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.**

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

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.