PhD in Information Technology and Electrical Engineering
Università degli Studi di Napoli Federico II

PhD Student: Pellegrino Casoria

Cycle: XXXIX

Training and Research Activities Report

Academic year: 2024-25 - PhD Year: Second

Pellegrino Casoria

Tutor: prof. Simon Pietro Romano

Simon Pietro Romano

Date: October 31, 2025

Training and Research Activities Report

PhD in Information Technology and Electrical Engineering

Cycle: XXXIX **Author:** Pellegrino Casoria

1. Information:

- **PhD student: Pellegrino Casoria** **PhD Cycle: XXXIX**
- **DR number: DR997204**
- **Date of birth: 17/05/1988**
- **Master Science degree: Computer Engineering**
- **University: University of Naples Federico II**
- **Scholarship type: no scholarship**
- **Tutor: Simon Pietro Romano**
- **Co-tutor: N/A**
- **Period abroad: N/A**

2. Study and training activities:

Activity	Type[1]	Hours	Credits	Dates	Organizer	Certificate[2]
GMLE - GIAC Machine Learning Engineer	Course	n.a.	3	01/01/25 - 31/01/25	SANS Institute - David Hoelzer	Y
How To Boost Your PhD	Course	18	5	08/01/25 - 29/01/25	UNINA, CNR-ISASI - Antigone Marino	Y
Emergent behaviors and collective decisions in cyber-physical-human systems	Seminar	1	0.2	13/02/25	KTH Royal Institute of Technology - Karl H. Johansson	Y
IEEE Xplore Webinar: Search Strategies to Supercharge Your Research Experience	Seminar	1	0.2	05/03/25	IEEE - Brian Ryckman, Michael Spada	Y
Micro-Credenziali Digital (Cybersecurity - Modulo Base e Avanzato)	Tutorship	40	3.2	04/06/25 - 23/07/25	Rome Technopole - Accenture	N
Innovation and Entrepreneurship	Course	16	3	05/06/25 - 23/07/25	UNINA, DII - Pierluigi Rippa	Y

1) Courses, Seminar, Doctoral School, Research, Tutorship

2) Choose: Y or N

2.1. Study and training activities - credits earned

	Courses	Seminars	Research	Tutorship	Total
Bimonth 1	**0**	**0**	**10**	**0**	**10**
Bimonth 2	**8**	**0.2**	**6**	**0**	**14.2**
Bimonth 3	**0**	**0.2**	**8**	**0**	**8.2**
Bimonth 4	**0**	**0**	**8**	**1.6**	**9.6**
Bimonth 5	**3**	**0**	**6**	**1.6**	**10.6**
Bimonth 6	**0**	**0**	**9**	**0**	**9**
Total *	**11** (35)	**0.4** (5.5)	**47** (80.2)	**3.2** (3.2)	**61.6** (123.9)
Expected	**30 - 70**	**10 - 30**	**80 - 140**	**0 – 4.8**	

* Values in parentheses indicate Y1+Y2 credits.

3. Research activity:

The cybersecurity landscape is undergoing a profound transformation, driven by increasingly interconnected technologies and rising sophistication of cyber threats. As traditional defenses struggle to keep pace, Generative AI (GenAI) and Large Language Models (LLMs) emerge as promising enablers to automate routine tasks and empower cybersecurity professionals. However, the current discourse on GenAI remains fragmented and disconnected from practical application, particularly in critical domains such as cybersecurity.

In this context, my second-year research has focused on modeling the complexity of **LLM-based cybersecurity applications**, with the goal of experimentally validating new use cases that demonstrate their practical relevance and limitations. The work aims to bridge the gap between prototyping and controlled adoption, establishing a foundation for systematic evaluation, safe integration, and progressive deployment of LLM-driven solutions within cybersecurity environments.

This study spans multiple application areas, which will be detailed in the following sections.

AREA 1: Modeling LLM complexity in cybersecurity landscape

This research area responds to the overwhelming informational noise around LLM and cybersecurity by proposing a structured, process-oriented approach to manage production complexity. Building on an umbrella **Systematic Literature Review** (SLR) of more than 1K unique publications from major bibliographic aggregators over the last 10 years - with 46 deeply inspected articles - the work led to the definition of BabeLLM, a unified **taxonomy of LLM applications for cybersecurity**, prioritizing operational applicability, alignment with existing standards, and readiness for agentic AI. The proposed methodology is grounded in established approaches from the literature and internationally recognized guidelines, including Webster & Watson, Brocke et al., PRISMA, and Nickerson et al.

Both State-of-the-Art analysis and taxonomy definition have been organized into three contributions: i. **NLP tasks**; ii. **Cybersecurity processes**; iii. **LLM applications within Cybersecurity**. We compared our results with existing works and discussed emerging trends and directions. To demonstrate practical

applicability of the proposed taxonomy, we also developed a **machine-readable OWL implementation** and used it to support **GPT-based use cases** for automated inference of data and object properties.

This work led to the following publication: *P. Casoria, S.P. Romano, "BabeLLM: Umbrella Systematic Review for a Standardized Taxonomy of LLM-based Cybersecurity Tasks", submitted to Computer Science Review (Elsevier), pending review as of October 2025.* The ontology, scripts, and datasets are made publicly available upon publication.

Beyond the ontological work, this research area expands into broader modeling challenges associated with LLM-driven cybersecurity. Ongoing work examines **performance and testing** models capable of characterizing LLM reliability in operational settings, as well as **cost-benefit frameworks** that determine when human execution is more effective than GenAI-enabled automation. In parallel, it investigates formal representations for **GenAI-supported workflows**, including augmentation pipelines and agentic-AI orchestration, to clarify integration patterns and operational boundaries.

These research directions will be further developed next year and are designed to converge toward a coherent modeling ecosystem for LLM use in cybersecurity. This effort provides a rigorous and actionable foundation that supports consistent alignment among researchers, practitioners, and institutions, enabling scalable and high-impact cybersecurity innovation in the GenAI era.

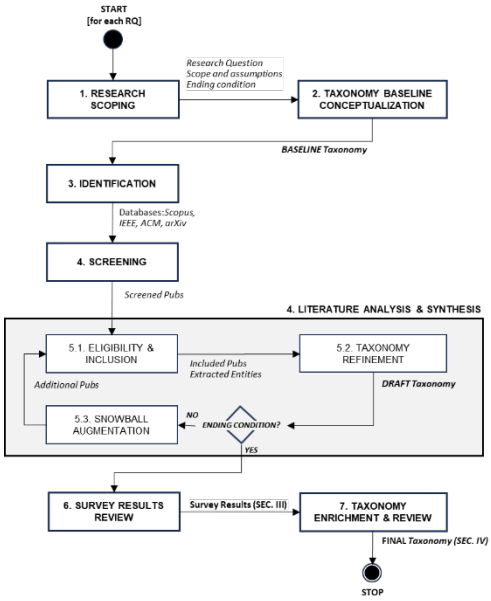
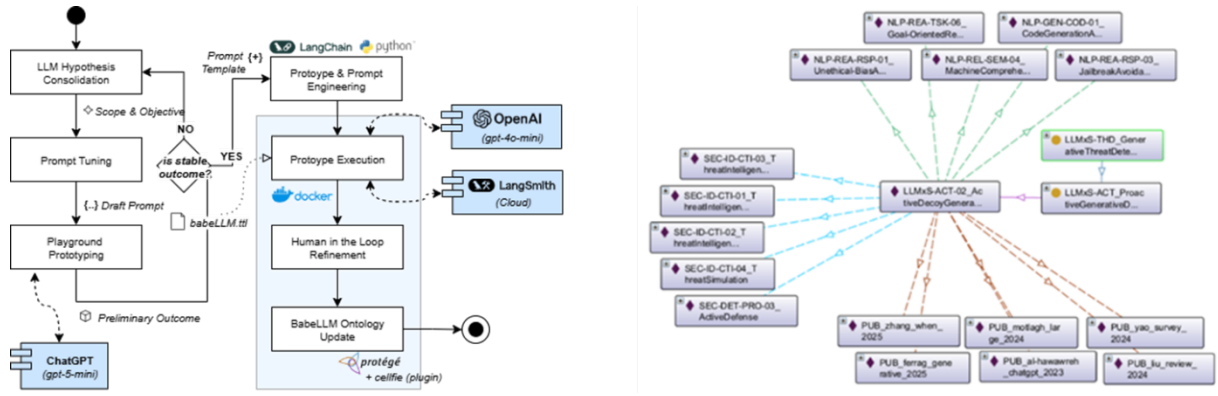


***Figure 1** - SLR Methodology*

AREA 2: Continuous LLM experimentation for Cybersecurity

Many research studies explored challenges that modern Security Operations Centers (SOCs) are facing as they strive to effectively respond to increasingly complex threats. From a high-level perspective, the main SOC concerns include: lack of resources, talents, and infrastructures to scale up operations; low-value repetitive tasks; disparities in skills to operate with heterogeneous technologies. To address these challenges, GenAI has the potential to revolutionize the security industry, significantly enhancing SOC capabilities to detect and respond to emerging threats, supporting humans in decision making processes.

This research area aims to translate the rationalization of LLM applications developed in the other research areas (such as BabeLLM ontology) into a systematic exploration of **advanced GenAI use cases for cybersecurity**. The objective is to progress from isolated proofs-of-concept toward robust, end-to-end solutions that can be rigorously evaluated and operationalized within SOCs. The area focuses on engineering complete workflows, designing and validating **prompting strategies**, and conducting comparative experiments across model families and configurations. The taxonomy will guide the investigation of underexplored scenarios - such as automated literature-review support, CTI workflow enrichment, compliance assistance, and strategic decision-making augmentation - with the overarching goal of establishing repeatable, scalable, and practically deployable LLM-enabled capabilities for SOCs.



a) b)

***Figure 2** - LLM-based ontology enrichment use case: a) experimentation methodology; b) sample enriched ontology*

To demonstrate the practical applicability of the proposed taxonomy, we developed a case study in which LLMs **enrich the descriptive metadata** associated with each BabeLLM ontological entity. The rationale is to extend the initial taxonomy entries with detailed and structured metadata that can support classification, knowledge representation, and advanced tasks such as context-aware prompting and semantic enrichment. The validated outputs were integrated into the ontology through the Cellfie plugin, ensuring seamless incorporation of the enriched taxonomy into a machine-readable format. To maximize the quality of GPT outputs, the prompting strategy combined known approaches including system prompting, role prompting, contextual prompting, and few-shots. Integration with LangSmith enabled centralized and continuous monitoring of token consumption, latency, and estimated invocation costs, providing a quantitative foundation for optimization and evaluation of the solution. The overall accuracy reached approximately 90%, with higher performance on security-related categories.

In addition, exploratory experimentation is progressing along several emerging directions. One line of work focuses on **Cyber-HUMINT** activities, where we developed an automated pipeline for collecting publicly accessible social-network data to generate user profiles through LLM-based interpretation. The prototype leverages LangFlow for orchestration and GPT-5 for semantic extraction and profiling. Parallel activities include the development of a triage automation workflow for the systematic analysis of phishing emails, enabling rapid classification and evidence extraction. We are also investigating automated approaches for threat hunting, active deception, and security-by-design analysis, as well as LLM-driven extraction of actionable intelligence from CTI sources. These efforts aim to identify high-value opportunities for GenAI integration and to establish reproducible methods for evaluating their effectiveness in realistic cybersecurity scenarios.

AREA 3: World of Untrusted

In recent years, the global cyber threat landscape has undergone a profound shift, moving from a peripheral consideration to a central element of cybersecurity strategy. Adversaries increasingly exploit these dynamics to operate with greater scale, speed, and sophistication, amplified by the offensive capabilities enabled by GenAI. Within this evolving “world of untrusted” in which digital content, identities, and interactions cannot be assumed to be legit, this research area focuses on anticipating next-generation **GenAI-powered attacks** by simulating realistic threat scenarios and assessing how defensive processes must adapt.

Current experimentation progresses along several complementary directions. Controlled pipelines for generating highly realistic **deepfake audio and video artefacts** have been designed to support structured awareness programs, scenario-driven tabletop exercises, and experimentally grounded training environments. These efforts are linked to systematic evaluations of deepfake-detection mechanisms and quantitative assessments of human and system-level susceptibility. In parallel, cryptographic **content-signing approaches**, such as C2PA, are being analyzed to assess how provenance-aware media workflows can strengthen verification processes and mitigate the operational impact of manipulated or ambiguous content. Additional work examines how SOC operations and **incident-response procedures** must be adapted to address synthetic-media threats, including the specification of required telemetry, analytical procedures, and decision-making boundaries in critical scenarios. Finally, GenAI-powered **exploit simulation** is being investigated through the use of LLMs and multimodal models to emulate attacker behaviors, generate persuasive social-engineering artefacts, and test both personnel and defensive controls within reproducible and high-fidelity experimental settings aligned with emerging offensive capabilities.

Overall, this area aims to establish a forward-looking framework for assessing and preparing for GenAI-enabled threats, enabling organizations to anticipate adversarial evolution, and to develop cybersecurity capabilities that remain effective as the boundary between real and synthetic artefacts continues to erode.

AREA 4: Responsible and trustworthy AI

The rapid diffusion of GenAI systems introduces new defensive opportunities but also novel classes of risks directly affecting the reliability, safety, and integrity of AI-enabled cybersecurity workflows. This evolution is accompanied by emerging regulatory frameworks and domain standards (e.g., EU AI Act, NIST AI Framework) that impose clearer requirements on model assurance, risk management, and operational controls. In this context, **protecting LLMs and AI platforms** becomes a strategic necessity, requiring systematic methods to understand their attack surface, evaluate systemic vulnerabilities, and implement guardrails capable of maintaining trustworthy behavior under adversarial conditions.

This year's research has focused on analyzing the security posture of AI systems along several dimensions. First, we examined **AI requirements** imposed by sector standards and regulatory obligations, with particular attention to risk categorization, control baselines, and assurance mechanisms relevant to high-impact AI systems. In parallel, we explored **Threat Modelling** approaches tailored to AI platforms (e.g., STRIDE) to characterize attack vectors specific to LLMs, including prompt manipulation, data poisoning, and model misuse. Additional work investigated **prompt guardrail** strategies, assessing the effectiveness of methodological and technological solutions in mitigating harmful or manipulative interactions. In this direction, a **prototype workflow** has been developed for detecting and mitigating anomalous LLM prompts - such as prompt injections or jailbreak attempts - through a combination of contextual and semantic filtering techniques. The prototype was presented at Accenture Cyber Hackademy, demonstrating early feasibility of integrated detection mechanisms for operational environments.

Overall, the area aims to establish a rigorous foundation for securing GenAI systems through structured threat analysis, adherence to evolving regulatory standards, and practical mechanisms for safeguarding model behavior, ultimately enabling trustworthy deployment of AI capabilities in high-stakes cybersecurity settings.

4. Research products:

- **(Publication) Systematic Literature Review:** P. Casoria, S.P. Romano, "BabeLLM: Umbrella Systematic Review for a Standardized Taxonomy of LLM-based Cybersecurity Tasks", Computer Science Review, Elsevier, pending review as of October 2025.
- **(Ontology) BabeLLM Ontology:** Machine-readable representation of LLM-based cybersecurity tasks, integrating concepts from natural language processing, cybersecurity processes, and model implementation modes. The outcome is available on GitHub and Mendeley Data upon publication.
- **(Prototype) LLM-powered Metadata Inference:** Langchain integration with GPT models to automatically infer BabeLLM ontology metadata (i.e., data and object properties) and enrich individuals for NLP tasks, cybersecurity processes and LLM cybersecurity applications.
- **(Prototype) Cyber-HUMINT Reconnissance:** Automated LangFlow pipeline for scraping social network data to build user profiles through LLM-based interpretation of publicly exposed information. The work has been presented at "Red Hot Cyber Conference 2025" in Rome.
- **(Prototype) Prompt Guardrails:** Integrated workflow for detecting and mitigating anomalous LLM prompts - such as prompt injections or jailbreak attempts - using contextual and semantic filtering techniques. The work has been presented at "Accenture Cyber Hackademy" (3rd edition).

5. Conferences and seminars attended

- (Attendee) ITASEC25 & SERICS, Joint National Conference on Cybersecurity, University of Bologna, Bologna, 3-8-Feb-2025.
- (Speaker) Doxing with Langflow: Are We Building the End of Privacy?, Red Hot Cyber Conference 2025, Rome, 08-09-Mag-2025.
- (Speaker) Social Engineering 2.0: Uncovering Emerging Deepfake Threats, Red Hot Cyber Conference 2025, Rome, 08-09-Mag-2025**.**

6. Periods abroad and/or in international research institutions

Research activities have been conducted in partnership with **Accenture**. Specifically, I am currently onboarded on Accenture Cybersecurity group, responsible to support Clients on identification, detection, and monitoring of security threats on IT services and infrastructures. I am currently holding the position of Security Consulting Senior Manager, coordinating the applied innovation team to develop innovative propositions, enforce collaborations with innovation stakeholders, and perform R&D on emerging security topics. This includes research in all areas described in the previous sections (Sec. 3).

During this second year, the **Public-Private Partnership (PPP)** between Accenture and UNINA has fostered productive knowledge exchange between academia and industry, creating a synergistic environment for research and innovation in cybersecurity while preserving the distinctive perspectives and roles of each domain. My activities in Accenture have focused on analyzing security needs across different markets, exploring the creation of many LLM-based prototypes in partnership with academia, vendors, technology partners, and corporate innovation centers.

Throughout these activities, Accenture has supported my research providing specialized knowledge and expertise in relevant security sectors, technologies and markets. Described studies were conducted independently within my academic activities, and Accenture had no involvement in the study design, data collection, analysis, interpretation, or decision to publish the results.

7. Tutorship

The Rome Technopole consortium brings together regional universities, public research institutions, local authorities, and industrial partners to strengthen the innovation ecosystem in strategic domains such as digital transformation, energy transition, and life sciences.

Within this framework, I contributed in **Micro-Credenziali Digital** program with **supplementary teaching** activities (40 hrs) focused on advanced cybersecurity topics. The course covered both foundational and advanced cybersecurity areas, in collaboration with Sapienza University of Rome, University of Rome Tor Vergata, Roma Tre University, and UNICAS.

The initiative provided students with a structured progression from foundational principles to applied competence on **cybersecurity**. The foundational module focused on core concepts of information security management, with emphasis on ISO/IEC 27001, risk identification, and internal audit practices, supported by practical exercises on assessing and mitigating cyber risk in accordance with the standard. The advanced module expanded on these foundations by addressing network security, incident response, and applied security engineering, culminating in a project-based exercise. Together, the two modules offered a comprehensive pathway for acquiring both theoretical understanding and practical skills aligned with contemporary cybersecurity practice.

8. Plan for year three

The third year of the PhD will focus on consolidating ongoing research lines and extending them toward mature scientific contributions. Priority will be given to publishing results in international conferences and journals, with particular emphasis on the modeling of AI security performance, cost, and operational trade-offs. Further experimentation will advance the investigation of GenAI-augmented SOC workflows, deepfake-driven threat scenarios, and trustworthy AI mechanisms, alongside exploratory work on IoT security and next-generation communication technologies.

A research period abroad is planned at an international institution to strengthen collaborations, refine methodological approaches, and broaden the impact of the research outcomes. Additional teaching and tutorship activities will be undertaken through advanced courses aligned with the doctoral program.

The objective for the third year is to finalize the experimental work, complete the dissertation, and position the research for meaningful impact within both academic and applied cybersecurity domains.