Secure Use of Large Language Models for Cybersecurity Decision Support

University of Turku
Department of Computing
Master of Science (Tech) Thesis
Cyber Security Engineering
June 2026
Veera Ollila

Supervisors:
Ismayil Hasanov
Petri Sainio

UNIVERSITY OF TURKU
Department of Computing

Veera Ollila: Secure Use of Large Language Models for Cybersecurity Decision Support

Master of Science (Tech) Thesis, 55 p.
June 2026

This thesis examines the secure use of large language models (LLMs) in cybersecurity decision-support contexts, with particular attention to reliability, privacy, and regulatory compliance. As LLMs are increasingly integrated into security workflows, concerns arise regarding their tendency to produce plausible but incorrect outputs, as well as their handling of sensitive data. These challenges are especially significant in environments governed by strict data protection regulations.

The thesis evaluates the suitability of LLMs as assistive tools for cybersecurity analysts by analysing their performance characteristics, limitations in reasoning, and susceptibility to errors. In parallel, it investigates privacy risks associated with data processing, including issues related to data locality, leakage, and third-party model deployment.

Based on this analysis, the thesis proposes a framework for the controlled and secure use of LLMs in cybersecurity settings. The framework emphasises human oversight, data minimisation, and alignment with existing governance models. The findings suggest that while LLMs can enhance efficiency and support analytical tasks, they should not be treated as authoritative decision-making systems. Instead, their value lies in augmenting human expertise within well-defined operational and regulatory boundaries.

Keywords: Artificial Intelligence, AI, Large Language Model, LLM, privacy, reliability, governance

Contents

Declaration on the Usage of Artificial Intelligence

I used the local Gemma 3 (12B) language model for proofreading this thesis, specifically to identify typographical and minor grammatical errors. AI was not used to generate content or contribute to the ideas, analysis, or conclusions.

1 Introduction

The rapid advancement of artificial intelligence (AI), particularly in the form of large language models (LLMs), has introduced new opportunities for enhancing cybersecurity operations. These models are capable of processing large volumes of unstructured data, generating natural language explanations, and assisting analysts in tasks such as threat interpretation, documentation, and incident analysis. As a result, LLMs are increasingly considered as decision-support tools within security environments.

However, their integration into cybersecurity workflows raises fundamental concerns. Unlike traditional software systems, LLMs operate based on probabilistic pattern recognition rather than deterministic logic. This can result in outputs that appear coherent and convincing but are factually incorrect or misleading. In cybersecurity contexts, where decisions may have significant operational and legal consequences, such behaviour introduces non-trivial risks.

In addition to reliability concerns, the use of LLMs introduces challenges related to privacy and data protection. Security-related data often contains sensitive or personally identifiable information, and its processing must comply with regulatory frameworks such as the General Data Protection Regulation (GDPR). The use of externally hosted models, in particular, raises questions about data locality, third-party access, and compliance with legal requirements.

Given these challenges, there is a pressing need for structured guidance on how

LLMs can be used safely and effectively in cybersecurity environments. This thesis addresses this need by developing a framework that defines appropriate use cases, outlines key risks, and proposes controls for mitigating them. The focus is not on replacing human analysts, but on understanding how LLMs can support their work without compromising reliability or compliance.

1.1 Thesis Structure

This thesis is organized as follows. Chapter 2 introduces the theoretical background, including the architecture of LLMs and key concepts such as reliability, trust, and explainability. Chapter 3 examines relevant governance models and regulatory frameworks, with a particular focus on GDPR and cybersecurity standards. Chapter 4 presents the framework for the secure use of LLMs in cybersecurity contexts. Chapter 5 discusses the findings in relation to the research questions, and Chapter 6 concludes the thesis by summarising key insights and identifying areas for future research.

1.2 Research Questions

RQ1: To what extent are large language models suitable as reliable and privacy-preserving decision support systems in a cybersecurity context?

RQ2: How do large language models meet the reliability requirements of typical cybersecurity support tasks when used in the role of a junior-level security analyst?

RQ3: How do constraints related to data location, data processing, and production costs affect the balance between reliability, privacy, and regulatory compliance in production environments utilising large language models?

1.3 Methodology

This research is based on a qualitative, literature-based approach, grounded in a structured review of recent academic and industry publications. Given the rapid evolution of the field, emphasis was placed on sources published after 2023, while selectively incorporating foundational works where necessary.

The primary data collection method involved literature search using platforms such as Google Scholar and IEEE Xplore. Search queries combined key terms including "large language models", "cybersecurity", "AI reliability", "privacy", and "LLM security". To ensure relevance, sources were selected based on their direct applicability to cybersecurity contexts, discussion of LLM limitations, or contribution to governance and risk management perspectives. Where possible, original sources were prioritised over secondary interpretations to ensure accuracy.

The analysis focuses on synthesising existing knowledge rather than conducting empirical experiments. Particular attention is given to identifying recurring patterns, limitations, and risks associated with LLM usage in cybersecurity contexts. These insights form the basis for the framework proposed in Chapter 4.

This study does not include empirical experimentation or quantitative evaluation of model performance. As a result, the findings are inherently dependent on the quality and scope of the selected literature. While this approach enables a broad and up-to-date overview of the field, it may be subject to selection bias and limited by the availability of publicly documented evidence. These limitations are addressed through careful source selection and critical comparison of multiple perspectives where possible.

2 Background

This chapter outlines the conceptual foundations for analysing LLMs in cybersecurity. It defines key terminology, introduces core architectural principles, and examines trust, reliability, and explainability. It also reviews security and privacy risks and situates LLM capabilities and limitations within cybersecurity practice. This chapter establishes the conceptual basis for evaluating whether LLMs can be reliably integrated into cybersecurity decision-support roles by examining how architectural design, training objectives, and interaction patterns contribute to limitations in reliability, explainability, and trust calibration.

AI	VS	LLM
Secure AI	VS	Securing AI
Trust	VS	Reliability
Transparency	VS	Explainability and Interpretability

Table 2.1: Key conceptual distinctions used in this thesis

The concepts introduced in this Chapter are closely related but refer to fundamentally different evaluation dimensions. This distinction is important because each term implies a different basis for assessing AI systems: some relate to empirical performance, while others involve human interpretation or organisational expecta-

tions. To clarify these differences, Table 2.1 summarises key conceptual contrasts used throughout this thesis.

AI is a broad field concerned with systems capable of tasks typically associated with human cognition, such as reasoning and learning. Machine learning (ML) is a subfield of AI that focuses on data-driven model development. LLMs are a class of ML systems based on deep neural networks trained on large-scale text corpora, optimised to predict and generate natural language.

It is also important to distinguish between *securing AI* and *secure AI*. Securing AI refers to protecting AI systems from external threats such as cyberattacks, data breaches, and model manipulation. This includes ensuring confidentiality, integrity, and availability of models and data [1]. In contrast, secure AI focuses on safe and responsible system behaviour, including alignment, risk mitigation, and harm reduction [1]. This thesis focuses on the latter, particularly in relation to reliability and decision-making.

Interpretability, explainability, and transparency capture different aspects of understanding AI systems. Interpretability concerns understanding how a model produces outputs, while explainability focuses on why a specific output was generated in a given instance. Transparency refers to visibility into system design, training data, and deployment processes [2]. In cybersecurity, these properties support error detection, bias identification, and regulatory compliance, and are therefore essential for justified reliance on model outputs [3].

2.1 Architecture of Large Language Models

LLMs are typically based on the transformer architecture, introduced by Vaswani et al. [4]. Transformers rely on a mechanism known as self-attention, which allows the model to weigh the importance of different words in a sequence when generating or interpreting text. This allows the model to capture contextual relationships across

long passages of text.

LLMs are trained using a next-token prediction objective, where the model learns to predict the most probable next word given a sequence of preceding words. Through training on large-scale text corpora, the model learns statistical patterns of language, enabling it to generate fluent and contextually appropriate responses [4].

This training paradigm is central to understanding the limitations of LLMs in security-critical contexts. Because the objective is to optimise for next-token prediction rather than truth or correctness, factual accuracy is not explicitly enforced during training. As a result, outputs may be fluent and contextually appropriate while still being incorrect or misleading. This misalignment between training objective and operational expectations is a fundamental source of unreliability in downstream applications.

2.2 Trust, Reliability and Explainability

Trust is a broad concept that encompasses different forms of dependence on systems or entities. In the context of technology, it is useful to distinguish between trust and reliance. Reliance refers to acting based on the expectation that a system will perform as intended, whereas trust may additionally involve normative or moral expectations [5], [6]. This thesis focuses primarily on reliance and reliability in the context of AI systems.

The concept of "trustworthy AI" itself has been criticised as conceptually ambiguous and potentially misleading, as it may conflate technical reliability with normative or ethical expectations that are difficult to formalise in practice [7].

An important distinction is that the possibility of trust does not imply its appropriateness. Trust is considered warranted only when it is grounded in the actual trustworthiness of the system, rather than in perceived competence or plausibility of

outputs [8]. Users may rely on systems without sufficient justification, particularly when outputs appear plausible. This creates the risk of overtrust, where confidence in a system exceeds its actual reliability.

In information security, this challenge can be addressed by approaches such as Zero Trust (discussed in Chapter 3.2), which assume that no system or component should be trusted by default. Instead, trust is treated as explicit, limited, and continuously re-evaluated [9], [10], [11]. Applying similar thinking to AI systems suggests that outputs should not be accepted at face value, but rather verified based on context, evidence, and risk.

In complex technological systems, complete understanding is rarely achievable, and some degree of reliance is unavoidable. However, the objective in safety-critical domains such as cybersecurity is not to eliminate reliance, but to ensure that it is justified. This requires minimising unjustified trust through improved transparency, reliability assessment, and continuous monitoring of system behaviour. Without such mechanisms, users may attribute unwarranted credibility to outputs that are not grounded in verifiable reasoning [12].

2.2.1 Reliability in LLMs

Reliability can be defined as the quality of being consistently dependable or trustworthy, whereas reliance refers to the act of depending on a system [13]. In the context of LLMs, reliability is challenged by several factors, including limited reproducibility, lack of transparency, and static training data [14].

Zhou et al. [15] demonstrate that modern language models do not fail in consistent or easily predictable ways. This phenomenon, referred to as difficulty mismatch, undermines the user's ability to form a reliable mental model of system performance.

Another critical issue is the shift from task avoidance to confident error generation. Earlier models were more likely to refrain from answering when uncertain,

whereas modern models tend to provide answers even when incorrect [16]. These errors are often presented with high confidence, making them difficult to detect. This behaviour is closely related to *ultracrepidarianism*, defined as the tendency to provide authoritative-sounding answers beyond the scope of the model's actual capabilities or training. In the context of LLMs, this manifests as confident responses in areas where the system lacks sufficient grounding, increasing the likelihood of misleading outputs.

Several factors contribute to these reliability issues. Reinforcement learning from human feedback (RLHF) tends to favour responses that are clear, complete, and confident, as these qualities are typically preferred by human evaluators. Consequently, answers that appropriately express uncertainty, such as "I'm not sure" or "this may be incorrect", are often rated less favourably, even when they are more accurate or intellectually honest. This introduces a bias in the training signal and implicitly encourages models to adopt an overconfident tone regardless of the actual correctness of the content. In addition, the underlying training objectives prioritise plausibility and linguistic fluency over factual accuracy. Errors that are simple or subtle, including minor factual inaccuracies or small flaws in reasoning, are not strongly penalised during training. As a result, models may generate outputs that are coherent and persuasive in form but unreliable in substance [15].

This behaviour is particularly problematic when LLMs are deployed as decision-support tools in analyst roles. Junior analysts are expected to recognise uncertainty, acknowledge limitations, and escalate unclear cases to more experienced personnel. In contrast, LLMs tend to provide definitive answers even when uncertain. This mismatch increases the risk that incorrect outputs are accepted without sufficient evaluation. This issue is directly related to RQ2: the tendency of LLMs to present uncertain information with unwarranted confidence raises concerns about their ability to fulfill the reliability expectations associated with such roles.

Another challenge related to reliability is prompt sensitivity. Model outputs may vary significantly depending on how a query is formulated [17], [18]. Minor changes in wording may lead to different or even contradictory responses, which complicates consistent use in operational environments. Zhou et al. [15] note that this behaviour can undermine user confidence, particularly when models fail on seemingly simple tasks but succeed when prompts are reformulated. This sensitivity remains an active area of research and poses a direct challenge to the standardisation of LLM-based workflows, particularly in environments where repeatability and consistency are required.

Taken together, these factors indicate that LLM reliability is structurally unstable. This instability makes it difficult for users to develop accurate mental models of system performance, particularly in operational environments where consistent behaviour is expected. In cybersecurity decision-support contexts, this creates a structural mismatch between system behaviour and human expectations of reliability.

A further complication in evaluating reliability is the absence of universally accepted measurement criteria for LLM behaviour in operational environments. Benchmark performance often provides only a partial indication of real-world reliability, as standardised evaluation datasets cannot fully capture the contextual variability and adversarial conditions present in cybersecurity workflows. Models that perform well on static benchmarks may still fail unpredictably when exposed to incomplete, ambiguous, or manipulated inputs.

This limitation is particularly significant in cybersecurity, where reliability is closely tied to contextual awareness and operational consistency rather than isolated task accuracy. Security analysts frequently operate under conditions characterised by uncertainty, incomplete information, and rapidly changing threat landscapes. In such environments, even infrequent model failures may produce disproportionately

large consequences if they occur during critical decision-making processes.

Another important issue concerns calibration. Reliability is not solely determined by whether a model produces correct outputs, but also by whether the confidence expressed by the system accurately reflects the probability of correctness. Poorly calibrated systems may produce answers that appear highly certain despite limited evidential grounding. In practice, this creates a greater operational risk than visible uncertainty, because users are more likely to accept outputs presented with authoritative language.

Research on calibration in machine learning demonstrates that modern neural systems frequently exhibit overconfidence, particularly in cases involving distributional shift or unfamiliar inputs. For LLMs, this problem is amplified by conversational interaction patterns that encourage fluent and definitive responses. As a result, users may interpret linguistic confidence as evidence of factual reliability, even though the underlying generation process is probabilistic rather than knowledge-based.

The challenge is further complicated by the difficulty of distinguishing between linguistic coherence and epistemic validity. LLMs are highly effective at producing text that resembles expert reasoning, but the appearance of expertise does not necessarily imply genuine understanding or accurate inference. In cybersecurity contexts, this distinction is critical because operational decisions often depend on subtle contextual relationships that cannot be reliably inferred through statistical association alone.

Reliability must therefore be understood as a socio-technical property rather than a purely technical metric. The practical dependability of an LLM depends not only on model architecture or benchmark performance, but also on deployment conditions, user expertise, validation processes, and organisational safeguards. A model that is sufficiently reliable in one operational context may become unsuitable

in another if surrounding controls are absent or if user assumptions change.

These observations reinforce the broader argument of this thesis: reliability in LLM systems cannot be assumed as an inherent property of the model itself. Instead, it emerges from the interaction between probabilistic model behaviour, human oversight, and the surrounding governance structure. Consequently, effective deployment depends less on eliminating model errors entirely and more on ensuring that failures remain observable, manageable, and recoverable within operational workflows.

2.2.2 Explainability in AI

Explainability in AI refers to the ability to provide understandable reasons for model outputs. It has been defined as providing insight into why a system produces a particular result in a specific instance [5], and more broadly as a set of methods that enable users to comprehend and trust model outputs [2].

There is ongoing debate regarding the role of explainability in supporting trust. Some researchers argue that explainability is not strictly necessary for trust, particularly when systems demonstrate consistent and reliable performance [5]. Others emphasise that a lack of interpretability limits accountability and complicates adoption in high-risk domains [19]. In cybersecurity contexts, explainability is therefore less about fostering trust and more about enabling verification, auditing, and responsibility assignment when systems influence critical decisions.

In cybersecurity environments, explainability also serves an operational purpose beyond transparency alone. Analysts must frequently justify why a particular recommendation, classification, or prioritisation decision was made. When AI-assisted systems contribute to these processes, explainability becomes necessary for documenting reasoning chains, supporting incident review, and enabling accountability during post-incident analysis.

Practical explainability in LLM systems is often achieved through indirect meth-

ods rather than full interpretability of internal model behaviour. These methods may include chain-of-thought prompting, retrieval-based citation mechanisms, confidence indicators, or external validation systems that provide supporting evidence for generated outputs. Although such approaches do not fully expose internal reasoning processes, they can improve the traceability and verifiability of outputs in operational settings.

Model cards and system documentation also contribute to explainability by clarifying intended use cases, training limitations, known weaknesses, and evaluation procedures. In organisational contexts, this documentation supports risk assessment and helps analysts understand the conditions under which model outputs may become unreliable. Without such contextual information, users may incorrectly generalise model capabilities beyond their intended scope.

Another important aspect is the distinction between explainability for developers and explainability for operators. Developers may require insight into training behaviour, architecture, or optimisation procedures, whereas operational users primarily require understandable justifications for specific outputs. In cybersecurity workflows, explainability must therefore be adapted to the needs of analysts, auditors, and decision-makers rather than focusing solely on technical transparency.

However, explainability mechanisms themselves introduce limitations. Generated explanations may not accurately reflect the actual internal processes of the model and can instead function as plausible post hoc rationalisations. Consequently, explainability should not be interpreted as proof of correctness. Rather, it should be viewed as a supporting mechanism that enables verification, critical evaluation, and responsible oversight.

For this reason, explainability alone cannot resolve the broader reliability challenges associated with LLMs. Instead, it should operate alongside validation procedures, human oversight, and governance controls as part of a broader trust man-

agement strategy. Within cybersecurity environments, the objective is not complete interpretability, but sufficient transparency to support safe operational use and informed human judgment.

The Black Box

Many AI systems based on deep learning are often described as "black boxes" due to limited transparency in their internal decision-making processes. While these models may achieve high performance, the mechanisms underlying their outputs are difficult to inspect or validate [8], [14], [19]. This opacity makes it difficult to determine why a specific output was produced or whether it is grounded in reliable reasoning. As a result, concerns arise regarding reliability, bias, and accountability, since incorrect or systematic errors may be difficult to detect or verify.

Addressing the black box problem requires progress on several areas. Explainability techniques aim to provide human-understandable justifications for model outputs, often through feature attribution or retrospective analysis. Transparency can be improved through practices such as documenting training data, maintaining model cards, and ensuring data provenance and lineage.

Accountability is another critical aspect. In real-world deployments, it must be clearly defined who is responsible for decisions influenced by AI systems. This often requires monitoring mechanisms and human-in-the-loop approaches, where human oversight is maintained in critical decision-making processes.

Addressing this opacity requires not only explainability techniques, but also broader transparency practices such as documenting data provenance, maintaining model cards, and providing insight into model confidence and feature importance. Together, these measures aim to reduce the opacity of AI systems and support their safe and responsible use.

2.3 Risks and Security Concerns

The risks associated with LLMs can be grouped into three main categories: technical security risks, privacy risks, and organisational risks. Technical risks include prompt injection, data poisoning, model exploitation, and data leakage [20], [21]. These arise from the fact that natural language interfaces allow adversarial manipulation through seemingly benign inputs, expanding the traditional attack surface of software systems.

Privacy risks arise from sensitive user inputs, third-party integrations, and contextual inference of user attributes [22]. Models may also memorise and reproduce training data, creating potential regulatory issues under frameworks such as GDPR.

Organisational risks include "shadow AI," where employees use external tools without oversight, potentially exposing confidential data [23]. Together, these risks highlight the need for strict governance, monitoring, and access controls in LLM deployment.

2.3.1 Privacy Issues with LLMs

LLMs introduce several privacy-related risks due to their reliance on large-scale data processing and their interaction with user-provided inputs. Yan et al. [22] identify issues such as sensitive user queries, third-party integrations, contextual leakage, and inference of personal preferences.

Sensitive queries involve users providing personally identifiable information (PII) or confidential data as input. When this information is processed by external services, it may be collected or shared in ways that are not fully transparent to users. This lack of clarity can lead to potential violations of privacy policies and user expectations [24], [25].

Another key issue is contextual leakage. Over time, repeated interactions can allow systems to infer who a user is or where they are located, even if they never ex-

plicitly share that information. In addition, AI models can build detailed profiles of users by learning their preferences, habits, and behavioural patterns. While this can improve personalisation, it also increases the risk of profiling and targeted content that users may not have knowingly agreed to, making strong privacy protections and clear disclosures essential [26], [27], [28], [29], [30], [31].

These dynamics blur the boundary between input data and inferred personal data. Even when explicit identifiers are absent, contextual information and repeated interactions can enable the reconstruction of identity, location, and behavioural profiles. As a result, privacy risks extend beyond direct data disclosure and include inference-based profiling, which complicates both regulatory compliance and meaningful user consent.

2.3.2 Data Leakage

LLMs can unintentionally memorise and reproduce portions of their training data, including sensitive or proprietary information [22]. This behaviour poses potential compliance risks under data protection regulations such as GDPR. In addition, prompt leakage can occur when sensitive inputs are revealed through model outputs or recorded in system logs [14], [32].

A related concern is "shadow AI," which refers to the unauthorised use of AI tools within organisations [23]. In practice, employees may input confidential or sensitive information into external systems without organisational oversight. This creates a governance gap, where data handling practices fall outside formal security controls. Unlike traditional leakage risks, shadow AI is difficult to detect and regulate because it emerges from everyday user behaviour rather than system-level vulnerabilities.

Overall, both training data memorisation and prompt leakage highlight the need for careful data governance when deploying LLMs. Ensuring sensitive information is not embedded in inputs or outputs, and monitoring the use of external AI services,

can mitigate these risks [14].

2.3.3 Data Poisoning

Data poisoning refers to the manipulation of training data or fine-tuning processes to introduce biases, vulnerabilities, or hidden backdoors into a model [14]. Even small amounts of carefully crafted malicious data can influence model behaviour, particularly when embedded in large datasets where detection is difficult [33], [34], [35].

Importantly, increasing dataset size does not inherently mitigate this risk. On the contrary, larger datasets may obscure targeted manipulations, making them less visible during curation and evaluation. Empirical evidence across pre-training, fine-tuning, and preference learning stages demonstrates that even a minimal fraction of poisoned data can reliably induce harmful or unintended behaviours. Consequently, data poisoning represents a subtle yet high-impact threat in the development and deployment of LLMs [34], [36], [37].

2.4 Implications for Cybersecurity Practice

LLMs function as high-leverage but high-risk tools in cybersecurity environments. Their ability to process unstructured data enables efficiency gains across analytical workflows. However, these advantages are inseparable from inherent limitations related to reliability, explainability, and data handling. As a result, LLMs should be treated as assistive tools whose outputs require validation.

From a practical standpoint, the integration of LLMs reshapes how analysts interact with information. Rather than replacing human expertise, these systems extend analytical capabilities while simultaneously introducing new forms of uncertainty. Analysts therefore face two competing requirements: exploiting efficiency

gains while preventing overreliance. Taken together, these architectural properties, reliability limitations, and security risks define the conditions under which LLMs can be applied in practice.

2.4.1 Capabilities

LLMs support a range of cybersecurity functions, particularly those involving interpretation and transformation of unstructured data. They are effective in tasks such as summarising vulnerability reports, correlating threat intelligence, and extracting relevant insights from large datasets. These capabilities are especially valuable in environments where analysts must process high volumes of heterogeneous information under time constraints.

In addition, LLMs can assist in areas such as phishing detection, incident response, and digital forensics by providing contextual interpretations of ambiguous inputs. Their ability to generate natural language explanations also makes them useful for documentation and reporting, reducing the cognitive burden on analysts.

Beyond analytical support, LLMs contribute to automation in tasks such as vulnerability assessment, malware classification, and protocol analysis. They are also increasingly used in training and simulation contexts, including penetration testing scenarios and educational environments, where they can model attacker behaviour or explain defensive strategies.

In integrated systems, LLMs often interact with external services through APIs and authentication mechanisms such as single sign-on and delegated authorisation protocols. While this enhances functionality, it also introduces additional security considerations, particularly in the handling of credentials and access permissions.

Overall, these capabilities position LLMs as tools that enhance scalability and efficiency in cybersecurity operations, particularly in data-intensive tasks.

2.4.2 Limitations

Despite their utility, LLMs exhibit structural limitations that constrain their reliability in security-critical contexts. One of the most prominent issues is hallucination, where the model generates plausible but incorrect information. In cybersecurity, where accuracy is essential, such behaviour can lead to incorrect assessments or flawed decision-making.

Another key limitation is the absence of grounded reasoning. LLM outputs are based on statistical correlations rather than verified logical processes, which can result in misinterpretation of complex or context-dependent data, such as system logs or attack patterns.

Consistency is also a concern. Model outputs are highly sensitive to prompt formulation, meaning that small variations in input can produce different or even contradictory results. This variability complicates standardisation and reduces predictability in operational workflows.

Privacy and data protection introduce further constraints. In cloud-based deployments, sensitive information may be processed outside organisational boundaries, raising concerns related to data sovereignty, regulatory compliance, and exposure to third-party systems.

Taken together, these limitations highlight that LLM outputs cannot be treated as inherently reliable. Instead, they require validation, contextual interpretation, and integration within controlled workflows.

2.4.3 LLMs in Security Research

The use of LLMs in cybersecurity research is an emerging and rapidly evolving field. Current research explores applications such as automated vulnerability detection, malware analysis, and the automation of security workflows. These efforts demonstrate the potential of LLMs to augment both defensive and analytical capabilities.

However, most implementations remain experimental, and large-scale deployment in operational environments is still limited. This is largely due to unresolved challenges related to reliability, evaluation methodologies, and risk management.

Research in this area increasingly focuses not only on capabilities but also on limitations, including adversarial vulnerabilities, prompt manipulation, and data leakage. As a result, there is growing recognition that technical performance alone is insufficient. Robust evaluation frameworks and governance models are equally necessary.

In this context, LLMs should be viewed as tools for exploration and augmentation rather than fully mature solutions. Their role in research is to extend analytical possibilities, while also highlighting the need for structured approaches to safe and controlled deployment.

3 Governance and Security Frameworks for Evaluating LLM Use

The deployment of LLMs in cybersecurity environments is influenced by both regulatory requirements and organisational security practices. This is particularly important when LLMs process personal, confidential, or otherwise sensitive information. In such situations, compliance obligations extend beyond conventional data processing activities and may affect how models are trained, deployed, and used in operational workflows.

Within the European Union, the General Data Protection Regulation (GDPR) provides the primary legal framework governing the processing of personal data. However, regulatory compliance alone does not guarantee that an LLM can be deployed safely or effectively. Organisations must also consider broader questions related to governance, accountability, risk management, and system resilience.

When LLMs are integrated into cybersecurity workflows, these requirements extend beyond traditional data processing to include model training, inference, and system interaction phases. This creates a complex compliance landscape in which technical design decisions directly influence legal obligations.

From an organisational perspective, these regulatory requirements are closely tied to broader risk management practices, which aim to identify, assess, and mitigate risks associated with data processing and system deployment [38].

3.1 Security Governance Frameworks

Regulatory compliance alone is not sufficient to ensure secure deployment of LLMs. Organisations must also adopt structured governance frameworks that define how cybersecurity risks are identified, managed, and mitigated. These frameworks provide a systematic approach to integrating LLMs into existing security operations.

3.1.1 GDPR

The General Data Protection Regulation (GDPR) establishes a comprehensive legal framework governing the processing of personal data within the EU. Its primary objective is to ensure a high level of protection for individuals while enabling the free movement of personal data under harmonised conditions across member states [39], [40].

A defining characteristic of the GDPR is its broad and technology-neutral definition of personal data. In addition to direct identifiers such as names or identification numbers, personal data includes indirect identifiers such as IP addresses, location data, and online behavioural patterns that can be linked to an individual [39], [41]. In the context of LLMs, this definition is particularly significant, as inputs that appear non-sensitive in isolation may become identifiable when combined with contextual information or through inference.

The regulation is structured around a set of core principles that govern all data processing activities. These include lawfulness, fairness, and transparency, along with purpose limitation, data minimisation, accuracy, storage limitation, and integrity and confidentiality [39], [42]. The principle of accountability requires organisations not only to comply with these principles but also to demonstrate such compliance through appropriate documentation and controls.

Collectively, these principles establish a lifecycle-oriented approach to data protection, shaping how personal data is collected, used, stored, and ultimately disposed

of across processing activities.

For LLM-based systems, these principles primarily constrain data handling practices rather than directly regulating model behaviour. The principle of data minimisation limits the amount of input data that can be provided to models, particularly when such data contains personal or sensitive information. Purpose limitation restricts the reuse of data across different tasks, which may conflict with practices such as prompt logging, dataset reuse, or iterative improvement processes.

The GDPR further defines lawful bases for data processing, including consent, performance of a contract, legal obligation, and legitimate interest [40]. In cybersecurity contexts, legitimate interest is commonly applied; however, it requires a balancing test to ensure that organisational objectives do not override the rights and freedoms of individuals. This is particularly relevant when LLMs are used to analyse user-generated or operational data.

A particularly challenging aspect of GDPR compliance in AI systems relates to automated decision-making. Where processing leads to decisions that produce legal or similarly significant effects on individuals, the regulation requires that individuals are provided with meaningful information about the logic involved, as well as the possibility of human intervention [41]. For LLMs, which often operate through opaque and probabilistic mechanisms, providing such explanations in a meaningful and verifiable manner can be difficult.

Another important aspect is the treatment of inferred data. The GDPR recognises that information derived from data processing may still qualify as personal data if it relates to an identifiable individual [39]. For LLMs, this is significant because model outputs may reveal behavioural patterns, preferences, or risk indicators, even when the original input does not contain explicit identifiers.

Overall, the GDPR does not regulate the internal functioning or risk classification of AI systems, but it imposes strict requirements on how data is collected,

processed, stored, and reused. As a result, compliance primarily influences data handling practices, including input design, logging policies, and data governance mechanisms. Broader requirements related to system behaviour, transparency, and risk management are addressed by complementary regulatory frameworks such as the EU AI Act discussed in Chapter 3.1.3.

3.1.2 NIST Cybersecurity Framework

The National Institute of Standards and Technology (NIST) Cybersecurity Framework (CSF) provides a practical structure for examining how LLM-based systems fit into existing cybersecurity programmes. Rather than prescribing specific technologies, the framework focuses on outcomes and risk management activities, making it adaptable to emerging technologies such as generative AI. Framework's functions are Identify, Protect, Detect, Respond, Recover, and Govern. [43]

When viewed through the NIST lens, an LLM should not be treated solely as a productivity tool. It is simultaneously an organisational asset, a potential attack surface, and a source of operational risk. This perspective encourages organisations to consider not only how the model is protected, but also how its outputs are monitored and validated.

The CSF functions provide a structured approach to managing risks associated with LLM deployment. When applied to AI systems, the framework emphasises asset identification, protection mechanisms, monitoring, incident response, recovery, and governance. Its outcome-oriented nature makes it adaptable to emerging technologies such as LLMs.

The NIST framework also provides a useful structure for understanding the operational lifecycle of LLM-based systems. Within the Identify function, organisations must evaluate not only technical assets but also the informational dependencies associated with model deployment. This includes identifying where prompts originate,

what categories of information are processed, and which external services participate in inference or storage operations.

Within the Protect function, controls must address both conventional cybersecurity risks and AI-specific vulnerabilities. Traditional protections such as authentication, encryption, and access control remain essential, but additional safeguards are required for prompt filtering, model isolation, and prevention of unauthorised data exposure. Because LLMs interact through natural language interfaces, attack vectors may emerge through ordinary user interaction rather than direct technical exploitation.

The Detect function becomes particularly important in probabilistic AI systems. Unlike conventional software failures, problematic LLM behaviour may not produce obvious technical errors. Instead, failures may appear as subtle hallucinations, misleading recommendations, or inconsistent reasoning patterns. Detection mechanisms must therefore include qualitative monitoring of outputs in addition to traditional infrastructure telemetry.

The Respond and Recover functions similarly require adaptation. Incident response procedures must account for the possibility that harmful outputs originate from model behaviour rather than external compromise alone. Organisations may therefore need rollback procedures for model versions, mechanisms for disabling unsafe functionality, and escalation pathways for AI-related operational anomalies.

The Govern function integrates these activities into broader organisational strategy. Governance structures must define responsibility for model oversight, acceptable use policies, validation standards, and regulatory compliance obligations. Without clear governance, organisations risk inconsistent deployment practices and fragmented accountability structures.

From this perspective, the NIST framework provides more than a general cybersecurity model; it offers a conceptual foundation for operationalising AI governance

within existing security programs. Its flexibility allows organisations to incorporate AI-specific controls without abandoning established cybersecurity practices, making it particularly valuable for environments adopting LLM-based decision-support systems.

3.1.3 EU AI Act

The European Union Artificial Intelligence Act (EU AI Act) establishes a risk-based regulatory framework for the development and deployment of AI systems within the European Union. Unlike the GDPR, which focuses on data protection, the AI Act addresses the broader lifecycle of AI systems, including design, deployment, and use.

The regulation classifies AI systems into four risk categories: unacceptable risk, high risk, limited risk, and minimal risk. Systems classified as unacceptable risk are prohibited, while high-risk systems are subject to strict requirements related to risk management, transparency, human oversight, and technical robustness.

In cybersecurity contexts, LLM-based systems may fall into the high-risk category when they are used in critical infrastructure, decision-support roles, or contexts that significantly affect individuals or organisations. This classification introduces obligations such as documentation of system behaviour, traceability of outputs, and implementation of human oversight mechanisms.

A key requirement of the AI Act is the establishment of risk management systems that operate throughout the lifecycle of the AI system. This aligns closely with cybersecurity practices, where continuous monitoring, validation, and incident response are already standard. For LLMs, this implies that risks such as hallucination, prompt injection, and data leakage must be explicitly identified, assessed, and mitigated.

Transparency requirements under the AI Act also have direct implications for LLM deployment. Users must be informed when they are interacting with an AI

system, and, in certain cases, meaningful information about system capabilities and limitations must be provided. This reinforces the need to avoid presenting LLM outputs as authoritative or fully reliable.

Compared to GDPR, which focuses on data, the AI Act focuses on system behaviour and risk. Together, the GDPR and the EU AI Act establish complementary constraints: the former governs data processing, while the latter governs system-level risk and behaviour. For organisations deploying LLMs in cybersecurity, compliance requires alignment with both regulatory regimes, integrating data protection with system-level risk management. [44], [45], [46]

3.2 System Resilience and Trust Architecture

Resilience is a central concept in modern cybersecurity, referring to the ability of systems to withstand, adapt to, and recover from disruptions. Rather than assuming that all threats can be prevented, resilient systems are designed to maintain acceptable levels of operation even under adverse conditions.

In practice, resilience is closely tied to prioritisation and alignment with organisational risk. As noted by Leppänen [47], cybersecurity failures often do not stem from a lack of visibility or tooling, but from an inability to focus on what matters most. When all risks are treated as equally critical, resources are dispersed inefficiently, and genuinely high-impact vulnerabilities may remain unaddressed. From this perspective, resilience is not achieved by maximising coverage, but by ensuring that critical assets and attack paths are consistently protected and monitored.

In the context of LLM-based systems, resilience must also account for inherent uncertainty. Unlike traditional software, LLMs do not fail deterministically; instead, they may produce plausible but incorrect outputs, behave inconsistently across similar inputs, or degrade under adversarial interaction. As a result, resilience cannot be achieved solely through preventive controls or static safeguards, but must be

engineered through mechanisms that limit the impact of such behaviour.

This shifts the focus from correctness to failure management. The objective is not to eliminate errors, but to ensure that errors are detectable, containable, and recoverable. In practice, this includes validating outputs, restricting high-risk actions, and preventing the propagation of incorrect information into downstream processes.

Continuous adaptation is another defining characteristic of resilient systems. As highlighted in continuous resilience models, effective cybersecurity requires a feedback loop in which detection, response, and risk assessment inform one another over time [47]. For LLM systems, this implies ongoing monitoring of outputs, identification of failure patterns, and iterative refinement of prompts, controls, and deployment strategies.

An additional dimension of resilience involves graceful degradation. In resilient systems, failures should not propagate uncontrollably or cause disproportionate operational disruption. Instead, system capability should degrade in a controlled and predictable manner. For LLM-based systems, this may involve limiting autonomous functionality during periods of uncertainty, restricting access to sensitive actions, or falling back to deterministic rule-based systems when reliability thresholds are not met [48], [49].

This principle is particularly important because LLM failures are often difficult to detect immediately. Unlike conventional system outages, probabilistic failures may continue producing seemingly functional outputs while gradually introducing incorrect assumptions into analytical workflows. Consequently, resilience mechanisms must focus not only on maintaining availability but also on preserving decision quality under degraded conditions.

Redundancy also plays an important role in resilient AI deployment. Independent verification systems, secondary analytical tools, or parallel review mechanisms

can reduce the likelihood that a single incorrect output propagates directly into operational decisions. In cybersecurity environments, where false positives and false negatives both carry significant consequences, layered validation provides an important safeguard against model instability [50].

Overall, resilience in LLM systems is achieved through layered safeguards and continuous adjustment rather than static protection. This perspective aligns with broader cybersecurity practices and provides a foundation for integrating probabilistic AI systems into environments where reliability and risk control are critical.

Trust architectures define how systems manage assumptions about reliability, authenticity, and correctness across interactions. In cybersecurity contexts, these architectures are shifting from static assumptions of trustworthiness to dynamic, continuously evaluated models of risk. This transition is particularly relevant for AI systems, where both inputs and outputs may be uncertain, adversarial, or context-dependent.

Zero Trust

Zero Trust represents a shift from traditional perimeter-based security models to an approach based on continuous verification. It operates under the assumption that no system component should be trusted by default.

In practical terms, all interactions must be authenticated, authorised, and evaluated based on context. Trust is treated as dynamic and conditional, requiring continuous reassessment.

Applying Zero Trust principles to LLMs implies that model outputs should not be accepted without validation. Instead, outputs must be treated as potentially unreliable and subject to verification based on their intended use. Similarly, inputs must be controlled to prevent prompt injection and data leakage.

This approach aligns closely with cybersecurity best practices and reinforces the

role of LLMs as assistive tools rather than authoritative systems.

3.3 AI System Autonomy and Agent Models

The increasing autonomy of AI systems represents a significant shift in how decision-support technologies are designed and deployed. While earlier systems operated primarily as passive tools, modern LLM-based systems can perform multi-step reasoning, interact with external services, and act with limited or delayed human intervention. This evolution introduces new governance challenges, as system behaviour becomes less predictable and more difficult to constrain.

In cybersecurity contexts, autonomy must be carefully balanced against risk. Systems that are capable of independent action can improve efficiency, but they also increase the potential impact of errors, misuse, or adversarial manipulation. As a result, understanding different models of autonomy, particularly agent-based architectures, is essential for evaluating how LLMs can be safely integrated into operational environments.

This section examines the implications of AI autonomy for security, focusing on agent-based systems and their governance requirements. It explores how increasing system capability necessitates stronger controls, clearer accountability, and more robust mechanisms for oversight and validation.

3.3.1 Agent-Based Systems

The increasing autonomy of AI systems introduces new challenges for governance and security. LLM-based agents are capable of performing multi-step tasks, interacting with external systems, and making decisions with limited human intervention. While this enhances efficiency, it also increases the risk of unintended or harmful outcomes.

Governance of such systems can be understood along three dimensions: technical safeguards, process controls, and organisational structures. Technical safeguards include mechanisms such as access restrictions, sandboxing, and fail-safe controls. Process controls involve monitoring, auditing, and risk-based permissions. Organisational structures define accountability and ensure that responsibility for system behaviour is clearly assigned.

One critical mechanism in this context is session isolation. By restricting each interaction to a controlled environment, organisations can reduce the risk of data leakage and cross-session contamination. This is particularly important in multi-user systems where sensitive data may be processed.

As system autonomy increases, requirements for transparency and traceability also become more stringent. It must be possible to reconstruct decisions, identify contributing factors, and assign responsibility. Without such capabilities, effective governance becomes difficult.

Robustness and Conservative Decision-Making

Recent developments in multi-agent systems emphasise robustness over optimal performance. Instead of maximising expected outcomes under ideal conditions, these systems prioritise strategies that remain effective under uncertainty and adversarial conditions.

This perspective is directly applicable to LLMs in cybersecurity contexts. In high-risk environments, it is often preferable for a system to avoid making a decision rather than provide an incorrect one with high confidence. This aligns with the expectations placed on human analysts, who are trained to recognise uncertainty and escalate when necessary.

Accordingly, LLM-based systems should be designed to reflect conservative decision-making principles. This includes mechanisms for expressing uncertainty, deferring

decisions, and incorporating human oversight. Such an approach reduces the risk of critical errors and improves overall system reliability.

DeepNash

Perolat et al. [51] introduced DeepNash, a model-free multi-agent system designed to operate under uncertainty and incomplete information. Rather than optimising for maximal expected reward under ideal assumptions, it prioritises strategies that remain stable against exploitation and adversarial conditions.

This perspective is particularly relevant for LLM-based decision support in cybersecurity. In such contexts, overconfident but incorrect recommendations can have significant consequences. Systems that prioritise conservative and reliable behaviour may therefore be preferable to those optimised for aggressive performance.

This suggests a design philosophy in which uncertainty is treated as a fundamental constraint. Rather than attempting to model all possible variables or predict adversarial intent, systems should rely on observable signals and incorporate mechanisms that discourage speculative or unsupported outputs. When confidence is insufficient, deferring decisions or escalating to human operators is often the safer course of action.

This approach aligns with expectations for junior-level cybersecurity practitioners, who are expected to recognise uncertainty and escalate when necessary. Similarly, LLM systems that acknowledge uncertainty and defer decisions when confidence is insufficient may provide more reliable support than systems optimised for definitive answers.

Avoiding overreach is therefore an important design objective. Systems that recognise the limits of their knowledge contribute to safer and more trustworthy operations.

Together, these governance, regulatory, and resilience perspectives provide the

foundation for the framework presented in Chapter 4. They translate the technical limitations of LLMs into practical requirements for their safe use in cybersecurity decision-support environments.

4 Framework for Secure LLM Usage

This chapter presents a structured framework for the secure integration of LLMs into cybersecurity environments. Grounded in the technical limitations identified in Chapter 2 and the governance requirements outlined in Chapter 3, this framework translates abstract risks into concrete operational controls.

The framework assumes that LLMs should be treated as probabilistic and potentially unreliable components rather than trusted decision-making systems. Because uncertainty cannot realistically be removed from LLM behaviour, the focus shifts toward containing its effects through monitoring, validation, and layered safeguards. The proposed model follows a defence-in-depth approach, where controls are applied across input handling, system architecture, output validation, and organisational governance.

4.1 Framework Overview

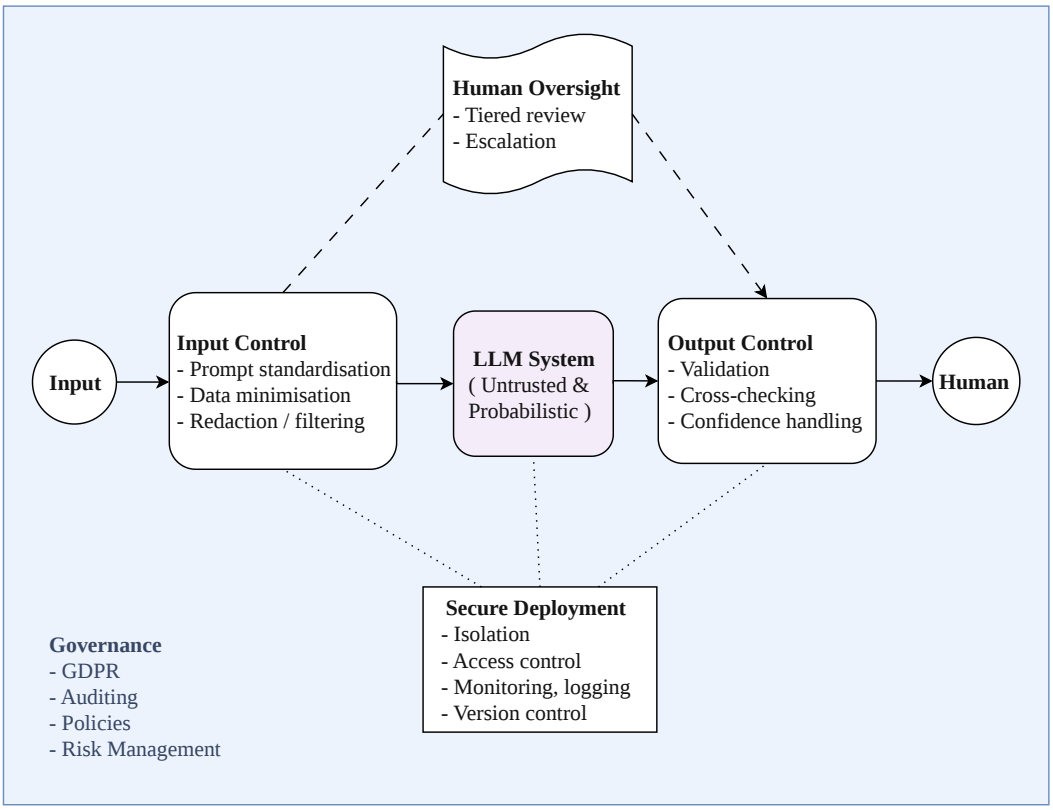

Figure 4.1: Framework

The proposed framework is illustrated in Figure 4.1, which presents a layered control model for the secure integration of LLMs in cybersecurity contexts. The model conceptualises the LLM not as a trusted decision-making entity, but as an untrusted, probabilistic component embedded within a controlled system.

The framework combines a linear processing pipeline with cross-cutting control layers. At its core, the system follows a structured flow in which inputs are first subject to pre-processing controls, then processed by the LLM, and finally validated before being exposed to human operators. This pipeline reflects the fundamental assumption that LLM outputs cannot be treated as inherently reliable and must

therefore be systematically verified.

Surrounding this core process, the framework introduces multiple layers of control that operate at different stages of the system lifecycle. Input control mechanisms constrain the data and instructions provided to the model, reducing the risk of prompt injection, ambiguity, and unintended data exposure. Output control mechanisms ensure that generated responses are treated as provisional results, subject to validation and cross-checking before use.

Human operators remain central to the model. Rather than acting as passive recipients of model outputs, human operators are responsible for validating, interpreting, and, where necessary, overriding system behaviour. The inclusion of feedback loops from human oversight to earlier stages of the pipeline reflects the adaptive nature of the framework, enabling continuous refinement of prompts, policies, and validation strategies.

In addition to these operational controls, the framework incorporates a secure deployment layer that underpins the entire system. This layer includes architectural safeguards such as isolation, access control, monitoring, and logging. Unlike the primary data flow, these mechanisms do not directly process inputs or outputs but instead constrain the environment in which the LLM operates, limiting the impact of potential failures or compromises.

Beyond the technical controls, the framework also incorporates governance constraints representing regulatory, organisational, and risk management requirements. This includes compliance with data protection regulations, auditing requirements, and internal security policies. By positioning governance as an outer layer, the framework emphasises that technical controls alone are insufficient. Secure LLM integration requires alignment with legal and organisational requirements.

4.2 Reliability Requirements

Due to their probabilistic nature, LLMs cannot guarantee consistent or factually accurate outputs. As established in Chapter 2, their behaviour is shaped by statistical patterns rather than grounded reasoning, leading to risks such as hallucination, overconfidence, and prompt sensitivity. To address these limitations, reliability must be enforced externally through structured controls.

Human oversight represents a central reliability requirement. A tiered oversight model is particularly suitable. Low-risk activities, such as text summarisation or document support, may require lighter review, while high-impact tasks, including incident classification or analytical recommendations, should be evaluated by more experienced personnel. Clear escalation procedures are equally important. When outputs appear uncertain, inconsistent, or difficult to verify, responsibility should shift to human reviewers rather than allowing the system to proceed autonomously.

Prompt engineering standards provide another important mechanism for improving reliability. Prompt construction should follow standardised practices that emphasise clarity, sufficient contextual information, and explicit output requirements. Standardisation reduces variability caused by ambiguous phrasing and helps produce more predictable behaviour across comparable tasks. In addition, organisations should document common prompt failure patterns, including vague instructions, incomplete contextual framing, and leading questions, in order to reduce avoidable sources of error.

Output validation methods are necessary because LLM responses cannot be assumed to be inherently correct. Generated outputs should therefore be treated as provisional results requiring verification before operational use. Validation mechanisms may include cross-checking responses against trusted sources, applying rule-based verification methods, or comparing outputs across multiple prompts, models, or independent tools. For higher-risk applications, confirmation through external

datasets, established analytical methods, or expert review becomes especially important.

An important practical objective is proper calibration rather than the maximisation of confidence. A useful decision-support system should not merely provide correct answers, but should align its expressed certainty with its actual likelihood of correctness. In cybersecurity environments, poorly calibrated confidence may be more harmful than visible uncertainty, as it increases the probability that incorrect recommendations are accepted without sufficient scrutiny or review.

This consideration is particularly relevant for LLMs, which frequently present uncertain or incomplete conclusions in authoritative language. Reliability therefore depends not only on factual accuracy, but also on the system's ability to appropriately communicate uncertainty and operational limitations.

An additional reliability requirement concerns awareness of uncertainty. Systems should communicate when outputs may be unreliable rather than presenting all responses with equal confidence. This can be achieved through cautious wording, confidence indicators, or mechanisms that explicitly acknowledge ambiguity or incomplete information. When confidence is insufficient, outputs should be delayed, escalated, or subjected to additional review. Such behaviour aligns more closely with expectations placed on human analysts, who are expected to recognise uncertainty and seek validation when appropriate.

Beyond operational validation, reliability also requires continuous evaluation throughout the deployment lifecycle. Static testing performed during initial implementation is insufficient because model behaviour may change over time due to updates, modified prompts, environmental changes, or evolving user interaction patterns. Organisations should therefore establish recurring evaluation procedures that monitor both technical performance and operational suitability.

One important mechanism is adversarial testing, commonly referred to as red

teaming. In this context, systems are intentionally exposed to manipulative, ambiguous, or malicious inputs in order to identify weaknesses before operational exploitation occurs. For LLMs, red teaming may include prompt injection attempts, attempts to bypass safeguards, or scenarios designed to provoke hallucinations and policy violations. Such testing is particularly important because vulnerabilities may emerge through ordinary language interaction rather than through conventional software exploitation techniques.

Reliability assessment should also incorporate scenario-based testing aligned with actual operational workflows. Generic benchmark datasets often fail to capture the contextual complexity of cybersecurity environments. Evaluations should therefore reflect realistic analyst tasks, including incident triage, threat interpretation, vulnerability assessment, and ambiguous event analysis. By grounding evaluation procedures in operational use cases, organisations can better understand how systems behave under practical conditions.

Another important consideration is output consistency. In operational settings, the same or semantically similar inputs should not produce radically different conclusions without clear justification. Excessive variability complicates standardisation and reduces analyst confidence in system behaviour. Consistency testing can therefore provide insight into prompt sensitivity and identify situations where outputs become unstable or contradictory.

Organisations should additionally establish explicit reliability thresholds that define acceptable operational use. Lower-risk functions may tolerate moderate inaccuracies if outputs remain subject to human review, whereas high-risk functions may require significantly stricter validation criteria. These thresholds should be determined through risk assessment rather than purely technical performance metrics.

The role of human expertise remains central throughout this process. Reliability cannot be fully automated because the operational significance of outputs depends

heavily on contextual interpretation. Human analysts provide the domain understanding necessary to evaluate whether generated recommendations are plausible, relevant, and proportionate to the observed situation.

4.3 Privacy and Data Protection Requirements

Using LLMs in cybersecurity settings requires compliance with GDPR principles and related data protection obligations. This framework prioritises minimising privacy risks, ensuring data sovereignty, safeguarding sensitive information, and maintaining compliance.

A central requirement is the *operationalisation of GDPR principles*, particularly data minimisation and pseudonymisation. Only the minimum amount of information necessary for a specific task should be provided to the model. Personally identifiable information (PII) should be excluded unless explicitly authorised and operationally justified. Where possible, information should be pseudonymised or anonymised prior to processing. In practical terms, this may involve hashing IP addresses, replacing sensitive identifiers with tokens, or separating identifying information from operational datasets.

Another important consideration is *data residency and sovereignty*. Deployment choices directly affect regulatory exposure and organisational control over information processing. Preference should therefore be given to models hosted within the European Union or within environments that provide equivalent data protection guarantees. When data transfers outside the EU are unavoidable, organisations must rely on appropriate contractual safeguards and comply with applicable adequacy requirements and transfer mechanisms.

Input filtering and redaction provide an additional protective layer before information reaches the model. Automated mechanisms should proactively identify and remove sensitive content from prompts, including names, email addresses, phone

numbers, financial information, and other potentially identifying data. While automated filtering is valuable as a first-line safeguard, it should not be regarded as sufficient in isolation. Manual review processes remain necessary to detect false negatives, assess edge cases, and ensure comprehensive protection. These reviews should be documented and periodically audited to maintain accountability and demonstrate compliance.

These measures reduce unnecessary data exposure while supporting compliant deployment.

4.4 Secure Deployment Architecture

As previously established, a secure deployment architecture is essential for mitigating risks associated with LLMs and supporting their responsible use in cybersecurity contexts. Three design principles are particularly important: isolated deployment, comprehensive monitoring and auditing, and version control combined with security hardening.

Isolated deployment forms the foundation of a secure operational environment. LLM infrastructure should operate through dedicated inference endpoints separated from production systems and protected within hardened network boundaries. Such separation reduces the available attack surface and limits the potential impact of system compromise. Access to the environment should be restricted through strong authentication mechanisms, including multi-factor authentication, role-based permissions, and network segmentation. Outbound connectivity should also be constrained to minimise opportunities for unauthorised communication or data leakage.

Comprehensive monitoring and auditing are necessary for maintaining accountability, detecting abnormal behaviour, and supporting incident response. Logging mechanisms should capture relevant operational events, including prompts, responses, user interactions, system activity, and configuration changes. Because

these records may themselves contain sensitive information, logged data should be de-identified where possible and protected through strict access controls and secure storage practices. Regular auditing of logs and monitoring systems is essential for identifying anomalies, validating policy adherence, and detecting emerging vulnerabilities.

The third principle concerns *version control and security hardening*. Stable operation requires systematic management of models, configurations, and deployment environments throughout the system lifecycle. Version control mechanisms should track model revisions and support rapid rollback when issues arise. Updates must undergo structured evaluation, including regression testing, adversarial testing, and validation against known operational requirements before production deployment. In parallel, the surrounding environment should be hardened using secure configuration practices, vulnerability management procedures, regular scanning, and penetration testing.

Deployment architecture also significantly influences the balance between privacy, performance, and operational control. Organisations choose between externally hosted cloud-based services, hybrid deployments, or fully local inference environments. Each model introduces distinct security and governance implications.

Externally hosted systems provide scalability and access to highly capable models without requiring extensive local infrastructure. However, they also increase dependency on third-party providers and reduce direct control over data handling practices. Sensitive information processed through such systems may become subject to external logging, retention policies, or cross-border data transfer requirements.

Hybrid architectures attempt to balance these concerns by separating workloads according to sensitivity. Less sensitive analytical tasks may be processed through external services, while high-risk or regulated information remains within local infrastructure. This approach can improve operational flexibility while reducing exposure

of sensitive datasets.

Fully local deployment provides the greatest degree of control over data residency, monitoring, and system configuration. Such environments are particularly relevant for highly regulated sectors or organisations handling classified information. However, local deployment also introduces significant infrastructure requirements, including GPU resources, maintenance overhead, and specialised expertise for model management and security hardening.

Another important architectural consideration involves API mediation layers. Rather than allowing users to interact directly with the model, requests can be routed through controlled intermediary systems that perform filtering, logging, validation, and policy enforcement. This approach reduces the likelihood of unsafe interactions and enables more consistent governance across organisational workflows.

Deployment architecture is therefore more than an infrastructure choice. The chosen deployment model directly shapes organisational exposure to privacy risks, operational failures, and regulatory obligations.

4.5 Appropriate Use Cases

The proposed framework provides a structured approach for integrating LLMs in cybersecurity. However, not all applications are equally suitable. Some applications tolerate uncertainty better than others. This section explores appropriate use cases for the framework, differentiating between applications where LLMs can effectively augment human expertise and those where their limitations necessitate a more cautious approach, or outright exclusion.

LLMs are particularly well-suited for tasks that involve the processing and transformation of unstructured information. In cybersecurity contexts, this includes activities such as summarising threat intelligence reports and *drafting* documentation. These tasks benefit from the model's ability to synthesise large volumes of text and

present information in a coherent and accessible format. Importantly, they are also relatively low risk, as inaccuracies can typically be identified and corrected through routine human review.

LLMs can also assist analysts when working with incomplete or ambiguous information. For example, they may help analysts identify patterns in textual data and provide contextual explanations of known attack techniques. In these roles, the model functions as a supportive tool that enhances efficiency without assuming responsibility for final decisions.

However, the use of LLMs becomes significantly more problematic in high-risk or decision-critical applications. Tasks such as automated incident response, legal or compliance assessments, and forensic analysis require a high degree of reliability, accountability, and contextual understanding. In these scenarios, even minor inaccuracies may lead to significant operational or legal consequences. Given the limitations discussed earlier, including overconfidence and lack of grounded reasoning, LLMs cannot be relied upon to perform such tasks independently.

For these higher-risk use cases, any integration of LLMs must be accompanied by strict controls, including mandatory human validation and clearly defined boundaries for system autonomy. In some cases, the risks may outweigh the benefits entirely, and the use of LLMs should be avoided. This distinction between suitable and unsuitable applications reinforces the broader principle that *LLMs should augment human expertise rather than replace it*, particularly in security-critical environments.

4.6 Refinement and Considerations

The framework presented in this chapter should be understood as an evolving model rather than a fixed solution. The rapid pace of development in LLMs, combined with the emergence of new threats and regulatory changes, requires continuous reassessment of both risks and controls. As a result, the secure use of LLMs must be

supported by ongoing refinement and adaptation.

One of the central challenges in this context is balancing competing priorities, particularly reliability, privacy, and cost. Enhancing model performance often involves the use of larger or more complex systems, which may increase operational costs and introduce additional data protection concerns. Conversely, prioritising privacy through local deployment or strict data minimisation may limit model capabilities and reduce analytical effectiveness. Organisations must therefore evaluate these trade-offs in relation to their specific operational requirements and risk tolerance.

Continuous monitoring and evaluation are essential for maintaining the effectiveness of the framework. This includes regular audits of system behaviour, validation processes, and data handling practices. Feedback from users, particularly cybersecurity analysts, plays a key role in identifying weaknesses and improving system design. Over time, such feedback can inform the refinement of prompt structures, validation mechanisms, and governance policies.

Another important consideration is the need for organisational alignment. The successful integration of LLMs is not solely a technical challenge, but also involves human and procedural factors. Clear policies, defined responsibilities, and appropriate training are required to ensure that users understand both the capabilities and limitations of the system. Without this alignment, there is a risk of inconsistent use, overreliance, or unintended misuse.

In summary, the framework provides a structured foundation for secure LLM integration, but its effectiveness depends on continuous adaptation and critical evaluation. By incorporating iterative improvements and aligning technical controls with organisational practices, it is possible to maintain a balance between leveraging the benefits of LLMs and managing their inherent risks.

5 Results and Discussion

This chapter synthesises the findings of the thesis in relation to the research questions and evaluates the proposed framework within the broader context of cybersecurity practice. The analysis demonstrates that while LLMs offer clear operational benefits, limitations related to reliability, transparency, and data handling introduce constraints that must be actively managed.

5.1 Suitability of LLMs as Decision-Support Systems

The findings indicate that LLMs are suitable as assistive tools in cybersecurity contexts, but not as authoritative decision-making systems. Their primary value lies in enhancing productivity by supporting tasks such as information retrieval, summarisation, and preliminary analysis. In these roles, LLMs can reduce cognitive load and improve efficiency, particularly when dealing with large volumes of unstructured data.

However, their probabilistic nature and tendency to produce plausible but incorrect outputs limit their suitability for independent decision-making. The analysis in earlier chapters highlights that LLMs optimise for linguistic coherence rather than factual correctness. As a result, outputs may appear credible even when they are inaccurate or incomplete. This creates a risk of overreliance, particularly in envi-

ronments where users may not have sufficient expertise to critically evaluate model responses.

From a security standpoint, LLM outputs are better treated as material for review than as trusted conclusions. The framework proposed in Chapter 4 addresses these challenges through human oversight, output validation, and controlled use cases. These controls align with the broader principle that LLMs should augment, rather than replace, human expertise.

5.2 Evaluation of Reliability in Analyst Roles

When evaluated against the expectations associated with junior-level cybersecurity analyst roles, LLMs exhibit a mixed performance profile. They are effective in structured and low-risk tasks but struggle with activities requiring judgement, uncertainty management, and contextual understanding.

LLMs perform well in tasks such as summarising threat intelligence reports and vulnerability descriptions, providing general explanations of known attack techniques, and identifying patterns or recurring elements in textual data. These capabilities align with entry-level analytical tasks that involve information processing rather than critical decision-making. In this sense, LLMs can function as productivity-enhancing tools that support routine workflows.

However, significant limitations emerge in tasks that require deeper reasoning or contextual understanding. The limitations are directly linked to the structural properties of LLMs discussed in Chapter 2, including prompt sensitivity, difficulty mismatch, and overconfident error generation. Unlike human analysts, who are trained to recognise uncertainty and escalate unclear cases, LLMs tend to provide definitive answers regardless of confidence. This behavioural mismatch increases the risk of incorrect conclusions being accepted without sufficient examination.

LLMs can reproduce parts of junior analyst work, but they do not yet satisfy the

reliability expectations associated with independent analytical roles. Their role is better understood as supportive rather than substitutive. The framework's emphasis on escalation and human-in-the-loop validation directly addresses this gap. Overall, LLMs remain insufficiently reliable for independent operation in security-critical environments. Their role should remain supportive and tightly controlled.

5.3 Trade-offs between Reliability, Privacy, and Cost

A central finding of this thesis is that the deployment of LLMs in production environments involves a fundamental trade-off between reliability, privacy, and cost. These dimensions are interdependent, and optimising one often introduces constraints on the others.

From a reliability perspective, larger and more advanced models generally provide higher-quality outputs. However, these models are often hosted externally, which raises concerns related to data privacy and regulatory compliance. Transferring sensitive data to third-party systems may conflict with data protection requirements, particularly under GDPR.

Conversely, deploying smaller or locally hosted models improves data control and compliance, but may reduce performance and accuracy. This can limit their usefulness in complex analytical tasks, thereby affecting overall system effectiveness.

Cost considerations further complicate this balance. High-performance models typically involve greater computational and financial resources, especially when deployed at scale. Organisations must therefore evaluate whether the marginal gains in reliability justify the increased operational costs and potential privacy risks.

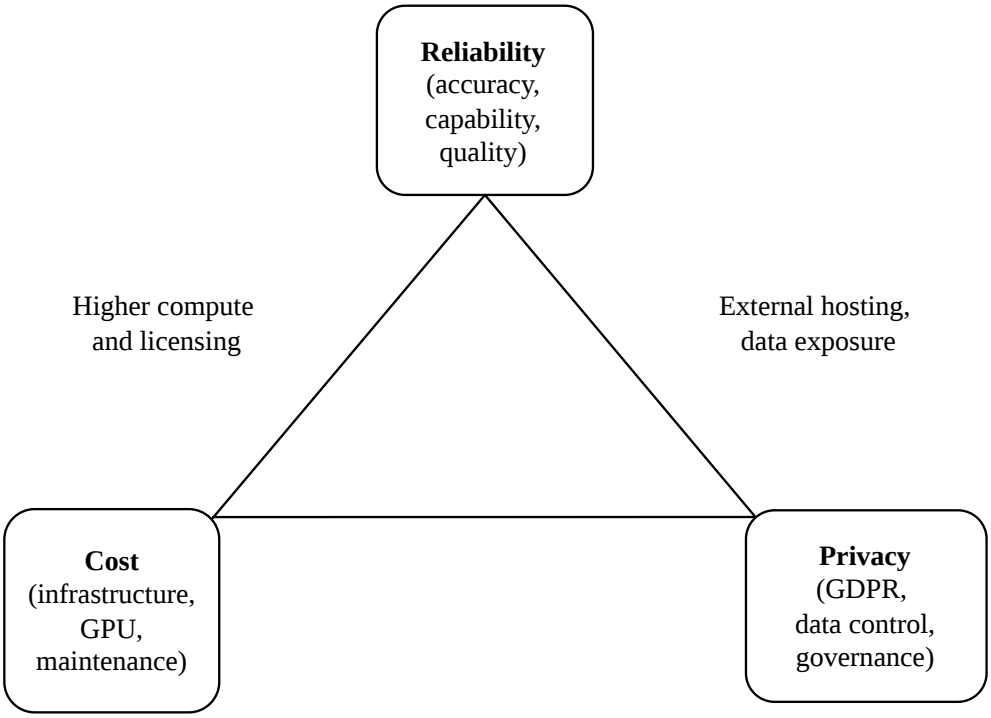

Figure 5.1: Trade-off triangle

These trade-offs can be conceptualised as a triangular relationship presented in Figure 5.1. The first dimension, reliability, is primarily driven by model capability, reasoning performance, and output quality. The second dimension, privacy and compliance, is shaped by deployment architecture, data governance practices, and regulatory obligations. The third dimension, cost, reflects infrastructure requirements, licensing expenses, maintenance overhead, and computational consumption.

The relationship between these dimensions is not strictly linear. Increasing reliability through the use of large commercial models may simultaneously increase compliance complexity and operational expenditure. Similarly, prioritising privacy through fully local deployment may reduce external exposure but require significant investment in infrastructure while accepting lower analytical performance. In prac-

tice, organisations cannot maximise all three dimensions simultaneously and must instead determine an acceptable balance based on their operational context and risk tolerance.

The appropriate balance varies depending on organisational requirements. Highly regulated sectors such as finance, healthcare, or critical infrastructure may prioritise privacy and compliance even at the expense of model performance or increased operational costs. By contrast, organisations with lower regulatory exposure or less sensitive datasets may accept greater use of externally hosted models in exchange for improved reliability and reduced infrastructure burden.

Organisations processing relatively low-sensitivity operational data may prioritise access to highly capable external models in order to maximise analytical performance and scalability. In such cases, the organisation may accept increased dependence on third-party infrastructure because the operational benefits outweigh the associated privacy risks.

By contrast, organisations operating in heavily regulated environments often face substantially different constraints. Healthcare providers, financial institutions, and operators of critical infrastructure may be legally or contractually restricted from transferring sensitive data outside tightly controlled environments. For these organisations, local deployment or highly restricted hybrid architectures may be preferable despite increased operational cost and potentially lower model capability.

Public sector organisations face additional challenges related to transparency, accountability, and procurement. Decisions supported by AI systems may require traceability and explainability beyond what many commercial systems currently provide. Consequently, public institutions may adopt more conservative deployment strategies, prioritising governance and auditability over maximal model performance.

Smaller organisations encounter a different set of constraints. While local deployment may improve privacy and regulatory control, the associated infrastructure

and maintenance requirements may exceed available resources. Such organisations may therefore depend more heavily on externally hosted services despite the resulting governance challenges. In practice, this often creates asymmetry in AI adoption, where larger organisations possess greater capacity to implement privacy-preserving deployments.

The trade-off between cost and reliability is also dynamic rather than static. Advances in model optimisation, open-source systems, and specialised hardware may gradually reduce the computational requirements associated with local deployment. Similarly, regulatory developments may alter the relative attractiveness of different deployment models over time.

These findings suggest that LLM adoption should not be approached as a purely technical optimisation problem. Instead, deployment decisions should be framed as governance decisions involving risk management, legal considerations, and operational priorities. The framework presented in Chapter 4 provides a structured approach to navigating these trade-offs. By prioritising data minimisation, controlled deployment, and appropriate use cases, organisations can achieve a balanced configuration that aligns with both operational needs and regulatory constraints.

5.4 Implications for Practice

The results of this study suggest that successful integration of LLMs into cybersecurity workflows depends less on model capability alone and more on the surrounding governance and system design. Technical performance must be complemented by organisational controls that ensure safe and appropriate use.

First, LLM outputs must be subject to validation before operational use. Treating them as unverified ensures that errors and inconsistencies are caught early. This reframes their role from answer providers to exploratory analytical tools.

Second, data handling practices should align with regulatory requirements such

as GDPR. Sensitive data should be reduced, anonymised, or masked before use, and strict controls should prevent unauthorised access (especially when using external or cloud-based models).

Third, systems should be designed to tolerate and contain errors rather than assume output correctness. Since mistakes are unavoidable, systems must detect and limit their impact. This includes monitoring, logging activity, and having fallback procedures.

Finally, effective user training remains essential. Analysts need to understand both the strengths and weaknesses of LLMs, including their tendency to produce convincing but incorrect answers. Training should focus on critical thinking, proper use, and when to escalate issues.

Overall, the findings suggest that technical safeguards alone are insufficient. Safe deployment depends equally on organisational practice, human oversight, and regulatory constraints.

5.5 Limitations

This thesis has several limitations. First, the thesis adopts a qualitative literature-based methodology and does not include empirical experimentation. Consequently, the findings depend on the scope, quality, and availability of existing literature.

Second, the rapid evolution of LLM technologies presents an inherent limitation. Model capabilities, deployment architectures, and regulatory requirements continue to evolve quickly. Some technical observations or implementation practices may therefore change over time, particularly as models become more capable, more efficient, or subject to new governance requirements.

Third, the proposed framework is conceptual rather than empirically validated. While it is grounded in current literature, governance principles, and cybersecurity practices, its effectiveness has not been tested through organisational deployment,

simulation, or longitudinal evaluation. Practical implementation may reveal constraints or trade-offs not fully captured in the present analysis.

Finally, the thesis focuses primarily on cybersecurity decision-support contexts and does not attempt to generalise findings across all AI deployment domains. Different sectors may operate under distinct threat models, regulatory environments, and reliability expectations.

5.6 Summary

In summary, LLMs offer clear benefits as assistive tools in cybersecurity, particularly in tasks involving information processing and analysis. However, their limitations in reliability and transparency prevent them from functioning as autonomous decision-makers. The trade-offs between reliability, privacy, and cost further constrain their deployment in real-world environments.

The framework proposed in this thesis addresses these challenges by embedding LLM usage within a structured system of controls, emphasising human oversight, data protection, and risk-aware deployment. This approach enables organisations to leverage the advantages of LLMs while mitigating their inherent risks.

6 Conclusion

This thesis examined how LLMs can be used as decision-support tools in cybersecurity environments, with particular attention to reliability, privacy, and regulatory compliance. The analysis identified several limitations associated with LLMs, including probabilistic behaviour, limited explainability, prompt sensitivity, and the tendency to produce plausible but incorrect outputs. These characteristics complicate their use in environments where decisions may carry operational, legal, or security consequences.

At the same time, the findings demonstrate that LLMs provide meaningful practical value when applied to appropriate tasks. Their ability to process large volumes of unstructured information, generate summaries, and support analytical workflows makes them useful tools for cybersecurity operations. However, their usefulness is conditional rather than universal. The benefits of LLMs depend heavily on deployment context, task selection, and the safeguards surrounding their use.

The framework proposed in this thesis addresses these challenges by translating abstract concerns into operational controls. Rather than treating LLMs as trusted decision-makers, the framework positions them as untrusted probabilistic components operating within a layered governance structure. Human oversight, output validation, controlled deployment, and privacy-preserving data handling form the primary mechanisms for reducing risk.

The findings further suggest that deploying LLMs in cybersecurity is not solely a

technical optimisation problem. Decisions regarding model selection, hosting strategy, data processing, and validation mechanisms involve trade-offs between reliability, privacy, and cost. As a result, organisations must approach LLM integration as a governance and risk-management challenge rather than a question of model capability alone.

Overall, LLMs appear most suitable as constrained analytical assistants that support human decision-making without replacing human judgment. Their role in cybersecurity should therefore remain supportive, supervised, and bounded by clear operational and regulatory controls.

6.1 Future Work

A natural next step would be empirical evaluation of the proposed framework in realistic cybersecurity environments. This could include testing LLM performance across common analyst tasks, assessing failure patterns, and measuring the effectiveness of controls such as prompt standardisation, validation pipelines, and human escalation procedures.

Future work could also examine reliability calibration in greater depth. While this thesis discusses overconfidence and uncertainty, additional research is needed on mechanisms that help LLM systems communicate confidence more appropriately and support safer analyst interaction.

Another promising direction concerns deployment trade-offs. Comparative studies between locally hosted and externally hosted models could provide clearer evidence regarding the balance between privacy, compliance, operational cost, and analytical capability in production environments.

The increasing autonomy of LLM-based systems also warrants further investigation. Agentic architectures, tool-using systems, and multi-step automated workflows introduce new governance and accountability challenges that extend beyond tradi-

tional chatbot-style interaction models. Understanding how these systems can be constrained, monitored, and audited in cybersecurity contexts remains an important research problem.

Finally, future research should continue to explore how regulatory frameworks such as GDPR and the EU AI Act interact with rapidly evolving AI technologies. As technical capabilities change, governance models must be reassessed to ensure that legal requirements, operational practices, and system behaviour remain aligned.

References

[1] A. McGrath and A. Jonker, *What Is AI Safety?*, Nov. 2024. Accessed: Feb. 13, 2026. [Online]. Available: https://www.ibm.com/think/topics/ai-safety.

[2] A. Jonker, A. Gomstyn, and A. McGrath, *What Is AI Transparency?*, Sep. 2024. Accessed: Feb. 17, 2026. [Online]. Available: https://www.ibm.com/think/topics/ai-transparency.

[3] A. McGrath and A. Jonker, *What Is AI Interpretability?* Accessed: Feb. 17, 2026. [Online]. Available: https://www.ibm.com/think/topics/interpretability.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. u. Kaiser, and I. Polosukhin, "Attention is All you Need", in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30, Curran Associates, Inc., 2017.

[5] S. Baron, "Trust, Explainability and AI", *Philosophy & Technology*, vol. 38, no. 1, p. 4, Jan. 2025, ISSN: 2210-5441. DOI: 10.1007/s13347-024-00837-6.

[6] S. Blanco, "Human trust in AI: A relationship beyond reliance", *AI and Ethics*, vol. 5, no. 4, pp. 4167–4180, Aug. 2025, ISSN: 2730-5961. DOI: 10.1007/s43681-025-00690-z.

[7] O. Freiman, "Making sense of the conceptual nonsense 'trustworthy AI'", *AI and Ethics*, vol. 3, no. 4, pp. 1351–1360, Nov. 2023, ISSN: 2730-5953, 2730-5961. DOI: 10.1007/s43681-022-00241-w.

[8] A. Ferrario and M. Loi, "How Explainability Contributes to Trust in AI", in *2022 ACM Conference on Fairness Accountability and Transparency*, Seoul Republic of Korea: ACM, Jun. 2022, pp. 1457–1466, ISBN: 978-1-4503-9352-2. DOI: 10.1145/3531146.3533202.

[9] S. Rose, O. Borchert, S. Mitchell, and S. Connelly, "Zero Trust Architecture", Aug. 2020. DOI: 10.6028/NIST.SP.800-207.

[10] R. Tomsett, A. Preece, D. Braines, F. Cerutti, S. Chakraborty, M. Srivastava, G. Pearson, and L. Kaplan, "Rapid Trust Calibration through Interpretable and Un-certainty-Aware AI", *Patterns*, vol. 1, no. 4, Jul. 2020, ISSN: 2666-3899. DOI: 10.1016/j.patter.2020.100049.

[11] P. Laplante and J. Voas, "Zero-Trust Artificial Intelligence?", *Computer*, vol. 55, no. 2, pp. 10–12, Feb. 2022, ISSN: 1558-0814. DOI: 10.1109/MC.2021.3126526.

[12] V. Leppänen, S. Rauti, K. Rindell, and J. Holvitie, "Cyber security and securing data transfer in the development and operation of autonomous vessels",

[13] *Cambridge English Dictionary: "Reliability"*. Accessed: Feb. 23, 2026. [Online]. Available: https://dictionary.cambridge.org/dictionary/english/reliability.

[14] H. F. Atlam, "LLMs in Cyber Security: Bridging Practice and Education", *Big Data and Cognitive Computing*, vol. 9, no. 7, p. 184, Jul. 2025, Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 2504-2289. DOI: 10.3390/bdcc9070184.

[15] L. Zhou, W. Schellaert, F. Martínez-Plumed, Y. Moros-Daval, C. Ferri, and J. Hernández-Orallo, "Larger and more instructable language models become less reliable", *Nature*, vol. 634, no. 8032, pp. 61–68, Oct. 2024, ISSN: 1476-4687. DOI: 10.1038/s41586-024-07930-y.

[16] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu, "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions", *ACM Trans. Inf. Syst.*, vol. 43, no. 2, 42:1–42:55, Jan. 2025, ISSN: 1046-8188. DOI: 10.1145/3703155.

[17] B. Cao, D. Cai, Z. Zhang, Y. Zou, and W. Lam, "On the Worst Prompt Performance of Large Language Models", Oct. 2024. DOI: 10.48550/arXiv.2406.10248.

[18] A. Razavi, M. Soltangheis, N. Arabzadeh, S. Salamat, M. Zihayat, and E. Bagheri, "Benchmarking Prompt Sensitivity in Large Language Models", Feb. 2025. DOI: 10.48550/arXiv.2502.06065.

[19] W. Knight, *The Dark Secret at the Heart of AI.* Accessed: Jan. 22, 2026. [Online]. Available: https://www.technologyreview.com/2017/04/11/5113/the-dark-secret-at-the-heart-of-ai/.

[20] M. A. Ferrag, F. Alwahedi, A. Battah, B. Cherif, A. Mechri, N. Tihanyi, T. Bisztray, and M. Debbah, "Generative AI in cybersecurity: A comprehensive review of LLM applications and vulnerabilities",

Internet of Things and Cyber-Physical Systems, vol. 5, pp. 1–46, Jan. 2025, ISSN: 2667-3452. DOI: 10.1016/j.iotcps.2025.01.001.

[21] Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun, and Y. Zhang, "A survey on large language model (LLM) security and privacy: The Good, The Bad, and The Ugly", *High-Confidence Computing*, vol. 4, no. 2, Jun. 2024, ISSN: 2667-2952. DOI: 10.1016/j.hcc.2024.100211.

[22] B. Yan, K. Li, M. Xu, Y. Dong, Y. Zhang, Z. Ren, and X. Cheng, "On protecting the data privacy of Large Language Models (LLMs) and LLM agents: A literature review",
High-Confidence Computing, vol. 5, no. 2, p. 100 300, Jun. 2025, ISSN: 26672952. DOI: 10.1016/j.hcc.2025.100300.

[23] T. Krantz, A. Jonker, and A. McGrath, *What Is Shadow AI?* Accessed: Feb. 17, 2026. [Online]. Available: https://www.ibm.com/think/topics/shadow-ai.

[24] Z. Liu, L. Yu, T. Dong, Y. Meng, S. Li, G. Chen, and H. Zhu, "Inferring Activities and Profiles of Users Based on Trajectory Leakage in Mobile Ad Network", in *2024 20th International Conference on Mobility, Sensing and Networking (MSN)*, ISSN: 2994-3523, Dec. 2024, pp. 545–552. DOI: 10.1109/MSN63567.2024.00080.

[25] X. Ma, F. Ding, K. Peng, Y. Yang, and C. Wang, "CP-Link: Exploiting Continuous Spatio-Temporal Check-In Patterns for User Identity Linkage", *IEEE Transactions on Mobile Computing*, vol. 22, no. 8, pp. 4594–4606, Aug. 2023, ISSN: 1558-0660. DOI: 10.1109/TMC.2022.3157292.

[26] H. Schultz and W. Freedman, "Plugins and POPI: A Critical Discussion into the Legal Implications of Social Plugins and the Protection of Personal

Information", *Potchefstroom Electronic Law Journal*, vol. 27, Mar. 2024, ISSN: 1727-3781. DOI: 10.17159/1727-3781/2023/v26i0a15758.

[27] E. Purificato, L. Boratto, and E. W. D. Luca, *User Modeling and User Profiling: A Comprehensive Survey*, Feb. 2024. DOI: 10.48550/arXiv.2402.09660.

[28] H. Shen, Z. Gu, H. Hong, and W. Han, *PII-Bench: Evaluating Query-Aware Privacy Protection Systems*, Feb. 2026. DOI: 10.48550/arXiv.2502.18545.

[29] M. Windl, R. Amberg, and T. Kosch, "The Illusion of Privacy: Investigating User Misperceptions in Browser Tracking Protection", in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, ser. CHI '25, New York, NY, USA: Association for Computing Machinery, Apr. 2025, ISBN: 979-8-4007-1394-1. DOI: 10.1145/3706598.3713912.

[30] N. Lukas, A. Salem, R. Sim, S. Tople, L. Wutschitz, and S. Zanella-Béguelin, *Analyzing Leakage of Personally Identifiable Information in Language Models*, Apr. 2023. DOI: 10.48550/arXiv.2302.00539.

[31] M. R. Sanfilippo, N. Apthorpe, K. Brehm, and Y. Shvartzshnaider, "Privacy governance not included: Analysis of third parties in learning management systems", *Information and Learning Sciences*, vol. 124, no. 9-10, pp. 326–348, Sep. 2023, ISSN: 2398-5348. DOI: 10.1108/ILS-04-2023-0033.

[32] A. Reyes, *Keeping secrets out of logs*. Accessed: Feb. 13, 2026. [Online]. Available: https://slideslive.com/39021794/keeping-secrets-out-of-logs?ref=og-meta-tags.

[33] D. A. Alber, Z. Yang, A. Alyakin, E. Yang, S. Rai, A. A. Valliani, J. Zhang, G. R. Rosenbaum, A. K. Amend-Thomas, D. B. Kurland, C. M. Kremer,

A. Eremiev, B. Negash, D. D. Wiggan, M. A. Nakatsuka, K. L. Sangwon, S. N. Neifert, H. A. Khan, A. V. Save, A. Palla, E. A. Grin, M. Hedman, M. Nasir-Moin, X. C. Liu, L. Y. Jiang, M. A. Mankowski, D. L. Segev, Y. Aphinyanaphongs, H. A. Riina, J. G. Golfinos, D. A. Orringer, D. Kondziolka, and E. K. Oermann, "Medical large language models are vulnerable to data-poisoning attacks", *Nature Medicine*, vol. 31, no. 2, pp. 618–626, Feb. 2025, ISSN: 1546-170X. DOI: 10.1038/s41591-024-03445-1.

[34] A. Souly, J. Rando, E. Chapman, X. Davies, B. Hasircioglu, E. Shereen, C. Mougan, V. Mavroudis, E. Jones, C. Hicks, N. Carlini, Y. Gal, and R. Kirk, *Poisoning Attacks on LLMs Require a Near-constant Number of Poison Samples*, Oct. 2025. DOI: 10.48550/arXiv.2510.07192.

[35] I. K. Egbuna, H. M. Olayinka, A. B. Tijani, S. H. Aliyu, A. O. Felix, O. A. Elijah, and M. A. Alabi, "Training Time Vulnerabilities in Large Language Models: Data Poisoning and Backdoor Attacks", *International Journal of Future Engineering Innovations*, vol. 2, no. 3, pp. 60–68, 2025, ISSN: 30491215. DOI: 10.54660/IJFEI.2025.2.3.60-68.

[36] A. Wan, E. Wallace, S. Shen, and D. Klein, "Poisoning Language Models During Instruction Tuning", May 2023. DOI: 10.48550/arXiv.2305.00944.

[37] T. Fu, M. Sharma, P. Torr, S. B. Cohen, D. Krueger, and F. Barez, "PoisonBench: Assessing Large Language Model Vulnerability to Data Poisoning", Jun. 2025. DOI: 10.48550/arXiv.2410.08811.

[38] A. McGrath and A. Jonker, *What Is Risk Management?*, Jan. 2025. Accessed: Feb. 17, 2026. [Online]. Available: https://www.ibm.com/think/topics/risk-management.

[39] European Parliament,
General Data Protection Regulation (GDPR) – Legal Text.
Accessed: Mar. 17, 2026. [Online]. Available: https://gdpr-info.eu/.

[40] B. Wolford, *What is GDPR, the EU's new data protection law?*, Nov. 2018.
Accessed: Mar. 17, 2026. [Online]. Available:
https://gdpr.eu/what-is-gdpr/.

[41] European Parliament,
General Data Protection Regulation (GDPR) Compliance Guidelines.
Accessed: Mar. 17, 2026. [Online]. Available: https://gdpr.eu/.

[42] G. Sartor and F. Lagioia, *The impact of the General Data Protection Regulation (GDPR) on artificial intelligence.*
Publications Office of the European Union, 2020. DOI: 10.2861/293.

[43] National Institute of Standards and Technology,
"The NIST Cybersecurity Framework (CSF) 2.0",
National Institute of Standards and Technology, Gaithersburg, MD,
Tech. Rep. NIST CSWP 29, Feb. 2024, NIST CSWP 29.
DOI: 10.6028/NIST.CSWP.29.

[44] V. Tiple, *Recommendations on the European Commission's WHITE PAPER on Artificial Intelligence - A European approach to excellence and trust*,
SSRN Scholarly Paper, Rochester, NY, Jun. 2020.
DOI: 10.2139/ssrn.3706099.

[45] European Parliament, *EU AI Act: first regulation on artificial intelligence*,
Jun. 2023. Accessed: Apr. 16, 2026. [Online]. Available: https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence.

[46] M. Kosinski and M. Scapicchio, *What is the EU AI Act? | IBM*, Sep. 2024. Accessed: Apr. 16, 2026. [Online]. Available: https://www.ibm.com/think/topics/eu-ai-act.

[47] M. Leppänen, *How continuous resilience transforms cybersecurity*, Oct. 2025. [Online]. Available: https://cybersecuritynordic.messukeskus.com/continuous-resilience-in-cybersecurity/.

[48] U. K. Agarwal, A. Chan, and K. Pattabiraman, "Resilience assessment of large language models under transient hardware faults", in *2023 IEEE 34th International Symposium on Software Reliability Engineering (ISSRE)*, 2023, pp. 659–670. DOI: 10.1109/ISSRE59848.2023.00052.

[49] Y. Sun, Z. Coalson, S. Chen, H. Liu, Z. Zhang, S. Hong, B. Fang, and L. Yang, "Demystifying the Resilience of Large Language Model Inference: An End-to-End Perspective", Nov. 2025, pp. 1127–1144. DOI: 10.1145/3712285.3759803.

[50] L. Myllyaho, M. Raatikainen, T. Männistö, T. Mikkonen, and J. Nurminen, "Systematic Literature Review of Validation Methods for AI Systems", 2021. DOI: 10.1016/j.jss.2021.111050.

[51] J. Perolat, B. De Vylder, D. Hennes, E. Tarassov, F. Strub, V. de Boer, P. Muller, J. T. Connor, N. Burch, T. Anthony, S. McAleer, R. Elie, S. H. Cen, Z. Wang, A. Gruslys, A. Malysheva, M. Khan, S. Ozair, F. Timbers, T. Pohlen, T. Eccles, M. Rowland, M. Lanctot, J.-B. Lespiau, B. Piot, S. Omidshafiei, E. Lockhart, L. Sifre, N. Beauguerlange, R. Munos, D. Silver, S. Singh, D. Hassabis, and K. Tuyls, "Mastering the game of Stratego with model-free multiagent reinforcement learning",

Science, vol. 378, no. 6623, pp. 990–996, Dec. 2022.

DOI: 10.1126/science.add4679.