generation-quality independently by an auxiliary grounding loss. Grounding loss The grounding loss encourages the model to output tokens for which it can find supporting evidence, and is computed as approximate token-level overlap checks of generated outputs with retrieved text. This double goal provides the generator with a strong incentive for generating short and fact-supported suggestions. At prediction time, the generator has to satisfy some extra constraints: it is requested to explicitly

| ISSN: 2455-1864 | www.ijrai.org|editor@ijrai.org | A Bimonthly, Scholarly and Peer-Reviewed Journal |

||Volume 6, Issue 3, May-June 2023||

DOI:10.15662/IJRAI.2023.0603003

include prov enance markers in its output strings that connect Advisory sentences with Support sentences' IDs. This method is used to enhance traceability and the credibility of automated CTI generation as hallucination, unverifiable claims in such generation will be eliminated or reduced.

3.4 Datasets and Evaluation Design

A test dataset of 2,400 CTI documents including open-source intelligence reports and vendor advisories was assembled to assess the system. Following the annotation, each document was marked for incident-level labels and multi-step relational structures, and thus can be taken as an appropriate benchmark dataset both for classification and relation understanding tasks. To simulate operational data there is also a set of 1,200 synthetic alert sequences that were created with instances inspired from the corpus for this same purpose. These alert sequences are similar to raw signals an analyst might see in their position.

Three baselines were created for comparison. The first baseline is based on keyword and TTP matching, which results in a rule-based security alert used by SIEM pipelines. The second baseline is based on vector retrieval without KG boosting, in which the results are fed back to the same generator structure as built in our system. The third baseline uses only the generator to function without retrieval, thus focusing on the effect of grounding by external evidences.

Evaluation measures were chosen to reflect system effectiveness and analyst-oriented utility. These are precision, recall and F1 score for incident classification and multi-stage correlation accuracy i.e. the proportion of correctly mapped stages, Incident records in INI were classified into three categories: Normal Traffic, Suspicious Activity (SA) and Anomaly or Threat. To evaluate operational gains, simulated analyst sessions were performed to determine the time of triage and a workload reduction factor was calculated. Finally, two more generation-centric scores were used: the grounding rate (the fraction of all generated statements compatible with retrieved evidence), and hallucination rate (the extent of generating unsupported or falsified claims in advisories).

3.5 Implementation

The prototype system was developed by leveraging different open-source tools, guarantees reproducibility and scalability. A bi-encoder fine-tuned with contrastive loss on CTI sentence pairs was used as a source of semantic embeddings. FAISS was used as the vector index with fast similarity search on hundreds of thousands of passages. The knowledge graph was stored in a Neo4j graph database for large-scale storage and query support of the extracted entities and relations. The generator was realized as a transformer-based sequence-to-sequence model further finetuned over an in-house compiled CTI question-answer summarization dataset. Retrieval, KG queries and generation were orchestrated using a lightweight microservice managing the flow of ranked passages and relation summaries over to the generator. Thus, what remained was an extensible and modular pipeline that could deliver CTI advisories in semi real-time with structured evidence enablers.

IV. RESULTS AND ANALYSIS

4.1 Detection performance

Table 2 summarizes detection metrics across baselines and proposed system.

Table 2 — Detection performance (2,400 test documents)

System	Precision	Recall	F1
Keyword/TTP baseline	0.71	0.62	0.66
Vector retrieval + gen (no KG)	0.78	0.68	0.73
Proposed RAG + KG (hybrid)	**0.86**	**0.80**	**0.83**

| ISSN: 2455-1864 | www.ijrai.org|editor@ijrai.org | A Bimonthly, Scholarly and Peer-Reviewed Journal |

||Volume 6, Issue 3, May-June 2023||

DOI:10.15662/IJRAI.2023.0603003

The benchmark compared three CTI generation systems: a keyword/TTP baseline, a vector retrieval and generator model without knowledge graph integration, and the proposed hybrid RAG with knowledge graph. Performance was evaluated in terms of precision, recall, and F1 score that indicated accuracy of incident detection and quality of generated advisories.

The precision of the keyword/TTP baseline was 0.71, recall was 0.62 and F1 measure was 0.66. Although this rules-based method does indeed successfully identify a fair number of indicators, it failed to cover many relevant threats by a limitation of lexical and domain restrictions. Its lower recall also highlights the inefficacy of a static matching approach to capture polymorphic or synonym-rich threat descriptions, which is a typical issue with unstructured CTI feeds.

The vector retrieval plus generator model, but without KG integration, also achieved performance exceeding all baselines with precision 0.78, recall 0.68 and F1 0.73. Making use of semantic embeddings, it could capture synonyms and contextual meaning, which are not covered by just matching keywords to each other. But the absence of relational reasoning meant some relevant relationships among entities, for example relating across multi-stage attack campaigns, could be missed and that in turn reduced correlation precision.

The RAG + KG hybrid network yielded the best results: P=0.86, R = 0.80, F1=0.83. These gains illustrate the synergy from semantic retrieval and KG-boosted relevance. The retriever did not only retrieve contextually related passages, but also exploited graph relations to score evidence that is connected by confident paths. This fusion enhanced multi-stage attack detection and helped in abating false positives by anchoring generation in both, semantic similarity as well as structured relationships.

In general we notice that semantic text retrieval plus knowledge graph reasoning improves CTI workflows, contributing to more reliable advisories supported by evidence in comparison to rule-based baselines and retrieval only.

V. CONCLUSION AND FUTURE WORK

We present a RAG-driven cybersecurity intelligence pipeline that incorporates semantic search (dense retrieval), a structured cybersecurity knowledge graph, and a fine-tuned language model to generate grounded and actionable advisories. Empirical evaluation on a narrow domain corpus demonstrated significant improvements: 25.7% higher F1 over keyword baselines, 22% increase in multi-stage correlation accuracy with knowledge-graph augmentation, and 31% reduction in average analyst triage time; model hallucinations are largely suppressed as we enforce generator provenance.

The major contributions include hybrid retrieval-KG scoring mechanism, the grounding loss during generator fine-tuning and an evaluation methodology that includes not only standard NLP metrics but also analyst efficiency and correlation quality. The findings can be seen as a further validation of the benefit of semantic retrieval and relational reasoning for enhancing CTI workflows.

In future, three aspects will be studied. Secure retrieval: the adoption of secure protocols for encrypted indexing, as well as per query access control to contra adverse privacy effects and leakage internal signals. Second, adversarial robustness: how to harden retrieval index against poisoning attacks provenance aware indexing, adversarial example filtering. Operational integration, and human factors: Deploy in production SOCs to collect analyst adoption rates, trust calibration practices, and incident response times with real-world data feeds; integrate a feedback loop from analysts to refine KG generator behaviors. Further research should also consider automation of the reconciliation of conflicting reports, confidence-aware advisory generation and standard for CTI provenance metadata to facilitate interoperability between vendors and platforms.

| ISSN: 2455-1864 | www.ijrai.org|editor@ijrai.org | A Bimonthly, Scholarly and Peer-Reviewed Journal |

||Volume 6, Issue 3, May-June 2023||

DOI:10.15662/IJRAI.2023.0603003

REFERENCES

[1] L. F. Sikos, "Cybersecurity knowledge graphs," *Knowledge and Information Systems*, vol. 65, pp. 3511–3531, Apr. 2023. [Online]. Available: SpringerLink.
[2] Abid A, Jemili F (2020) Intrusion detection based on graph oriented big data analytics. Procedia Comput Sci 176:572–581.
[3] Chen X, Shen W, Yang G (2021) Automatic generation of attack strategy for multiple vulnerabilities based on domain knowledge graph. In: 47th Annual Conference of the IEEE Industrial Electronics Society. IEEE.
[4] C. Shin, I. Lee, and C. Choi, ''Towards GloVe-based TTP embedding with ATT&CK framework,'' in Proc. Korea Inst. Military Sci. Technol., Daejeon, South Korea, 2023, pp. 1606–1607.
[5] Noor, U.; Anwar, Z.; Amjad, T.; Choo, K.-K.R. A machine learning-based FinTech cyber threat attribution framework using high-level indicators of compromise. *Future Gener. Comput. Syst.* **2019**, *96*, 227–242
[6] Husák, M.; Bartoš, V.; Sokol, P.; Gajdoš, A. Predictive methods in cyber defense: Current experience and research challenges. *Future Gener. Comput. Syst.* **2021**, *115*, 517–530.
[7] Tang, B.; Wang, J.; Yu, Z.; Chen, B.; Ge, W.; Yu, J.; Lu, T. Advanced Persistent Threat intelligent profiling technique: A survey. *Comput. Electr. Eng.* **2022**, *103*, 108261
[8] Garrido JS, Dold D, Frank J (2021) Machine learning on knowledge graphs for context-aware security monitoring. In: 2021 IEEE International Conference on Cyber Security and Resilience. IEEE, pp 55–60
[9] Grojek AE, Sikos LF (2022) Ontology-driven artificial intelligence in IoT forensics. In: Daimi K, Francia G III, Encinas LH (eds) Breakthroughs in digital biometrics and forensics. Springer, Cham, pp 257–286
[10] Homayoun, S.; Dehghantanha, A.; Ahmadzadeh, M.; Hashemi, S.; Khayami, R.; Choo, R.; Newton, D.E. Deep Dive into Ransomware Threat Hunting and Intelligence at Fog Layer. *Future Gener. Comput. Syst.* **2018**, *90*, 94–104
[11] Lekkala, C. (2020). Leveraging Lambda Architecture for Efficient Real-Time Big Data Analytics. European Journal of Advances in Engineering and Technology, 7(2), 59–64.
[12] Islam R, Refat RUD, Yerram SM et al (2022) Graph-based intrusion detection system for controller area networks. IEEE Trans Intell Transp Syst 23(3):1727–1736.
[13] Kang JJ, Sikos LF, Yang W (2021) Reducing the attack surface of edge computing IoT networks via hybrid routing using dedicated nodes. In: Ahmed M, Haskell-Dowland P (eds) Secure edge computing: applications, techniques and challenges. CRC Press, Boca Raton, pp 97–111.
[14] Wagner, T.D.; Palomar, E.; Mahbub, K.; Abdallah, A.E. A Novel Trust Taxonomy for Shared Cyber Threat Intelligence. *Secur. Commun. Netw.* **2018**, *2018*, 9634507.
[15] Khan, T.; Alam, M.; Akhunzada, A.; Hur, A.; Asif, M.; Khan, M.K. Towards augmented proactive cyberthreat intelligence. *J. Parallel Distrib. Comput.* **2019**, *124*, 47–59.
[16] Tatam, M.; Shanmugam, B.; Azam, S.; Kannoorpatti, K. A review of threat modelling approaches for APT-style attacks. *Heliyon* **2021**, *7*, e05969