

POLITECNICO DI TORINO

Master's Degree in Cybersecurity



Master's Degree Thesis

AI Explainability in Cybersecurity: a framework for explainability evaluation of language models in cybersecurity decision-making

Supervisors:

Prof. Cataldo Basile

Dr. Alessandro Cortina

Dr. Giovanni Ferraris

Candidate:

Alessio M. Palermo

Academic Year 2025/2026
Torino

Abstract

The increasing adoption of sophisticated AI systems in high-stakes domains such as cybersecurity is driven by the need to manage rapidly growing volumes of heterogeneous data, increasingly complex infrastructures, and persistent shortages of specialised human resources. Organisations rely on data-driven, learning-based solutions to automate detection, prioritisation, and response activities, aiming to improve the speed, consistency, and coverage of security operations. At the same time, the shift towards deep neural architectures and language-based models has been accompanied by a substantial increase in model complexity and opacity. As these systems are deployed to maximise predictive performance, their internal decision-making processes often become less accessible to human stakeholders, thereby widening the gap between effectiveness and interpretability.

This thesis addresses this tension through a structured analysis of the Explainable Artificial Intelligence (XAI) paradigm. The initial research phase critically reviews the main explainability methodologies, examining their strengths, limitations, and consistency with the transparency requirements emerging from the regulatory landscape.

On this basis, the thesis investigates Small Language Models (SLMs) as efficient and domain-oriented alternatives to Large Language Models (LLMs) in cybersecurity, and develops a compliance-oriented evaluation framework for their assessment in security-sensitive settings. The framework combines model selection criteria, deployment-regime analysis, explainability techniques, and audit-oriented reporting, thereby linking technical performance to governance requirements such as local deployability, data residency, and verifiable transparency.

The experimental phase instantiates this framework for two critical cybersecurity tasks, phishing detection and vulnerability severity classification. It evaluates lightweight open-weight models through a common explainability pipeline based on attention visualisation, integrated gradients, and LIME, in order to analyse how predictive behaviour and explanation quality vary across tasks. Through this explainability analysis, the study examines the decision-making behaviour of the evaluated models, highlighting reasoning patterns and potential biases, and assessing their alignment with operational and transparency requirements.

Ringraziamenti

Al termine di questo percorso, il mio grazie più sincero va a chi mi ha accompagnato con affetto e costante incoraggiamento, rendendo possibile questo traguardo.

Alla mia famiglia, il mio porto sicuro. Grazie ai miei genitori e a mia sorella: siete stati la mia forza nei momenti più duri, pronti a tirarmi su e instancabili nel non farmi mai sentire la mancanza di casa, nonostante la distanza. Un grazie speciale alle mie nonne, esempi viventi di determinazione e saggezza; grazie per i vostri consigli preziosi e per quel calore che solo voi sapete dare. Grazie a tutto il resto della mia famiglia: sapere che avete sempre creduto in me è stata la spinta più grande.

Ai miei amici di sempre, grazie per essere stati la mia valvola di sfogo. Grazie per ogni risata e per ogni momento di svago che mi ha permesso di staccare dallo stress ogni volta che tornavo a casa. Senza di voi, questo percorso sarebbe stato decisamente più grigio.

Ai miei coinquilini, Lorenzo, Dani e Carlo. Non è facile trovare persone come voi. Se in questi due anni ho vissuto così bene e ho raggiunto questo obiettivo, è soprattutto merito vostro. Grazie per avermi spronato quando ne avevo bisogno e per aver creato un ambiente in cui mi sono sentito sempre a mio agio. Tornare a casa, che fosse dopo le lezioni o dopo un viaggio dalla Sicilia, e trovare persone rispettose, simpatiche e con cui poter condividere tutto, ha reso questa esperienza indimenticabile.

Ai ragazzi di JEToP. Chi l'avrebbe mai detto? Io in questa associazione non volevo nemmeno entrare, e invece, grazie a Dani e Lore che mi hanno quasi “forzato” a farlo, ho scoperto un mondo nuovo. Grazie per avermi dato l'opportunità di conoscere persone fantastiche e vivere esperienze inimmaginabili. Grazie perché questa realtà mi ha permesso di crescere, di mettermi in gioco e, sotto certi aspetti, di rivoluzionarmi completamente, portandomi a fare cose che mai avrei pensato fossero alla mia portata.

Infine, un pensiero a chi non c'è più fisicamente, ma la cui presenza resta costante nel mio cuore. Grazie ai miei nonni e alla mia sorellina: so che da lassù mi avete sempre protetto, dandomi la forza e la determinazione per non arrendermi mai.

In particolare, oggi ricorre il sesto anniversario della scomparsa di mio **nonno Melo**. A te dedico questa tesi e questo intero traguardo. Se in questo giorno, tipicamente grigio per le nostre vite, da oggi splenderà un raggio di sole, so che è merito del tuo “zampino”. Porterò sempre dentro di me i tuoi insegnamenti e la tua saggezza; sei e sarai per sempre il mio esempio da seguire. Spero di rendervi sempre orgogliosi: ogni mio futuro traguardo sarà dedicato a voi.

Table of Contents

List of Figures	VIII
Acronyms	X
1 Introduction	1
1.1 Background and Motivation	1
1.2 Explainable AI and the Transparency Problem	2
1.3 Regulatory Drivers for AI Transparency	4
1.4 Problem Statement and Research Gap	6
1.5 Research Questions and Objectives	8
1.6 Contributions of this Thesis	10
1.7 Thesis Structure	11
2 State of the Art in XAI	13
2.1 Fundamental dimensions of XAI	13
2.2 Method centric taxonomies of XAI techniques	15
2.2.1 Functioning-based classifications	15
2.2.2 Result-based classifications	16
2.2.3 Mixed and operational taxonomies	17
2.3 Conceptually oriented taxonomies and socio-technical perspectives	17
2.3.1 Stakeholders and explanation needs	18
2.3.2 Purposes of explanations: trust, control and compliance	18
2.3.3 The Tower of Babel problem	19
2.4 Limitations of existing taxonomies and open challenges	19
2.5 Towards a practice oriented framework	20
2.5.1 Synthesis of relevant dimensions	21
2.5.2 Alignment with trustworthy and regulated AI requirements	21
2.5.3 Implications for high-impact domains	22

3	The Regulatory Landscape for AI Transparency	24
3.1	From Opacity to Explainability: A Regulatory Problem Framing	25
3.2	Data Protection, Profiling, and Automated Decision-making: GDPR as a Baseline	27
3.2.1	GDPR Principles Relevant to AI Systems	27
3.2.2	Profiling, Article 22 GDPR, and the Debate on a Right to Explanation	28
3.2.3	Accountability, DPIA, and Algorithmic Traceability	29
3.3	The EU AI Act: Risk-based Transparency and Explainability Duties	30
3.3.1	Risk Taxonomy and AI System Categories	30
3.3.2	Transparency, Documentation, and Human Oversight for High-risk AI	30
3.3.3	General Purpose and Generative AI: Transparency for LLMs and SLMs	31
3.4	Cybersecurity Oriented Regulation: NIS2, Cyber Resilience Act, and DORA	32
3.4.1	NIS2: Security by Design, Logging, and Incident Reporting	32
3.4.2	The Cyber Resilience Act: Secure and Transparent Digital Products	33
3.4.3	DORA and ICT Risk Management in the Financial Sector	33
3.5	Soft Law Frameworks and Technical Standards for Trustworthy and Ex- plainable AI	34
3.5.1	EU Ethics Guidelines for Trustworthy AI	34
3.5.2	OECD Principles, NIST AI RMF, and ISO and IEC Standards . . .	34
3.5.3	The Tower of Babel Problem: Misaligned Notions of Explainability	35
3.6	From Legal Requirements to Design Principles for XAI SLM Systems in Cybersecurity	35
3.6.1	Regulatory Derived Design Principles	35
3.6.2	Stakeholders and Explanation Recipients in Security Operations . .	36
3.6.3	Gaps, Tensions, and Open Issues	37
3.7	Interim Conclusions and Research Questions	37
4	Small Language Models in Cybersecurity	39
4.1	Defining Small Language Models for Security Operations	39
4.2	General-purpose and agentic SLMs: lessons for cybersecurity	42
4.3	Cybersecurity expert SLMs: the case of CyberPal 2.0	43
4.4	Deployment regimes and constraints for SLMs in cybersecurity	46
4.5	Interim conclusions and selection of SLMs for the experimental study . . .	48
5	Methodology for XAI–SLM Experimentation	50
5.1	Research design and overall approach	51
5.2	Selection of cybersecurity use cases and datasets	52
5.3	Selection of SLMs and model selection criteria	53
5.4	Portfolio of XAI techniques and evaluation dimensions	55

5.4.1	Selected families of XAI methods	55
5.4.2	Excluded families and rationale	56
5.4.3	Evaluation dimensions for explanations	57
5.5	Evaluation metrics and experimental protocol	58
5.6	Threats to validity and limitations	60
6	Implementation and Results	62
6.1	Scope and Goals of the Artifact	62
6.2	Framework Overview	63
6.2.1	Configuration Parser	64
6.2.2	Dataset Loader	64
6.2.3	Model Wrapper and Evaluation Adapter	64
6.2.4	Explainability Orchestrator	64
6.2.5	Reporting Engine	65
6.3	Configuration and Execution	67
6.4	Outputs and Generated Artifacts	68
6.5	Results Presentation and Analysis	72
6.5.1	Comparative Overview	72
6.5.2	Predictive Performance Analysis	72
6.5.3	Explainability Results Across Methods	73
6.5.4	Representative Local Explanations	74
6.6	Operational and Audit Discussion	75
7	Conclusions and Future Work	81
7.1	Summary of Key Contributions	82
7.1.1	Conceptual and taxonomical contribution	82
7.1.2	Regulatory-technical mapping	82
7.1.3	Empirical evaluation of XAI on lightweight open-weight security models	82
7.1.4	Framework and prototype contribution	83
7.2	Revisiting the Research Questions	83
7.2.1	RQ1 – XAI methods for SLMs in cybersecurity	83
7.2.2	RQ2 – Effectiveness and limitations of XAI on lightweight open-weight SLMs, with implications for constrained deployments	84
7.2.3	RQ3 – Regulatory alignment of XAI explanations for SLM-based systems	84
7.2.4	RQ4 – Framework design for XAI-oriented SLM compliance	85
7.3	Operational Implications for Security Operations and Audit Readiness	85
7.4	Limitations	86
7.5	Future Work	86

7.5.1	Extending the empirical scope of cybersecurity tasks	87
7.5.2	Comparison with general-purpose open-weight baselines	87
7.5.3	Systematic comparison across deployment regimes	87
7.5.4	Richer explanation modalities for security text	87
7.5.5	Human-centred validation	88
7.5.6	Continuous evaluation and drift-aware governance	88
7.5.7	Compliance-by-design reporting	88
7.6	Final Remarks	88
Bibliography		90

List of Figures

6.1	Configuration Parser Flowchart	65
6.2	Dataset Loader Flowchart	65
6.3	Model Wrapper and Evaluation Adapter Flowchart	66
6.4	Explainability Orchestrator Flowchart	67
6.5	Reporting Engine Flowchart	67
6.6	Command-line execution for the use cases	68
6.7	Example of an artifact directory	71
6.8	Representative extracts from generated reports	77
6.9	Confusion matrices for phishing and vulnerability tasks	78
6.10	Comparison of XAI metrics across methods	79
6.11	Representative token-level local explanations	80

Acronyms

AI

Artificial Intelligence

XAI

Explainable Artificial Intelligence

LLM

Large Language Model

SLM

Small Language Model

SOC

Security Operations Centre

API

Application Programming Interface

GDPR

General Data Protection Regulation

AI Act

European Union Artificial Intelligence Act

NIS2

Network and Information Security Directive 2

CRA

Cyber Resilience Act

EU

European Union

RQ

Research Question

ML

Machine Learning

SIEM

Security Information and Event Management

DORA

Digital Operational Resilience Act

DPIA

Data Protection Impact Assessment

CLI

Command Line Interface

CVE

Common Vulnerabilities and Exposures

Chapter 1

Introduction

The growing reliance on data-driven and learning-based technologies in security-sensitive environments raises fundamental questions about how complex AI systems can be made transparent, accountable, and amenable to scrutiny. This thesis addresses these questions by focusing on the role of XAI in understanding and governing the behaviour of SLMs deployed in cybersecurity workflows. Building on recent advances in language modelling, the emergence of SLMs as efficient and adaptable alternatives to large-scale models, and the parallel evolution of regulatory frameworks that emphasise transparency and human oversight, the work develops a structured analysis spanning technical, organisational, and legal perspectives. The remainder of this chapter introduces the background and motivation for the study, formulates the core problem and research questions, outlines the methodological approach, and summarises the main contributions and overall structure of the thesis.

1.1 Background and Motivation

Over the last decade, advances in machine learning and, in particular, deep neural networks have led to the widespread deployment of opaque, high-capacity AI models in safety- and security-critical domains such as healthcare, finance, industrial control, and cybersecurity [1, 2]. The increasing deployment of sophisticated AI systems, particularly in high-stakes domains such as cybersecurity, is driven by the need to cope with rapidly growing volumes of heterogeneous data, increasingly complex infrastructures, and persistent shortages of specialised human resources [3, 4]. Organisations adopt data-driven, learning-based solutions to automate detection, prioritisation, and response activities, aiming to improve coverage, speed, and consistency in security operations. At the same time, this shift towards deep neural architectures and large-scale language models has been accompanied by a marked rise in model complexity and opacity. As organisations adopt these models to maximise predictive and generative capabilities, the internal decision processes of the resulting systems often become less accessible to human stakeholders, widening the gap between effectiveness and interpretability [5, 6, 7].

In the cybersecurity domain, the trade-off between performance and transparency is particularly acute. Organisations must cope with increasingly sophisticated adversaries, large volumes of heterogeneous security data (e.g., logs, alerts, vulnerability reports, threat intelligence feeds), and stringent requirements on detection speed, coverage, and resilience [3]. As a consequence, machine-learning- and AI-based tools are progressively being embedded into SOC’s, intrusion detection systems, vulnerability management platforms, and incident response workflows, where they support analysts in tasks such as anomaly detection, alert correlation, and incident triage [3, 8]. Recent surveys on XAI for cybersecurity emphasise that explainability is not merely a desirable property but a prerequisite for operational trust, effective human–AI collaboration, and meaningful auditing of AI-assisted decisions [3, 2]. If an AI-based system flags a vulnerability as critical, correlates disparate alerts into a single incident, or recommends specific mitigation actions, security professionals must be able to understand why this decision was produced, which evidence and assumptions underlie it, and under which conditions the recommendation might fail. Without such explanations, it becomes difficult to calibrate trust in AI-assisted workflows, to identify systematic failure modes and biases, and to integrate AI into wider risk management, compliance, and governance processes in a responsible manner [9, 2].

In parallel with these developments, language models have rapidly evolved from research prototypes to central components of many AI-based systems. While LLMs have attracted most of the public and scientific attention, recent surveys argue that SLMs are increasingly relevant as a more efficient and controllable alternative, particularly in resource-constrained or domain-specific settings [10, 11]. SLMs offer an attractive trade-off between performance, resource consumption, and deployment flexibility: they can be executed on-premises, integrated with existing security tooling, and fine-tuned on proprietary datasets, thereby reducing latency and privacy risks compared to fully cloud-based solutions [10]. Initial works have started to explore the use of SLMs as cybersecurity-oriented assistants and “expert” models for tasks such as log analysis, threat triage, and vulnerability reasoning [12, 8]. However, it remains unclear to what extent existing XAI methods, originally developed for generic deep learning or large-scale models [5, 1, 13], are applicable, reliable, and sufficient for explaining SLMs in realistic cybersecurity scenarios. Furthermore, current XAI taxonomies and evaluation practices seldom take into account the specific needs of cybersecurity stakeholders (e.g., SOC’s analysts, incident responders, auditors) and the constraints imposed by emerging transparency and accountability requirements [13, 14, 2]. These gaps motivate a focused investigation into XAI techniques tailored to SLMs in cybersecurity, with particular attention to their practical utility and their potential role in supporting regulatory-compliant AI system design and governance.

1.2 Explainable AI and the Transparency Problem

XAI has gradually emerged as an umbrella term for a heterogeneous set of concepts, methods, and design principles that aim to make complex AI systems more understandable to human stakeholders. Despite the apparent consensus on the need for explainability, there is still no widely agreed definition of what constitutes an “explanation” in AI, nor of how it should be evaluated in different contexts [5, 1, 9]. The literature frequently distinguishes

between related notions such as *interpretability*, *explainability*, and *transparency*, yet these terms are often used inconsistently across disciplines and even within the same community [13, 14]. For instance, some authors reserve “transparency” for models whose internal mechanisms are inherently understandable (e.g., sparse linear models or small decision trees), whereas “explainability” is associated with post-hoc techniques that generate additional artefacts (feature attributions, visualizations, counterfactuals) to render opaque models more comprehensible [7, 6]. Others adopt a broader socio-technical view, framing explainability as a form of communication between an AI system and its users, in which explanations must be tailored to the goals, background knowledge, and constraints of different stakeholders [1, 13, 9]. This conceptual diversity reflects the richness of the field but also contributes to a “Tower of Babel” effect, where technical, ethical, and legal discourses about transparency risk talking past one another [14].

To bring structure to this landscape, several surveys and reviews have proposed taxonomies of XAI methods based on different classification criteria [5, 1, 13]. A common distinction is between *intrinsic* (or ante-hoc) and *post-hoc* approaches. Intrinsic methods seek to use inherently interpretable models or architectures, such as rule-based systems, generalized additive models, attention mechanisms, or concept-based representations, so that explanations can be derived directly from model structure and parameters. Post-hoc methods, by contrast, take an already trained model, often a highly non-linear black box, and generate explanations of its outputs through auxiliary techniques, including feature attribution (e.g., gradient-based saliency, SHAP, LIME), surrogate models, example-based and prototype-based explanations, or counterfactual reasoning [7, 15]. Other taxonomies further refine this picture by distinguishing between global and local explanations, model-specific versus model-agnostic techniques, and different explanation modalities, such as visual, textual, or example-based artefacts [1, 4]. Together, these works highlight that explanations are inherently multi-dimensional objects, whose adequacy depends not only on their technical properties but also on their alignment with user needs and task requirements.

Despite this substantial body of work, systematic critiques of XAI emphasise that many existing techniques exhibit important limitations when scrutinised from the perspective of transparency and accountability in real-world systems [6, 9, 2]. For example, feature attribution methods may produce explanations that are highly sensitive to small perturbations in the input or to implementation details, undermining their stability and reliability as tools for debugging and auditing [7]. Surrogate models can offer intuitive summaries of a black-box model’s behaviour, but only within a restricted data manifold and at the cost of introducing an additional layer of approximation whose fidelity is often difficult to quantify rigorously. Moreover, explanations that are formally faithful to the underlying model may still be unusable in practice if they are too complex, too technical, or poorly aligned with the mental models and time constraints of their intended users. Several authors have warned that XAI artefacts can foster a false sense of understanding or trust, leading to what has been called “explanation theatre”: explanations that satisfy superficial demands for transparency without genuinely improving oversight, contestability, or error detection [9, 14]. These issues are particularly critical in high-stakes domains, where explanations may have to support not only technical debugging but also legal justification, documentation, and accountability.

Another important challenge concerns the evaluation of explanations. While there is growing agreement that multiple dimensions, such as fidelity, stability, completeness, complexity, and user satisfaction, should be taken into account, there is no consensus on standard metrics or benchmarks, especially across heterogeneous application domains [5, 6, 4]. Many studies focus on synthetic tasks or small-scale user studies, limiting the generalisability of their conclusions. In addition, comparatively little work has explored how explanations interact with domain-specific workflows, organizational practices, and regulatory requirements. For instance, explanations that might be acceptable for internal debugging by data scientists could be insufficient for external auditors, regulators, or affected individuals who require a different level of detail, justification, or documentation. This misalignment between technical XAI proposals and the needs of diverse stakeholders is at the core of the broader *transparency problem*: the difficulty of ensuring that explanations meaningfully support understanding, oversight, and responsibility in the socio-technical environments where AI systems are deployed [2, 8].

In the context of this thesis, these observations have two main implications. First, they call for a critical rather than merely appreciative use of XAI methods: explanations generated for SLMs in cybersecurity should not be taken at face value but examined with respect to their fidelity, robustness, and user-centred usefulness. Second, they highlight the need to align XAI techniques with explicit requirements arising from governance and regulatory frameworks, which increasingly demand transparency, documentation, and accountability from AI-based systems. Existing taxonomies and surveys seldom integrate such normative constraints into their classification schemes or evaluation criteria [13, 14]. As a result, there is limited guidance on how to select, configure, and deploy XAI techniques in a way that systematically supports compliance, especially for SLM-based systems operating in cybersecurity settings. Addressing this gap requires an integrated perspective that connects the technical properties of explanations, the characteristics of SLMs, and the evolving expectations of regulators and other institutional stakeholders.

1.3 Regulatory Drivers for AI Transparency

The rapid diffusion of AI systems into socio-technical infrastructures, including critical digital services, public administration, healthcare, and financial systems, has been accompanied by an evolving legal and regulatory landscape that increasingly emphasizes transparency, explainability, and accountability. In the European Union, several instruments form a layered framework: Regulation (EU) 2024/1689 on artificial intelligence (the Artificial Intelligence Act, AI Act), Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (the General Data Protection Regulation, GDPR), Directive (EU) 2022/2555 on measures for a high common level of cybersecurity across the Union (the NIS2 Directive), Regulation (EU) 2022/2554 on digital operational resilience for the financial sector (DORA), and Regulation (EU) 2024/2847 on horizontal cybersecurity requirements for products with

digital elements (the Cyber Resilience Act, CRA)¹. While their scopes and objectives differ, they collectively articulate expectations regarding documentation, risk management, human oversight, and the intelligibility of automated decision-making processes, especially where fundamental rights, critical infrastructures, or essential services are at stake [2, 9].

Within this landscape, the AI Act plays a central role as the first comprehensive attempt to regulate AI systems at Union level. It establishes a risk-based taxonomy that distinguishes, among other things, prohibited practices, high-risk systems, certain limited-risk systems subject to specific transparency duties, and minimal-risk uses. For high-risk AI systems, including many applications in critical infrastructure management, access to essential services, and certain security and fraud-detection tools, the Act lays down horizontal requirements on risk management, data and data governance, technical documentation, logging, transparency, and human oversight [16]. Among these requirements, transparency occupies a prominent position: high-risk systems must be designed and developed in such a way that their operation is sufficiently transparent to enable deployers to interpret the system’s outputs and use them appropriately, and providers must supply clear, complete, and correct instructions describing the system’s intended purpose, performance characteristics, and limitations². These obligations are complemented by specific rules for general-purpose AI models that pose systemic risk, which must comply with additional requirements for technical documentation, model evaluation, and incident reporting. In this sense, the expectation that AI behaviour should be intelligible and traceable is not confined to a narrow subset of high-risk applications, but extends to a broader class of systems that shape decision-making in socio-technical environments [8, 2].

Although the AI Act does not mandate particular XAI techniques, it implicitly presupposes that deployers must be able to understand, document, and justify the behaviour of AI systems in a manner that is commensurate with the risks they pose. From the perspective of SLM-based cybersecurity tools, this implies that explanations cannot be purely informal or anecdotal: they must provide evidence that links model outputs to documented data, design choices, and controls, so that they can be scrutinised by internal and external auditors and incorporated into the technical documentation required by the Act [8, 9].

Other instruments reinforce these expectations at different layers. The NIS2 Directive strengthens cybersecurity obligations for entities that are classified as essential or important, including financial institutions, digital infrastructure providers, and certain public administrations, by requiring them to implement risk management measures, incident reporting procedures, and governance arrangements for their network and information systems. In particular, organisations must document security policies, maintain appropriate logging and monitoring capabilities, and report significant incidents within strict timeframes [17]. Although NIS2 is technology-neutral and does not refer explicitly to AI,

¹See Regulation (EU) 2024/1689 (Artificial Intelligence Act), Regulation (EU) 2016/679 (GDPR), Directive (EU) 2022/2555 (NIS2), and Regulation (EU) 2024/2847 (Cyber Resilience Act).

²See Articles 13 and 14 (*transparency and human oversight*) and Article 50 (*transparency obligations for providers and users of certain AI systems*) of Regulation (EU) 2024/1689 (Artificial Intelligence Act) [16].

AI-enabled security tools deployed for detection, monitoring, or incident handling become part of the documented security posture and may themselves be subject to supervisory scrutiny. The CRA, in turn, introduces mandatory cybersecurity requirements for products with digital elements, including obligations relating to secure-by-design development, vulnerability handling processes, and the provision of information enabling users to understand and manage the security properties of products throughout their life cycle [18]. For security products embedding AI components, this implies that transparency is not limited to exposing vulnerabilities or update practices, but also extends to clarifying the conditions under which embedded models may fail, behave unexpectedly, or require human oversight [2]. Taken together, these frameworks steer technical practice towards forms of transparency and explainability that are operational, auditable, and closely tied to risk governance, rather than purely conceptual or user-experience-oriented [9].

Data protection law adds another layer of transparency obligations, particularly in relation to automated decision-making that significantly affects individuals. Article 22 of the GDPR [19] grants data subjects the right not to be subject to decisions based solely on automated processing, including profiling, that produce legal or similarly significant effects, while Articles 13–15 and Recital 71 are commonly interpreted as requiring controllers to provide meaningful information about the logic involved in such processing and about the envisaged consequences for the data subject. Recent case law has begun to clarify the practical implications of these provisions. In *CK v Dun & Bradstreet Austria GmbH*, the Court of Justice of the EU held that, in the context of automated credit scoring, data subjects are entitled to an explanation that enables them to understand which of their personal data were used, how they were processed, and how variations in those data would have affected the outcome, and that trade secrecy cannot be invoked in a way that systematically undermines this right [20, 21]. This jurisprudence reinforces the idea that transparency obligations are not satisfied by the mere disclosure of source code or generic algorithmic descriptions: regulators expect explanations that are intelligible to affected individuals and capable of supporting contestation and human review.

However, as discussed in Section 1.2, there remains a significant mismatch between the abstract regulatory language used to define transparency or explainability and the concrete form taken by most XAI techniques. Surveys of the field point to a proliferation of heterogeneous taxonomies, evaluation criteria, and user assumptions, a “Tower of Babel” that complicates both technical progress and regulatory alignment [13, 14]. From the viewpoint of SLM-based cybersecurity applications, this means that meeting regulatory expectations cannot simply be reduced to selecting a popular attribution method: explanations must be embedded into end-to-end AI workflows in ways that support documentation, monitoring, incident investigation, and external audit. Clarifying how to move from high-level regulatory requirements to actionable, domain-specific XAI practices therefore constitutes a central challenge and a key motivation for the present work.

1.4 Problem Statement and Research Gap

The developments outlined in the previous sections converge towards a common tension: on the one hand, organisations increasingly rely on AI-based systems, including language

models, to support critical cybersecurity functions; on the other hand, both technical and regulatory considerations demand that the behaviour of such systems be sufficiently transparent, explainable, and accountable. Existing work on XAI has produced a rich catalogue of methods, taxonomies, and conceptual frameworks [5, 1, 7, 13], and there is a growing body of research that applies XAI techniques to security-relevant tasks such as intrusion detection, malware classification, and anomaly detection [3, 4]. In parallel, regulatory instruments such as the AI Act, the GDPR, NIS2, and the CRA articulate increasingly concrete transparency and documentation obligations for AI systems deployed in high-risk or security-sensitive settings [8, 9]. Nevertheless, there remains a significant gap between these high-level requirements and the concrete XAI practices that can be deployed to support them, particularly for SLMs integrated into operational cybersecurity workflows.

From a technical perspective, most XAI research relevant to cybersecurity has focused on models trained on structured or low-level data, such as tabular features, network flows, or byte-level representations of binaries [3]. Explanations typically take the form of feature attributions, rule lists, or prototype examples that highlight which input variables contributed most to a detection outcome. While these approaches are valuable for traditional security analytics, they do not readily transfer to language-based representations and reasoning, where inputs and outputs are expressed in natural or domain-specific text. In particular, SLM-based systems for cybersecurity, for instance, assistants that summarise logs, prioritise alerts, or reason about vulnerability reports [12, 10], operate on high-dimensional, context-dependent embeddings and generate free-form textual outputs. Explaining such systems requires methods that can capture not only which tokens or concepts influence a prediction, but also how the model structures its internal reasoning and how this reasoning maps to security-relevant constructs. Existing XAI techniques for natural language processing have begun to address some of these challenges, using attention visualisation, gradient-based saliency, or example-based explanations; however, their applicability to SLMs in security-specific tasks has rarely been examined in a systematic and domain-aware manner [5, 15].

From a governance and regulatory perspective, there is also a clear mismatch between the way transparency and explainability are conceptualised in legal and policy documents and the way they are operationalised in technical XAI work. As discussed earlier, regulatory instruments often invoke terms such as transparency, explicability, interpretability, and traceability without providing precise technical definitions or guidance on which types of explanation are appropriate in specific contexts [13, 14]. At the same time, XAI taxonomies and evaluation frameworks rarely incorporate explicit regulatory or compliance-oriented criteria into their design. For example, many studies evaluate explanations solely in terms of fidelity, stability, or user preference, without considering whether they provide the kind of evidence that auditors, regulators, or affected individuals would require to assess compliance with obligations under the AI Act or the GDPR [6, 9, 2]. This disconnect is particularly problematic in cybersecurity, where explanations may need to satisfy heterogeneous and sometimes conflicting requirements: they must be sufficiently detailed to support accountability and incident investigation, yet not so revealing as to compromise sensitive information about detection logic, threat intelligence sources,

or defensive capabilities [8]. To date, there is limited work that explicitly links XAI methods for security systems with concrete regulatory transparency and documentation requirements, or that proposes systematic frameworks for evaluating explanations in compliance-aware terms.

A further gap concerns the heterogeneity of language model deployment models and their implications for explainability. In practice, organisations may rely on a mix of open-weight, semi-open, and closed SLMs: open-weight models whose weights and sufficient technical documentation are available; semi-open models that expose partial documentation or APIs but not full internals; and proprietary closed models accessed via remote services [10, 11]. Each of these categories raises different opportunities and constraints for XAI. Open SLMs, in principle, allow for more intrusive inspection and the use of gradient-based or concept-based explanation methods; semi-open and closed models may restrict explanations to model-agnostic techniques or to logs and documentation provided by the vendor. Yet, existing XAI studies seldom differentiate between these deployment regimes in a systematic way, nor do they investigate how explanation strategies should adapt to the varying degrees of access, control, and responsibility that they entail. In cybersecurity, where regulatory obligations often fall on the deployer rather than the model provider, this raises practical questions about how organisations can fulfil transparency, documentation, and oversight requirements when using externally hosted SLMs as part of their security tooling [8]. Addressing these questions requires an integrated view that considers not only the technical properties of explanations but also the organisational and contractual arrangements under which SLMs are deployed.

Taken together, these observations define the core problem addressed by this thesis. There is a need for a structured investigation into how XAI techniques can be selected, adapted, and evaluated for SLMs used in cybersecurity scenarios, in a way that explicitly accounts for regulatory transparency and accountability requirements. More concretely, the research gap can be articulated along three dimensions. First, there is a lack of systematic, domain-specific analyses of XAI methods for SLMs in security tasks such as vulnerability detection, alert triage, and incident analysis. Second, there is no widely adopted framework that connects the properties of explanations (e.g., fidelity, stability, user comprehensibility) to the obligations emerging from instruments such as the AI Act, GDPR, NIS2, and the CRA, particularly in the context of SLM-based security systems. Third, empirical evidence on the effectiveness and limitations of XAI methods across different categories of SLMs (open, semi-open, closed) in realistic cybersecurity workflows remains scarce. This thesis addresses these gaps by combining a critical review of XAI and regulatory literature with targeted experimentation on representative SLMs and by proposing a novel XAI-oriented framework for analysing and guiding the design of SLM-based cybersecurity systems in a compliance-aware manner.

1.5 Research Questions and Objectives

The problem and research gap outlined above can be made more concrete by formulating a set of research questions that guide the theoretical and empirical components of this thesis. At a high level, the overarching goal is to understand how XAI techniques can be

systematically leveraged to make SLMs used in cybersecurity more transparent, accountable, and aligned with emerging regulatory requirements. To achieve this, the thesis focuses on four main research questions (RQs), which are interrelated but address distinct dimensions of the problem.

RQ1 - XAI methods for SLMs in cybersecurity. *Which families of explainable AI methods are most suitable for providing meaningful explanations of SLMs used in cybersecurity-relevant tasks, and how can existing XAI taxonomies be adapted to capture their specific properties in this context?* This question aims to connect the general XAI literature with the particular characteristics of SLM-based systems and security workflows. It requires identifying relevant explanation techniques (e.g., feature attribution, example-based, surrogate, concept-based, and hybrid methods), analysing their assumptions and outputs, and assessing their potential applicability to SLM-driven tasks such as vulnerability analysis, log summarisation, and alert triage.

RQ2 - Effectiveness and limitations of XAI on lightweight open-weight SLMs, with implications for constrained deployments. *How do selected XAI techniques perform when applied to lightweight open-weight SLMs in concrete cybersecurity use cases, in terms of fidelity, stability, and practical usefulness, and what implications do these findings suggest for more constrained semi-open and closed deployment regimes?* This question introduces an empirical dimension centred on SLMs that can be executed and inspected locally, while treating semi-open and closed models as analytical reference cases. It calls for the design of an experimental setup in which representative lightweight open-weight SLMs are evaluated on security-related tasks under different explanation methods, with metrics that capture technical properties and qualitative aspects of operational usefulness. The broader deployment regimes are then used to interpret which findings are likely to transfer, and which explanation options become unavailable under black-box constraints.

RQ3 - Regulatory alignment of XAI explanations for SLM-based systems. *To what extent can existing XAI methods, when applied to SLMs in cybersecurity scenarios, support compliance with transparency, documentation, and accountability requirements emerging from instruments such as the AI Act, GDPR, NIS2, and the CRA?* This question addresses the interface between technical XAI artefacts and legal or governance expectations. It requires mapping explanation properties (e.g., level of detail, traceability, stability over time, documentation affordances) to specific obligations and principles articulated in the relevant regulatory frameworks, and identifying gaps where existing methods may be insufficient or misaligned.

RQ4 - Framework design for XAI-oriented SLM compliance. *How can we design an XAI-oriented framework that systematically links SLM behaviour, explanation techniques, and regulatory requirements in cybersecurity settings, and how can this framework be instantiated in a practical methodology or prototype?* This question synthesises the theoretical and empirical insights of the previous RQs into a structured model or framework. The aim is to provide guidance on how organisations can select, configure, and evaluate explanation strategies for SLM-based security systems in a way that is both technically grounded and compliance-aware, and, where feasible, to demonstrate this via a proof-of-concept implementation.

These research questions are reflected in the concrete objectives of the thesis, which can be summarised as follows:

- **O1 - Critical synthesis of XAI and SLM literature in cybersecurity.** Develop a structured and critical review of existing XAI taxonomies, methods, and evaluation approaches, with a specific focus on their applicability to SLMs in cybersecurity-relevant tasks. This includes identifying strengths, limitations, and open issues that are particularly salient for security workflows.
- **O2 - Analysis of regulatory requirements for AI transparency in cybersecurity.** Extract and systematise transparency-, documentation-, and accountability-related requirements from relevant regulatory instruments (AI Act, GDPR, NIS2, CRA), highlighting their implications for AI-based cybersecurity systems and for the design and evaluation of explanations produced by SLMs.
- **O3 - Design and execution of an XAI-SLM experimental study.** Define an experimental methodology in which lightweight open-weight SLMs are applied to one or more cybersecurity use cases, select and implement suitable XAI techniques, and evaluate the resulting explanations using quantitative metrics centred on fidelity and stability, together with qualitative analysis of their operational plausibility for security workflows. Semi-open and closed deployments are treated as analytical reference cases for discussing transferability and black-box constraints.
- **O4 - Proposal of an XAI-oriented framework for SLM compliance.** On the basis of the literature review, regulatory analysis, and experimental results, propose a novel framework that connects SLM behaviour, XAI methods, and regulatory requirements in cybersecurity settings, and outline (or implement) a prototype workflow or toolchain that instantiates this framework in practice.

By articulating the research questions and objectives in this way, the thesis sets the stage for a coherent progression from conceptual foundations and regulatory analysis (Chapters 2 and 3) to methodological design (Chapter 5) and empirical investigation (Chapter 6).

1.6 Contributions of this Thesis

Situated at the intersection of XAI, SLMs, cybersecurity, and emerging AI governance frameworks, this thesis aims to provide contributions that are both conceptually grounded and practically oriented. Building on the problem statement, research questions, and methodological approach outlined above, the main contributions can be summarised along four dimensions.

- **Conceptual and taxonomical contribution.** The thesis develops a critical synthesis of existing XAI taxonomies and methods with a specific focus on their applicability to SLMs in cybersecurity-relevant tasks. By revisiting and reorganising

XAI categories in light of SLM characteristics (e.g., transformer-based architectures, textual inputs and outputs, deployment regimes) and security workflow requirements, it proposes a refined conceptual mapping that highlights which families of explanation techniques are *prima facie* suitable for SLM-based systems and which aspects remain underexplored.

- **Regulatory-technical mapping.** The thesis conducts a structured analysis of key regulatory instruments that shape expectations of transparency, documentation, and accountability for AI systems in security-sensitive settings, including the AI Act, the GDPR, the NIS2 Directive, and the CRA. It then translates these high-level legal and policy requirements into a set of technical and procedural criteria that explanations and associated governance practices should satisfy in the context of SLM-based cybersecurity systems. This mapping provides a bridge between normative concepts of transparency and the operational properties of XAI artefacts.
- **Empirical assessment of XAI techniques on SLMs in cybersecurity use cases.** Through a targeted experimental study on lightweight open-weight SLMs applied to representative cybersecurity use cases, the thesis evaluates selected XAI methods through fidelity- and stability-oriented metrics and qualitative analysis of explanation plausibility. The empirical results shed light on the practical strengths and limitations of different explanation strategies for locally deployable SLMs, while the broader discussion clarifies how deployment regimes constrain the choice and effectiveness of XAI techniques in real-world security settings.
- **XAI-oriented framework for SLM compliance in cybersecurity.** Drawing on the conceptual synthesis, regulatory analysis, and empirical findings, the thesis proposes a novel framework that systematically links (i) the characteristics of SLM-based cybersecurity applications, (ii) the design space of XAI methods, and (iii) the transparency and accountability requirements emerging from the relevant regulatory landscape. The framework provides guidance on how to select, configure, and evaluate explanation strategies for SLM-based security systems in a compliance-aware manner, and it is instantiated as a methodological workflow and, where feasible, as a proof-of-concept prototype integrating XAI into security tooling.

Taken together, these contributions aim to advance the understanding of how XAI can be operationalised for SLMs in cybersecurity, not only from a technical perspective but also in relation to the governance and regulatory constraints that increasingly shape the design and deployment of AI-based security systems.

1.7 Thesis Structure

The remainder of this thesis is organised as follows.

Chapter 2 reviews the state of the art on XAI, including core concepts, taxonomies, and evaluation dimensions, with particular emphasis on security-relevant settings. This

chapter provides the conceptual basis for structuring the XAI design space and supports Research Question RQ1.

Chapter 3 analyses the EU regulatory and governance landscape relevant to transparency and accountability in AI-enabled cybersecurity, with focus on the GDPR, the AI Act, NIS2, the CRA, and DORA. It clarifies how legal and compliance expectations shape explainability requirements, addressing Research Question RQ3.

Chapter 4 focuses on SLMs in cybersecurity, discussing deployment regimes and access constraints (open-weight, semi-open, and closed models) and their implications for explainability and evaluation. This chapter provides the technical and operational context for Research Question RQ2.

Chapter 5 details the research methodology and experimental protocol, including use-case and dataset selection, model and XAI technique selection, and evaluation metrics. This chapter operationalises Research Questions RQ1 and RQ2.

Chapter 6 introduces the proposed evaluation-oriented XAI framework for SLM-based cybersecurity applications and reports on the implementation and empirical results of the study. It presents the experimental setup, model configurations, and the comparative assessment of XAI techniques, including discussion of trade-offs and practical constraints, providing direct evidence to address Research Question RQ2 and informing the framework validation.

Finally, Chapter 7 concludes the thesis. It summarises the key contributions, revisits the research questions in light of the results, and outlines limitations and directions for future work on trustworthy and responsible AI.

Chapter 2

State of the Art in XAI

This chapter organises the main approaches to XAI into a conceptual framework that will be used throughout the rest of the thesis. While Chapter 1 introduced the motivation for explainability and positioned XAI within the broader landscape of AI systems, the focus here is on structuring the diverse set of methods, definitions and perspectives that have emerged in the literature. The goal is not to propose yet another taxonomy, but to clarify the dimensions along which existing contributions can be compared and combined in a way that is meaningful for both research and practice, with a view towards security-critical and regulated applications.

Several surveys and reviews already provide overviews of XAI methods, often proposing their own taxonomies and classification schemes [5, 1, 7, 13, 2, 3, 4]. At the same time, more recent work has highlighted how the rapid growth of the field has produced a certain fragmentation of terminology and conceptual foundations, sometimes described as a Tower of Babel effect [14]. This situation complicates both scientific communication and the translation of regulatory requirements into concrete design choices for XAI in cybersecurity and other high-impact domains.

The remainder of the chapter is structured as follows. Section 2.1 introduces a set of fundamental dimensions for classifying XAI methods. Section 2.2 reviews method-centric taxonomies that focus on how explanations are generated and which artefacts they produce. Section 2.3 discusses conceptual taxonomies and socio-technical perspectives that emphasise stakeholders and purposes of explanations. Section 2.4 analyses the main limitations and open challenges of current taxonomies. Finally, Section 2.5 presents a practice-oriented framework that will be used in the rest of the thesis, with particular focus on high-impact and regulated application domains.

2.1 Fundamental dimensions of XAI

The literature on XAI proposes a large number of definitions, categories and taxonomies, often developed from different disciplinary perspectives and driven by heterogeneous use cases [5, 1, 7, 13, 2]. Despite this heterogeneity, a set of recurring conceptual dimensions

Dimension	Main categories	Typical questions and implications
Position in the model life cycle	Ante-hoc (intrinsically interpretable models); post-hoc (explanations for trained models)	Can the architecture be redesigned to improve interpretability, or must explainability wrap an existing system [5, 1]?
Transparency of the model	Intrinsically interpretable; black-box or opaque	Do stakeholders need insight into internal mechanisms, or is it sufficient to describe input output relationships [2]?
Scope of explanations	Local; global; hybrid	Is the main need to justify individual decisions, monitor behaviour over time, or support both tasks [5, 7]?
Dependence on the model family	Model-agnostic; model-specific	How easily can the method be ported across different ML models, and what are the computational or integration costs [5, 1]?

Table 2.1: Summary of fundamental dimensions for classifying XAI methods.

emerges that can be used to organise explanation methods and highlight their main trade-offs. These dimensions do not constitute a single, universally accepted taxonomy, but rather a set of orthogonal axes along which different techniques can be positioned, making the comparison between approaches developed in different contexts and time periods more systematic.

Table 2.1 summarises the main dimensions considered in this thesis and illustrates the types of design questions they raise for practitioners.

A first dimension concerns the point in the model life cycle at which the explanation is produced. Several works distinguish between *ante-hoc* methods, which aim to design models that are intrinsically interpretable, and *post-hoc* methods, which provide an explanation after a model has been trained, typically treating it as a black-box [5, 1]. A second dimension relates to the internal transparency of the algorithm and distinguishes between intrinsically interpretable models and opaque models, for which the explanation only describes the relationship between inputs and outputs without revealing the internal functioning of the system [2].

A third widely discussed axis separates local and global explanations. Local explanations focus on a single instance or on a small neighbourhood in the data space, and are particularly useful to support activities such as contesting an automated decision or reviewing a specific case. Global explanations instead aim to describe the overall behaviour of the model, for example by highlighting which features most strongly influence predictions across the entire dataset [5, 7]. Finally, a fourth dimension concerns the dependence of an explanation technique on the underlying model, distinguishing between *model-agnostic* methods, which

can be applied to a broad class of ML algorithms, and *model-specific* methods, which are designed to exploit structural properties of a particular model family [5, 1].

From a critical perspective, these dimensions are useful as a first approximation but they can also be misleading if interpreted as strict and independent categories. Speith notes that the same method is often described differently across taxonomies, and that boundaries between local and global, or between intrinsic and post-hoc, are frequently blurred in practice [13]. Moreover, the emphasis on a small set of dimensions can obscure other relevant aspects such as data quality, robustness or governance requirements, which are central in security-critical and regulated settings [14, 2]. In this thesis these dimensions are therefore used as flexible reference axes rather than as a rigid classification scheme.

In security and financial applications, these dimensions also interact with domain specific constraints. For instance, post-hoc and model-agnostic methods can be attractive when organisations rely on proprietary or third party models that cannot be modified, while ante-hoc approaches may be preferable when regulatory requirements or internal governance frameworks demand full control over the architecture [8, 3]. Likewise, the choice between local and global explanations is not merely technical but shapes how analysts investigate incidents, how managers monitor risk and how auditors reconstruct decision processes ex post. Treating these dimensions as design levers rather than static categories is therefore essential in high-impact domains.

2.2 Method centric taxonomies of XAI techniques

Building on the high level dimensions introduced in Section 2.1, several surveys propose taxonomies that focus more directly on how explanations are generated and which properties of the underlying models they exploit [5, 1, 7, 6, 13]. In these method centric taxonomies, XAI techniques are grouped according to their functioning, their algorithmic building blocks or the type of artefact they produce, for example saliency maps, feature attributions, rules, counterfactuals or example based explanations [5, 7, 15]. The aim of this section is to summarise the main patterns that emerge from these classifications and to identify a small set of categories that will be reused in later chapters.

2.2.1 Functioning-based classifications

Several authors group XAI methods according to the internal mechanism used to construct explanations [5, 1, 6, 7]. A first broad class comprises gradient or backpropagation-based methods, which propagate relevance or sensitivity information through the layers of a neural network to obtain an importance score for each input feature. Techniques such as saliency maps, guided backpropagation and layer wise relevance propagation exemplify this family and are particularly common in computer vision settings, where explanations can be visualised as heatmaps over images.

A second class is formed by perturbation-based methods, which estimate the contribution of a feature or a group of features by observing how the model output changes when these features are modified or masked [5, 6]. This family ranges from simple occlusion strategies,

where parts of the input are systematically removed, to more sophisticated approaches that approximate Shapley values or construct local surrogate models around the instance to be explained. These methods are typically model-agnostic and can be applied to a wide variety of ML models, at the cost of additional computational overhead.

A third group of methods builds explanations by training an interpretable surrogate that approximates the behaviour of a more complex model. In this case the explanation is provided by the structure of the surrogate itself, for instance a decision tree, a sparse linear model or a set of decision rules [5, 1]. Surrogate based approaches can be intrinsically global, when the interpretable model approximates the black-box on the whole input space, or local, when it is fitted only in a neighbourhood of the instance under analysis. Finally, example and prototype-based methods explain predictions by reference to representative instances, counterfactual examples or case-based reasoning schemes, which are often easier to interpret for non technical stakeholders [7, 4].

These functioning-based taxonomies are particularly useful when comparing the assumptions, strengths and limitations of different XAI techniques from an algorithmic standpoint. They also provide a bridge to implementation concerns, since classes such as gradient based or perturbation-based methods often correspond to different computational profiles and integration requirements in real systems.

From an operational viewpoint, these families of methods entail different trade-offs. Gradient based techniques are often efficient once access to internal model gradients is available, but they can be unstable and sensitive to small perturbations, which raises concerns when explanations need to be used as evidence in audits or investigations. Perturbation-based approaches are conceptually simple and broadly applicable, yet they can be prohibitively expensive for high volume, real time detection pipelines and may require strong assumptions about feature independence [6]. Surrogate based methods promise more compact and human-friendly artefacts, but they introduce an additional layer of approximation whose faithfulness (often referred to as fidelity) to the original model must itself be evaluated. These considerations suggest that functioning-based taxonomies should be complemented with explicit discussion of computational budgets, latency constraints and evidential requirements in specific domains.

2.2.2 Result-based classifications

Other taxonomies focus less on how explanations are computed and more on the form in which they are presented to users [5, 15, 4]. In these result-based classifications, methods are grouped by the type of artefact they produce. Common categories include visual attributions such as saliency maps or attention maps, numeric attributions in the form of feature importance scores, symbolic artefacts such as rules, decision trees or logical constraints, counterfactual explanations that describe how inputs would need to change to obtain a different outcome, and example-based explanations that select representative or influential instances.

This perspective is particularly relevant in application domains where different stakeholders may require different explanation formats. For instance, security analysts or developers might benefit from detailed feature attributions or counterfactuals that support

debugging and model refinement, while managers, auditors, or regulators may prefer higher-level rule-based explanations or summary statistics that can be more easily documented and communicated in reports [2, 3]. Result-based taxonomies therefore highlight the need to align XAI methods not only with the properties of the underlying models, but also with the expectations and capabilities of their intended audience.

2.2.3 Mixed and operational taxonomies

Some surveys explicitly combine multiple dimensions to construct more operational taxonomies. Das and Rad, for example, propose a classification that jointly considers the scope of explanations, the methodology behind the algorithms and the level at which explanations are used, leading to a matrix that blends functioning-based and result-based criteria [6]. Adadi and Berrada present a structured overview in which distinctions such as local versus global, intrinsic versus post-hoc, and specific versus agnostic are combined with categories based on explanation artefacts and model families [5]. Arrieta and co-authors organise recent work around a layered taxonomy that differentiates between interpretable models, post-hoc techniques, and application-specific contributions, while also connecting explainability to broader notions of responsible AI [1].

More domain-oriented surveys adopt similar strategies when focusing on specific sectors. In cybersecurity, for instance, taxonomies often group XAI techniques by both the type of security task and the underlying mechanism used to generate explanations, reflecting the need to integrate interpretability within complex detection and response pipelines [3, 4]. Speith takes a step back and reviews these different taxonomies themselves, highlighting their complementary strengths and pointing out the difficulties of relying on a single classification scheme to chart a rapidly evolving field [13].

Mixed taxonomies are particularly useful for practitioners, as they provide concrete decision criteria for choosing XAI methods under real-world constraints. At the same time, their proliferation reinforces the impression of a fragmented landscape. This motivates the need for a more conceptually grounded perspective, which takes into account stakeholders, purposes and socio-technical considerations and which will be discussed in Section 2.3.

2.3 Conceptually oriented taxonomies and socio-technical perspectives

The method-centric taxonomies reviewed in Section 2.2 provide a useful lens on XAI techniques by focusing on how explanations are generated and which artefacts they produce. However, several authors argue that explainability cannot be fully understood without considering who the explanations are for, why they are needed and in which organisational and regulatory context they are used [9, 14, 2]. Conceptually oriented taxonomies therefore complement method centric classifications by organising XAI along dimensions related to stakeholders, purposes and broader socio-technical considerations.

For regulators and standardisation bodies, this Tower of Babel effect is more than a terminological nuisance. If concepts such as transparency, interpretability or accountability

are used inconsistently across technical and legal documents, requirements risk becoming either too vague to be enforceable or too prescriptive to accommodate future methodological progress [14]. From the perspective of organisations deploying AI systems, this ambiguity translates into uncertainty about what counts as an adequate explanation and how to demonstrate compliance in practice. Conceptual taxonomies that aim to bridge technical and legal vocabularies are therefore not merely academic exercises, but prerequisites for workable governance frameworks.

2.3.1 Stakeholders and explanation needs

A first conceptual dimension concerns the stakeholders who interact with or are affected by AI systems and their explanations. Reviews of XAI research highlight that different groups often have distinct needs and capabilities, which translate into different expectations about explanations [9, 14]. Typical stakeholder categories include developers and data scientists, who use explanations primarily for debugging, model selection and performance monitoring; domain experts and operators, who rely on explanations to integrate model outputs into existing workflows and to calibrate their trust in automated recommendations; managers and decision makers, who may need higher level justifications that support accountability and risk management; and external actors such as auditors, regulators or affected individuals, who require explanations to verify compliance with legal requirements and to contest decisions when necessary [9, 2].

From this perspective, the suitability of a given XAI method cannot be evaluated solely by its technical properties. The same explanation artefact, for example a feature importance vector or a set of decision rules, may be useful for one group but opaque or misleading for another. Conceptual taxonomies that explicitly map methods to stakeholder groups therefore emphasise the need for audience aware explainability, where explanation design and evaluation are grounded in the actual practices, constraints and responsibilities of the people involved.

2.3.2 Purposes of explanations: trust, control and compliance

A second conceptual dimension concerns the purposes that explanations are meant to serve. Gerlings and co authors identify several recurring themes in XAI scholarship, including the role of explanations in fostering appropriate trust, enabling control and oversight, supporting learning and understanding and satisfying external requirements such as fairness or accountability [9]. Chamola and colleagues connect explainability with the broader agenda of trustworthy AI, where transparency, robustness and human agency are seen as intertwined properties rather than isolated features [2].

In regulated domains these purposes are closely linked to legal and ethical obligations. For instance, explanations may be required to support data subject rights under GDPR, to document risk management processes for critical AI systems under the European AI Act, or to justify automated decisions in high-impact sectors such as finance and healthcare. Conceptual taxonomies that focus on purposes therefore highlight how the same technical method can play different roles depending on whether it is used to debug a model during

development, to justify a decision to an affected individual or to provide evidence during an audit. This reinforces the idea that explainability should be treated as a means to specific ends rather than as an abstract property of models.

2.3.3 The Tower of Babel problem

The rapid growth of XAI has led to a proliferation of terms, definitions and conceptual frameworks. Schneeberger and co authors describe this situation as a Tower of Babel effect, where different communities, including ethics, law, standardisation and computer science, use overlapping but not fully compatible vocabularies to talk about explainability and related notions such as transparency, interpretability and accountability [14]. Speith similarly shows how existing taxonomies of XAI methods are built on partially divergent assumptions about what counts as an explanation and which criteria matter most when classifying techniques [13].

This fragmentation has practical consequences. If researchers, practitioners and regulators use the same terms with different meanings, it becomes harder to compare methods, to translate requirements into technical specifications and to evaluate whether a given system meets expectations. Conceptually oriented work on XAI therefore advocates for greater clarity about definitions and for explicit mapping between vocabularies used in different communities [14]. Rather than attempting to enforce a single universal definition, these contributions suggest building unified lexicons and reference models that can serve as a bridge between disciplines.

In this thesis, these socio-technical perspectives are used to complement method centric taxonomies. When selecting and assessing XAI techniques for security-critical and regulated settings, the analysis will explicitly consider which stakeholders are targeted, which purposes explanations are meant to serve and how the adopted terminology relates to ongoing discussions in law, standardisation and trustworthy AI.

2.4 Limitations of existing taxonomies and open challenges

Existing taxonomies of XAI provide useful entry points into a rapidly growing field, but several limitations have been identified in recent reviews [13, 14]. A first issue is that many taxonomies rely on categories that are not mutually exclusive. For example, a method can be both local and global in different configurations, or it can be used as an intrinsic component of a model and as a post-hoc tool depending on how it is integrated into the pipeline [5, 1]. This overlap is not necessarily problematic in itself, but it means that taxonomies should be interpreted as a flexible map rather than a rigid classification scheme, as also highlighted by the conceptual fragmentation discussed in Section 2.3. For practitioners this can create uncertainty about how to label a technique and which category to consult when searching for suitable methods.

A second limitation concerns the fact that many taxonomies are largely model centric and only weakly connected to concrete application contexts. Dimensions such as ante-hoc versus

post-hoc, local versus global or model-agnostic versus model-specific are important from an algorithmic perspective, but they do not directly capture constraints and requirements that arise in specific domains, for example in healthcare, finance or cybersecurity [7, 3, 4]. Conceptual work on stakeholders and purposes of explanations partially addresses this gap, yet the link between high level socio-technical considerations and concrete design choices for XAI methods often remains implicit [9, 2].

A third challenge relates to the evaluation and comparison of XAI methods. Several surveys highlight that there is no widely accepted set of metrics or benchmarks for explanation quality, and that different taxonomies implicitly endorse different criteria when grouping methods [6, 7, 13]. Some evaluations focus on faithfulness to the underlying model, others on human centred properties such as usefulness, comprehensibility or impact on decision-making, and many combine these aspects in ad hoc ways. As a result, it is difficult to use taxonomies to make robust claims about which methods are preferable in a given setting, beyond broad qualitative statements.

The lack of shared evaluation protocols is particularly problematic in domains where explanations are expected to support high stakes decisions. In cybersecurity, for example, explanation methods are often demonstrated on small case studies or synthetic datasets, with limited evidence about how they perform when embedded in large scale monitoring infrastructures [3]. Taxonomies rarely capture this gap between proof of concept demonstrations and deployable solutions, and they typically do not distinguish between methods that have been evaluated with human users and those assessed only via proxy metrics such as sparsity or stability. As a result, practitioners may overestimate the maturity of certain techniques and underestimate the organisational effort required to turn them into reliable tools.

Finally, the field of XAI is evolving quickly. New techniques and application scenarios continue to emerge, and taxonomies that are well aligned with the state of the art at a given time can become outdated in a few years [5, 4, 13]. This dynamism reinforces the need for taxonomies that are not only descriptive snapshots, but that also provide stable conceptual anchors which can accommodate future developments. It also suggests that taxonomies should be used with caution as prescriptive tools for design, especially in high-impact domains where regulatory and ethical requirements play a central role.

In this thesis, these limitations are treated as a motivation to combine multiple perspectives rather than to rely on a single taxonomy. The conceptual framework proposed in the next section is therefore designed to be lightweight and practice oriented, focusing on a subset of dimensions that are particularly relevant for security-critical and regulated applications.

2.5 Towards a practice oriented framework

The previous sections have shown that XAI can be organised along several complementary dimensions, ranging from algorithmic functioning and explanation artefacts to stakeholder needs and socio-technical considerations. Building on these insights, this section introduces a practice oriented framework that will guide the analysis and selection of XAI methods in

the remainder of the thesis. The framework is intentionally modest in scope. Rather than proposing a new general taxonomy, it identifies a subset of dimensions that are particularly salient for the design of explainable solutions in security-critical and regulated domains.

2.5.1 Synthesis of relevant dimensions

For the purposes of this work, five groups of dimensions are considered especially relevant. The first group captures the position of an explanation method within the model life cycle, distinguishing between ante-hoc approaches, which rely on intrinsically interpretable models, and post-hoc approaches, which wrap explainability around an existing system. This dimension is important for understanding which parts of a pipeline can be modified and which components are constrained by legacy systems or external vendors.

The second group concerns the scope and granularity of explanations, with a focus on the local versus global distinction and on whether explanations target individual predictions, classes of predictions, or the overall model behaviour. This dimension directly influences which tasks explanations can support, for example, contesting a specific automated decision, monitoring model drift, or documenting system behaviour for audits.

The third group relates to the dependence of methods on particular model families. Model-agnostic techniques such as many perturbation-based approaches offer flexibility and can be applied to diverse ML models at the cost of computational overhead, whereas model-specific techniques can exploit structural properties of neural networks, tree ensembles or other architectures to provide more targeted or efficient explanations [5, 1, 6]. This dimension is crucial when integrating explainability into existing technology stacks.

The fourth group concerns the mechanisms and artefacts of explanation, combining function-based and outcome-based perspectives. In this thesis, particular attention is given to feature attribution methods, rules and surrogate based explanations, counterfactuals and example based approaches, since these categories recur across multiple surveys and have concrete relevance for security related tasks [7, 4, 3]. The final group of dimensions focuses on stakeholders and purpose-related dimensions are treated as explicit design constraints: for each case study, candidate XAI methods will be assessed not only for their algorithmic properties, but also for their ability to serve specific explanation goals for clearly identified audiences [9, 2].

2.5.2 Alignment with trustworthy and regulated AI requirements

Trustworthy AI frameworks emphasise properties such as transparency, robustness, fairness and human oversight as preconditions for the responsible deployment of AI systems. In parallel, regulatory initiatives in the European context, including GDPR, the AI Act and sector specific rules, increasingly require organisations to document how automated decision-making systems operate and to provide meaningful information about their logic to different stakeholders. Surveys that explicitly connect XAI with trustworthy AI show that explainability should be understood not as an isolated technical feature, but as a component of broader governance and risk management processes [2, 14].

In security-critical domains such as cybersecurity and financial services, these requirements are further compounded by the need to detect and respond to complex threats under time and resource constraints. Work on XAI for cybersecurity illustrates how explanation methods can support tasks such as intrusion detection, anomaly analysis and incident investigation, while also contributing to compliance and auditability [3, 4]. Domain specific studies in financial cybersecurity highlight how techniques like feature attribution, counterfactual reasoning and surrogate models can be embedded into fraud detection and access control systems to provide traceable justifications for alerts and decisions [8].

Against this background, the practice oriented framework adopted here explicitly links the selection of XAI methods to regulatory and governance considerations. When evaluating candidate techniques, the analysis will therefore consider not only their algorithmic properties, but also how they contribute to documentation, audit trails and human-in-the-loop decision processes in environments shaped by GDPR, the AI Act and related regulations.

2.5.3 Implications for high-impact domains

The dimensions discussed above provide a structured way to reason about XAI choices in high-impact domains, where decisions must satisfy not only technical performance criteria but also stringent regulatory and organisational constraints. For example, in a security operations context it may be necessary to combine local, instance level explanations that help analysts understand individual alerts with more global insights that support model monitoring and governance. Model-agnostic techniques can facilitate the integration of explainability into heterogeneous detection pipelines, while domain specific surrogate models and rule based explanations may be better suited for communicating with non technical stakeholders or for satisfying documentation requirements.

In regulated environments, explanation artefacts and formats must be chosen with particular care. Counterfactuals and example based explanations can be useful to communicate options and to support the contestation of decisions, whereas feature attribution vectors may be more appropriate for internal debugging and risk analysis. The stakeholder and purpose oriented dimensions help ensure that these different needs are explicitly considered rather than assumed.

Consider, for instance, a fraud detection system in a financial institution that combines gradient boosted trees with deep learning models for transaction risk scoring. From the perspective of the framework introduced in this chapter, most practical deployments will rely on post-hoc and largely model-agnostic explanation techniques, such as feature attribution methods or local surrogate models, because the underlying models are chosen for predictive performance and must integrate with existing risk engines [8]. Local explanations are required to justify individual alerts to analysts and, in some cases, to customers, while more global summaries are needed to monitor model behaviour and to document risk controls for internal and external audits. Stakeholders span security analysts, fraud investigators, compliance officers and external regulators, each with different explanation needs. Analysing this scenario through the lens of the proposed dimensions makes explicit which XAI methods are even plausible candidates and which trade-offs they entail in terms

of latency, robustness and evidential value.

In the rest of the thesis, this framework will serve as a reference for analysing concrete XAI techniques and for assessing their suitability in cybersecurity scenarios that are subject to stringent regulatory and organisational constraints. By keeping both method centric and socio-technical dimensions in view, the aim is to move from abstract taxonomies to design oriented guidance that can inform the development of explainable, trustworthy and compliant security solutions.

Chapter 3

The Regulatory Landscape for AI Transparency

The previous chapters introduced the motivation for explainable and trustworthy AI in cybersecurity and surveyed the main families of XAI methods and taxonomies, with a specific focus on their applicability to SLMs.[5, 1, 4, 13] Technical advances alone, however, do not fully explain why transparency and explainability have become central requirements for contemporary AI systems. A second driver is the evolving regulatory environment, which increasingly constrains how opaque model based decisions may be in practice, especially when they affect individuals, organisations, or critical infrastructures [2, 9].

This chapter analyses that environment with a focus on the European Union, where a layered set of horizontal and sectoral instruments jointly shape expectations about transparency, traceability, and accountability for AI systems in security-critical infrastructures. Rather than providing an exhaustive legal commentary, the goal is to extract those obligations and principles that are most relevant for the design and governance of SLM based solutions in cybersecurity, and to connect them to the types of explanation artefacts discussed in Chapter 2.[3, 8] Building on the overview of regulatory drivers in Chapter 1, the chapter maps legal requirements to design oriented constraints and opportunities for XAI.

The discussion proceeds as follows. Section 3.1 reframes algorithmic opacity as a regulatory problem. Section 3.2 presents the General Data Protection Regulation (Regulation (EU) 2016/679, GDPR)¹ as a baseline framework for data protection, profiling, and automated decision-making. Section 3.3 examines the risk-based transparency obligations introduced by Regulation (EU) 2024/1689 (Artificial Intelligence Act, AI Act)². Section 3.4

¹See in particular Articles 5, 13-15, 22 and 35 and Recital 71 of Regulation (EU) 2016/679 (General Data Protection Regulation, GDPR) [19]

²See Regulation (EU) 2024/1689 laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act, AI Act), in particular Articles 6-9 on risk classification and Articles 9-15 on risk management, transparency, human oversight and robustness.[22, 16]

focuses on cybersecurity oriented instruments such as the NIS2 Directive³ (hereafter NIS2), the Cyber Resilience Act⁴ (CRA) and the Digital Operational Resilience Act (DORA).⁵ Section 3.5 complements this picture with soft law frameworks and technical standards. Finally, Section 3.6 synthesises these sources into regulatory derived design principles for XAI oriented SLM systems in cybersecurity and formulates the questions that guide the empirical and framework oriented chapters.

3.1 From Opacity to Explainability: A Regulatory Problem Framing

Modern AI systems are frequently deployed as part of complex socio-technical infrastructures, where automated components interact with human operators, organisational processes, and legal responsibilities. The dominant engineering paradigm of highly parameterised, data driven models, including transformer based architectures that underpin SLMs, has reintroduced an old concern: it is often difficult for humans to understand how and why a particular output was produced [5, 1, 6]. This opacity is not merely a technical inconvenience. When model outputs inform decisions that affect individuals, organisations, or critical infrastructures, opacity can undermine values such as accountability, non-discrimination, procedural fairness, and due process [2, 9, 15].

Regulators have become increasingly sensitive to these risks. In domains such as credit scoring, insurance pricing, fraud detection, and cybersecurity incident response, automated assessments can influence access to services, exposure to liability, or the allocation of limited defensive resources [3, 8]. If such assessments are based on models that cannot be meaningfully interrogated, several problems arise. Affected stakeholders may be unable to contest decisions or understand how to adapt their behaviour. Supervisory authorities may struggle to verify compliance with substantive norms. Organisations themselves may find it difficult to diagnose systematic errors, concept drift, or adversarial manipulation [2, 4].

In response, legal instruments have progressively moved from a purely outcome oriented view of automation, focused only on whether a decision is lawful, to a more process oriented view, in which the way decisions are produced also matters. Transparency and explainability are central enablers of this shift.[1, 9] At a minimum, transparency requires that relevant information about the system and its operation be available to those who are affected by, or responsible for, its outputs. Explainability is typically understood as the capacity to offer intelligible reasons for specific outputs or behaviours of an AI system

³Directive (EU) 2022/2555 on measures for a high common level of cybersecurity across the Union (NIS2 Directive).[17]

⁴Regulation (EU) 2024/2847 on horizontal cybersecurity requirements for products with digital elements (Cyber Resilience Act, CRA).[23, 18]

⁵Regulation (EU) 2022/2554 of the European Parliament and of the Council of 14 December 2022 on digital operational resilience for the financial sector (Digital Operational Resilience Act, DORA). Relevant provisions include, in particular, Articles on ICT risk management, incident reporting, testing and oversight of third party service providers.

in a form that can be understood and acted upon by humans [5, 2]. While exact meanings vary across instruments, there is broad convergence on the idea that opaque black-box models are increasingly difficult to justify when significant risks are at stake.

For XAI research, this regulatory turn has two consequences. First, explainability becomes a potential compliance requirement, not only a usability or debugging feature. Organisations may need to demonstrate that their systems provide adequate explanations to users, auditors, or regulators, and that these explanations are grounded in actual model behaviour rather than generic rationalisations [7, 6]. Second, explanations must be tailored to heterogeneous addressees. End users, security analysts, internal auditors, and supervisory authorities may require different types and levels of detail. This multi stakeholder context complicates the design of explanation mechanisms, especially for SLMs integrated into larger workflows in security operations centres and other high stakes environments [3, 24, 25].

The cybersecurity domain amplifies these challenges. As outlined in Chapter 2, SLMs can be used to process unstructured security data, support vulnerability and threat analysis, or assist in incident triage and response.[26, 27, 28, 29] In these use cases, incorrect or misleading outputs may not only degrade model performance but also propagate through security operations, for instance by prioritising the wrong alerts or misclassifying critical vulnerabilities. From a regulatory perspective, such failures can translate into violations of obligations to ensure the confidentiality, integrity, and availability of systems and data, or into deficiencies in required risk management and incident handling processes.[3, 8] Explanations are therefore needed not only to satisfy abstract transparency duties but also to enable timely verification and justification of security relevant model behaviour.

At the same time, legal texts, ethical guidelines, and technical literature often use a similar vocabulary, such as transparency, interpretability, explainability, traceability, and human oversight, while attaching different expectations to these terms.[14, 13] Lawyers may be mainly concerned with whether a decision can be reconstructed ex post in a way that supports accountability and contestability. Engineers may focus on local feature attributions or surrogate models. Security practitioners might value explanations that highlight operationally meaningful factors, such as which log sources or vulnerability descriptors most influenced a triage decision. Without explicit alignment between these perspectives, systems that are formally explainable can still fail to meet regulatory expectations or operational needs.

In the remainder of this chapter, the regulatory landscape is treated not simply as background context, but as a source of concrete constraints and design targets for XAI enabled SLM systems in cybersecurity. The analysis focuses on obligations related to information provision and documentation, requirements for human oversight, logging, and traceability, and expectations about risk-based proportionality in transparency measures.[2, 9] These elements will later be translated, in Section 3.6, into design principles and evaluation criteria for the experimentation with SLMs and the proposal of a novel XAI oriented framework.

3.2 Data Protection, Profiling, and Automated Decision-making: GDPR as a Baseline

Concerns about opacity and explainability intersect most directly with the GDPR, which remains the central framework governing data driven systems in Europe.[2, 8] Any AI system that processes personal data, including SLMs used in security operations, must comply with GDPR principles and obligations. These set both substantive limits on how data can be used and procedural duties to demonstrate control, oversight, and accountability. This section highlights three aspects that are particularly relevant for explainable AI in cybersecurity: core principles, provisions on profiling and automated decision-making, and accountability tools such as the Data Protection Impact Assessment (DPIA).

3.2.1 GDPR Principles Relevant to AI Systems

Article 5 of the GDPR articulates principles that govern any processing of personal data: lawfulness, fairness, and transparency; purpose limitation; data minimisation; accuracy; storage limitation; integrity and confidentiality; and accountability. Although not drafted with modern AI in mind, they have direct implications for learning based systems.⁶

Lawfulness and fairness require that processing rests on a valid legal basis and does not produce unjustified or disproportionate effects on individuals. In the context of AI, this implies that the use of large scale datasets and automated profiling must be tied to clearly defined purposes and legal grounds, and that systematic biases or discriminatory outcomes are incompatible with fairness.[6, 4] Transparency complements these duties by requiring that data subjects receive meaningful information about how their data are used and, in certain cases, about the logic involved in automated processing. For SLMs in cybersecurity, this does not imply exposing full internal mechanics to every data subject, but it does require the ability to explain at a high level how the system uses personal data, what inferences it draws, and what impact these inferences may have.⁷

Purpose limitation and data minimisation constrain the reuse of data collected for security monitoring for broader analytics without proper justification. Models trained on logs or monitoring data must still respect the original purposes for which data were collected, unless a compatible purpose or new legal basis can be established. This is challenging for SLMs, which often rely on heterogeneous corpora and can be repurposed for multiple tasks. XAI techniques that help characterise which features and data sources contribute to a given prediction can support assessments of whether the model in practice respects these constraints or implicitly exploits information that should not be used.[1, 5, 7]

The principles of accuracy and integrity and confidentiality further reinforce the need

⁶See Article 5 GDPR.[19]

⁷On information duties and the notion of “meaningful information about the logic involved”, see Articles 13-15 and 22 and Recital 71 GDPR.[19]

for robust and controllable models. Accuracy affects both training data quality and the reliability of inferences. Integrity and confidentiality require appropriate security of personal data, including protection against unauthorised processing and accidental loss. For SLMs in cybersecurity, this creates a dual responsibility: models must themselves be protected against attacks such as model inversion or poisoning, and their use must not introduce new vulnerabilities into the processes they support.[3, 4]

Finally, the accountability principle requires controllers not only to comply with the GDPR but also to be able to demonstrate such compliance. This demonstration increasingly depends on the ability to document and reconstruct how data are processed through learning based systems, which models were used, what data they were trained on, and how their behaviour is monitored.[2, 9] In this sense, the GDPR incentivises technical and organisational measures that make opaque models more transparent and traceable, providing a natural entry point for XAI.[1, 8]

3.2.2 Profiling, Article 22 GDPR, and the Debate on a Right to Explanation

Beyond general principles, the GDPR contains specific provisions on profiling and automated decision-making. Profiling is defined as any form of automated processing of personal data that uses those data to evaluate personal aspects relating to a natural person, in particular to analyse or predict aspects such as performance at work, economic situation, health, preferences, reliability, behaviour, or movements. Many AI applications in cybersecurity, such as user and entity behaviour analytics or anomaly detection based on log data, fall within this broad notion.[3, 4]

Article 22 establishes a right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning the individual or similarly significantly affects the individual. Subject to limited exceptions, such decisions are only permissible if appropriate safeguards are in place, including the right to obtain human intervention, to express a point of view, and to contest the decision. While Article 22 does not explicitly mention a right to explanation, it is commonly interpreted in light of transparency provisions in Articles 13-15 and Recital 71 as requiring controllers to provide meaningful information about the logic involved in automated processing and the envisaged consequences for data subjects.⁸

The precise scope of this requirement has been debated in legal scholarship and regulatory guidance.[9, 21] Some commentators argue that the GDPR implies a strong right to an individualised explanation of automated decisions. Others stress that the text only guarantees information about general logic and effects, rather than full disclosure of source code or model parameters. Recent case law has begun to clarify these expectations. In *CK v Dun & Bradstreet Austria GmbH*, the Court indicated that, at least in the context of automated credit scoring, data subjects are entitled to an explanation that enables them to understand which of their personal data were used, how they were processed, and how

⁸See Articles 13-15 and 22 and Recital 71 GDPR.[19]

variations in those data would have affected the outcome.⁹ Trade secrecy cannot be used in a way that systematically undermines this right.

For SLM-based cybersecurity tools, these developments imply that explanations aimed at affected individuals or regulators should go beyond generic statements about the use of AI. Where automated processing significantly affects individuals, controllers should be able to describe, in clear terms, which categories of data were considered, which factors influenced the decision, and, at least at a high level, how alternative inputs might have changed the outcome.[2, 1] At the same time, not every AI-supported security decision will fall within the strict scope of Article 22. Many uses of SLMs in cybersecurity are advisory or supportive rather than decisive: they assist analysts in triaging alerts, summarising incidents, or generating reports.[24, 25] Even in such cases, GDPR principles of transparency and fairness, as well as accountability obligations, support a softer but still meaningful requirement to explain how SLM outputs are generated and used in decision-making pipelines. In this sense, GDPR sets a baseline expectation of explainability for SLM-based security analytics, even where a formal right not to be subject to automated decisions does not apply in a strict sense.[8]

3.2.3 Accountability, DPIA, and Algorithmic Traceability

The accountability principle is operationalised through obligations that require controllers to assess, document, and monitor the risks associated with their processing operations. The Data Protection Impact Assessment plays a central role for data intensive and high-risk uses of AI. A DPIA is required where a type of processing, in particular using new technologies, is likely to result in a high-risk to the rights and freedoms of natural persons.¹⁰ Typical examples include systematic and extensive evaluation of personal aspects based on profiling, large scale processing of special categories of data, or systematic monitoring of publicly accessible areas.

For SLMs in cybersecurity, a DPIA may be triggered when the system performs large scale monitoring of network traffic, logs, or user behaviour, especially in critical infrastructures or financial institutions.[3, 29] Conducting a DPIA in such contexts requires a detailed understanding of how the model operates, what data it uses, what risks it may create, and what measures mitigate those risks. XAI techniques can contribute by offering systematic ways to inspect feature importance, decision boundaries, and failure modes, which can be integrated into the DPIA as part of the justification for the chosen design.[7, 6, 4]

Algorithmic traceability is another dimension of accountability that resonates with both GDPR obligations and cybersecurity practice. Controllers must maintain records of processing activities, implement appropriate logging, and be able to investigate incidents involving personal data. When SLMs are integrated into security monitoring pipelines, their

⁹See Court of Justice of the European Union, Judgment of 27 February 2025, Case C-203/22, *CK v Dun & Bradstreet Austria GmbH and Magistrat der Stadt Wien*, ECLI:EU:C:2025:117.[20]

¹⁰See Article 35 GDPR.[19]

outputs may feed into alerts, dashboards, or automated responses. To satisfy both data protection and security requirements, it becomes important to log not only conventional technical events but also relevant aspects of model behaviour, such as model versions, configuration parameters, and key input output pairs for critical decisions.[3, 8] Without such traceability, it is difficult to reconstruct how a given decision was made or to provide convincing explanations to affected individuals or supervisory authorities.[2]

3.3 The EU AI Act: Risk-based Transparency and Explainability Duties

Alongside the GDPR, the AI Act is the most significant attempt to create a comprehensive regulatory framework for AI in the European Union. While the GDPR focuses on data protection, the AI Act targets the broader risks posed by AI systems, including safety, fundamental rights, and the functioning of critical infrastructures.[16, 14] A key feature is its risk-based approach, which differentiates between prohibited, high-risk, and lower risk systems and assigns corresponding obligations. Among these obligations, transparency, documentation, and human oversight have direct implications for the design of explainable systems[2, 9].

3.3.1 Risk Taxonomy and AI System Categories

The AI Act organises obligations around a graded taxonomy of risk. Certain practices are prohibited altogether, such as specific forms of manipulative or exploitative AI. Minimal risk systems are essentially left to voluntary codes of conduct. In between, the Act identifies a broad class of high-risk systems that are subject to detailed ex ante and ex post obligations, and a set of limited risk systems that mainly face transparency duties[16, 22].

The classification of a system as high-risk depends primarily on its function and deployment context, not on a specific technique. This is particularly relevant for SLMs, which may appear generic from a technical perspective but can play critical roles in supporting decisions on incident handling, fraud detection, or operational resilience.[26, 28, 24] A standalone model used as an internal assistant in a security operations centre may not by itself qualify as a high-risk system, but an SLM embedded in an automated transaction monitoring tool for a financial institution could form part of a high-risk system subject to stringent obligations. The same underlying model may therefore be tied to different transparency requirements depending on how it is integrated.

3.3.2 Transparency, Documentation, and Human Oversight for High-risk AI

High-risk AI systems under the AI Act are subject to cumulative requirements that aim to ensure controllable, predictable, and auditable behaviour. Providers must implement a risk management system, maintain technical documentation, ensure data governance,

design logging and record-keeping functions, and enable effective human oversight. They must also demonstrate an appropriate level of accuracy, robustness, and cybersecurity.¹¹

From an XAI perspective, these obligations create a structural demand for mechanisms that make model behaviour intelligible. Technical documentation must include information about the model architecture, training process, intended purpose, performance, and known limitations. Logging must capture events and outputs in a way that allows ex post analysis of failures or unexpected behaviour. Human oversight requires that operators can understand when and how to rely on automated outputs and when to override them.[1, 7] Purely opaque models that produce outputs without any supporting rationale are difficult to reconcile with these expectations.

For SLMs integrated into high-risk systems in cybersecurity, providers must decide what level of documentation is feasible, how to expose relevant information to integrators and users, and how to structure logs that are both privacy aware and rich enough to support audits.[3, 8] XAI techniques can support these goals by providing structured artefacts that can be incorporated into documentation, logging, and user interfaces.[4, 5] However, the AI Act formulates transparency and oversight duties in broad terms, without specifying what counts as an adequate explanation or how explanation quality should be assessed. This leaves room for misalignment between legal expectations and technical practices.[9, 14]

3.3.3 General Purpose and Generative AI: Transparency for LLMs and SLMs

The AI Act introduces a specific regime for general purpose AI models, including large language models and, by extension, many families of SLMs. These models are not tied to a single predefined use case and can be adapted to a wide range of downstream applications, sometimes in ways that the original provider cannot fully foresee.[10, 11] This generality raises concerns for transparency and control, since behaviour may change significantly when models are fine tuned, combined with other components, or exposed to different prompts and data.

Providers of general purpose models must prepare technical documentation, provide information to downstream providers and deployers, and, for certain models, implement measures to address systemic risks. Generative models also face transparency duties regarding synthetic content, for example informing users when they interact with AI or when content has been generated by AI.[16, 22] For SLMs in cybersecurity, these obligations imply that upstream providers should disclose sufficient information about capabilities, limitations, and known risks to allow downstream users to perform adequate risk assessments. They also encourage the development of interfaces and tooling that expose model behaviour in ways that support safe integration into security workflows, for instance through confidence indicators or structured rationales.[26, 27, 28]

¹¹See in particular Articles 9-15 AI Act on risk management, data governance, technical documentation, record-keeping, transparency, human oversight, and robustness and cybersecurity.[22, 16]

At the same time, there are tensions between transparency and other objectives. In cybersecurity, detailed disclosure of model internals or prompt handling logic can increase the attack surface for adversaries. Commercial providers may resist extensive disclosure on intellectual property grounds. The AI Act acknowledges some of these concerns but leaves considerable discretion in practice. For XAI solutions around SLMs, this creates a complex space of trade-offs between regulatory transparency, security by design, and protection of proprietary information.[3, 15]

3.4 Cybersecurity Oriented Regulation: NIS2, Cyber Resilience Act, and DORA

Horizontal instruments such as the GDPR and the AI Act are complemented by sectoral frameworks that focus directly on cybersecurity and operational resilience. Among these, the NIS2, the CRA, and DORA are particularly relevant for SLM based systems deployed in critical sectors.[3, 8] They rarely mention explainability explicitly, but they contain requirements on logging, monitoring, vulnerability management, and incident handling that implicitly demand transparency and traceability in the behaviour of complex digital systems, including AI components.[2]

3.4.1 NIS2: Security by Design, Logging, and Incident Reporting

The NIS2 imposes obligations on essential and important entities in sectors such as energy, transport, banking, health, and digital infrastructure. Entities must implement risk management measures that ensure the security of network and information systems, including policies on incident handling, business continuity, supply chain security, and monitoring.¹² These obligations are technology neutral but clearly encompass AI enabled tools when they form part of an organisation’s detection, monitoring, or response capabilities.[3]

When SLMs are used to triage alerts, summarise threat intelligence, or prioritise vulnerabilities, their behaviour becomes part of the security posture that NIS2 seeks to safeguard.[26, 25] While NIS2 does not explicitly mandate particular explainability techniques, its emphasis on logging, incident reporting, and reconstructability makes explainability a practical enabler for evidence collection and review when SLMs mediate detection and analysis.[17, 2, 9] Explanations of model decisions can help analysts understand why certain events have been flagged as critical and whether specific classes of threats might be systematically under detected.[3, 8] In line with this, NIS2 reinforces the importance of logging. For forensic analysis and post incident review, entities must be able to reconstruct how detection and analysis systems operated. Logs that capture only final outputs may not suffice when SLMs are involved. Audit friendly explanation artefacts that link specific decisions to relevant evidence can support compliance.

¹²See in particular Chapter IV NIS2 on cybersecurity risk management and reporting obligations.[17]

3.4.2 The Cyber Resilience Act: Secure and Transparent Digital Products

The CRA addresses the security of products with digital elements, including software and hardware placed on the EU market. It introduces essential cybersecurity requirements that manufacturers must meet throughout the life cycle of their products, from design to vulnerability handling.¹³ These requirements include secure by design principles, vulnerability management processes, and obligations to provide information and updates to users.[2]

For products that integrate AI components such as SLMs, the CRA raises questions about how to demonstrate that the product does not introduce avoidable cybersecurity risks. Security relevant SLMs may, for example, be embedded in endpoint detection solutions or vulnerability management platforms. Explainability plays a dual role. Internally, explanation mechanisms can help developers and security teams understand model failure modes and adversarial behaviour. Externally, documentation that describes capabilities and limitations, including known blind spots, can inform customers and integrators about safe usage patterns and residual risks.[8, 3]

3.4.3 DORA and ICT Risk Management in the Financial Sector

DORA focuses on the digital operational resilience of financial entities, including banks, insurance companies, and critical third party providers. It requires comprehensive frameworks for managing ICT risk, covering governance structures, incident reporting, testing of ICT tools, and oversight of third party service providers. Financial institutions are early adopters of AI and data driven techniques in fraud detection, anti money laundering, credit scoring, and cybersecurity monitoring.[4, 29] SLMs can support analysts by processing unstructured logs or automating parts of incident response workflows.[26, 24] In these roles, model outputs can influence decisions with significant financial and regulatory consequences. Although DORA does not prescribe specific explainability techniques, explanation artefacts can make ICT risk reporting and supervisory evidence more concrete by linking alerts to the underlying data and assumptions captured in logs.[2, 9] DORA therefore reinforces the need for transparency and explainability at multiple levels: management must understand risks associated with ICT tools, risk functions must assess and monitor those risks, and supervisors must evaluate whether institutions have adequate control. XAI techniques that provide traceable explanations for model outputs can facilitate this process, for example by documenting which factors contributed to specific alerts or patterns of misclassification.[8, 2]

¹³See Regulation (EU) 2024/2847 (Cyber Resilience Act, CRA), notably Annex I on essential cybersecurity requirements for products with digital elements.[23, 18]

3.5 Soft Law Frameworks and Technical Standards for Trustworthy and Explainable AI

Beyond binding regulations, a rich ecosystem of soft law instruments and technical standards guides the development and deployment of trustworthy and explainable AI. These include ethical guidelines, high level principles by international organisations, risk management frameworks, and standards issued by standardisation bodies. While they lack the legal force of regulations or directives, they influence both regulatory interpretation and industry practice, and they offer more detailed guidance on what transparency and explainability can entail.[1, 2, 9]

3.5.1 EU Ethics Guidelines for Trustworthy AI

The EU High Level Expert Group on AI published the Ethics Guidelines for Trustworthy AI, which propose that AI systems should be lawful, ethical, and robust. The Guidelines identify seven requirements for trustworthy AI: human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity and non-discrimination, societal and environmental well being, and accountability.[2]

Transparency covers several dimensions: traceability of model development, communication of system capabilities and limitations, and availability of explanations for individual decisions where appropriate. Explanations should be tailored to different stakeholders, such as lay users, domain experts, and regulators.[1, 4] For SLMs in cybersecurity, this supports the notion that analysts, risk officers, and affected individuals require different explanatory content and formats.

3.5.2 OECD Principles, NIST AI RMF, and ISO and IEC Standards

The OECD AI Principles and the NIST AI Risk Management Framework call for AI systems to be transparent and explainable, robust, secure, and subject to appropriate human oversight. Although these instruments are non binding, they shape expectations about reasonable organisational practices. Surveys on trustworthy and explainable AI emphasise that such frameworks treat explainability as one dimension of broader governance and risk management practices, rather than as a purely technical property of models.[2, 9]

Standardisation bodies such as ISO and IEC are beginning to produce standards related to AI management systems, transparency, and impact assessment. While many are still in early stages, they point toward a future in which conformity assessment and certification schemes play a larger role in demonstrating that AI systems meet recognised criteria for transparency and explainability.[14, 13] These standards generally do not mandate a particular XAI technique, yet they emphasise demonstrable processes and documentation; explanation artefacts can therefore serve as the evidence layer that connects model behaviour to logged records for audits or certification.[14, 2, 9] For providers and users of SLMs in cybersecurity, aligning with such standards can help structure internal

processes and documentation in a way that is legible to auditors and regulators.[8]

3.5.3 The Tower of Babel Problem: Misaligned Notions of Explainability

Despite the proliferation of guidelines and frameworks, there is no single, universally accepted definition of explainability. Legal, ethical, and technical communities often use terms such as transparency, interpretability, explainability, traceability, and human oversight in overlapping but not identical ways. This lack of alignment has been described as a Tower of Babel problem for XAI: stakeholders talk about explainability using the same words, but they refer to different properties, artefacts, and evaluation criteria.[14, 13]

For example, a technical XAI paper may consider a model explainable if it is accompanied by saliency maps or feature importance rankings. From a legal perspective, however, explainability might require that individuals can understand the reasons for a decision in language that relates to their situation and rights, and that decision makers can justify their actions in line with procedural fairness. In cybersecurity, operators may prioritise explanations that are actionable in operational terms, such as indicating which log sources or events triggered an alert.[3, 4] These divergent expectations can lead to systems that are technically explainable according to one set of metrics but fail to satisfy other stakeholders. This is particularly problematic for SLMs, which can generate convincing but sometimes unfaithful explanations of their own outputs.[6, 15]

3.6 From Legal Requirements to Design Principles for XAI SLM Systems in Cybersecurity

The regulatory and soft law landscape reviewed in this chapter conveys a consistent message: transparency, explainability, and traceability are becoming structural expectations for AI systems, especially in high-risk or security-critical contexts.[2, 1] However, these expectations are articulated at a high level and leave significant discretion regarding implementation. For practitioners working with SLMs in cybersecurity, there is a need to translate abstract legal and ethical requirements into concrete design principles and system level properties.[9, 8]

3.6.1 Regulatory Derived Design Principles

Several recurring themes from Sections 3.2 to 3.5 can be distilled into design principles for SLM based systems.

First, traceability by design is central. Both data protection law and cybersecurity regulation demand that organisations can reconstruct how data have been processed, which systems were involved, and how specific decisions were reached. For SLMs, this implies version control over models and prompts, documentation of training and fine tuning procedures, and logging strategies that capture relevant inputs, outputs, and context for

critical decisions.[3, 8] XAI techniques that generate structured explanation artefacts can be integrated into this traceability layer.[7, 6, 4]

Second, contestability and human oversight require explanations that support meaningful challenge and review. Individuals affected by automated decisions, as well as auditors and regulators, should be able to question outputs and consider alternative outcomes. In the context of SLMs in cybersecurity, this may involve counterfactual or example-based explanations that show how changes in behaviour or system state would alter risk assessments or alert prioritisation.[5, 1, 15]

Third, proportionality and risk awareness suggest that transparency efforts should be calibrated to the level of risk posed by the system and to the sensitivity of the domain. High-risk applications demand more thorough documentation, monitoring, and explanation than low risk internal tools. At the same time, transparency must be balanced with security, avoiding disclosures that could aid adversaries.[3, 14] Design principles must therefore account for different levels of disclosure for different audiences.

Fourth, explanations should be multi layered and audience sensitive. Different stakeholders require different kinds of information: analysts may need fine grained, feature level explanations, risk and compliance teams may prioritise model level documentation, and regulators may focus on whether governance processes and safeguards are adequately supported by explanations.[4, 9, 8] A single explanation format is unlikely to satisfy all these needs.

Finally, design must acknowledge the limits and potential risks of explanations. Post-hoc explanations may be unfaithful; detailed disclosures may increase the attack surface; and overly complex explanation interfaces may overwhelm users. A regulatory aligned design philosophy should therefore combine explainability with other controls, such as robust evaluation, red teaming, and constraints on admissible uses, and be explicit about residual uncertainties and limitations.[7, 6, 15]

3.6.2 Stakeholders and Explanation Recipients in Security Operations

The preceding analysis also highlights the diversity of explanation recipients in security operations. Security analysts working in a security operations centre need explanations that help them understand why specific alerts were raised, what indicators of compromise were most influential, and how to prioritise further investigation. System architects and engineers require explanations that reveal model failure modes, corner cases, and interactions with other components, enabling them to improve system design.[3, 24, 25]

Risk officers and compliance teams focus on explanations that demonstrate control and alignment with regulatory expectations, often in the form of model cards or system level documentation.[4, 8] Supervisory authorities and external auditors require explanations that fit within broader narratives about system design, risk management, and incident handling. Finally, in some cases, data subjects or end users may be indirectly affected by SLM based security analytics and need non technical explanations about the main reasons for a decision and available avenues for contestation.[2, 21]

Designing a single explanation mechanism that satisfies all these audiences is neither realistic nor desirable. Instead, XAI for SLMs in cybersecurity should be conceived as a layered system of explanations, where core explanatory artefacts can be presented and contextualised differently depending on the recipient.

3.6.3 Gaps, Tensions, and Open Issues

Several gaps and tensions remain. A recurring tension arises between transparency and security. Detailed disclosure of model internals or explanation logic can help stakeholders understand and trust SLM based systems, but it can also equip adversaries with information that facilitates targeted attacks or evasion strategies. In cybersecurity, where threat actors actively probe defensive systems, this tension is acute.[3, 27]

A second tension concerns the reliability and faithfulness of explanations. Many XAI techniques provide post-hoc explanations that approximate model behaviour rather than reflecting its true internal logic, and SLMs can generate fluent but unfaithful textual explanations.[5, 6, 15] From a regulatory perspective, misleading explanations may undermine accountability rather than support it. There is therefore a need for methods and metrics that provide assurance about the fidelity and stability of explanations in high stakes settings.[7, 1]

Finally, there is uncertainty about how far explainability can realistically go for large and complex models such as SLMs. Some aspects of their behaviour may remain difficult to interpret, especially when they emerge from high dimensional interactions across vast training corpora.[4, 9] This suggests that explainability should be complemented by other forms of control, such as rigorous evaluation, red teaming, and constraints on admissible uses.

3.7 Interim Conclusions and Research Questions

This chapter has examined the regulatory and soft law landscape that shapes expectations about transparency, explainability, and traceability for AI systems in cybersecurity. It has shown how GDPR principles, provisions on profiling and automated decision-making, and accountability tools such as DPIA and algorithmic traceability already create a demand for explainable and controllable data driven systems.[2, 1] The AI Act builds on this foundation with a risk-based framework that links system criticality to transparency and oversight obligations, including for general purpose models such as SLMs.[16, 22] Cybersecurity specific instruments such as NIS2, the CRA, and DORA further reinforce the need for logging, monitoring, and incident ready documentation in security-critical environments.[17, 23, 18] Soft law frameworks and standards add more detailed, although sometimes inconsistent, guidance on how to operationalise explainability.[14, 9]

At the same time, the vocabulary of explainability remains fragmented across communities, and tensions persist between transparency, security, proprietary interests, and the technical limits of XAI methods.[13, 6, 15] There is no simple checklist that can be

mechanically applied. Organisations must interpret and operationalise regulatory expectations in ways that are sensitive to their specific use cases, models, and organisational contexts.[8, 3]

For the purposes of this thesis, the analysis conducted in this chapter serves two main functions. First, it provides a normative backdrop against which different XAI techniques for SLMs can be evaluated: explanations are not only judged by their technical properties, but also by their capacity to support documentation, oversight, and accountability in regulatory constrained environments.[2, 4] Second, it motivates the design of an XAI oriented framework for SLM compliance that explicitly connects model behaviour, explanation artefacts, and governance requirements.[8, 14]

The next chapters build on this foundation. Chapter 5 sets out the methodology for evaluating XAI techniques applied to selected SLMs in cybersecurity relevant tasks. Chapter 6 presents the implementation, results, and analysis of the experimental study, synthesising the findings into a novel XAI oriented framework for SLM compliance in cybersecurity.

Chapter 4

Small Language Models in Cybersecurity

This chapter examines the role of SLMs in cybersecurity, with a particular focus on security operations and the transparency requirements discussed in the previous chapters. It builds on the problem formulation and RQs in Chapter 1; the conceptual and evaluative landscape of XAI presented in Chapter 2 [5, 1, 13, 14, 2, 4]; and the regulatory obligations derived from the GDPR, the AI Act, NIS2, the CRA and related instruments in Chapter 3 [19, 17, 22, 23, 8]. Against this background, the chapter narrows the focus to SLMs that are either general-purpose or explicitly specialised for cybersecurity. The broader literature on language models in cybersecurity has so far concentrated on LLMs and their applications in threat intelligence, malware analysis, incident response and security monitoring [26, 29, 27, 25, 28, 30, 24]. These works highlight both promising capabilities and significant risks, but they do not systematically address the efficiency and deployment constraints that are central for SLMs. In contrast, this chapter concentrates on SLMs as efficiency oriented counterparts to LLMs, arguing that their properties are better aligned with the operational and regulatory constraints of SOC. The analysis prepares the ground for the experimental methodology in Chapter 5, and the implementation and evaluation in Chapter 6, by clarifying which types of SLMs are relevant, what assumptions they embody, and under which deployment regimes they can satisfy both operational and regulatory constraints.

4.1 Defining Small Language Models for Security Operations

Recent survey work has introduced the term SLMs for architectures that operate in the same autoregressive or encoder-decoder paradigm as LLMs, but are designed to run under more stringent resource constraints and with tighter control over latency, memory usage and deployment costs [10]. Nguyen et al. propose a taxonomy of SLMs based on their construction method, distinguishing models trained from scratch, distilled models,

compressed models obtained through pruning or quantisation, and hybrid approaches that combine several techniques [10]. Across these categories, the common goal is to achieve acceptable task performance under limited computational budgets, for example on edge devices, on-premises servers or low-cost accelerators.

Industry oriented sources converge on a similar notion. Technical overviews and vendor guides describe SLMs as models that are small enough to run efficiently on commodity hardware, while still providing useful language understanding and generation capabilities, and often emphasise their suitability for embedded and on-device scenarios [31, 32]. The exact parameter threshold is not fixed, but SLMs are typically situated in the single digit to low double-digit billion parameter range, in contrast to frontier LLMs that exceed one hundred billion parameters. The key distinction is therefore functional rather than purely numerical: SLMs are optimised for efficiency, deployability and integration into downstream systems rather than for maximising raw benchmark performance at any cost [10].

Recent SLMs families illustrate these design choices in practice. SmolLM2 is a data centric SLM with around 1.7 billion parameters, trained on approximately eleven trillion tokens using a multi stage curriculum that mixes web text with specialised math, code and instruction data [33]. The authors show that careful data curation and training strategies allow SmolLM2 to approach or match the performance of significantly larger models on a range of benchmarks, while remaining inexpensive to run [33, 34]. Nemotron H, by contrast, explores architectural innovation through hybrid Mamba Transformer models that replace most self attention layers with state space components to reduce computation and memory per token [35]. The associated research artefacts and engineering blog emphasise that Nemotron H is intended as a scalable family that can support reasoning and tool use at lower inference costs than traditional Transformer-only models [36, 37]. Both cases support the general claim that efficiency and deployability can be treated as first-class design objectives in modern language models.

Nguyen et al. also highlight that the SLM landscape is populated by many open-weight families released by major providers, such as compact variants of LLaMA, Gemma and Phi, alongside lighter specialised models derived through distillation or compression [10]. These models blur the boundary between traditional LLMs and SLMs, since some were originally introduced as general-purpose LLMs but are now used in practice as small or medium sized backbones due to their parameter counts in the single or low double-digit billion range. This variety is important for cybersecurity because it means that organisations can choose from a spectrum of open-weight and semi-open models, rather than being limited to a few proprietary options.

From a systems perspective, these characteristics make SLMs attractive for SOC environments. Many security tasks are latency sensitive and interactive. Analysts need to triage alerts, summarise incident tickets, correlate threat intelligence reports and support vulnerability management within time constrained workflows. High parameter counts, very large context windows and sophisticated decoding strategies can significantly increase response times and resource consumption, which is problematic in high volume settings where thousands of events may be processed per hour. In contrast, SLMs typically target smaller context windows and more efficient architectures, which in turn reduce memory

footprint and inference latency [10, 33]. Even when their raw accuracy on open domain benchmarks is lower than that of frontier LLMs, the operational benefits can be decisive in security contexts where responsiveness and stability are crucial [3].

A second reason why SLMs are appealing in cybersecurity relates to deployment flexibility. Surveys on LLMs in cybersecurity agree that most current deployments in threat intelligence, malware analysis and SOC automation rely on large, often proprietary models offered through remote services [26, 29, 27, 25, 28, 30, 24]. These models are effective in many tasks, but they raise persistent concerns about data residency, vendor lock in, limited observability and the difficulty of meeting transparency and audit requirements. Due to their reduced computational requirements, SLMs can often be deployed on commodity servers or moderate on-premises clusters instead of highly specialised accelerators in cloud environments [10, 31]. This makes it more realistic to host models within infrastructure controlled by the organisation, for instance in the same data centres where SOC tooling, security information and event management platforms and ticketing systems already reside. Local or EU based deployment reduces data exfiltration risks and gives security teams more control over logging, access control and model updates, which are central both for security posture and for compliance with the data protection and operational resilience requirements examined in Chapter 3 [8, 2].

Nguyen et al. also note that many SLMs are explicitly designed to support quantisation and compression, for example by constraining activation ranges or using training regimes that are robust to reduced numerical precision [10]. These properties further facilitate deployment in edge and near edge environments, such as gateways that collect network telemetry or security appliances that monitor specific segments of an infrastructure. While such scenarios are common in industrial or Internet of Things settings, the same pattern applies to security appliances in corporate networks, which increasingly embed AI components for anomaly detection and automated response [3]. In these settings, SLMs offer a better fit than very large models that would require offloading sensitive data to external providers.

In this thesis, the notion of deployment regimes introduced in Chapter 3 provides a useful lens for classifying both LLMs and SLMs. At a high level, three regimes are distinguished: open-weight models whose parameters can be downloaded and hosted by the organisation, semi-open models where weights are available but subject to licensing or technical restrictions, and closed models that are only accessible through an API offered by an external provider. For SLMs, the open-weight regime is particularly important because the smaller size makes it feasible to host the model locally, analyse its behaviour and integrate it into existing security tooling [10, 31]. Semi-open and closed regimes remain relevant, but the main motivation for SLMs in security operations is precisely the possibility of moving away from purely closed, black-box services.

Operationalising this deployment perspective for concrete security use cases requires a closer look at how SLMs are architected, how they can be combined with tools and agents, and how domain specialisation changes their behaviour. The next sections deepen this analysis by first examining general-purpose and agentic SLMs, then focusing on cybersecurity expert models exemplified by CyberPal 2.0.

4.2 General-purpose and agentic SLMs: lessons for cybersecurity

The recent literature on agentic AI argues that powerful behaviour in complex environments emerges not from single monolithic models alone, but from systems where language models orchestrate tools, memory and external services as part of multi step workflows. Within this paradigm, SLMs are proposed as a natural backbone because they can be executed repeatedly, invoked in parallel and deployed close to the data sources without incurring prohibitive costs [11]. Agentic SLMs are therefore designed not only as text generators, but as components that plan, call tools, interpret intermediate results and revise their own actions.

Belcak et al. analyse several disadvantages of relying exclusively on very large models for agentic systems, including inference cost, latency variability and difficulties in deployment at scale [11]. They argue that many agentic applications involve repetitive, structured tasks that do not require the full generality of frontier LLMs, and propose that well designed SLMs can perform these tasks more reliably and efficiently. The paper introduces design patterns for SLMs that are explicitly optimised for agentic settings, including tool augmented inference, programmatic scaffolding and hierarchical planning. Instead of maximising single turn benchmark scores, these models are tuned for robust function calling, structured output generation and the ability to operate as part of a larger control loop with external verification [11].

Concrete model families illustrate these ideas. SmolLM2 is presented as a small but capable backbone that can serve as the reasoning engine in tool augmented systems, with training choices that emphasise robustness and data quality [33, 34]. Nemotron H explores hybrid architectures that aim to preserve reasoning performance while improving throughput, and is explicitly advertised as a family suitable for applications that require frequent calls to reasoning models with structured outputs and function calls [35, 37]. The SLM survey and vendor guides further note that many open-weight or semi-open models released by major providers, such as compact Gemma and Phi variants, can be configured as tool centric backbones in agentic systems [10, 31, 32]. Belcak et al. also discuss agentic systems that rely on distilled versions of Qwen and related models, emphasising that small, carefully tuned models are often sufficient for the repetitive, structured tasks typical of agents [11]. These developments support the broader claim that small, efficient models can underpin agentic systems when combined with appropriate scaffolding.

These design goals are closely aligned with typical workflows in a SOC, which are often described as multi step, tool centric and knowledge intensive in the XAI for cybersecurity literature [3]. Many security tasks are naturally decomposable into sequences of tool invocations: querying log management or security information and event management platforms for events related to an alert, enriching indicators of compromise with threat intelligence feeds, checking vulnerability information against asset inventories, and creating or updating tickets in issue trackers. Surveys of LLMs in cybersecurity report early prototypes where models are used to orchestrate such workflows, but also emphasise challenges in controllability, hallucination and integration with existing tools [26, 29, 27, 24]. An agentic SLM can act as a coordinator that translates analyst intents into

concrete tool calls, interprets the results and maintains a structured reasoning trace. This decomposition not only leverages existing security tools, it also offers opportunities for explanation because intermediate steps, queries and responses can be logged and inspected.

The agentic SLM literature also highlights robustness and governance issues. Because general-purpose models are trained on broad domains, they may exhibit undesired behaviours, such as hallucinating non-existent vulnerabilities or misclassifying benign events as malicious. The surveys by Jaffal et al. and Yazigholami report such issues in threat intelligence and SOC settings, where hallucinated indicators or misclassified events can have operational consequences [29, 27]. Belcak et al. argue that scaffolding, constrained decoding and external validation steps can mitigate these behaviours, and that SLMs integrated into robust agentic pipelines can be easier to monitor and govern than monolithic frontier models [11]. In a security context, similar mechanisms can be used to enforce guardrails, such as limiting the set of tools that the model may call, imposing strict schemas on its outputs, or requiring explicit confirmations before certain actions are executed. These techniques interact with XAI because they produce structured artefacts that explanations can refer to, such as the sequence of tool invocations or the intermediate hypotheses considered by the model [2, 4].

The focus of this thesis is not on any specific general-purpose agentic SLM, and the experimental work in Chapter 5 does not attempt to reproduce sophisticated multi-agent architectures. Nevertheless, the agentic SLM literature is relevant in two respects. First, it supports the claim that small, efficient models are sufficient to orchestrate complex workflows when appropriately combined with tools and scaffolding [11, 32]. Second, it suggests that explanations in such systems will often be hybrid, combining natural language rationales produced by the model with logs and artefacts produced by external tools [3, 5, 1]. These insights motivate the later discussion of SLMs as components of broader security workflows, even though the implemented experiments evaluate them through controlled classification tasks rather than full agentic pipelines.

To move from conceptual lessons to concrete case studies, it is necessary to examine specialised cybersecurity models rather than general-purpose ones. The next section introduces CyberPal 2.0 as a state-of-the-art theoretical reference for cybersecurity expert SLMs, clarifying what domain specialisation can offer even though the model family is not included in the empirical evaluation of this thesis.

4.3 Cybersecurity expert SLMs: the case of CyberPal 2.0

While general-purpose LLMs and SLMs have demonstrated impressive capabilities across a wide range of tasks, their behaviour in security operations remains problematic. Surveys on LLMs in cybersecurity report issues such as hallucinated indicators, incomplete coverage of recent threats, inconsistent reasoning about vulnerabilities and significant costs associated with high volume API usage in SOC and threat intelligence workflows [26, 29, 27, 25, 28, 30, 24]. Prior work on XAI for cybersecurity further shows that domain knowledge, high stakes and adversarial environments make it difficult to trust black-box models that are

trained on generic corpora and exposed via opaque interfaces [3, 8]. Moreover, using generic frontier models often conflicts with data minimisation and data residency requirements when logs or incident data must be sent to third-party providers [2].

Levi et al. propose CyberPal 2.0 as a response to these limitations [12]. CyberPal 2.0 is a family of cybersecurity expert SLMs that are designed as a unified backbone for a broad set of security operations tasks, including cyber threat intelligence summarisation, alert triage, investigative reasoning, vulnerability analysis and security decision support. The models range from approximately four to twenty billion parameters and are trained on a dedicated security corpus, which positions CyberPal 2.0 firmly within the SLM regime while still providing sufficient capacity for complex reasoning [12]. In contrast to generic LLMs that are later adapted to security tasks, CyberPal 2.0 is designed from the outset as a security focused backbone.

A central ingredient in CyberPal 2.0 is the SecKnowledge 2.0 data pipeline, which constructs a multi source, multi stage security knowledge base tailored to model training [12]. The pipeline collects data from heterogeneous sources, such as vulnerability databases, security advisories, malware analysis reports, threat intelligence blogs and community knowledge bases. These raw inputs are cleaned, de duplicated and normalised, for example by mapping vulnerabilities to canonical identifiers, extracting structured fields and aligning related documents. The pipeline also enriches the data with derived artefacts, such as extracted indicators of compromise, inferred relationships between entities and simplified summaries. Compared to the largely undocumented training corpora of many general-purpose LLMs, this explicit curation is an important step towards domain relevance and data governance.

Beyond basic curation, SecKnowledge 2.0 employs multi step grounding and expert in the loop supervision. Grounding refers to the process of transforming raw text into structured prompts and target outputs that reflect realistic SOC interactions, such as question answer pairs, multi turn investigations or decision support scenarios. For example, a long threat report may be converted into a sequence of tasks where the model must extract key tactics, techniques and procedures, assess the impact on a given infrastructure, or propose mitigation steps. Expert annotators then review and refine these transformations, providing chain of thought style rationales that guide the model towards explicit intermediate reasoning [12]. This contrasts with generic instruction tuning, where supervision is often crowd sourced and not domain-specific, and aligns with recommendations from the XAI literature on aligning explanations with expert reasoning processes [5, 1].

The empirical evaluation reported for CyberPal 2.0 covers a range of benchmarks designed to approximate real-world security tasks. These include cyber threat intelligence question answering, classification of alerts and events, vulnerability description understanding, and scenario based investigative reasoning [12]. Across these tasks, CyberPal 2.0 variants achieve performance that is competitive with or superior to general-purpose LLMs and SLMs of comparable size, particularly on specialised benchmarks that require precise understanding of security terminology and procedures. The authors report improvements in calibration and reductions in hallucination rates compared to generic baselines, especially in settings where incorrect assertions about vulnerabilities or misinterpretation of indicators could have operational consequences [12]. This pattern is consistent with

observations in surveys of domain adapted LLMs, which often report improved reliability when models are trained on curated, expert annotated corpora [25, 28].

From a SOC perspective, CyberPal 2.0 offers several strengths. The domain-specific training data and expert supervision reduce the semantic gap between the model and the tasks that analysts perform daily, which increases the likelihood that the model will produce contextually appropriate outputs, such as correctly prioritised vulnerability summaries or plausible investigation hypotheses. The chain of thought supervision provides an implicit form of explanation: even when explicit XAI techniques are not applied, the model tends to express structured reasoning steps, which can support human understanding and post-hoc analysis. The availability of multiple sizes within the same family allows organisations to select models that match their deployment constraints, from resource constrained settings to larger clusters with more abundant compute [12]. These features make CyberPal 2.0 a useful state-of-the-art reference for reasoning about how a domain-specialised SLM could support SOC workflows, in contrast to generic LLMs that may offer stronger open domain performance but less control, transparency and domain alignment.

At the same time, CyberPal 2.0 has limitations that are relevant for a thesis on trustworthy and explainable AI. One concern is data transparency. While the SecKnowledge 2.0 pipeline is described at a high level, some details about the underlying data sources, filtering criteria and annotation processes may not be fully disclosed for security, privacy or commercial reasons [12]. This complicates assessments of dataset bias, coverage and potential contamination with benchmark data, which are key issues in the XAI literature on evaluation and trust [7, 13, 15]. A second issue is reproducibility. Training specialised SLMs requires substantial computational resources and access to curated security corpora, which may not be available to all practitioners. Even if the authors release pre trained checkpoints and parts of the pipeline, replicating the full training process, including expert in the loop supervision, remains challenging.

Another limitation relates to evaluation scope. Although the benchmarks used in the CyberPal 2.0 study are more security oriented than generic natural language understanding datasets, they still represent a finite snapshot of tasks and scenarios. There is a risk that the model may overfit to certain benchmark patterns or that its performance may degrade when confronted with novel attack techniques, emerging vulnerabilities or organisation specific log formats. The dynamic nature of cyber threats implies that any static evaluation will have limited validity over time, unless it is complemented by continuous monitoring and re evaluation [3, 27]. These concerns echo broader critiques of benchmark driven evaluation in XAI, which emphasise the need for task specific, longitudinal assessments [14, 4].

Despite these caveats, CyberPal 2.0 remains highly relevant for this thesis as a theoretical reference model. It exemplifies a family of truly security specialised SLMs that are small enough to be discussed in relation to realistic SOC constraints, while still covering a broad set of operations tasks. The model family is therefore used in this chapter to frame the state of the art in domain specialisation and to clarify what a security expert SLM might offer in terms of task alignment, reasoning support and explainability requirements. Furthermore, the chain of thought supervision and explicit grounding in security data help to illuminate the interaction between intrinsic model rationales and external XAI methods

[2, 8]. However, the empirical study developed in Chapters 5 and 6 does not evaluate CyberPal 2.0 directly; instead, it focuses on lighter open-source models that were more readily implementable and instrumentable within the operational constraints of the thesis.

4.4 Deployment regimes and constraints for SLMs in cybersecurity

As anticipated in Chapter 3, the way SLMs are deployed in practice has significant implications for transparency, controllability and compliance. In the context of cybersecurity, it is useful to distinguish three deployment regimes that apply to both general-purpose and specialised models: open-weight models, semi-open models and closed API based services. Each regime offers different degrees of control over the model and its environment, and thus enables or constrains different XAI strategies [2, 8].

Open-weight SLMs are models whose parameters and, in some cases, training scripts are publicly available under licences that allow local hosting. When an organisation deploys such a model within its own infrastructure, it gains extensive control over how the model is integrated into security workflows, how inputs and outputs are logged, and how updates are handled. From an explainability standpoint, open-weight SLMs enable access to internal representations and gradients, which allows families of XAI methods that rely on model internals, such as gradient based feature attributions or representation probing, as discussed in Chapter 2 [5, 1, 7]. In addition, the organisation can instrument the model by collecting detailed telemetry about prompts, intermediate states and outputs, which is valuable for incident investigation and for meeting audit obligations under the GDPR, the AI Act and sectoral regulations on digital operational resilience [19, 22, 23, 8].

Semi-open SLMs occupy an intermediate position. Their weights may be available, but subject to licensing restrictions that limit certain uses, such as commercial deployment or further training, or the provider may offer only restricted interfaces for fine-tuning or instrumenting the model. In such settings, some XAI methods that require internal access remain feasible, but the organisation may have less freedom to modify the model or integrate custom monitoring and logging. Compliance with data protection and cybersecurity regulations is still possible, but it may depend on contractual arrangements with the provider and on guarantees about where inference takes place and which logs are retained [2, 28].

Closed models, by contrast, are only available as API services operated by external providers. The organisation can send prompts and receive outputs, but has no direct access to the model parameters or training data. This regime severely constrains the range of XAI methods that can be applied, since internal gradients and representations are inaccessible. In practice, explanations must rely on model-agnostic techniques that treat the model as a black-box, such as perturbation based methods that probe the sensitivity of outputs to changes in inputs, or on external artefacts such as logs and intermediate computations performed by surrounding systems. Surveys on LLMs for cybersecurity emphasise that, in such settings, observability depends heavily on the provider logging and auditing facilities, which may not be sufficient to meet regulatory expectations for

transparency and accountability [26, 29, 27, 30].

These deployment regimes interact with practical constraints in SOC environments. Many organisations, particularly in regulated sectors such as finance or critical infrastructure, require that sensitive logs and incident data remain within the EU or even within specific national jurisdictions, and that detailed logs of all automated decisions and recommendations be available for a specified retention period, as mandated by NIS2 and financial sector regulation on digital operational resilience. Open-weight or suitably configured semi-open SLMs are well suited to these requirements, since they can be deployed on-premises or in controlled private clouds, with full control over logging and access control [8]. Closed API based models, particularly those hosted outside the EU, pose greater challenges and may require additional technical and contractual safeguards, as emphasised in recent analyses of LLMs for SOC and threat intelligence [24, 30].

From a XAI perspective, deployment regimes also determine the granularity and type of explanations that can be realistically offered. With open-weight SLMs, it is possible to combine intrinsic explanations, such as chain of thought outputs or attention visualisations, with post-hoc methods that use gradients or internal representations. This enables richer analyses, for example tracing how specific terms in a log file influenced the model decision that an event is suspicious, or how particular features of a vulnerability description affected its prioritisation. In semi-open and closed regimes, explanations may need to focus more on input output relationships and on the broader system context, such as which data sources were consulted and which rules were applied by surrounding tools [3, 4]. These differences are particularly important when mapping technical explanations to legal notions of transparency and meaningful information for affected data subjects, as discussed in Chapter 3 [8, 2].

CyberPal 2.0 occupies an interesting position in this landscape because it shows how a domain-specialised, open-weight security model could in principle support a rich explainability portfolio. Levi et al. report the release of model checkpoints, data pipeline components and evaluation code to support research on cybersecurity specialised SLMs [12]. From a technical perspective, this makes CyberPal 2.0 a useful reference point for reasoning about open-weight deployments: in principle, gradient based analyses, representation probing and detailed logging are feasible. At the same time, the specialised training on SecKnowledge 2.0 raises questions about dataset governance, update cadence and alignment with organisational threat landscapes, which must be taken into account when designing an XAI framework for its use in a SOC. For the purposes of this thesis, these issues are treated conceptually rather than empirically. CyberPal 2.0 helps illustrate the opportunities and constraints of state-of-the-art domain specialisation, while the implemented study later narrows its scope to lightweight open-source models that could be executed locally and evaluated under controlled conditions. More generally, general-purpose models may fall into different regimes, from open-weight SLMs that can be hosted on-premises and instrumented, to closed API-based services that remain convenient for prototyping but less compatible with stringent regulatory and operational requirements [10, 31, 26, 29]. These contrasts remain analytically important, even when they are not all instantiated in the empirical study.

4.5 Interim conclusions and selection of SLMs for the experimental study

This chapter has provided a focused examination of SLMs in cybersecurity, connecting the general XAI and regulatory considerations of Chapters 2 and 3 to the concrete experimental design in Chapter 5. Starting from the definition of SLMs as efficiency oriented counterparts to larger LLMs, it has argued that their lower latency, reduced resource requirements and greater deployability make them particularly attractive for SOC settings where responsiveness, controllability and data residency are critical [10, 33, 31]. The agentic SLM literature further shows that small models, when combined with tools and scaffolding, are sufficient to orchestrate complex workflows and to provide structured artefacts that can support explanations [11, 32]. At the same time, surveys on LLMs in cybersecurity highlight limitations related to hallucinations, data governance and cost, which SLMs can partly mitigate through more controllable deployment regimes [26, 29, 27, 25, 28, 30, 24].

The discussion of CyberPal 2.0 illustrates how domain specialisation can be instantiated in a state-of-the-art cybersecurity SLM family. By training on the SecKnowledge 2.0 corpus with expert in the loop, chain of thought supervision, CyberPal 2.0 achieves strong performance on cyber threat intelligence, alert triage and vulnerability reasoning tasks, while offering more appropriate behaviour for security operations than generic models of similar size [12]. In this thesis, however, it serves as a theoretical benchmark rather than as an experimental subject. Its main value lies in clarifying the benefits and governance challenges that specialised security SLMs raise, including data transparency, reproducibility and benchmark coverage, all of which are central to any claim of trustworthiness or regulatory compliance [7, 13, 3, 27, 15].

The analysis of deployment regimes emphasises that open-weight, semi-open and closed models afford different levels of control and observability, which in turn determine which XAI methods are technically feasible and which regulatory obligations can be satisfied in practice [2, 8]. Open-weight SLMs, including specialised families such as CyberPal 2.0 and compact general-purpose backbones from ecosystems such as Gemma and LLaMA, are attractive because they can be hosted on-premises, instrumented and combined with a wide range of XAI techniques. Semi-open and closed models remain relevant, especially for organisations without the capacity to host models locally, but they impose stronger constraints on explainability, logging and auditability, which need to be explicitly addressed in any trustworthy AI framework [24, 30].

On the basis of this analysis, the experimental study in Chapters 5 and 6 deliberately narrows its empirical scope to lightweight open-source models that can be hosted locally and instrumented end-to-end on the phishing detection and vulnerability severity tasks. CyberPal 2.0 is retained as a conceptual state-of-the-art reference for domain specialisation, while compact general-purpose backbones, semi-open models and closed services remain analytical reference points rather than separately benchmarked model families. The exact hyperparameters and training procedures of the implemented models are outside the scope of this chapter and will be detailed in Chapter 5.

The rationale for this scoping decision is threefold. First, lightweight open-source models offer direct access to weights, gradients and execution logs, which is essential for a controlled evaluation of XAI methods. Second, they are compatible with the operational goal of local deployment in security-sensitive environments, where data residency and auditability matter. Third, keeping CyberPal 2.0, semi-open models and closed services as theoretical comparison points still makes it possible to discuss how domain specialisation and deployment constraints reshape the explainability space, without claiming empirical evidence that the implemented study does not provide.

Chapter 5

Methodology for XAI–SLM Experimentation

This chapter details the methodological approach adopted to investigate how explainable artificial intelligence (XAI) techniques can be applied to small language models (SLMs) in cybersecurity workflows. Building on the conceptual and taxonomic analysis of XAI methods developed in Chapter 2 and on the regulatory problem framing and design principles derived in Chapter 3, the goal here is to translate these insights into a reproducible experimental protocol. The methodology is designed to bridge three complementary dimensions of the thesis: the organisation of the XAI design space for language models, the legal and governance constraints that shape transparency duties in security sensitive settings, and the empirical assessment of explanation strategies for SLM-based security tools.

The research design follows a mixed method logic that combines conceptual synthesis, regulatory analysis, and empirical evaluation. Conceptual work on XAI taxonomies and evaluation criteria provides the basis for selecting candidate explanation techniques, identifying relevant dimensions such as faithfulness, robustness, and user-centred interpretability, and positioning methods with respect to stakeholders and purposes of explanations. Regulatory analysis clarifies which properties of explanations are required or expected in different deployment scenarios, for example in high-risk uses of AI systems or in monitoring operations that involve personal data, and thus constrains the design space of acceptable XAI configurations. The experimental component then instantiates this combined perspective in concrete cybersecurity use cases, where selected SLMs are instrumented with multiple XAI techniques and evaluated along quantitative and qualitative dimensions.

Within this overall logic, Chapter 5 serves three main functions. First, it specifies the criteria for choosing representative cybersecurity tasks and datasets that are suitable for SLM-based analysis and for systematic inspection of explanations. Second, it defines how model selection is structured in light of deployment regimes and explainability constraints, while making explicit why the implemented study focuses on lightweight open-weight models that can be executed and inspected locally. Third, it describes the evaluation

protocol used to compare explanation strategies, including the metrics and procedures for both model centred and human-centred assessment. By making these design choices explicit, the chapter supports the reproducibility of the study and clarifies how the experimental workflow addresses the research questions introduced in Chapter 1 and refined in the interim conclusions of Chapter 3.

5.1 Research design and overall approach

The methodological approach adopted in this thesis is structured around a sequence of interconnected phases that mirror the research questions and the layered nature of the problem. At a high level, the design can be seen as a pipeline that moves from conceptual mapping to empirical validation and framework construction, with feedback loops between these stages. Rather than treating XAI techniques, SLMs, and regulatory requirements as separate objects of study, the methodology explicitly couples them, so that each empirical choice is grounded in the conceptual and legal analysis developed in earlier chapters.

The first phase consolidates the conceptual work on XAI presented in Chapter 2 into a practice-oriented design space for explanations in SLM-based systems. Existing taxonomies are reinterpreted in terms of a small number of operational dimensions, such as local versus global scope, post-hoc versus ante-hoc integration, model-agnostic versus model-specific dependence, and the type of explanation artefact produced. These dimensions are then aligned with the stakeholders, purposes, and evaluation challenges discussed in the literature, with specific attention to security operations and compliance scenarios. The outcome of this phase is a structured catalogue of XAI methods and properties that can be instantiated in concrete cybersecurity tasks.

The second phase connects this design space with the regulatory landscape analysed in Chapter 3. Transparency, documentation, and human oversight duties are translated into technical and procedural requirements that explanations should satisfy in different deployment contexts, distinguishing for example between internal support tools for security analysts and AI systems that contribute to decisions with potential legal or similarly significant effects on individuals. This mapping informs which combinations of SLM architectures, explanation techniques, and usage patterns are in principle compatible with the obligations identified under the AI Act, the GDPR, the NIS2, and the CRA.

The third phase implements an empirical study where selected lightweight open-weight models are applied to one or more cybersecurity use cases and instrumented with multiple XAI techniques. For each combination of model, task, and explanation method, the study collects quantitative indicators of explanation quality, such as measures of faithfulness and stability, and qualitative evidence about interpretability and usefulness for security analysts working in a SOC or similar environments. The evaluation protocol is designed to compare explanation strategies under controlled local conditions, while using the broader distinction between open, semi-open, and closed deployments as an analytical lens rather than as a full empirical benchmarking dimension. The empirical findings produced in this phase therefore provide direct evidence about XAI for locally deployable open-source models, and indirect evidence about how far such findings may transfer to more constrained

deployment regimes.

Finally, the insights from the conceptual, regulatory, and empirical phases are integrated into the design of a novel XAI oriented framework for SLM compliance in cybersecurity, which will be presented in Chapter 6. From a methodological perspective, this framework is treated as a design science artefact that synthesises the lessons learned from the experimental study into a reusable workflow. The present chapter therefore not only documents the steps required to reproduce the experiments, but also sets the stage for turning the resulting methodology into a prototype toolkit or reference process that can be adopted and adapted by practitioners.

5.2 Selection of cybersecurity use cases and datasets

The empirical study requires concrete cybersecurity tasks and datasets that are sufficiently realistic to stress SLM-based solutions, yet sufficiently constrained to allow systematic evaluation of explanations. The selection of use cases is therefore guided by three main criteria derived from the problem framing in Chapter 1, the state of the art analysis in Chapter 2, and the regulatory discussion in Chapter 3.

First, use cases must be aligned with the strengths and limitations of SLMs as discussed in Chapter 4. In practice, this means focusing on tasks where the input is primarily textual and structured around security relevant artefacts, such as vulnerability descriptions, security advisories, incident reports, or log snippets. These tasks leverage the capacity of SLMs to encode domain-specific language patterns and to perform classification or scoring over natural language inputs, without requiring full scale autonomous decision-making. At the same time, they expose models to realistic ambiguity and noise, which is essential for testing explanation methods beyond toy examples.

Second, use cases must raise non trivial questions of transparency and accountability, so that explanations are not merely a cosmetic add on but a necessary element of responsible system design. From the perspective of Chapter 3, this includes scenarios where model outputs affect how security teams prioritise investigations, where they can indirectly influence the exposure of systems to attacks, or where they contribute to decisions that may have consequences for organisations or individuals. Examples include the prioritisation of vulnerability reports based on natural language descriptions, the triage of alerts in a SOC setting, or the classification of potentially malicious emails for further inspection. In these scenarios, explanations are needed both to support day to day decision-making and to provide an auditable account of why certain items were flagged or deprioritised.

Third, datasets must be amenable to a rigorous experimental protocol. This requires the availability of enough labelled instances to train or adapt SLMs when needed, to estimate performance with reasonable statistical confidence, and to compute explanation metrics on representative subsets. It also requires that the labelling scheme captures properties that security analysts care about in practice, such as severity, exploitability, affected asset type, or high level category of a threat. Where possible, the study relies on publicly available corpora that include natural language descriptions of security relevant events and associated labels, so that the methodology can be reproduced or adapted by other

researchers. When this is not feasible, synthetic or semi synthetic labelling procedures are clearly documented, including their limitations.

Within this methodological frame, the thesis concentrates on two complementary text classification case studies: phishing email detection and the analysis of vulnerability related texts. The phishing task captures a common SOC workflow in which language cues, formatting patterns, and social-engineering indicators must be inspected rapidly. The vulnerability task instead focuses on descriptions of software weaknesses, advisories, and related security records, where the goal is to infer severity-relevant information from more technical language. Taken together, these use cases provide one binary and one multiclass setting, allowing the methodology to stress the explanation pipeline under different linguistic and decision structures.

For each selected dataset, the methodology specifies a standardised pre processing pipeline, which is detailed in Chapter 6. In this chapter, it is sufficient to highlight the main design choices. Texts are cleaned in a minimal way, preserving technical tokens, identifiers, and domain-specific terminology whenever possible, since these elements often carry crucial information for both model predictions and explanations. Instances are partitioned into training, validation, and test splits, using stratified sampling when class imbalance is present, to ensure that evaluation metrics on explanations are not dominated by a small set of frequent classes. Where the volume of data permits, an additional hold out set is reserved for qualitative analysis of explanations, so that the examples used for case studies are not inadvertently optimised during model selection.

The selection of use cases and datasets is not intended to cover the full diversity of AI applications in cybersecurity. Instead, it provides a focused yet flexible test bed that is rich enough to expose the behaviour of different XAI methods on SLMs, while remaining tractable for detailed analysis. The limitations of this choice, and possible extensions to other tasks such as log anomaly explanation or incident report triage, are discussed in Section 5.6. For the purposes of the present chapter, the important point is that each subsequent methodological choice, from model selection to metric definition, is made with explicit reference to the characteristics of these use cases and datasets.

5.3 Selection of SLMs and model selection criteria

The choice of models plays a central role in the experimental design, since it determines both what kinds of explanations can be computed in practice and how the results of the study can be generalised. In line with the discussion of SLMs and deployment regimes in Chapter 4, the model selection strategy aims to define an empirical portfolio that is realistic for local cybersecurity deployment and, at the same time, amenable to systematic XAI analysis.

From a methodological perspective, three criteria guide the selection. First, models should occupy a plausible point in the trade-off between size, performance, and deployability. This implies including compact SLMs that are realistic candidates for on-premises or edge deployment, without exceeding the computational budget typical of an academic setting. Second, models should maximise controllability and reproducibility, which in practice

favours open-weight checkpoints with stable local execution over API-mediated or licence-restricted alternatives. Third, models should remain sufficiently aligned with cybersecurity language to make the selected tasks meaningful without requiring the full engineering overhead of larger state-of-the-art specialised systems such as CyberPal 2.0.

To satisfy the first criterion, the experimental setup includes only the two open-weight models used in the empirical study, both of which can be fully instrumented for XAI. These task-specific checkpoints are chosen to remain compatible with the hardware resources available for the thesis while still capturing the practical characteristics of lightweight transformer-based cybersecurity tools. Their local executability allows stable logging of the full inference process and direct computation of the explanation methods implemented in Chapter 6. When necessary, these open-weight models can also be adapted to the selected datasets, which is important for generating meaningful explanations in the target tasks.

The broader conceptual framework of the thesis still accounts for semi-open and closed models, because they mark important boundaries for explainability. However, these regimes are not instantiated in the empirical setup. Semi-open models can limit redistribution, fine-tuning, or monitoring freedom, while closed API-based services hide gradients, internal representations, and parts of the execution environment. Including them would have reduced reproducibility, weakened local control over logging and data flows, and made the comparison with gradient-based XAI methods structurally uneven. For this reason, semi-open and closed deployments are treated in this chapter as analytical reference cases for discussing black-box constraints, not as directly evaluated models.

The third criterion concerns domain relevance. Rather than reproducing a larger state-of-the-art cybersecurity family such as CyberPal 2.0 or introducing a separate benchmark family of compact general-purpose baselines, the implemented study relies on lightweight open-source transformer checkpoints already adapted to the target tasks, namely phishing detection and vulnerability severity classification. This choice preserves the link with security-specific language patterns while keeping the experiments within the local computational and operational constraints of the thesis. It also allows the study to observe how XAI behaves on models that are realistic candidates for immediate deployment in controlled organisational environments.

Intrinsically interpretable baselines remain methodologically important as comparison points in the literature, because they remind us that the value of post-hoc explanations must ultimately be assessed against simpler and more directly transparent alternatives. In this thesis, however, they are discussed only as conceptual reference points and are not included in the implemented comparison, whose main objective is to evaluate XAI techniques on lightweight transformer-based models under a uniform execution pipeline.

Overall, the empirical portfolio is intentionally narrow: two lightweight open-weight models, two cybersecurity classification tasks, and three explanation methods that can be executed end-to-end under local control. CyberPal 2.0, compact general-purpose baselines, semi-open deployments, closed services, and intrinsically interpretable baselines remain relevant comparative references at the conceptual level, but they are outside the practical scope of the implemented experiments. For the purposes of the methodological discussion, this scoping choice is not a weakness to hide but a condition for making the evaluation reproducible, operationally realistic, and coherent with the local deployment requirements

that motivate the thesis.

5.4 Portfolio of XAI techniques and evaluation dimensions

The experimental study requires a concrete yet diverse portfolio of XAI methods that can be applied to the selected models on the cybersecurity tasks defined in Section 5.2. The aim is not to exhaust the variety of techniques surveyed in Chapter 2, but to select a subset of families that are technically mature, implementable for text-based classification with transformer models, and representative of the main explanation paradigms discussed in the literature. At the same time, it is important to be explicit about which methods are not included in the empirical evaluation and why, so that the scope and limitations of the proposed framework are transparent.

5.4.1 Selected families of XAI methods

Based on the conceptual analysis in Chapter 2 and on the requirements of the vulnerability and phishing tasks, the implemented experiments focus on three explanation families that are instantiated directly in Chapter 6.

Feature attribution for transformer based models. The first family comprises gradient-based and backpropagation-based feature attribution methods adapted to transformer architectures. These methods assign an importance score to each input token with respect to a given prediction, by computing gradients or variations of activation values. In practice, the implemented runs rely on Integrated Gradients via Captum, which is directly supported by the locally executed transformer checkpoints used in the study. This family captures a central notion of explanation in the LLM literature, namely the idea that explanations should identify the words or phrases that contribute most to a prediction.

Local surrogate and perturbation-based methods. The second family is formed by local surrogate approaches, where simple interpretable models are fitted in the neighbourhood of a specific input instance by probing the black-box model with perturbations. The most prominent example in this category is LIME for text classification, which trains a sparse linear model on perturbed versions of the input and uses its coefficients as an explanation of the original prediction. For the tasks considered in this thesis, local surrogate methods are attractive because they are model-agnostic and can be applied directly to the open-weight models evaluated here, while remaining in principle applicable to closed models accessed only via API. They also provide explanations in the form of weighted features that can be compared to gradient-based attributions.

Attention-based explanations. The third family exploits the internal self attention mechanisms of transformer based SLMs. Attention weights are extracted and aggregated

to produce heatmaps that highlight which tokens were most attended to by the model when producing a prediction. While there is an ongoing debate on whether attention can be treated as a faithful explanation, attention visualisation remains an important diagnostic tool in practice and is particularly natural for language tasks. In this study, attention-based explanations are not treated as a stand alone ground truth, but as an additional perspective that can be compared with feature attribution and surrogate-based explanations in order to understand where they agree or diverge.

No standalone counterfactual explanation module is instantiated in the reported experiments. Perturbation enters the study through the evaluation metrics themselves, for example when deletion- and sufficiency-based tests are used to assess whether the tokens highlighted by an explanation materially influence the model output.

5.4.2 Excluded families and rationale

Several other families of XAI methods discussed in Chapter 2 and in the broader literature are not included in the empirical evaluation. This does not imply that they are irrelevant for SLM-based cybersecurity, but rather that their implementation and assessment fall outside the feasible scope of the present study.

First, Shapley value based approaches such as SHAP are not included in the implemented comparison, even though they remain relevant in the broader XAI literature and could in principle be integrated into the framework. For the scope of this thesis, they were excluded because their computational cost would have significantly reduced the number of instances and cross-task comparisons that could be executed under the available resources. Restricting the study to Integrated Gradients, LIME, and attention therefore made it possible to keep the empirical comparison consistent across both tasks.

Second, intrinsically interpretable models such as generalised additive models, rule lists, or shallow decision trees are considered only as methodological comparison points, as discussed in Section 5.3, and not as XAI methods in their own right for the models implemented in this study. While their transparency is valuable as a conceptual reference, the core research questions focus on post-hoc explanations for black-box and grey box models, which are the dominant paradigm for modern language based security tools.

Third, global function based techniques such as partial dependence plots, individual conditional expectation plots, and accumulated local effects are not employed. These methods are well suited to tabular data with a limited number of features, but they become difficult to interpret and computationally expensive in high dimensional discrete input spaces such as tokenised text. In the vulnerability and phishing tasks considered here, where the number of distinct tokens and positional combinations is large, global function plots would either require unrealistic simplifications or produce artefacts that are hard to relate to the actual decision logic.

Fourth, concept based methods such as testing with concept activation vectors and representation level visualisations based on techniques like t-SNE or UMAP are not integrated in the experimental pipeline. These approaches require the definition and curation of high level concepts through labelled examples and the analysis of internal representation spaces, which is a promising direction for future work but would demand a

dedicated methodological effort. Given the focus of this thesis on instance level explanations that can be directly consumed by security analysts and used as evidence in transparency and accountability processes, concept based analyses are treated as beyond scope.

Fifth, example-based methods in the strict sense, such as prototype and criticism selection or advanced nearest neighbour retrieval in embedding space, are not systematically evaluated. While nearest neighbour examples can be informative in illustrating typical and atypical behaviour, turning them into a fully fledged explanation modality would require careful design of similarity measures, clustering procedures, and user interfaces. In order to keep the experimental design focused, example-based explanations are used, if at all, only in an illustrative way in the qualitative case studies of Chapter 6, and not as a primary target of metric based evaluation.

Finally, sophisticated counterfactual generation methods for text, which search for minimal input modifications that change the model prediction under semantic and grammatical constraints, are not implemented. Existing approaches often rely on additional generation models, complex optimisation procedures, or task specific constraints, and their integration with SLMs for security tasks would significantly increase the complexity of the experimental setup. Instead, the thesis relies on simpler deletion- and sufficiency-based perturbation tests as evaluation instruments, rather than presenting counterfactual generation as a separate explanation family.

5.4.3 Evaluation dimensions for explanations

The choice of these three implemented explanation families is tightly coupled with the evaluation dimensions that will be used to assess explanations. Drawing on the discussion in Chapter 2 and on the regulatory requirements analysed in Chapter 3, the study focuses on three main groups of properties.

- **Fidelity or faithfulness**, that is, the extent to which an explanation reflects the actual behaviour of the underlying model. For feature attribution and surrogate-based methods, fidelity will be probed through perturbation tests and deletion or sufficiency metrics that measure how much predictions change when the features identified as important are removed or isolated.
- **Stability or robustness**, understood as the sensitivity of explanations to small perturbations of the input. For the selected methods, stability can be quantified by computing similarity measures between explanations generated for slightly modified versions of the same vulnerability description or email, for example by replacing non critical tokens or introducing benign noise.
- **Task and stakeholder alignment**, which concerns the usefulness and adequacy of explanations for security analysts and for governance processes. While fully fledged user studies are beyond the scope of this thesis, proxy indicators are used, such as whether highlighted tokens and patterns correspond to security-meaningful elements (for example vulnerability identifiers, affected components, or typical phishing cues) and whether explanations can be logged and reused as evidence in documentation and audit contexts.

Complexity and sparsity remain relevant interpretive concerns, but they are not reported in this thesis as standalone quantitative metrics. To keep the comparison consistent across methods, explanations are generated under common sampling and top- k settings, while issues of compactness or diffuseness are discussed qualitatively where they materially affect analyst usability.

By restricting the empirical evaluation to a portfolio of XAI methods that can be meaningfully assessed along these dimensions for the selected tasks and models, the methodology aims to produce evidence that is both technically grounded and relevant for the compliance-oriented framework developed in Chapter 6. At the same time, being explicit about excluded families clarifies that the proposed framework is open to extension, and that future work may incorporate additional explanation paradigms as long as they can be evaluated with comparable rigour.

5.5 Evaluation metrics and experimental protocol

The portfolio of XAI methods outlined in Section 5.4 must be coupled with an evaluation procedure that is both rigorous and feasible for the selected tasks and models. The purpose of this section is to define how the abstract properties introduced in Chapter 2 and recalled in Section 5.4 are operationalised into concrete metrics, and how these metrics are computed within a controlled experimental protocol. The focus is on instance level explanations for text classification tasks, where explanations take the form of token-level importance scores or simple surrogate models.

From a high-level perspective, the evaluation protocol proceeds in three stages. First, for each combination of model, task, and XAI method, explanations are generated on a representative subset of instances from the test split of the dataset. Second, quantitative metrics are computed from these explanations, capturing fidelity and stability in a manner comparable across methods. Third, selected instances and their explanations are examined qualitatively to assess task and stakeholder alignment and identify patterns that may not be captured by aggregate statistics. The following paragraphs specify the main metrics and procedures used in each stage.

Fidelity metrics. Fidelity, or faithfulness, is assessed through perturbation-based tests that measure how model predictions change when features identified as important by an explanation are removed or isolated. For the token attribution methods implemented in this thesis, the protocol uses deletion and sufficiency style scores. In the deletion setting, for each instance and explanation, the top k tokens by importance score are removed from the input (for example replaced with a neutral token or masked), and the change in predicted probability for the original class is recorded. High fidelity explanations are expected to lead to a significant drop in confidence when their most important tokens are removed. In the sufficiency setting, only the top k tokens are retained, the rest of the input is masked, and the probability of the original class given this reduced input is measured. If explanations capture the truly decisive features, the model should remain relatively confident under sufficiency. For local surrogate models, fidelity is approximated

by the agreement between the surrogate and the original model on perturbed instances in the local neighbourhood of the explanation point. These metrics provide complementary views on whether explanations reflect the decision logic of the underlying models.

Stability metrics. Stability, or robustness, is measured by generating explanations for slightly perturbed versions of the same input and quantifying how similar they are. Perturbations are designed to be semantically benign for the task, for example by introducing variations in whitespace, changing non critical tokens such as stopwords, or inserting neutral markers where appropriate. For each original instance and its perturbed variant, explanations are computed with the same XAI method and represented as vectors of token importance scores. Stability is then quantified through similarity measures such as rank correlation between these vectors, aggregated over the evaluation subset. High stability implies that small, non substantive changes in the input do not produce radically different explanations, which is important if explanations are to be used as evidence in incident investigation or audit processes.

Complexity as an interpretive consideration. Although explanation compactness is relevant for analyst usability, it is not operationalised in this thesis as a standalone reported metric. Instead, explanation size is controlled indirectly through common top- k settings and is discussed qualitatively when explanations appear excessively diffuse or overly sparse for the task at hand. This keeps the empirical comparison focused on faithfulness- and robustness-oriented metrics while still acknowledging the practical importance of explanation size.

Task and stakeholder alignment. Task and stakeholder alignment is more difficult to capture through purely automatic metrics, since it involves questions of meaning and usefulness that depend on domain knowledge and organisational roles. In this thesis, alignment is assessed through a combination of proxy indicators and qualitative analysis. For vulnerability related tasks, one indicator is whether explanations systematically highlight tokens that correspond to security-meaningful elements, such as affected components, vulnerability types, or impact descriptors, rather than superficial lexical artefacts. For phishing detection, explanations are examined for the presence of typical phishing cues, such as urgency phrases, suspicious links, or requests for sensitive information. In addition, the experimental setup is configured so that explanations can be logged and associated with model decisions, creating artefacts that could, in principle, be integrated into documentation or audit trails as required by the instruments discussed in Chapter 3. While no formal user study with practitioners is conducted, these indicators provide a first approximation of whether explanations are aligned with the informational needs of analysts and governance functions.

Experimental protocol. The practical execution of the evaluation follows a structured protocol. For each dataset and task defined in Section 5.2, models selected according to Section 5.3 are trained or adapted using the training split, with hyperparameters and

training procedures reported in Chapter 6. The validation split is used for model selection and tuning, without computing explanation metrics. Once model configurations are frozen, predictions and explanations are generated on the test split. For computational reasons, the full test set may be used for model performance metrics, while a stratified subset is sampled for explanation metrics, ensuring coverage of all relevant classes and input lengths. Explanations from different methods are computed on the same subset of instances to allow direct comparisons.

Metrics for fidelity and stability are then computed for each method and model, aggregated at the level of the task. In the reported results, these metrics are presented primarily as task-level mean values over the evaluation subset, complemented by qualitative case studies that surface method-specific strengths, weaknesses, and failure modes. Finally, a small number of instances per task are selected for more detailed qualitative analysis in Chapter 6, where explanations are visualised and discussed with reference to the typical workflows of security analysts.

By combining these quantitative and qualitative components within a single protocol, the methodology aims to strike a balance between reproducibility and contextual relevance. The quantitative metrics provide structured evidence that can support comparisons between XAI methods and models, while the broader discussion in later chapters situates these findings within the wider landscape of open, semi-open, and closed deployments. The qualitative analysis ensures that such comparisons are interpreted in the light of domain-specific considerations and regulatory expectations, rather than in purely abstract terms.

5.6 Threats to validity and limitations

No empirical study is free from limitations, and the validity of the conclusions that can be drawn from the proposed experimental setup depends on how these limitations are recognised and addressed. This section briefly discusses the main threats to validity that affect the methodology outlined in this chapter and indicates how they are mitigated where possible.

A first threat concerns external validity, that is, the extent to which the results generalise beyond the selected tasks, datasets, and models. The focus on vulnerability related text and phishing emails provides a coherent and practically relevant test bed, but it does not cover the full spectrum of cybersecurity applications of AI, such as log based anomaly detection, malware analysis, or user behaviour analytics. Similarly, the set of lightweight open-weight models included in the study represents only a subset of the models available in the fast evolving ecosystem described in Chapter 4. To mitigate this, the methodological discussion explicitly treats the chosen tasks and models as instances of broader classes and emphasises patterns and trade-offs that are likely to be stable across similar settings, rather than claiming universal rankings of methods.

A second threat relates to construct validity, that is, whether the metrics used in the study actually capture the properties they are intended to measure. Fidelity and stability are inherently multi faceted concepts, and the specific operationalisations adopted here, such as deletion based scores or rank-correlation-based robustness checks, provide

only partial views of these constructs. There is a risk that explanations may perform well according to these metrics while still being misleading or unhelpful in practice, or conversely that methods with poor metric scores might still offer valuable insights in some contexts. To reduce this risk, the study combines multiple metrics where feasible and complements quantitative analysis with qualitative case studies. Explanation compactness and stakeholder usefulness are treated as qualitative interpretive dimensions rather than as standalone reported metrics. Moreover, the limitations of each metric are discussed alongside the results in Chapter 6, to avoid over interpretation.

A third threat concerns internal validity, in particular the possibility that implementation details, hyperparameter choices, or data preprocessing steps inadvertently favour some methods or models over others. Since different XAI methods have different sensitivity to hyperparameters and to the distribution of inputs, careless implementation could introduce systematic biases. To mitigate this, the experimental protocol adopts consistent preprocessing pipelines across models for each task, uses common evaluation subsets for all explanation methods, and documents key implementation choices in Chapter 6. When possible, default configurations recommended in established libraries are used as starting points, and sensitivity to major hyperparameters is checked on a small scale.

A fourth threat involves the absence of direct human-centred evaluation. The thesis does not include controlled user studies with security analysts or compliance officers, which limits the strength of claims about usability, trust, or decision support. Proxy measures of task and stakeholder alignment, while informative, cannot fully substitute for systematic human feedback. This is partly mitigated by grounding the qualitative analysis in typical workflows of a SOC as described in Chapter 1 and by linking explanation properties to concrete regulatory expectations in Chapter 3. Nevertheless, any conclusions about human factors must be considered tentative and subject to refinement in future work that includes direct interaction with practitioners.

Finally, there are practical limitations related to computational resources and time. Some XAI methods, especially local surrogate approaches and heavier attribution techniques considered as possible extensions, are computationally expensive when applied to language models on sizeable datasets. This constrains the number of instances for which explanation metrics can be computed and may limit the exploration of more complex tasks or larger models. To manage this, the protocol restricts explanation intensive computations to stratified subsets of the test data and prioritises depth of analysis over breadth in the selection of cases. While this choice is reasonable for a master level thesis, it implies that the empirical findings should be seen as evidence from well chosen samples rather than exhaustive coverage.

Recognising these threats to validity does not invalidate the methodology, but clarifies the scope within which its results can be interpreted and the directions in which the work can be extended. In particular, the compliance-oriented framework developed in Chapter 6 is presented not as a definitive standard, but as a structured proposal grounded in this specific set of experiments and open to refinement as new models, methods, and evaluation practices emerge.

Chapter 6

Implementation and Results

This chapter details the implementation of the experimental methodology described in Chapter 5 and presents the findings of the empirical study. The central artefact developed for this purpose is the *XAI-SLM Cyber Eval* framework, a modular, configuration-driven toolchain designed to execute, evaluate, and explain SLM predictions on cybersecurity classification tasks. By bridging technical performance metrics with structured explainability assessments, the framework provides a concrete instantiation of the principles discussed in previous chapters, particularly the need for reproducible, auditable, and regulatory-aligned AI operations.

The chapter is structured as follows. Section 6.1 outlines the scope and goals of the implemented artefact. Section 6.2 describes the overarching architecture and its key components. Section 6.3 details the configuration mechanisms and execution flow, providing concrete examples of the setups used for the use cases. Section 6.4 catalogues the generated outputs and their role in establishing an auditable trail. Section 6.5 presents the results of the experimental study across the selected cybersecurity scenarios, focusing on both predictive performance and the evaluation of the generated explanations. Finally, Section 6.6 contextualises these results within the broader operational and regulatory landscape of security operations.

6.1 Scope and Goals of the Artifact

The primary objective of the *XAI-SLM Cyber Eval* framework is to operationalise the comparative assessment of different explanation techniques applied to lightweight open-weight models within realistic cybersecurity workflows. In line with the requirements identified in Chapter 3 and the methodological constraints outlined in Chapter 5, the framework was designed to satisfy four core goals:

- **Configurability and Reproducibility:** Security experiments must be easily configurable and highly reproducible. The framework relies on a declarative YAML-based configuration system, allowing precise specification of datasets, models, task definitions, and XAI parameters without requiring code modifications. This ensures that

every experimental run can be deterministically repeated and peer-reviewed.

- **Modularity and Extensibility:** The diversity of possible SLM deployment regimes and explanation methods requires a decoupled architecture. By relying on abstract interfaces and dynamic registries, the framework cleanly separates data handling, inference, and explainability, enabling the seamless integration of new models or XAI algorithms, even though the present experiments focus on locally executable open-weight models.
- **Systematic XAI Evaluation:** Beyond merely generating explanations, the artefact quantitatively assesses their quality. It implements standardized metrics, namely comprehensiveness, sufficiency, and stability, to evaluate the faithfulness and robustness of feature attributions, providing structured evidence rather than subjective interpretations.
- **Audit Readiness and Reporting:** To align with regulatory expectations under instruments such as the AI Act and NIS2, the framework automatically generates comprehensive run artefacts, including machine-readable metrics, visual plots, and structured reports. These outputs are intended to serve as tangible documentation for internal review, incident triage contexts, and external audits.

An important scoping decision concerns deployment regimes. Although Chapter 5 framed the broader methodological discussion against open-weight, semi-open, and closed models, the implemented experiments deliberately focus on two lightweight open-source models that could be run locally and instrumented end-to-end. This restriction was necessary for three concrete reasons. First, the evaluation depends on direct control over weights, gradients, tokenisation, and intermediate artefacts so that attention-based and gradient-based explanations can be computed under identical and inspectable conditions. Second, the thesis requires a reproducible evidence trail with stable logging of inputs, predictions, and explanation outputs; this would have been weakened by API rate limits, provider-side model updates, opaque serving stacks, or licence restrictions on instrumentation. Third, the targeted cybersecurity setting assumes strong data residency and auditability constraints, which are easier to guarantee when inference remains on infrastructure controlled by the evaluator. For these reasons, CyberPal 2.0 is retained as a state-of-the-art theoretical benchmark rather than as an experimental subject, while semi-open and closed services are excluded from the empirical runs. As a consequence, RQ2 is addressed empirically only for open-weight deployments; for black-box or closed-source models, XAI remains conceptually relevant but is discussed here only through the lens of post-hoc, model-agnostic techniques and system-level controls.

6.2 Framework Overview

The architecture of the *XAI-SLM Cyber Eval* framework follows a pipeline design, orchestrated by a central runner that sequentially executes discrete stages. This flow is

managed through abstract base classes and internal registries, ensuring loose coupling between components.

The pipeline consists of the following sequential modules, each responsible for a distinct phase of the evaluation process:

The orchestration logic of these modules is summarised in Figures 6.1, 6.2, 6.3, 6.4, and 6.5, which provide a visual overview of how configuration, data ingestion, model execution, explanation generation, and reporting are linked within the framework.

6.2.1 Configuration Parser

The Configuration Parser is the entry point of the pipeline. It validates the input YAML file against strictly typed schemas, coercing parameters and setting sensible defaults for the experiment. This ensures that the configuration is complete and correctly formatted before any computationally expensive steps are initiated, guaranteeing reproducibility across runs.

6.2.2 Dataset Loader

The Dataset Loader module (`datasets/`) manages the ingestion of cybersecurity text data. It handles dataset splitting, stratification, and sample limits. By interfacing with sources like the Hugging Face Hub, it retrieves vulnerability descriptions or phishing emails and packages them into standardized data structures ready for model ingestion.

6.2.3 Model Wrapper and Evaluation Adapter

The Model Wrapper (`models/`) abstracts the underlying ML backends. Models are wrapped in classes exposing explicit *capabilities* (e.g., `supports_gradients`, `supports_tokenization`)}. This dynamically determines which XAI methods can be applied. Following inference, the Evaluation and Label Adapter (`eval/`) computes standard classification metrics. Crucially, it incorporates a label adaptation layer that maps model logits to the target label space using configurable aggregation strategies (e.g., mapping granular sub-categories to a binary legitimate/phishing task).

6.2.4 Explainability Orchestrator

The Explainability Orchestrator (`explain/`) applies the requested XAI methods based on the model’s capabilities. It routes requests to specific implementations, supporting ante-hoc attention extraction as well as post-hoc methods such as Integrated Gradients via Captum and perturbation-based methods like LIME. The framework can be extended with slower explainers such as SHAP, but the experiments reported in this chapter instantiate only the three methods listed in the configuration files below. This phase computes token-level importance scores to interpret the model’s decisions.

6.2.5 Reporting Engine

Finally, the Reporting Engine (`reporting/`) constructs the final artifacts. It synthesises the metrics and explanations, using Jinja2 templates to compile Markdown reports. It also utilizes HTML-to-PDF converters to render presentation-ready documents, seamlessly embedding computed visually interpretable plots and tables to facilitate audit readiness.

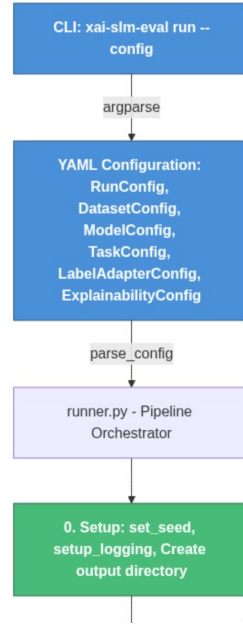


Figure 6.1: Flowchart detailing the Configuration Parser phase.

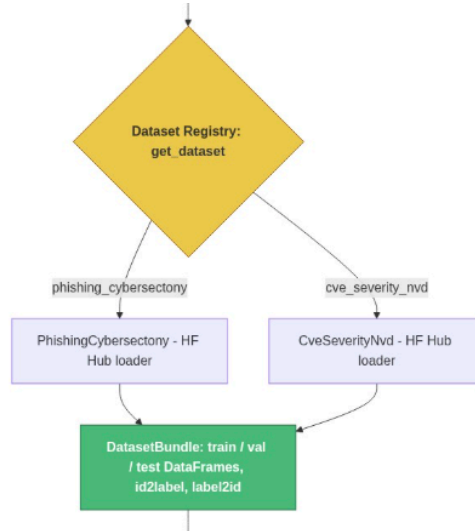


Figure 6.2: Flowchart detailing the Dataset Loader phase.

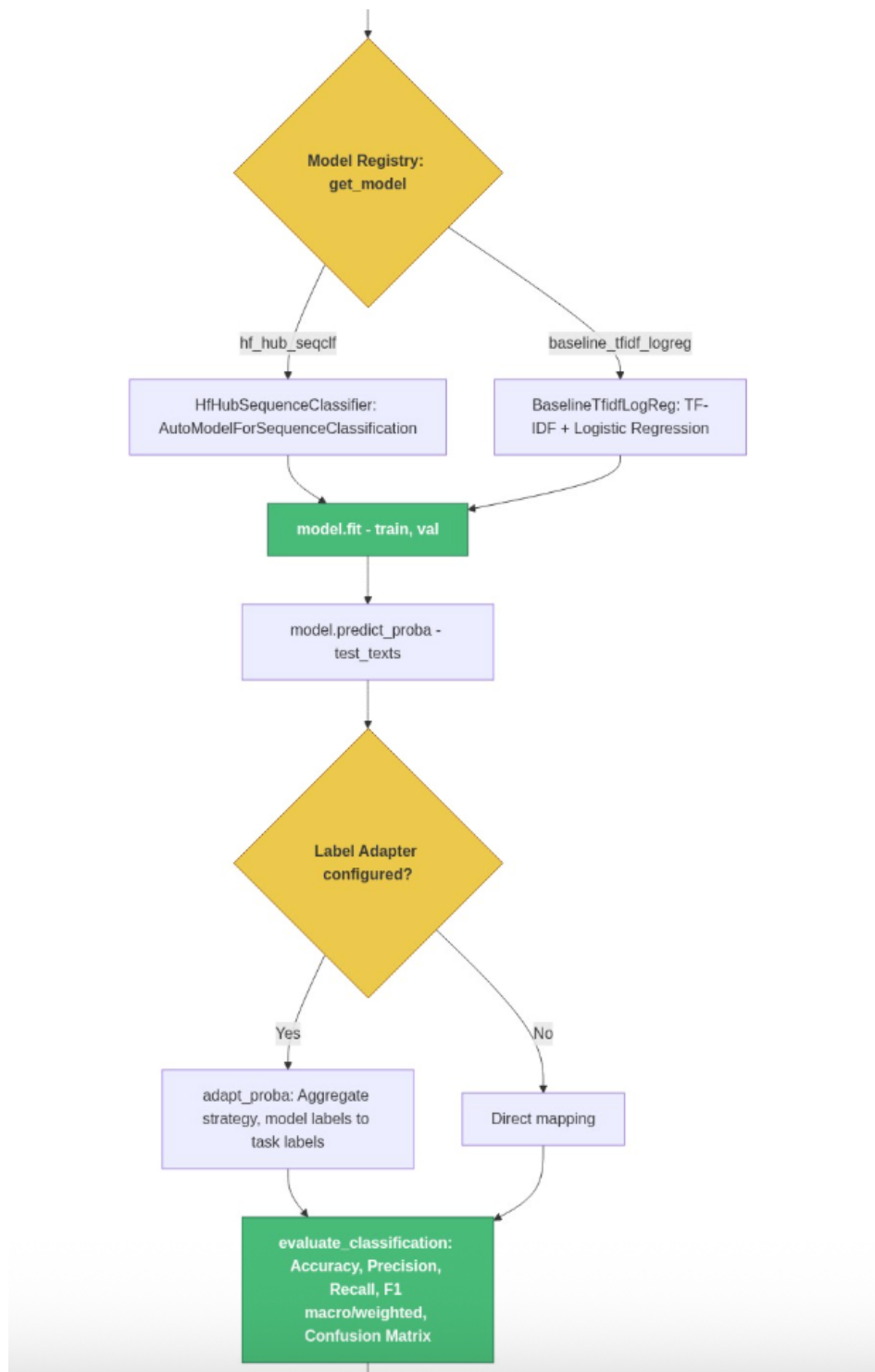


Figure 6.3: Flowchart detailing the combined Model Wrapper and Evaluation Adapter phases.



Figure 6.4: Flowchart detailing the Explainability Orchestrator phase.

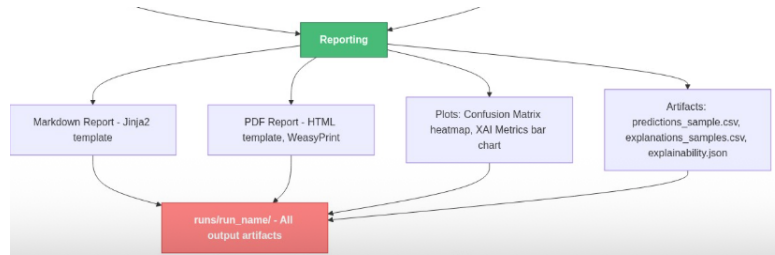


Figure 6.5: Flowchart detailing the Reporting Engine phase.

6.3 Configuration and Execution

The framework’s operational execution is entirely driven by YAML configuration files and a unified CLI entrypoint. This approach abstracts the underlying complexity of PyTorch, Hugging Face Transformers, and the explainability libraries from the end-user.

To execute a full pipeline run, the user invokes the framework with the desired configuration:

Figure 6.6 illustrates the command-line progression of the pipeline for the two evaluated use cases, while Figure 6.7 shows the structure of the artifact directory generated after execution.

```

(.venv) alessiopalemo@MacBookPro Thesis_Project & xai-slm-eval run --config configs/run_phishing_hf.yaml
[19:06:41] INFO Dataset=phishing_cybersectony | sizes train/val/test = 13493/2352/6799 runner.py:113
[19:06:41] INFO Task=Binary | Labels: ['LEGITIMATE_EMAIL', 'PHISHING_EMAIL'] runner.py:117
/Users/alessiopalemo/Desktop/Thesis_Project/.venv/lib/python3.11/site-packages/huggingface_hub/file_download.py:949: FutureWarning: 'resume_download' is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you want to force a new download, use 'force_download=True'.
warnings.warn(
[19:06:43] INFO Model=hf_hub_seqclf loaded runner.py:123
[19:22:43] INFO Running explainability module .. runner.py:167
[19:22:43] INFO Starting explainability evaluation .. runner.py:188
WARNING Stratified sampling failed - falling back to random sampling. runner.py:182
[19:22:43] INFO Running XAI method: attention runner.py:115
Explaining (attention): 100% | 25/25 [00:13<00:00, 1.85it/s]
[19:22:57] INFO Running XAI method: captum ig runner.py:115
Explaining (captum_ig): 100% | 25/25 [40:28<00:00, 97.14s/it]
[20:03:25] INFO Running XAI method: lime runner.py:115
Explaining (lime): 100% | 25/25 [30:16<00:00, 72.68s/it]
[20:33:51] WARNING Ignored 'word-break: break-word' at 279:5, invalid value.
[20:33:51] WARNING Ignored 'word-break: break-word' at 294:5, invalid value.
[20:33:51] INFO PDF report generated with weasyprint: runs/phishing_hf_email_only/report.pdf pdf_builder.py:187
[20:33:51] INFO PDF report: runs/phishing_hf_email_only/report.pdf pdf_builder.py:212
[20:33:51] INFO Run completed: runs/phishing_hf_email_only runner.py:218

```

(a) Phishing Detection Execution

```

(.venv) alessiopalemo@MacBookPro Thesis_Project & xai-slm-eval run --config configs/run_vuln_hf.yaml
[20:40:19] INFO Dataset=vulnerability_nvd_cve_severity | sizes train/val/test = 28999/3008/6001 runner.py:113
[20:40:19] INFO Task=multiclass | Labels: ['LOW', 'MEDIUM', 'HIGH', 'CRITICAL'] runner.py:117
/Users/alessiopalemo/Desktop/Thesis_Project/.venv/lib/python3.11/site-packages/huggingface_hub/file_download.py:949: FutureWarning: 'resume_download' is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you want to force a new download, use 'force_download=True'.
warnings.warn(
[20:40:19] INFO Model=hf_hub_seqclf loaded 499M/499M [00:13<00:00, 36.1MB/s]
[20:40:19] INFO Running explainability module .. runner.py:167
[21:21:37] INFO Starting explainability evaluation .. runner.py:188
WARNING Stratified sampling failed - falling back to random sampling. runner.py:182
[21:21:37] INFO Running XAI method: attention runner.py:115
Explaining (attention): 100% | 25/25 [00:12<00:00, 1.94it/s]
[21:21:50] INFO Running XAI method: captum ig runner.py:115
Explaining (captum_ig): 100% | 25/25 [09:39<00:00, 23.18s/it]
[21:31:30] INFO Running XAI method: lime runner.py:115
Explaining (lime): 100% | 25/25 [23:51<00:00, 57.25s/it]
[21:55:26] WARNING Ignored 'word-break: break-word' at 279:5, invalid value.
[21:55:26] WARNING Ignored 'word-break: break-word' at 294:5, invalid value.
[21:55:30] INFO PDF report generated with weasyprint: runs/vuln_severity_hf_20260307_204617/report.pdf pdf_builder.py:187
[21:55:30] INFO PDF report: runs/vuln_severity_hf_20260307_204617/report.pdf pdf_builder.py:212
[21:55:30] INFO Run completed: runs/vuln_severity_hf_20260307_204617 runner.py:218

```

(b) Vulnerability Severity Execution

Figure 6.6: Command-line execution of the evaluation pipeline showing phase progression for both use cases.

The YAML files define every aspect of the experiment. Below are the exact configurations used for the two primary use cases investigated in this study. Listing 6.1 describes the setup for evaluating a Hugging Face transformer on a phishing email detection task. Notably, it demonstrates the use of the `label_adapter` to map a multiclass model output into a binary legitimate/phishing evaluation space, and it enables three distinct XAI methods.

Listing 6.2 details the setup for vulnerability severity classification. Here, a model maps CVE descriptions directly mapped to a four-class severity scale, requiring no label adaptation, while computing the same suite of XAI metrics.

6.4 Outputs and Generated Artifacts

A central design choice of the *XAI-SLM Cyber Eval* framework is that each run must leave behind a structured and inspectable evidence trail. For this reason, every execution produces a timestamped output directory under `runs/<run_name>_<timestamp>/`, where both machine-readable artifacts and human-readable reports are stored. This design operationalises the reproducibility and auditability principles discussed in Chapter 5,

Listing 6.1: Configuration file (run_phishing_hf.yaml) for the phishing use case.

```

1 run:
2   name: phishing_hf_email_only
3   seed: 42
4   output_dir: runs
5   log_level: INFO
6
7 dataset:
8   ref: phishing_cybersectony
9   params:
10    hf_dataset_id: cybersectony/PhishingEmailDetectionv2.0
11    mode: email_only
12    max_samples: 30000
13    split:
14      train: 0.7
15      val: 0.1
16      test: 0.2
17      stratify: true
18
19 model:
20   ref: hf_hub_seqclf
21   params:
22     hf_model_id: cybersectony/phishing-email-detection-distilbert_v2.4.1
23     batch_size: 32
24     max_length: 256
25     device: auto
26
27 # Evaluation task definition (label space)
28 task:
29   type: binary
30   label_names: ["LEGITIMATE_EMAIL", "PHISHING_EMAIL"]
31
32 # Config-driven mapping from model output labels to task labels.
33 # This avoids hard-coded behaviour in the model wrapper, and keeps the
34   framework versatile.
35 label_adapter:
36   type: aggregate
37   groups:
38     LEGITIMATE_EMAIL: ["LABEL_0", "LABEL_2"]
39     PHISHING_EMAIL: ["LABEL_1", "LABEL_3"]
40
41 # Explainability (Step 2)
42 explainability:
43   enabled: true
44   methods: ["attention", "captum_ig", "lime"] # attention=ante-hoc,
45     captum_ig/lime=post-hoc, shap also supported but slow
46   sample_size: 25
47   per_class: true
48   top_k: 15
49   seed: 42
50   save_artifacts: true

```

Listing 6.2: Configuration file (run_vuln_hf.yaml) for the vulnerability severity use case.

```

1 run:
2   name: vuln_severity_hf
3   seed: 42
4   output_dir: runs
5   log_level: INFO
6
7 dataset:
8   ref: vulnerability_nvd_cve_severity
9   params:
10    hf_dataset_id: stasvinokur/cve-and-cwe-dataset-1999-2025
11    text_col: DESCRIPTION
12    label_col: SEVERITY
13    allowed_labels: [LOW, MEDIUM, HIGH, CRITICAL]
14    max_samples: 30000
15    split:
16      train: 0.7
17      val: 0.1
18      test: 0.2
19      stratify: true
20
21 model:
22   ref: hf_hub_seqclf
23   params:
24     hf_model_id: CIRCL/vulnerability-severity-classification-roberta-
25     base
26     batch_size: 32
27     max_length: 256
28     device: auto
29
30 task:
31   type: multiclass
32   label_names: ["LOW", "MEDIUM", "HIGH", "CRITICAL"]
33
34 # Explainability (Step 2)
35 explainability:
36   enabled: true
37   methods: ["attention", "captum_ig", "lime"]
38   sample_size: 25
39   per_class: true
40   top_k: 15
41   seed: 42
42   save_artifacts: true

```

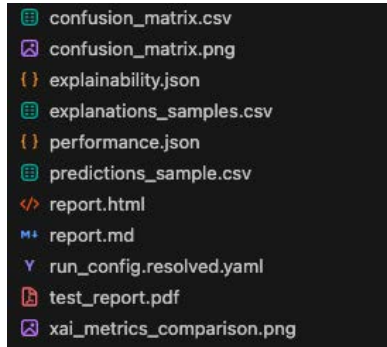


Figure 6.7: Example of the artifact directory generated after a framework execution.

moving the framework beyond a mere experimental prototype and toward a compliance-aware evaluation workflow.

More specifically, the generated artifacts can be grouped into four functional categories:

- **Run reconstruction artifacts**, such as `run_config.resolved.yaml`, preserve the exact configuration used during execution, including default values resolved at runtime. This allows the experiment to be reproduced without ambiguity.
- **Predictive performance artifacts**, such as `performance.json`, `confusion_matrix.csv`, and `predictions_sample.csv`, document the classifier behaviour from both an aggregate and an instance-level perspective.
- **Explainability artifacts**, notably `explainability.json` and `explanations_samples.csv`, record the numerical evaluation of the explanation methods and the token-level attributions produced for the analysed samples.
- **Presentation artifacts**, namely `report.md`, `report.html`, and the final PDF report, consolidate metrics, plots, and representative local explanations into a format suitable for expert review, archival, and audit documentation.

This artifact structure is particularly relevant in cybersecurity settings, where model outputs should not be considered in isolation. Rather, they should be contextualised through supporting evidence that allows an analyst to reconstruct *what* the model predicted, *why* it produced that prediction, and *how reliable* the selected explanation method proved to be for that run.

The report generation layer deserves special attention. In both use cases analysed in this chapter, the reporting engine aggregated dataset metadata, model configuration, predictive performance, confusion matrices, explainability metrics, and representative local explanations into a compact final report. This is an important practical outcome: the framework does not merely calculate metrics internally, but transforms them into a traceable narrative that can be consumed by technical stakeholders, security analysts, and compliance personnel. Representative excerpts of the resulting reports are shown in Figure 6.8.

Use case	Labels	Accuracy	F1 (macro / weighted)
Phishing email detection	2	97.73%	97.73% / 97.73%
Vulnerability severity classification	4	82.67%	76.45% / 82.77%

Table 6.1: Overall predictive results for the two evaluated cybersecurity tasks.

6.5 Results Presentation and Analysis

The empirical evaluation was conducted on two distinct cybersecurity classification tasks: phishing email detection and vulnerability severity classification. Although both runs were executed through the same pipeline and with the same explainability configuration (`attention`, `captum_ig`, and `lime`; sample size of 25; top- $k = 15$; per-class sampling enabled), the two scenarios exhibit markedly different predictive and explanatory behaviours. This makes them suitable for assessing not only model effectiveness, but also the extent to which explanation quality varies with task structure, class granularity, and linguistic characteristics of the input.

6.5.1 Comparative Overview

Table 6.1 summarises the main predictive results obtained in the two use cases. The phishing scenario is a binary classification problem with highly separable classes, whereas vulnerability severity classification is a four-class task characterised by greater semantic overlap and a less sharply separable decision space.

The performance gap between the two tasks is expected and methodologically meaningful. Binary phishing detection relies on lexical and structural regularities that are often strong indicators of malicious intent, whereas severity classification requires finer semantic discrimination across neighbouring severity categories. In other words, the second task is not only more granular, but also more sensitive to contextual ambiguity in vulnerability descriptions.

6.5.2 Predictive Performance Analysis

Phishing email detection. The phishing model achieved very strong performance, with 97.73% accuracy and balanced class-wise behaviour. Precision, recall, and F1-score were all above 0.97 for both classes, with `LEGITIMATE_EMAIL` reaching an F1-score of 0.9769 and `PHISHING_EMAIL` reaching 0.9777. The confusion matrix further confirms this behaviour: among 3,370 legitimate emails, 3,262 were correctly classified and only 108 were flagged as phishing; among 3,429 phishing emails, 3,383 were correctly detected and only 46 were missed. The asymmetry is operationally relevant, because the number of false negatives remains lower than the number of false positives, which is preferable in many defensive settings where missed phishing attempts are costlier than over-flagging legitimate communications.

Vulnerability severity classification. The vulnerability severity task yielded lower but still solid results, with 82.67% accuracy, 76.45% macro-F1, and 82.77% weighted-F1. As expected, performance varies across classes. **MEDIUM** is the best-performing class (F1-score 0.8778), followed by **HIGH** (0.8182), while **CRITICAL** (0.6998) and especially **LOW** (0.6621) are more difficult to separate. The confusion matrix reveals that most errors occur between adjacent severity categories rather than across distant ones. For example, a non-negligible portion of **LOW** samples is mapped to **MEDIUM**, and **HIGH** samples are often confused with both **MEDIUM** and **CRITICAL**. This suggests that the model captures the ordinal structure of the severity scale, but still struggles when the textual description does not contain explicit severity cues.

Such a pattern is particularly important from an operational standpoint. In cybersecurity triage, confusing **HIGH** with **MEDIUM** is materially different from confusing **CRITICAL** with **LOW**. The observed error structure therefore points to a model that is imperfect, yet not arbitrarily unreliable: its mistakes remain largely confined to semantically neighbouring classes.

This contrast between the two tasks is also visible in the confusion matrices reported in Figure 6.9.

6.5.3 Explainability Results Across Methods

Beyond predictive performance, the framework evaluates explanation quality using *comprehensiveness*, *sufficiency*, and *stability*, in line with the explainability evaluation criteria introduced in Chapter 5. In the XAI literature, these dimensions are commonly used to assess whether highlighted features are actually influential for the model decision, whether they are sufficient to preserve the decision signal, and whether explanations remain robust under minor perturbations [5, 1, 7]. This is especially relevant in cybersecurity, where explanations are not merely illustrative, but should support verifiable and reviewable decision-making [3, 2].

Phishing task. For phishing detection, all three methods produced relatively modest comprehensiveness values: **attention** scored 0.0517, **captum_ig** 0.0422, and **lime** 0.0400. This suggests that, in the evaluated sample, removing only the top-ranked tokens did not lead to very large confidence drops. One possible explanation is that the classifier is highly confident and distributes decision evidence across several lexical cues rather than a very small set of decisive tokens. On the sufficiency side, **captum_ig** is closest to zero (-0.0003), indicating that its top-ranked features preserve the model signal better than the other methods. With regard to stability, **attention** is clearly the best-performing method (0.6585), followed by **captum_ig** (0.5449), whereas **lime** drops to 0.1275, indicating markedly weaker robustness under perturbation.

Taken together, these results suggest that for the phishing use case **attention** is the most operationally stable explanation method, whereas **captum_ig** provides the most compact signal preservation. In contrast, **lime** appears less suitable when repeatability is an important requirement.

Vulnerability task. The vulnerability severity scenario exhibits a different pattern. Here the comprehensiveness values are much higher: **attention** reaches 0.3200, **captum_ig** 0.4668, and **lime** 0.4187. This indicates that, in this task, the selected tokens are substantially more decisive for the final prediction. The best result is obtained by **captum_ig**, suggesting that gradient-based attributions capture the most influential severity-bearing terms in the input descriptions. On the sufficiency axis, all methods remain negative, but **captum_ig** again performs best (-0.2182), followed by **lime** (-0.3021) and **attention** (-0.3552). In terms of stability, the ranking changes: **attention** becomes the most stable method (0.6346), with **captum_ig** (0.4327) and **lime** (0.2743) trailing behind.

This second pattern is highly informative. It suggests that no single method dominates across all criteria. **Captum_ig** is strongest when the goal is to identify tokens that are genuinely causal for the prediction, while **attention** is preferable when robustness and consistency across small textual variations are prioritised. In practice, this supports a multi-method approach to explainability, especially in security workflows where both faithfulness and operational repeatability matter.

Figure 6.10 visualises these trade-offs across comprehensiveness, sufficiency, and stability for the two use cases.

6.5.4 Representative Local Explanations

The local explanation samples included in the generated reports provide a qualitative counterpart to the aggregated XAI metrics. They are important because they reveal *which* lexical units the model considered important and whether those units are semantically plausible from a cybersecurity perspective.

Phishing examples. In the phishing report, representative positively classified samples show that the model relies on strongly suspicious lexical and stylistic cues. For instance, one phishing email contains high-attribution tokens such as “*tiffany*”, “*address*”, and “*hi*”, while another highlights terms such as “*come*”, “*marketing*”, “*fun*”, and “*easy*”. These explanations are coherent with the typical spam/phishing register, where promotional language, unusual senders, and informal engagement hooks act as useful risk indicators. Conversely, legitimate emails tend to activate more neutral organisational vocabulary, such as discussion- or process-oriented terms.

Vulnerability examples. The vulnerability report highlights a more technical and semantically dense attribution profile. In a correctly classified **MEDIUM** example (“*Missing Authorization vulnerability in WP CTA PRO*”), **captum_ig** attributes high relevance to tokens such as “*Missing*”, “*WP*”, “*PRO*”, and “*Authorization*”, which are indeed central to the vulnerability semantics. In a harder case (“*Out of bounds Read vulnerability in ssmis_get_frm*”), the model predicts **LOW** instead of the true class **MEDIUM**, yet the explanations still focus on highly meaningful tokens such as “*bounds*”, “*memory*”, “*read*”, “*Out*”, and “*attackers*”. This is a valuable finding: even when the final severity estimate is imperfect, the attribution map still indicates that the model is attending to conceptually relevant vulnerability features rather than to arbitrary noise.

Overall, the local explanations strengthen the interpretation of the quantitative XAI metrics. The phishing task tends to concentrate explanations on suspicious lexical patterns and stylistic hooks, whereas the vulnerability task leads the model to focus on technical descriptors, impact-related verbs, and software-specific identifiers. This cross-task difference confirms that explanation behaviour is not method-dependent only, but also shaped by the linguistic structure of the underlying security problem.

Representative examples of these token-level explanation artefacts are reported in Figure 6.11.

6.6 Operational and Audit Discussion

The results discussed above show why an evaluation framework of this kind is useful in operational cybersecurity contexts. Accuracy alone is not sufficient when AI-assisted outputs may influence analyst triage, alert prioritisation, or vulnerability assessment. What is needed is a combination of predictive evidence, explanation quality, reproducibility, and documentary traceability. This is precisely where the generated artifacts become relevant.

First, the framework creates a direct link between a model output and the exact conditions under which it was produced: dataset reference, split configuration, model identifier, runtime parameters, and explainability settings are all preserved in the run directory. Second, the explanation methods are not treated as decorative overlays, but as measurable components whose quality can be compared through explicit criteria. This is consistent with the broader movement toward trustworthy and accountable AI, where the emphasis is placed not only on performance, but also on transparency, reviewability, and structured human oversight [1, 2].

From a regulatory and governance perspective, such an evidence trail is increasingly important. The AI Act places strong emphasis on documentation, traceability, and the possibility of meaningful scrutiny of high-impact AI-supported processes, while GDPR-related debates and the recent *Dun & Bradstreet* judgment have further reinforced the practical relevance of providing meaningful information about automated logic and enabling contestability of consequential decisions [22, 19, 20, 21]. In cybersecurity governance, this expectation aligns with the broader accountability logic also visible in NIS2, where organisations are required to demonstrate structured control over security-relevant processes [17].

Operationally, the framework supports a calibrated use of model outputs. For example, a security analyst reviewing a phishing classification can rely not only on the predicted label, but also on the confusion matrix trends, on the per-class behaviour, and on the fact that **attention** explanations were the most stable in that specific run. Likewise, in vulnerability triage, the analyst can interpret **captum_ig** explanations as more faithful to decisive severity-bearing terms, while considering **attention** more robust under perturbation. This does not eliminate the need for human validation; rather, it creates a more informed basis for that validation.

This distinction also clarifies the place of black-box models within the thesis. Semi-open or closed services were not benchmarked directly, yet the conceptual framework remains

applicable to them. In those settings, XAI is chiefly post-hoc: explanations must be reconstructed from perturbation tests, surrogate models, prompt-response traces, and other external artefacts, because the internal architecture cannot be inspected directly. Consequently, the explanatory evidence obtainable from black-box deployments is narrower and depends more heavily on provider-side documentation and logging interfaces.

In summary, the main contribution of the framework is not limited to producing a better-looking report. Its real value lies in transforming SLM-based cybersecurity evaluation into a process that is reproducible, inspectable, and better aligned with the transparency expectations surrounding contemporary AI deployment. For high-stakes security environments, this shift from raw prediction to documented and explainability-aware evidence is a necessary condition for responsible adoption.

1. Dataset Overview

Property	Value
Reference	phishing_cybersectony
Training samples	13493
Validation samples	2352
Test samples	6799
Total	22644

2. Model Configuration

Parameter	Value
Model reference	hf_hub_seqclf
hf_model_id	cybersectony/phishing-email-detection-distilbert_v2.4.1
batch_size	32
max_length	256
device	auto

(a) Phishing run report extract

1. Dataset Overview

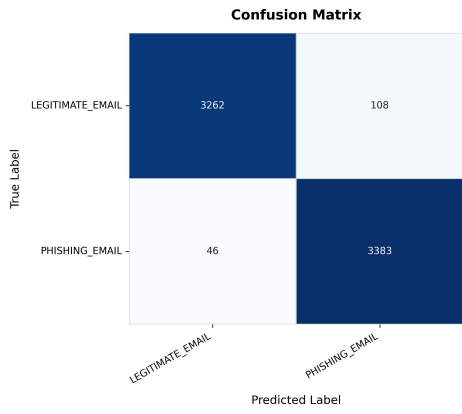
Property	Value
Reference	vulnerability_nvd_cve_severity
Training samples	20999
Validation samples	3000
Test samples	6001
Total	30000

2. Model Configuration

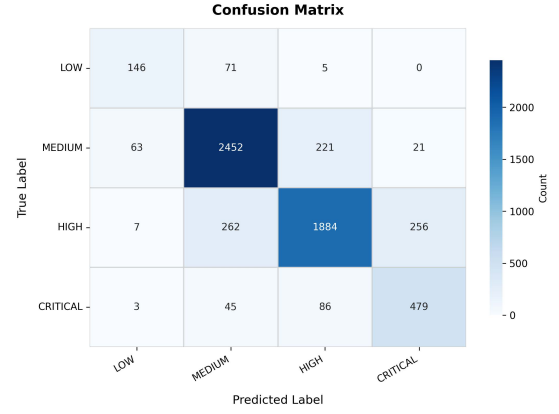
Parameter	Value
Model reference	hf_hub_seqclf
hf_model_id	CIRCL/vulnerability-severity-classification-roberta-base
batch_size	32
max_length	256
device	auto

(b) Vulnerability run report extract

Figure 6.8: Representative extracts from the automatically generated reports.



(a) Phishing confusion matrix



(b) Vulnerability confusion matrix

Figure 6.9: Confusion matrices for the two evaluated tasks. Their joint inspection highlights the difference between a sharply separable binary task and a more nuanced multi-class severity classification scenario.

4.3 Method Comparison

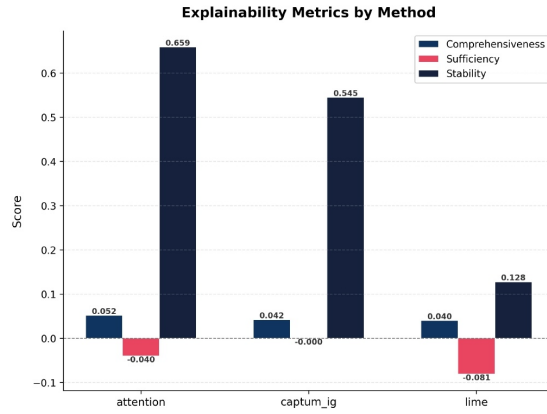


Figure 2 — Explainability Metrics by Method

Method	Comprehensiveness	Sufficiency	Stability
attention	0.0517	-0.0399	0.6585
captum Ig	0.0422	-0.0003	0.5449
lime	0.0400	-0.0808	0.1275

Interpretation: attention achieves the highest comprehensiveness (0.0517), better identifying tokens that drive predictions. attention shows the highest stability (0.6585), providing more consistent explanations.

(a) Phishing XAI comparison

4.3 Method Comparison

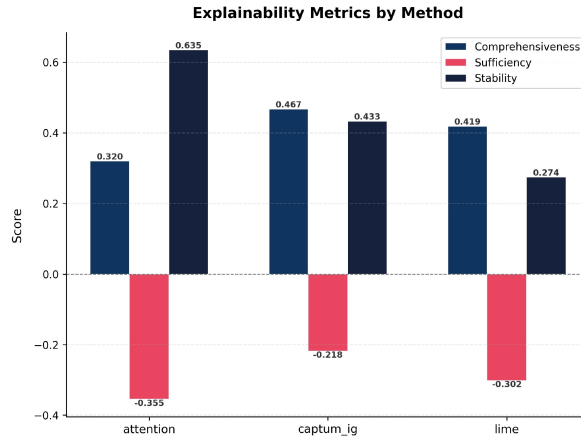


Figure 2 — Explainability Metrics by Method

Method	Comprehensiveness	Sufficiency	Stability
attention	0.3200	-0.3552	0.6346
captum Ig	0.4668	-0.2182	0.4327
lime	0.4187	-0.3021	0.2743

Interpretation: captum Ig achieves the highest comprehensiveness (0.4668), better identifying tokens that drive predictions. attention shows the highest stability (0.6346), providing more consistent explanations.

(b) Vulnerability XAI comparison

Figure 6.10: Comparison of comprehensiveness, sufficiency, and stability across explanation methods for the two tasks.

#13 — "sexually - explicit he puts his petah in these gal"

True label	PHISHING_EMAIL
Pred label	PHISHING_EMAIL
Confidence	1.000
Comprehensiveness	0.0003
Sufficiency	0.0000

(+0.672) sexually · (-0.205) ##aw · (+0.145) gal · (+0.106) la · (-0.091) / · (+0.089) explicit · (-0.077) - · (+0.075) evil · (-0.074) review · (-0.073) ##ah · (+0.057) ##gence · (+0.055) blow · (-0.051) policy · (-0.047) ##s · (-0.043) the

(a) Local explanation for a phishing sample

#0 — "Missing Authorization vulnerability in WP CTA PRO"

True label	MEDIUM
Pred label	MEDIUM
Confidence	0.934
Comprehensiveness	0.8016
Sufficiency	-0.1265

(+0.512) Missing · (+0.393) WP · (+0.353) PRO · (+0.231) Levels · (+0.205) . · (+0.197) / · (-0.180) 8 · (+0.177) Authorization · (+0.169) C · (+0.167) Expl · (+0.151) : · (-0.147) WordPress · (+0.143) C · (+0.106) n · (+0.097) ured

(b) Local explanation for a vulnerability sample

Figure 6.11: Representative token-level local explanations extracted from the generated reports.

Chapter 7

Conclusions and Future Work

This thesis set out to investigate whether Explainable Artificial Intelligence (XAI) can be employed not only to improve the interpretability of Small Language Models (SLMs) in cybersecurity tasks, but also to support broader requirements related to transparency, accountability, and compliance readiness. The starting point of the work was the growing gap between the operational attractiveness of AI-enabled security tools and the difficulty of justifying their outputs in environments where decisions may affect prioritisation, escalation, remediation, and audit trails. In cybersecurity, predictive accuracy alone is not sufficient. Models operate in dynamic and potentially adversarial contexts, and their outputs may influence time-sensitive actions. For this reason, the adoption of SLMs in security workflows requires not only satisfactory predictive performance, but also mechanisms to document how results are produced, how reliable the explanations are, and under which configuration a given decision can be reconstructed and reviewed.

Within this context, the thesis developed along three complementary directions. First, it examined the XAI landscape with specific attention to language-based cybersecurity tasks, identifying the main explanation families, their purposes, and their practical trade-offs, in line with the broader literature that frames explainability as a contextual and stakeholder-dependent property rather than as a universally valid output [5, 7, 13]. Second, it analysed the regulatory and governance dimensions surrounding the adoption of AI in the European context, showing that explainability becomes particularly relevant when linked to traceability, human oversight, documentation, and evidence preservation, especially in light of the emerging regulatory landscape shaped by instruments such as the GDPR, the AI Act, the NIS2, and the CRA [19, 22, 17, 23]. Third, it translated these conceptual and regulatory insights into an experimental and engineering contribution through the design of a reproducible, reporting-oriented framework for evaluating SLMs on representative cybersecurity use cases. Taken together, these elements support the central claim of this thesis: in cybersecurity, explainability should be treated as an operational and governance capability rather than as a merely illustrative add-on.

7.1 Summary of Key Contributions

The contributions of this thesis can be summarised along four main dimensions.

7.1.1 Conceptual and taxonomical contribution

A first contribution lies in the consolidation of the fragmented XAI literature into a practice-oriented perspective tailored to cybersecurity language tasks. As discussed in Chapter 2, explainability was not treated as a generic property of AI systems, but rather as a context-sensitive capability that must be analysed in relation to three main dimensions: what is being explained, to whom the explanation is addressed, and under which technical and organisational constraints it is produced. This perspective is especially relevant in cybersecurity, where the same model output may need to support different purposes, ranging from analyst triage and incident prioritisation to governance review and audit preparation. The thesis therefore showed that explainability is not a monolithic notion, but a situated property whose value depends on the stakeholder, the task, and the deployment setting.

7.1.2 Regulatory-technical mapping

A second contribution concerns the translation of high-level governance and regulatory expectations into concrete technical implications for SLM-based cybersecurity systems. Building on the analysis developed in Chapter 3, the thesis clarified that concepts such as transparency, traceability, accountability, and oversight cannot be satisfied by explanation artefacts alone. They require instead a broader combination of structured documentation, versioned configurations, reproducible execution, and explanation reliability assessment. In this sense, the work proposed a bridge between governance requirements and engineering artefacts, showing how explanation outputs, evaluation metrics, and configuration records can be combined into a documented evidence trail. This point is consistent with the broader debate on trustworthy and explainable AI, which increasingly emphasises that explanation quality must be assessed in relation not only to interpretability, but also to reliability, governance, and actual use conditions [2, 15].

7.1.3 Empirical evaluation of XAI on lightweight open-weight security models

The third contribution is empirical. Through the experiments presented in Chapter 6, the thesis assessed the behaviour of different explanation strategies on two representative cybersecurity tasks, namely phishing email detection and vulnerability severity classification, using lightweight open-source models that can be executed locally. These use cases were selected because they are both practically relevant and sufficiently different to highlight distinct challenges in the application of XAI to language-based security workflows. The evaluation showed that explanation methods may differ substantially in fidelity, stability, and practical usefulness. Some artefacts proved informative for local inspection and

debugging, while others were less useful for supporting stronger claims about the model’s underlying reasoning. A central result of the thesis is therefore that explanations should not be accepted at face value simply because they appear plausible. Rather, they need to be assessed through explicit reliability-oriented checks, particularly in cybersecurity contexts where inputs may be noisy, partially structured, or adversarial [3].

7.1.4 Framework and prototype contribution

The fourth contribution is methodological and engineering-oriented. As illustrated in Chapter 6, the thesis designed and implemented *XAI-SLM Cyber Eval*, a configuration-driven and reporting-oriented framework that operationalises the evaluation workflow developed throughout the dissertation. The framework connects model execution, explanation generation, explanation metric computation, and report production within a structured pipeline. Its practical relevance lies in the fact that reporting artefacts are treated as first-class outputs. Configuration files, performance metrics, explainability artefacts, and generated reports are preserved in a consistent form, thereby supporting repeatability and facilitating their reuse in governance or compliance-oriented processes. In this respect, the contribution is not limited to the experimental comparison of methods, but also provides a concrete template for evaluating SLM-based cybersecurity systems in a manner more aligned with organisational accountability needs.

7.2 Revisiting the Research Questions

The results presented in the previous chapters allow revisiting the research questions introduced in Chapter 1 and answering them in an integrated manner.

7.2.1 RQ1 – XAI methods for SLMs in cybersecurity

The thesis showed that, for typical cybersecurity language tasks, post-hoc and instance-level explanation methods constitute the most practical entry point. As discussed in Chapter 2 and operationalised in Chapter 5, feature attribution techniques and local surrogate approaches map naturally to questions that arise in operational environments, such as why a phishing email was flagged, why a vulnerability was assigned a specific severity class, or which parts of an input most strongly influenced the model output. These methods are attractive because they can often be integrated without redesigning the underlying model architecture.

At the same time, the work demonstrated that no explanation method can be considered universally suitable. The feasibility and usefulness of a method are strongly influenced by the deployment regime. Open-weight models enable deeper inspection through gradients, internal representations, or controlled perturbation analyses. Semi-open and closed services, by contrast, restrict the range of usable XAI techniques and increase dependence on model-agnostic approaches and system-level observability. Consequently, the selection of explanation methods should be understood not as a purely methodological preference, but

as a design decision shaped by access constraints, governance objectives, and operational needs.

7.2.2 RQ2 – Effectiveness and limitations of XAI on lightweight open-weight SLMs, with implications for constrained deployments

The empirical evaluation carried out in Chapter 6 addressed RQ2 in a deliberately scoped manner. Rather than benchmarking the full spectrum of open-weight, semi-open, and closed SLMs, the implemented study focused on lightweight open-source models that could be deployed locally, instrumented end-to-end, and analysed with both gradient-based and perturbation-based XAI methods. Within this empirical scope, the results confirmed that explanation methods exhibit non-trivial trade-offs across fidelity, stability, and practical interpretability.

This means that RQ2 is answered directly only for open-weight deployments. Semi-open and closed models were not evaluated experimentally, and CyberPal 2.0 served as a theoretical state-of-the-art reference rather than as a test subject. Nevertheless, the conceptual analysis developed in Chapters 4 and 5 supports a clear inference about such black-box settings: as access to weights, gradients, and internal representations decreases, XAI increasingly depends on post-hoc and model-agnostic techniques, as well as on surrounding controls such as logging, monitoring, provider documentation, and structured reporting.

The thesis also showed that attention-based visualisations can remain useful as diagnostic artefacts, particularly during exploratory analysis and debugging, but they should not automatically be interpreted as faithful representations of decision logic. More broadly, the results support the conclusion that explainability is not a binary property that a model simply possesses or lacks. It is better understood as a portfolio of explanation artefacts whose value depends on how reliably they track model behaviour under a given set of constraints. This is precisely why the thesis argued for accompanying explanations with quantitative indicators, such as fidelity- and stability-oriented proxies, rather than treating a single visualisation or score as sufficient evidence.

7.2.3 RQ3 – Regulatory alignment of XAI explanations for SLM-based systems

The regulatory analysis developed in Chapter 3 led to a central conclusion: in compliance-oriented settings, explanations are most valuable when they form part of a broader evidence trail. Their relevance does not stem solely from helping a human understand a single prediction, but from their ability to contribute to documentation, oversight, traceability, and contestability. An explanation becomes considerably more meaningful when it is linked to a known model version, a defined configuration, a documented dataset, and a reproducible evaluation context.

From this perspective, regulatory alignment should not be treated as a property of the

explanation method alone. It should instead be understood as a system-level property emerging from the interaction between model, explanation artefacts, lifecycle controls, and operational documentation. This aspect is particularly important for semi-open and closed deployments, where internal transparency is more limited and accountability depends more heavily on surrounding controls, such as logging, monitoring, provider assurances, and structured reporting. Accordingly, the thesis does not argue that XAI alone is sufficient to guarantee compliance. Rather, it shows that explainability can contribute to compliance readiness when it is embedded within a wider documentation and governance process.

7.2.4 RQ4 – Framework design for XAI-oriented SLM compliance

The thesis answered the fourth research question through the design and implementation of *XAI-SLM Cyber Eval*. As presented in Chapter 6, the framework operationalises a repeatable workflow for defining experiments, training or adapting models, generating explanations, computing reliability-oriented metrics, and producing structured reports. Its main contribution is to demonstrate that compliance readiness can also be approached through pipeline design choices. By making configurations explicit, versionable, and reproducible, the framework reduces ambiguity about what was executed, how it was executed, and what evidence supports the resulting claims.

In this sense, the framework is both a research artefact and a practical template for organisation-grade evaluation pipelines. It illustrates how technical evaluation can be aligned with documentation and governance needs without treating these dimensions as separate concerns. This is one of the core original outcomes of the thesis.

7.3 Operational Implications for Security Operations and Audit Readiness

One of the most relevant outcomes of this thesis is that the value of XAI in cybersecurity becomes clearer when it is framed as an operational control. In SOC workflows and vulnerability management processes, the central question is not whether an explanation can be generated in abstract terms, but whether the organisation can demonstrate how a model was evaluated, what its limitations are, and under which conditions its outputs may reasonably support human decision-making. This becomes especially important when model outputs influence alert prioritisation, classification decisions, vulnerability ranking, or remediation planning.

The reporting-oriented workflow proposed in this thesis addresses this need by linking predictive performance, explanation artefacts, reliability indicators, and configuration evidence within a unified process. This supports internal accountability because it enables teams to compare runs across model versions, datasets, or configuration changes and to assess whether behavioural changes are accompanied by corresponding shifts in explanations. At the same time, it strengthens audit readiness by preserving artefacts that can later be reused for review, governance, or post-incident analysis.

A further operational implication concerns deployment strategy. Even though only

open-weight models were tested directly, the broader analysis developed throughout the thesis makes it evident that deployment regimes strongly shape the attainable explainability portfolio. Open-weight deployments generally offer stronger possibilities for deep inspection and evidence generation, whereas closed services require a greater reliance on model-agnostic explanations and on controls external to the model itself. For this reason, the choice between open, semi-open, and closed solutions should not be made solely on the basis of predictive performance or cost. It should also be evaluated in relation to governance requirements such as traceability, data residency, reproducibility, and auditability.

7.4 Limitations

The conclusions of this thesis should be interpreted in light of several limitations.

First, the empirical evaluation focused on two representative cybersecurity classification tasks, namely phishing detection and vulnerability severity classification. These use cases provide a coherent and meaningful test bed, but they do not exhaust the range of security applications in which SLMs may be adopted. Other relevant contexts, such as log anomaly detection, incident narrative analysis, malware-related text processing, or broader SOC support tasks, may exhibit different explanation behaviours and different operational constraints. As a consequence, the external validity of the present findings should be considered with appropriate caution.

Second, explanation metrics remain proxies. Fidelity- and stability-oriented indicators provide structured evidence about explanation behaviour, but they cannot fully capture whether an explanation is misleading, whether it induces over-trust, or whether it genuinely improves decision quality in practice. This limitation is particularly relevant in cybersecurity, where time pressure, contextual knowledge, and analyst experience play a major role in the interpretation of model outputs. In other words, an explanation may be technically consistent while still proving insufficient from an operational or governance perspective.

Third, the thesis did not include a controlled human-centred validation study. Although qualitative inspection was used to assess whether explanations highlight security-meaningful cues, further work is needed to measure usability, calibration of trust, decision-time effects, and the actual value of explanation artefacts for analyst and governance stakeholders. Additional methodological limitations and threats to validity were already discussed in Chapter 5 and remain relevant when interpreting the experimental results.

7.5 Future Work

The results obtained in this thesis open several promising directions for future research and further development.

7.5.1 Extending the empirical scope of cybersecurity tasks

A first direction concerns extending the evaluation to a broader set of SOC and security operations tasks, including SIEM alert triage, incident report classification, log explanation, and analyst-assistance scenarios. This would strengthen the external validity of the findings and help determine whether the reliability indicators adopted in the thesis remain informative under more heterogeneous and operationally noisy conditions.

7.5.2 Comparison with general-purpose open-weight baselines

A second direction should benchmark the task-specific open-weight checkpoints used in this thesis against one or more compact general-purpose open-weight SLMs under the same execution and reporting pipeline. Such a study would make it possible to separate the contribution of cybersecurity specialisation from that of architecture or training setup, thereby clarifying whether the observed explanation patterns arise primarily from domain adaptation or from model-family differences. To be methodologically persuasive, this comparison should keep datasets, preprocessing, explanation methods, and evaluation metrics fixed, so that changes in fidelity, stability, and practical usefulness can be attributed to the model class rather than to surrounding implementation choices.

7.5.3 Systematic comparison across deployment regimes

A third line of research involves extending the open-weight focus of this thesis to a wider portfolio of semi-open and closed models. Such an analysis would enable more precise quantification of how access constraints reshape the feasibility of explanation and how much of the resulting transparency gap can be mitigated through surrounding controls, such as richer logging, stronger monitoring, and versioned reporting. It would also offer a direct empirical follow-up to the part of RQ2 that remains only conceptually addressed in the present work. Where access conditions permit, this line of research could also include a specialised state-of-the-art reference such as CyberPal 2.0 or a similar cybersecurity expert model, so as to contrast domain-specialised open-weight systems with provider-constrained alternatives under a common governance-oriented evaluation protocol.

7.5.4 Richer explanation modalities for security text

Another relevant direction is the integration of additional explanation modalities, such as example-based, concept-based, and counterfactual explanations. For practitioners, these forms of explanation may prove more actionable than raw token attributions, especially when they are linked to interpretable security concepts such as phishing cues, exploit patterns, or recurring vulnerability characteristics. At the same time, such modalities must be designed carefully to avoid plausible but unfaithful explanation artefacts.

7.5.5 Human-centred validation

Future work should also include controlled studies involving security analysts, governance roles, and, where appropriate, audit or compliance stakeholders. These studies would enable evaluation of whether explanations reduce decision time, improve triage quality, calibrate trust more appropriately, or support documentation tasks more effectively. Human-centred evidence is essential if XAI is to move from a technically promising capability to a practically validated one.

7.5.6 Continuous evaluation and drift-aware governance

Because phishing campaigns, vulnerability disclosures, and operational contexts evolve over time, another important line of development concerns longitudinal monitoring of both predictive performance and explanation behaviour. Drift-aware evaluation pipelines could help detect when explanation stability changes, even when accuracy remains apparently stable, thereby supporting more robust lifecycle governance and post-incident accountability.

7.5.7 Compliance-by-design reporting

Finally, the reporting perspective introduced in this thesis can be further extended by aligning the generated outputs more closely with organisational governance artefacts, such as model cards, risk assessment templates, internal control checklists, and audit preparation documentation. This would further strengthen the bridge between technical experimentation and practical compliance workflows, a key ambition of the work presented here.

7.6 Final Remarks

Overall, this thesis supports the conclusion that explainability in cybersecurity should be designed, evaluated, and documented as an integral component of trustworthy and accountable AI-enabled security systems. The work showed that XAI becomes most valuable when it is embedded into reproducible processes, linked to explicit model configurations, and assessed through reliability-oriented criteria rather than treated as a purely illustrative output.

The thesis also demonstrated that the deployment regime is a decisive factor in shaping what kinds of explanations can realistically be obtained and what strength of evidence they can provide. The empirical evidence developed here comes from open-weight deployments, which enable deeper inspection and stronger evidence generation, whereas closed services require a greater reliance on model-agnostic artefacts and system-level controls. For this reason, explainability cannot be separated from broader engineering and governance choices.

Although the field is evolving rapidly and important challenges remain, the methodological and engineering approach developed in this thesis offers a structured basis for

moving beyond ad hoc explanation practices. By combining conceptual analysis, regulatory interpretation, empirical evaluation, and framework design, this work contributes to a view of XAI that is more reproducible, more operationally grounded, and more closely aligned with the accountability demands of contemporary cybersecurity environments.

Bibliography

- [1] Alejandro Barredo Arrieta et al. *Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI*. Dec. 2019. DOI: 10.48550/arXiv.1910.10045. arXiv: 1910.10045 [cs]. (Visited on 11/12/2025).
- [2] Vinay Chamola, Vikas Hassija, A Razia Sulthana, Debshishu Ghosh, Divyansh Dhingra, and Biplab Sikdar. «A Review of Trustworthy and Explainable Artificial Intelligence (XAI)». In: *IEEE Access* 11 (2023), pp. 78994–79015. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2023.3294569. (Visited on 10/31/2025).
- [3] Gautam Srivastava et al. *XAI for Cybersecurity: State of the Art, Challenges, Open Issues and Future Directions*. June 2022. DOI: 10.48550/arXiv.2206.03585. arXiv: 2206.03585 [cs]. (Visited on 10/31/2025).
- [4] Khushi Kalasampath, K. N. Spoorthi, Sreeparvathy Sajeev, Sahil Sarma Kuppa, Kavya Ajay, and Angulakshmi Maruthamuthu. «A Literature Review on Applications of Explainable Artificial Intelligence (XAI)». In: *IEEE Access* 13 (2025), pp. 41111–41140. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2025.3546681. (Visited on 10/31/2025).
- [5] Amina Adadi and Mohammed Berrada. «Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)». In: *IEEE Access* 6 (2018), pp. 52138–52160. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2018.2870052. (Visited on 10/31/2025).
- [6] Arun Das and Paul Rad. *Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey*. June 2020. DOI: 10.48550/arXiv.2006.11371. arXiv: 2006.11371 [cs]. (Visited on 10/31/2025).
- [7] Erico Tjoa and Cuntai Guan. «A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI». In: *IEEE Transactions on Neural Networks and Learning Systems* 32.11 (Nov. 2021), pp. 4793–4813. ISSN: 2162-237X, 2162-2388. DOI: 10.1109/TNNLS.2020.3027314. (Visited on 10/31/2025).
- [8] Lokendra Singh Kushwah. «Designing Explainable AI (XAI) for Regulatory-Compliant Cybersecurity in Financial Systems». In: (Oct. 2025). ISSN: 2945-3585. DOI: 10.5281/ZENODO.17390468. (Visited on 10/31/2025).
- [9] Julie Gerlings, Arisa Shollo, and Ioanna Constantiou. «Reviewing the Need for Explainable Artificial Intelligence (xAI)». In: *Hawaii International Conference on System Sciences*. 2021. DOI: 10.24251/HICSS.2021.156. HDL: 10125/70768. (Visited on 10/31/2025).

-
- [10] Chien Van Nguyen et al. *A Survey of Small Language Models*. Oct. 2024. DOI: 10.48550/arXiv.2410.20011. arXiv: 2410.20011 [cs]. (Visited on 11/14/2025).
 - [11] Peter Belcak, Greg Heinrich, Shizhe Diao, Yonggan Fu, Xin Dong, Saurav Muralidharan, Yingyan Celine Lin, and Pavlo Molchanov. *Small Language Models Are the Future of Agentic AI*. Sept. 2025. DOI: 10.48550/arXiv.2506.02153. arXiv: 2506.02153 [cs]. (Visited on 11/14/2025).
 - [12] Matan Levi, Daniel Ohayon, Ariel Blobstein, Ravid Sagi, Ian Molloy, and Yair Allouche. *Toward Cybersecurity-Expert Small Language Models*. Oct. 2025. DOI: 10.48550/arXiv.2510.14113. arXiv: 2510.14113 [cs]. (Visited on 11/14/2025).
 - [13] Timo Speith. «A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods». In: *2022 ACM Conference on Fairness Accountability and Transparency*. Seoul Republic of Korea: ACM, June 2022, pp. 2239–2250. ISBN: 978-1-4503-9352-2. DOI: 10.1145/3531146.3534639. (Visited on 10/31/2025).
 - [14] David Schneeberger, Richard Röttger, Federico Cabitza, Andrea Campagner, Markus Plass, Heimo Müller, and Andreas Holzinger. «The Tower of Babel in Explainable Artificial Intelligence (XAI)». In: *Machine Learning and Knowledge Extraction*. Ed. by Andreas Holzinger, Peter Kieseberg, Federico Cabitza, Andrea Campagner, A Min Tjoa, and Edgar Weippl. Vol. 14065. Cham: Springer Nature Switzerland, 2023, pp. 65–81. ISBN: 978-3-031-40837-3. DOI: 10.1007/978-3-031-40837-3_5. (Visited on 10/31/2025).
 - [15] Nipuna Thalpage. «Unlocking the Black Box: Explainable Artificial Intelligence (XAI) for Trust and Transparency in AI Systems». In: *Journal of Digital Art & Humanities* 4.1 (June 2023), pp. 31–36. ISSN: 2712-8148. DOI: 10.33847/2712-8148.4.1_4. (Visited on 10/31/2025).
 - [16] European Parliament and Council of the European Union. *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance)*. Official Journal of the European Union, OJ L, 2024/1689, 12 July 2024. July 12, 2024. URL: <https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng> (visited on 03/15/2026).
 - [17] European Parliament and Council of the European Union. *Directive (EU) 2022/2555 of the European Parliament and of the Council of 14 December 2022 on Measures for a High Common Level of Cybersecurity across the Union, Amending Regulation (EU) No 910/2014 and Directive (EU) 2018/1972, and Repealing Directive (EU) 2016/1148 (NIS 2 Directive)*. Dec. 2022.
 - [18] European Parliament and Council of the European Union. *Regulation (EU) 2024/2847 of the European Parliament and of the Council of 23 October 2024 on horizontal cybersecurity requirements for products with digital elements and amending Regulations (EU) No 168/2013 and (EU) 2019/1020 and Directive (EU) 2020/1828 (Cyber Resilience Act) (Text with EEA relevance)*. Official Journal of the European

- Union, OJ L, 2024/2847, 20 November 2024. Nov. 20, 2024. URL: <https://eur-lex.europa.eu/eli/reg/2024/2847/oj/eng> (visited on 03/15/2026).
- [19] European Parliament and Council of the European Union. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation)*. 2016.
- [20] Court of Justice of the European Union. *Judgment of 27 February 2025, Case C-203/22, CK v Dun & Bradstreet Austria GmbH and Magistrat der Stadt Wien, ECLI:EU:C:2025:117*. First Chamber. Feb. 27, 2025.
- [21] Aída Ponce del Castillo. *The Right to Meaningful Explanation in Algorithmic Management: Commentary on the CJEU's Judgment in Dun & Bradstreet Austria, Case C-203/22*. ETUI Research Paper – Technical Brief 2025.02. European Trade Union Institute (ETUI), Nov. 17, 2025. DOI: 10.2139/ssrn.5821222. URL: <https://ssrn.com/abstract=5821222> (visited on 03/15/2026).
- [22] European Parliament and Council of the European Union. *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)*. June 2024.
- [23] European Parliament and Council of the European Union. *Regulation (EU) 2024/2847 of the European Parliament and of the Council of 23 October 2024 on Horizontal Cybersecurity Requirements for Products with Digital Elements and Amending Regulations (EU) No 168/2013 and (EU) 2019/1020 and Directive (EU) 2020/1828 (Cyber Resilience Act)*. Oct. 2024.
- [24] Ali Habibzadeh, Farid Feyzi, and Reza Ebrahimi Atani. *Large Language Models for Security Operations Centers: A Comprehensive Survey*. Sept. 2025. DOI: 10.48550/arXiv.2509.10858. arXiv: 2509.10858 [cs]. (Visited on 12/02/2025).
- [25] Ramanpreet Kaur, Tomaž Klobučar, and Dušan Gabrijelčič. «Harnessing the Power of Language Models in Cybersecurity: A Comprehensive Review». In: *International Journal of Information Management Data Insights* 5.1 (June 2025), p. 100315. ISSN: 26670968. DOI: 10.1016/j.jjime.2024.100315. (Visited on 12/02/2025).
- [26] Gabriel de Jesus Coelho da Silva and Carlos Becker Westphall. *A Survey of Large Language Models in Cybersecurity*. Feb. 2024. DOI: 10.48550/arXiv.2402.16968. arXiv: 2402.16968 [cs]. (Visited on 12/02/2025).
- [27] Yazi Gholami. «Large Language Models (LLMs) for Cybersecurity: A Systematic Review». In: *World Journal of Advanced Engineering Technology and Sciences* 13.1 (Sept. 2024), pp. 057–069. ISSN: 25828266. DOI: 10.30574/wjaets.2024.13.1.0395. (Visited on 12/02/2025).
- [28] Aristeidis Karras, Leonidas Theodorakopoulos, Christos Karras, Alexandra Theodoropoulou, Ioanna Kalliampakou, and Gerasimos Kalogeratos. «LLMs for Cybersecurity in the Big Data Era: A Comprehensive Review of Applications, Challenges, and Future Directions». In: *Information* 16.11 (Nov. 2025), p. 957. ISSN: 2078-2489. DOI: 10.3390/info16110957. (Visited on 12/02/2025).

- [29] Niveen O. Jaffal, Mohammed Alkhanafseh, and David Mohaisen. «Large Language Models in Cybersecurity: A Survey of Applications, Vulnerabilities, and Defense Techniques». In: *AI* 6.9 (Sept. 2025), p. 216. ISSN: 2673-2688. DOI: 10.3390/ai6090216. (Visited on 12/02/2025).
- [30] Tiago Dinis, Miguel Correia, and Roger Tavares. *Large Language Models for Explainable Threat Intelligence*. Nov. 2025. DOI: 10.48550/arXiv.2511.05406. arXiv: 2511.05406 [cs]. (Visited on 12/02/2025).
- [31] Hugging Face. *Small Language Models (SLMs): A Comprehensive Overview*. <https://huggingface.co/blog/jjokah/small-language-model>. Blog post, accessed 2 Dec. 2025. 2025.
- [32] Blue Prism. *The Small Language Models (SLMs) AI Guide: Big Impact with Small AI*. <https://www.blueprism.com/guides/ai/small-language-models-slms/>. Vendor guide on SLMs, accessed 2 Dec. 2025. 2025.
- [33] Loubna Ben Allal et al. *SmolLM2: When Smol Goes Big – Data-Centric Training of a Small Language Model*. Feb. 2025. DOI: 10.48550/arXiv.2502.02737. arXiv: 2502.02737 [cs]. (Visited on 12/02/2025).
- [34] Hugging Face. *SmolLM2 Models*. <https://huggingface.co/collections/HuggingFaceTB/smollm2-models-67a0e954601d6fca061dbae7>. Model collection page, accessed 2 Dec. 2025. 2025.
- [35] NVIDIA et al. *Nemotron-H: A Family of Accurate and Efficient Hybrid Mamba-Transformer Models*. Sept. 2025. DOI: 10.48550/arXiv.2504.03624. arXiv: 2504.03624 [cs]. (Visited on 12/02/2025).
- [36] NVIDIA. *Nemotron-H: A Family of Accurate and Efficient Hybrid Mamba Transformer Models*. <https://research.nvidia.com/labs/adlr/nemotronh/>. Research project page, accessed 2 Dec. 2025. 2025.
- [37] NVIDIA. *Introducing the Nemotron-H Reasoning Model Family*. <https://developer.nvidia.com/blog/nemotron-h-reasoning-enabling-throughput-gains-with-no-compromises/>. Developer blog post, accessed 2 Dec. 2025. 2025.