

Adaptive Retrieval Agents for Factual and Cost-Aware Cyber Threat Intelligence Summarization

Mona Rajhans*, Vishal Khawarey**

*Palo Alto Networks, Santa Clara, CA, USA

**Quicken Inc., Menlo Park, CA, USA

mrajhans@paloaltonetworks.com, vishal.sanfran@gmail.com

Abstract—Cyber threat intelligence (CTI) text is large, heterogeneous, and difficult for analysts to summarize accurately. While Retrieval-Augmented Generation (RAG) improves factual grounding in large language models (LLMs), current systems typically rely on a fixed retrieval depth, causing unnecessary latency or inconsistent factuality. This paper addresses this challenge by introducing an adaptive retrieval agent that adjusts context size in real time using a retrieval-aware factuality metric, the Grounded Entailed Claim Score (GECS). Retrieval selection is cast as a cost-utility optimization problem, $J = G - \lambda\tau$, enabling the agent to balance factual gain (G) against computational cost (τ) without modifying or retraining the underlying LLM. Experiments on two cybersecurity corpora—structured NVD vulnerabilities and semi-structured MITRE ATT&CK narratives—show that the agent matches optimal fixed-depth performance on structured text while avoiding costly retrieval when factual benefits are limited. These results demonstrate that factuality metrics can serve as lightweight internal feedback signals for self-calibrating and cost-efficient RAG systems in real-world CTI workflows.

Keywords—Retrieval-Augmented Generation, Large Language Models, Cyber Threat Intelligence, Factual Consistency, Adaptive Retrieval, Cost-Aware Inference

I. INTRODUCTION

Cyber threat intelligence (CTI) sources such as the National Vulnerability Database (NVD) and the MITRE ATT&CK framework provide large volumes of heterogeneous text describing vulnerabilities, adversarial behavior, and attack procedures. Summarizing this information accurately is challenging: human analysts face substantial cognitive load, while large language models (LLMs) often generate incomplete or hallucinated statements when used without retrieval support.

Retrieval-Augmented Generation (RAG) improves factual grounding by supplying the model with relevant evidence, yet most RAG systems employ a fixed retrieval depth. This static design ignores variation in query complexity, corpus structure, and computational constraints, leading to either (i) excessive latency from overly large contexts or (ii) factual degradation when insufficient evidence is retrieved. A principled mechanism for dynamically controlling retrieval depth is therefore essential for reliable and efficient CTI summarization.

To address this challenge, we introduce an *adaptive retrieval agent* that selects context depth based on real-time factuality feedback. The agent is driven by a retrieval-aware factuality metric, the Grounded Entailed Claim Score (GECS), and formulates retrieval selection as a cost-utility

optimization problem balancing factual gain against latency. This decision-theoretic formulation enables autonomous, per-query regulation of retrieval depth without retraining or modifying the underlying LLM. The remainder of this paper presents the design, implementation, and evaluation of this adaptive retrieval framework on structured (NVD) and semi-structured (MITRE ATT&CK) CTI corpora.

Contributions. This work makes the following key contributions:

- 1) **A retrieval-aware factuality metric (GECS).** We introduce the Grounded Entailed Claim Score (GECS), a reference-free metric that evaluates whether generated claims are logically supported by retrieved evidence, enabling fine-grained factuality assessment in RAG systems.
- 2) **A decision-theoretic adaptive retrieval agent.** We design an autonomous retrieval agent that uses GECS as an internal feedback signal to regulate context depth in real time. Retrieval selection is formulated as a cost-utility optimization problem $J = G - \lambda\tau$, allowing the agent to balance factual gain against latency without model retraining.
- 3) **Empirical evaluation on structured and semi-structured CTI corpora.** We provide a reproducible benchmark on the NVD (structured) and MITRE ATT&CK (semi-structured) datasets, showing that the adaptive agent matches optimal fixed-depth performance on structured data and behaves conservatively on conceptual text where additional retrieval offers limited factual benefit.

II. RELATED WORK

A. Retrieval-Augmented Generation and Adaptive Context Control

Retrieval-Augmented Generation (RAG) [1] has become a core technique for grounding large language models (LLMs) in external evidence by conditioning generation on retrieved documents. Subsequent research has optimized retrieval density and embedding quality using architectures such as Sentence-BERT [2] and FAISS [3], and explored hybrid sparse-dense retrieval pipelines. Recent extensions (e.g., adaptive context length [4]) propose dynamically truncating

retrieval based on query relevance, but they rely on heuristics or static thresholds rather than learned or feedback-driven control. In contrast, our work formulates retrieval selection as a *decision problem* guided by factuality feedback, allowing the retrieval process itself to behave as an autonomous agent that regulates context depth according to a quantitative objective.

B. Factuality and Hallucination Metrics

Factual consistency has been widely studied in generative models, with methods such as CheckList [5] and FactScore [6] assessing entailment or overlap with references. More recent surveys [7], [8] identify factual grounding as a persistent limitation in LLMs. However, these metrics are primarily used for post-hoc evaluation. The proposed **Grounded Entailed Claim Score (GECS)** extends this line of work by embedding factual consistency directly into the inference loop as a feedback signal, enabling dynamic control of retrieval behavior.

C. Decision-Theoretic and Agentic Approaches

Autonomous and multi-agent reasoning frameworks have long explored decision-making under cost constraints [9], [10]. Recent advances in AI agents [11] and tool-augmented LLMs [12] show how language models can act as controllers that adaptively query, plan, and retrieve information. Our *adaptive retrieval agent* follows this paradigm, operating as a lightweight decision-theoretic controller that optimizes the trade-off between factual gain and computational latency via $J = G - \lambda\tau$. This bridges retrieval-augmented generation with principles of bounded rationality and resource-aware reasoning central to agent research.

III. METHODOLOGY

Our framework extends the retrieval-augmented generation (RAG) paradigm [1] with quantitative factuality evaluation and adaptive retrieval control. The overall system architecture, shown in Figure 1, integrates a dual retriever, a Qwen2.5-7B-Instruct generator, and an adaptive controller driven by the *Grounded Entailed Claim Score (GECS)*. It operates in two stages: an offline phase that builds TF-IDF and MiniLM/FAISS indices from cybersecurity corpora, and an online phase where GECS and latency feedback guide real-time adjustment of retrieval depth k .

At inference time, retrieved documents are concatenated with the user query and passed to the open-weight **Qwen2.5-7B-Instruct** model via the `llama.cpp` Python bindings in quantized (GGUF) format. This configuration was selected for its strong factual precision and efficient CPU inference, enabling reproducible experiments on local hardware.

The adaptive controller uses GECS as a feedback signal to balance factual gain and latency through the cost-utility objective $J = G - \lambda\tau$. It incrementally increases retrieval depth until the marginal factuality gain falls below a tolerance ϵ or the latency budget $\tau_{\max}$ is reached. This policy enables *self-calibrating* retrieval selection without retraining or model modification.

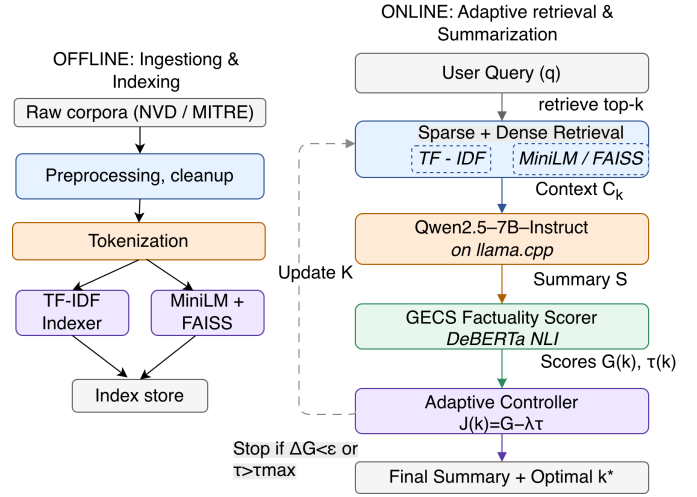


Fig. 1: Overview of the adaptive retrieval architecture.

Algorithm 1 GECS-Guided Adaptive Retrieval Controller

Require: Query q , corpus D , retrievers $\mathcal{R}_s$ (TF-IDF), $\mathcal{R}_d$ (MiniLM/FAISS), cost weight λ , tolerance ϵ , latency limit $\tau_{\max}$

- 1: Precompute indices for $\mathcal{R}_s, \mathcal{R}_d$ ▷ offline ingestion
- 2: Initialize $k \leftarrow 2, G_{prev} \leftarrow 0$
- 3: **while** True **do**
- 4: $C_k \leftarrow$ top- k retrieval from $\mathcal{R}_s \cup \mathcal{R}_d$
- 5: $S_k \leftarrow \text{QWENGGENERATE}(q, C_k)$
- 6: $(G_k, \tau_k) \leftarrow \text{GECSSCORE}(S_k, C_k)$
- 7: $J_k \leftarrow G_k - \lambda\tau_k$
- 8: **if** $(G_k - G_{prev}) < \epsilon$ **or** $\tau_k > \tau_{\max}$ **then**
- 9: **break**
- 10: **end if**
- 11: $G_{prev} \leftarrow G_k; k \leftarrow k + 2$
- 12: **end while**

Ensure: Final summary S_k , optimal depth k^* , and utility J_k

A. Retrieval-Augmented Generation Pipeline

The retrieval-augmented generation (RAG) component integrates retriever and generator modules to enable context-grounded text summarization for cybersecurity corpora. Given a user query q , the retriever selects a set of top- k relevant text segments $\{d_1, d_2, \dots, d_k\}$ from a corpus D , which are concatenated and provided as input to a generative model M . The resulting summary $\hat{S} = M(q, \{d_i\}_{i=1}^k)$ represents a synthesis conditioned on retrieved evidence.

We construct two complementary retrieval pipelines to capture both lexical and semantic similarity: (i) a **sparse retriever** based on TF-IDF weighting using `scikit-learn`, and (ii) a **dense retriever** that encodes text using Sentence-BERT [2] and performs vector similarity search via FAISS [3]. For dense retrieval, each document embedding $e_i = f_\theta(d_i)$ is precomputed using a MiniLM encoder, and retrieval scores are computed as cosine similarity:

$$r_i = \frac{\mathbf{e}_q \cdot \mathbf{e}_i}{\|\mathbf{e}_q\| \|\mathbf{e}_i\|}. \quad (1)$$

Top- k documents are selected according to the ranking r_i , forming the retrieval context C_k . This context is prepended

to the user query and passed directly to the open-weight **Qwen2.5-7B-Instruct** model via the `llama.cpp` Python bindings in quantized (GGUF) format. Qwen was chosen for its strong factual precision and efficient CPU inference capabilities, enabling all experiments to be executed on a local machine without GPU acceleration.

Each generated summary is evaluated using the Grounded Entailed Claim Score (GECS) described in Section III-D. All retrieval and generation modules are implemented in pure Python using open-source libraries and executed within a reproducible Jupyter environment.

B. Evaluation Metrics

a) 1) *Factual Consistency (Overlap-based Baseline)*: We first compute a lexical overlap measure as a baseline for factual consistency:

$$F = \frac{|S \cap G|}{|S|} \quad (2)$$

where S is the set of atomic facts extracted from the generated summary and G is the corresponding ground-truth fact set.

b) 2) *Hallucination Rate*: The proportion of unsupported or spurious claims is given by

$$H = 1 - F \quad (3)$$

C. Datasets

We evaluate the framework on two publicly available cybersecurity corpora that differ in structure and linguistic style:

- 1) **NVD CVE Feed** – structured vulnerability records with concise factual descriptions. [13]
- 2) **MITRE ATT&CK** – semi-structured narratives describing adversarial tactics and techniques. [14]

Table I summarizes dataset statistics. Both corpora were preprocessed into sentence-level segments, with identifiers and descriptions retained to preserve factual granularity for retrieval evaluation.

TABLE I: OVERVIEW OF EVALUATED DATASETS

Dataset	Entries	Avg. Length	Type
NVD CVE Feed	250K	120	Structured
MITRE ATT&CK	600	320	Semi-Structured

The NVD corpus represents short, fact-dense vulnerability descriptions where retrieval primarily reinforces lexical grounding. In contrast, MITRE ATT&CK contains longer, conceptually abstract entries that test the controller's behavior under looser semantic structure. Together, they provide a balanced benchmark for assessing how retrieval depth, factuality, and cost interact across structured and conceptual cybersecurity text.

D. Grounded Entailed Claim Score (GECS)

To quantify factual consistency between model-generated summaries and retrieved evidence, we introduce the **Grounded Entailed Claim Score (GECS)**—a retrieval-aware factuality metric inspired by entailment-based methods such as FactScore [6] and CheckList [5]. Unlike reference-based metrics, GECS directly measures whether individual claims in a generated summary are logically supported by retrieved context, enabling use as both an evaluation and control signal in retrieval-augmented generation (RAG).

Given a summary S consisting of n atomic claims $\{c_1, c_2, \dots, c_n\}$ and its retrieval context C_k , each claim is classified as *entailed*, *contradicted*, or *unsupported* using a DeBERTa-v3 [15] natural language inference (NLI) model, which yields entailment and contradiction probabilities $p_{\text{ent}}(c_i)$ and $p_{\text{con}}(c_i)$. For each claim, the *support score* is computed as:

$$u_i = \max(0, p_{\text{ent}}(c_i) - p_{\text{con}}(c_i)), \quad (4)$$

and the overall grounded entailed claim score is the mean support:

$$\text{GECS} = \frac{1}{n} \sum_{i=1}^n u_i. \quad (5)$$

Higher GECS values indicate stronger factual grounding of the generated text. From these claim-level signals, we also derive complementary statistics:

$$P_g = \frac{1}{N} \sum_{i=1}^N \mathbb{I}[u_i > \tau_s] \quad (\text{grounded precision}) \quad (6)$$

$$H = 1 - P_g \quad (\text{hallucination rate}) \quad (7)$$

$$X = \frac{1}{N} \sum_{i=1}^N \mathbb{I}[p_{ij}^c > \tau_x] \quad (\text{contradiction frequency}) \quad (8)$$

where τ_s and τ_x are empirical thresholds. Together, (GECS, P_g, H, X) provide a fine-grained factuality profile for retrieval-augmented summarization.

GECS differs from prior metrics in two key aspects: (i) it is *retrieval-aware*, grounding factuality in dynamically retrieved evidence rather than static references; and (ii) it is *operational*, serving as a real-time feedback signal for adaptive retrieval control described in Section III-E.

a) 4) *Component Metrics*: In addition to GECS, we report:

$$P_g = \frac{1}{N} \sum_{i=1}^N \mathbb{I}[u_i > \tau_s] \quad (\text{grounded precision}) \quad (9)$$

$$H = 1 - P_g \quad (\text{hallucination rate}) \quad (10)$$

$$X = \frac{1}{N} \sum_{i=1}^N \mathbb{I}[p_{ij}^c > \tau_x] \quad (\text{contradiction frequency}) \quad (11)$$

where τ_s and τ_x are empirical thresholds for supported and contradictory claims respectively. Together, (GECS, P_g, H, X) provide a fine-grained, model-agnostic view of factual reliability.

b) 5) *Efficiency Metrics*: Following [1], computational efficiency is measured by:

$$\tau = \frac{t}{|S|}, \quad \text{latency per generated token} \quad (12)$$

$$C = \frac{|x|}{|S|}, \quad \text{compression ratio}, \quad (13)$$

where t is inference time and $|x|$ is input length.

E. GECS-Guided Adaptive Retrieval Controller

Retrieval depth (k) strongly influences both factuality and efficiency in RAG systems, yet most pipelines use a fixed context size across all inputs. This is suboptimal: structured text (e.g., NVD) often saturates in factual gain at small k , while conceptual text (e.g., MITRE ATT&CK) may require broader retrieval. To address this, we propose a **GECS-guided adaptive retrieval controller** that adjusts k dynamically based on real-time factuality feedback.

Let $G(k)$ denote the grounded entailed claim score (GECS) and $\tau(k)$ the corresponding latency at retrieval depth k . The controller optimizes a cost-utility objective:

$$J(k) = G(k) - \lambda \tau(k), \quad (14)$$

where λ controls the trade-off between factual gain and computational cost. The goal is to select the smallest k that maximizes $J(k)$ without exceeding a latency budget $\tau_{\max}$.

Because evaluating $G(k)$ on full summaries is expensive, the controller estimates a proxy $\tilde{G}(k)$ from short probe generations. Retrieval depth is incrementally increased until the marginal gain falls below a threshold ϵ or latency exceeds $\tau_{\max}$:

$$\tilde{G}(k+\Delta k) - \tilde{G}(k) < \epsilon \quad \text{or} \quad \tau(k) \geq \tau_{\max}. \quad (15)$$

The selected depth k^* yields the final generation maximizing (14). This adaptive policy requires no retraining and generalizes to any RAG pipeline with retrieval and generation access. By transforming GECS from an evaluation metric into an operational feedback signal, it enables *self-calibrating* retrieval that balances factual grounding and latency per query in real time.

IV. EXPERIMENTS

We evaluate the proposed framework on two complementary cybersecurity corpora: the structured **National Vulnerability Database (NVD)** [13] and the semi-structured **MITRE ATT&CK** corpus [14]. These datasets represent realistic but linguistically distinct sources of threat intelligence, providing a natural test bed for retrieval-augmented summarization. The experiments address three questions: (i) how retrieval depth influences factuality and latency, (ii) how the GECS-guided adaptive controller compares with fixed-depth baselines, and (iii) how corpus structure affects the factuality-efficiency trade-off.

Each corpus was queried using five representative information needs covering privilege escalation, command and control, persistence, credential access, and defense evasion. For each query, we compute the grounded entailed claim

score (GECS), grounded precision P_g , hallucination rate H , contradiction frequency X , latency (τ), and cost-utility objective $J = G - \lambda\tau$ with $\lambda=0.02$. We compare fixed retrieval depths $k \in \{2, 4, 8\}$ against the proposed adaptive policy, which selects k dynamically according to the stopping rule in (15).

All experiments were conducted on a MacBook Pro (M3, 16 GB RAM) using the open-weight **Qwen2.5-7B-Instruct** model via `llama.cpp`, FAISS, and Hugging Face Transformers. All evaluations were implemented in reproducible Jupyter notebooks.

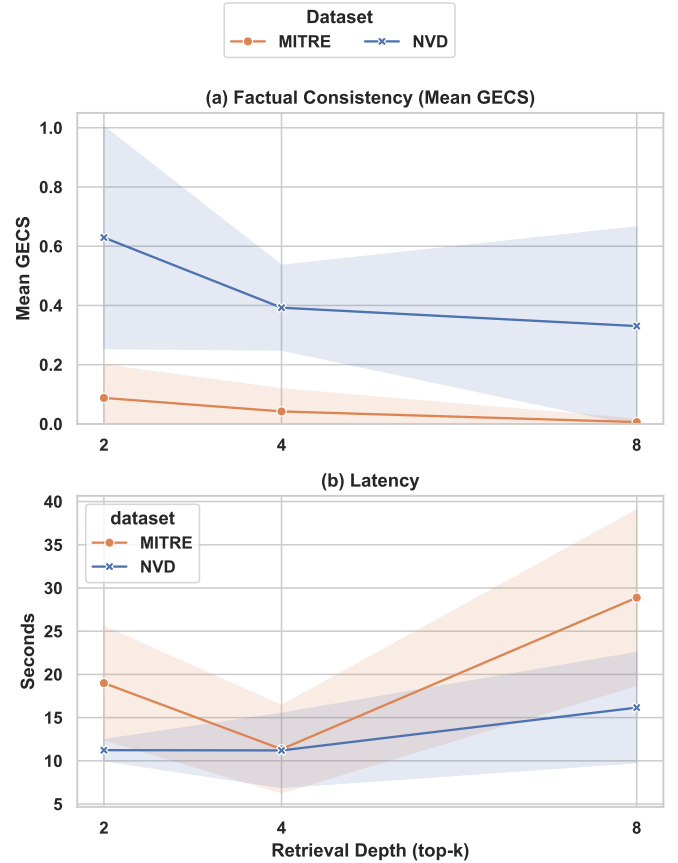


Fig. 2: Retrieval-Depth Effects on Factual Consistency and Efficiency Across NVD and MITRE Corpora

A. Results and Discussion

Figure 2 shows that retrieval depth affects structured and semi-structured sources differently. NVD entries, which contain concise factual descriptions, exhibit minimal benefit beyond small context sizes. MITRE ATT&CK narratives, in contrast, gain modest factuality improvements at moderate depths but display greater variability due to conceptual language. Overall, retrieval augmentation improves grounding while introducing a predictable latency overhead.

Table II reports the mean factuality (G), latency (τ), and objective ($J = G - \lambda\tau$) for both fixed and adaptive retrieval policies. On the structured **NVD corpus**, the adaptive controller converged to the same retrieval depth ($k=2$) as the best fixed

configuration, achieving comparable factuality ($G \approx 0.60$) and utility ($J \approx 0.41$). The higher latency reflects the sequential probing overhead inherent to adaptive evaluation rather than model inefficiency.

On the conceptual **MITRE ATT&CK corpus**, all configurations yielded lower GECS values ($G < 0.12$), highlighting the difficulty of grounding long, descriptive threat narratives. Both fixed and adaptive policies produced negative J values under $\lambda = 0.02$, showing that factual gains were outweighed by cost penalties. Nevertheless, the adaptive controller correctly selected a conservative context ($k=2$), avoiding further factual degradation at larger depths. This suggests that while GECS-guided adaptation remains stable across domains, the optimal cost weighting λ may need to vary with corpus complexity.

TABLE II: COMPARISON OF FIXED AND ADAPTIVE RETRIEVAL POLICIES

Dataset / Policy	Top- k	GECS	Latency (s)	$J = G - \lambda\tau$
NVD (Fixed)	2	0.629	11.19	0.406
	4	0.392	29.60	-0.200
	8	0.330	15.83	0.014
NVD (Adaptive)	2	0.597	39.76	0.406
MITRE (Fixed)	2	0.088	18.98	-0.292
	4	0.042	11.58	-0.189
	8	0.006	23.79	-0.470
MITRE (Adaptive)	2	0.115	37.02	-0.253

Note: Results shown for cost weight $\lambda = 0.02$.

Beyond external factuality metrics, we also examine the model's internal confidence dynamics using the supplemental uncertainty analysis described in Sec. IV-D. This analysis shows that retrieval grounding consistently reduces token-level entropy and induces moderate distributional shifts that correlate with higher GECS scores. To further illustrate how retrieval depth affects factuality and latency jointly, Fig. 3 visualizes the adaptive decision trajectories for both NVD and MITRE datasets.

B. Design Implications

In operational settings, the weighting factor λ can be tuned to reflect resource constraints: larger λ favors low-latency summarization, while smaller λ permits richer retrieval at higher computational cost. This cost-aware trade-off can be embedded directly into threat intelligence pipelines to maintain consistent factual quality under variable runtime budgets.

C. Limitations and Future Work

While the GECS-guided controller demonstrates stable behavior across structured and semi-structured corpora, several limitations remain. First, GECS is currently evaluated sequentially at each retrieval depth, introducing additional latency during inference. Parallelized or approximate scoring could substantially reduce this overhead. Second, λ was tuned manually; adaptive or learned calibration could yield better cost-accuracy balance. Third, experiments were conducted on a single open-weight model (Qwen2.5-7B-Instruct); future work

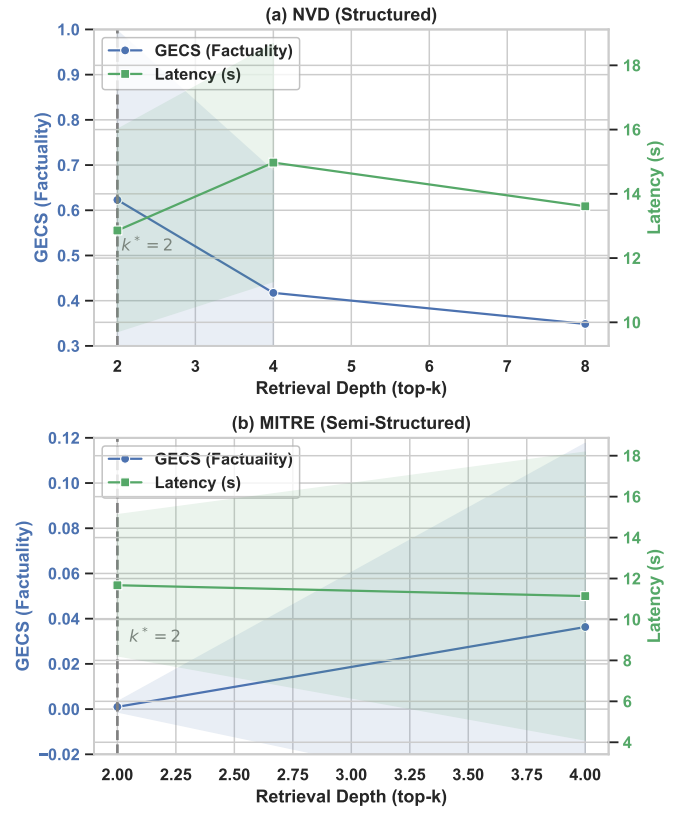


Fig. 3: Adaptive Retrieval Trajectories for Structured (NVD) and Semi-Structured (MITRE) Corpora

should evaluate larger models and diverse datasets, including APTnotes and dark web threat feeds. Finally, while GECS captures entailment-based factuality, it does not explicitly account for uncertainty or partial truthfulness. Integrating uncertainty-aware or token-level attribution methods could further improve robustness and interpretability in adaptive RAG systems.

D. Internal Uncertainty Analysis

To examine how retrieval grounding influences the internal confidence of the Qwen2.5-7B-Instruct model, we conduct an uncertainty analysis using token-level probability distributions collected from a subset of NVD vulnerability queries. Two complementary indicators are evaluated: (1) entropy-based confidence and (2) retrieval sensitivity.

1) *Entropy-Based Confidence*: For each generated token t_i with probability p_i , token-level uncertainty is approximated using:

$$H_i = -p_i \log p_i. \quad (16)$$

The sequence-level mean entropy is computed as:

$$\bar{H} = \frac{1}{N} \sum_{i=1}^N H_i, \quad (17)$$

where N is the number of generated tokens. Lower entropy indicates higher model confidence. Because quantized GGUF

models in `llama.cpp` do not expose full logits, we estimate entropy from the per-token log-probabilities returned by the inference API, which provides a stable relative measure across queries.

2) *Retrieval Sensitivity*: To measure how retrieved evidence affects the model's predictive distribution, we compare token probability sequences under two conditions: with retrieved context (p_i^{full}) and without context (p_i^{null}). Distributional shift is quantified using the Jensen–Shannon divergence:

$$S_{\text{JS}} = \sqrt{\frac{1}{2} D_{\text{KL}}(p^{\text{full}} \| m) + \frac{1}{2} D_{\text{KL}}(p^{\text{null}} \| m)}, \quad (18)$$

$$m = \frac{1}{2}(p^{\text{full}} + p^{\text{null}}).$$

Higher values indicate that retrieval grounding meaningfully shifts the predictive distribution, while still maintaining stable behavior.

3) *Findings*: Across all evaluated queries, retrieval grounding consistently reduces token-level entropy, indicating increased confidence when relevant evidence is present. Retrieval sensitivity values remain moderate (0.25–0.35), suggesting that context influences predictions without destabilizing them. A negative correlation (approximately $\rho = -0.46$) between mean entropy and GECS suggests that lower internal uncertainty aligns with stronger factual grounding.

V. CONCLUSION

This paper introduced an adaptive retrieval agent for RAG-based cyber threat intelligence summarization, driven by the GECS factuality metric and a cost–utility objective. The agent dynamically selects retrieval depth based on real-time evidence quality and latency. Experiments on the NVD and MITRE corpora demonstrate that the method replicates the optimal fixed-depth setting on structured text while adopting a conservative, cost-aware strategy on semi-structured descriptions. These results show that factuality metrics can function as effective internal signals for autonomous and efficient retrieval control. Future work includes automatic tuning of the cost weight λ , incorporation of model uncertainty signals, and extension of the framework to additional CTI sources, larger LLMs, and multilingual environments.

REFERENCES

- [1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, and S. Riedel, “Retrieval-augmented generation for knowledge-intensive nlp tasks,” *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- [2] N. Reimers and I. Gurevych, “Sentence-bert: Sentence embeddings using siamese bert-networks,” in *Proceedings of EMNLP-IJCNLP*, 2019.
- [3] J. Johnson, M. Douze, and H. Jégou, “Faiss: A library for efficient similarity search and clustering of dense vectors,” <https://github.com/facebookresearch/faiss>, 2017.
- [4] X. Liu, R. Chen, and C. Zhao, “Adaptive context length for efficient retrieval-augmented generation,” in *Proceedings of ACL*, 2024.
- [5] M. T. R. Wang *et al.*, “Checklist: A test suite for evaluating nlp models,” in *Proceedings of ACL*, 2020.
- [6] Y. Jiang, H. Trivedi, A. Pagnoni, H. Elshahar, Y. Cao, A. Fan, W.-t. Yih, and N. Rajani, “Evaluating the factual consistency of summaries with factscore,” *Transactions of the Association for Computational Linguistics (TACL)*, 2022.
- [7] K. Zhou, W. Shi, H. Zhang, and M. Chen, “A survey on hallucination in large language models: Causes and mitigation,” *arXiv preprint arXiv:2309.05922*, 2023.
- [8] R. Pang and K. Cho, “Factuality enhanced language models for knowledge-intensive tasks,” in *Proceedings of ACL*, 2023.
- [9] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*. Pearson, 2021.
- [10] L. Kaelbling, M. Littman, and A. Cassandra, “Planning and acting in partially observable stochastic domains,” *Artificial Intelligence*, vol. 101, no. 1, pp. 99–134, 1998.
- [11] Y. Shen, K. Qian, H. Zhang *et al.*, “A survey on llm-based autonomous agents,” *arXiv preprint arXiv:2401.13629*, 2024.
- [12] Y. Yang, M. Zhao, and Q. Liu, “Agentrag: Enabling adaptive retrieval through reasoning agents,” *arXiv preprint arXiv:2403.00292*, 2024.
- [13] NIST, “National vulnerability database (nvd) json feeds,” <https://nvd.nist.gov/vuln/data-feeds>, 2024.
- [14] B. E. Strom, A. Applebaum, D. Miller, K. Nickels, A. Pennington, and C. Thomas, “Mitre att&ck: A framework for describing adversary behavior,” *The MITRE Corporation Technical Report*, 2020.
- [15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of NAACL-HLT*, 2019.



Mona Rajhans received the B.Tech. degree in electronics and communication engineering from Rajasthan Technical University, Kota, Rajasthan, India, in 2012, and the M.S. degree in computer science from the Georgia Institute of Technology, Atlanta, GA, USA, in 2019.

Between 2012 and 2019, she held engineering and research roles in academia and industry in India and the United States. In May 2019, she joined Palo Alto Networks, Santa Clara, CA, USA, where she currently serves as a Senior Engineering Manager, focusing on generative AI-driven cybersecurity products and AI-assisted analyst workflows. Her research interests include human–computer interaction, explainable artificial intelligence, and cybersecurity.

Ms. Rajhans has authored peer-reviewed publications and holds multiple filed and pending patents in AI-assisted cybersecurity systems. She has served on technical program committees and as a reviewer and track chair for several IEEE- and ACM-sponsored conferences.



Vishal Khawarey received the B.Tech. degree in electronics and communication engineering from NIT Warangal, India, in 2007, and the M.S. degree in electrical engineering from North Carolina State University, Raleigh, NC, USA, in 2008.

He has over 15 years of industry experience designing and building large-scale, secure cloud and AI/ML platforms across multiple cloud environments, including AWS, Azure, and GCP. Since April 2025, he has been a Lead Software Engineer at Quicken Inc., Menlo Park, CA, USA, where he works on production generative AI systems for personal finance, including large language model (LLM) integration, retrieval-augmented generation, and cost-aware inference pipelines.

His research interests include artificial intelligence, explainable AI, cloud systems, and cybersecurity. Mr. Khawarey has authored peer-reviewed publications in the field of AI and serves as an industry practitioner bridging applied research and large-scale production systems.