

Leiden University

ICT in Business and the Public Sector

Leveraging Large Language Models to Support Cybersecurity Risk Estimation

Name:	Hafizh Haidar
Student ID:	s4150392

Date:	June 2, 2026

1st supervisor:	Prof. Dr. Ir. J.M.W. Visser
2nd supervisor:	Dr. C.J. Stettina MSc
Company supervisor:	B. Blaauwendraad

MASTER'S THESIS

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Einsteinweg 55
2333 CC Leiden
The Netherlands

Abstract

Background: Cyber risk assessments in consulting practice rely heavily on consultant judgement when translating organisational and policy documents, interview findings, and gap assessment results into impact and likelihood scores under non-prescriptive cybersecurity standards. Although these standards provide structured phases, they intentionally avoid prescribing detailed scoring methodologies, leaving much of the scoring process dependent on consultant interpretation. As a result, assessment outcomes may vary across consultants and engagements, and the reasoning behind scoring decisions often remains implicit and insufficiently documented. While Large Language Models (LLMs) and retrieval-augmented generation (RAG) approaches show potential for supporting automated analysis, their role in judgement-intensive risk scoring under non-prescriptive standards remains underexplored.

Aim: This research investigates how AI-assisted decision support can be designed and evaluated to improve consistency, transparency, and reproducibility of the judgement-intensive cyber risk scoring process in consulting practice, while preserving human judgement and accountability.

Method: This research follows the Design Science Research Methodology (DSRM). Problem understanding and solution objectives are derived from observed challenges in interpreting and documenting risk scoring decisions, validated through stakeholder interviews. Based on these objectives, this research designs and develops the Cyber Risk Estimation Support Tool (CREST), an AI-assisted decision-support artifact that combines an Excel-centred assessment workflow, a Python backend, retrieval-augmented generation, and LLM-assisted reasoning to support evidence-based risk scoring. CREST is evaluated through a multi-stage strategy consisting of a configuration comparison, controlled case evaluation, and consultant evaluation using TAM, SUS, and UEQ. The study also reflects on socio-technical risks such as over-reliance and shifts in professional judgement practices.

Results: The best-performing configuration was qwen3:30B with $k = 10$. In the controlled case evaluation, CREST achieved 77% scoring accuracy and an average traceability score of 60%, with all incorrect predictions underestimating the predefined risk level. Consultants perceived CREST positively across TAM, SUS, and UEQ, and valued it as a structured starting point for risk analysis, while highlighting the need for stronger traceability, human validation, clearer business-facing outputs, and better workflow integration.

Conclusion: This thesis demonstrates that AI-assisted decision support can support the cyber risk scoring process in consulting practice by helping consultants organise and retrieve relevant evidence, structure factor-level scoring reasoning, and make suggested scores more visible and reviewable. CREST does not replace consultant judgement, but provides a structured starting point for evidence interpretation and scoring justification. Further development should focus on improving traceability, reducing risk underestimation, and integrating the tool into real consulting workflows.

Contents

Abstract . . . 1

1 Introduction **3**
1.1 Problem statement . . . 3
1.2 Research questions . . . 4
1.3 Overview of the thesis . . . 5

2 Background and related work **6**
2.1 Cyber risk assessment . . . 6
2.1.1 Cyber risk assessment process . . . 6
2.1.2 Cyber risk assessment standards and guidance . . . 7
2.2 Risk scoring in cyber risk assessment . . . 8
2.2.1 Likelihood, impact, and risk scoring . . . 8
2.2.2 Judgement in risk scoring and associated challenges . . . 9
2.3 Decision support for judgement-intensive cyber risk assessment . . . 10
2.4 AI-assisted analysis and LLM applications . . . 10
2.4.1 Existing approaches . . . 11
2.4.2 Current Limitations . . . 11
2.5 Summary and research gap . . . 12

3 Method **14**
3.1 Research approach . . . 14
3.2 Problem identification and objective definition . . . 15
3.2.1 Literature review . . . 15
3.2.2 Stakeholder interviews . . . 15
3.3 Artifact design and development approach . . . 15
3.3.1 Research setting . . . 16
3.3.2 Artifact type . . . 16
3.3.3 Design requirements . . . 16
3.4 Demonstration . . . 16
3.4.1 Case construction and validation . . . 16
3.4.2 Case validation . . . 17
3.4.3 Data used for the demonstrator . . . 17
3.5 Evaluation method . . . 17
3.5.1 Evaluation strategy . . . 17
3.5.2 Participants . . . 18
3.5.3 Evaluation procedure . . . 18
3.5.4 Experimental setup for configuration comparison . . . 19
3.5.5 Evaluation metrics . . . 20
3.5.6 Socio-technical considerations . . . 23

4 Design **24**
4.1 Solution objectives . . . 24
4.2 Decision-support workflow . . . 24
4.2.1 Assessment model and scoring dimensions . . . 25
4.2.2 Data processing and extraction . . . 26
4.2.3 Likelihood and impact level assessment . . . 27

4.2.4 Risk scoring . . . 28
4.3 CREST architecture . . . 29
4.3.1 User interface . . . 29
4.3.2 Workbook structure . . . 30
4.3.3 Evidence retrieval architecture . . . 30
4.3.4 LLM-assisted reasoning and explanation generation . . . 31

5 Evaluation **33**
5.1 Configuration comparison results . . . 33
5.1.1 Comparison of tested combinations . . . 33
5.1.2 Selection of the final configuration . . . 35
5.2 Controlled case evaluation results . . . 35
5.2.1 Accuracy findings . . . 35
5.2.2 Traceability findings . . . 36
5.3 Consultant evaluation results . . . 37
5.3.1 Technology acceptance model . . . 37
5.3.2 System usability scale . . . 39
5.3.3 User experience questionnaire . . . 40
5.3.4 Consultant feedback results . . . 42

6 Discussion **45**
6.1 Interpretation of evaluation findings . . . 45
6.1.1 Configuration comparison and technical performance . . . 45
6.1.2 Controlled case evaluation . . . 47
6.1.3 Consultant evaluation findings . . . 48
6.2 Implications for cyber risk assessment practice . . . 49
6.2.1 Support for consistency and structured reasoning . . . 49
6.2.2 Traceability and reviewability in consulting practice . . . 50
6.2.3 Role of CREST as decision support rather than decision maker . . . 51
6.3 Limitations . . . 51
6.3.1 Evaluation scope and sample size . . . 51
6.3.2 Ground truth and scoring subjectivity . . . 52
6.3.3 Prototype and data constraints . . . 52

7 Conclusions **54**
7.1 Answers to the research questions . . . 54
7.2 Future work . . . 56
7.2.1 Performance optimisation . . . 56
7.2.2 Adoption and integration . . . 57

A Questionnaire item-level results **62**
A.1 TAM item-level results . . . 62
A.2 SUS item-level results . . . 63
A.3 UEQ item-level results . . . 64

Chapter 1

Introduction

In this chapter, we introduce the research context and problem addressed in this research. The research focuses on cyber risk assessment in consulting practice, where the translation of diverse assessment evidence into likelihood and impact scores remains highly reliant on consultant judgement. This creates challenges for process consistency, transparency, and reproducibility, particularly under flexible, non-prescriptive cybersecurity standards. Against this background, this chapter motivates the study and defines the research focus of the thesis.

The chapter is structured around three main parts. First, it presents the problem statement by outlining the practical challenges of judgement-intensive cyber risk scoring and the structural gap between flexible standards and the need for more explicit and defensible operationalisation in practice. Second, it formulates the research questions that guide the study. Finally, it provides an overview of the structure of the thesis.

1.1 Problem statement

The fast advancement of technology and the large-scale use of digital systems have fundamentally changed how businesses operate by improving efficiency and communication. However, this transformation has also made businesses more vulnerable, as it has become easier for malicious threat actors to compromise sensitive data by exploiting both human and information system vulnerabilities. Successful cyber attacks may have significant impact, including financial loss, reputational damage, non-compliance risks, privacy breaches, and regulatory fines. This has made cybersecurity one of the key priorities in enterprise risk management [27, 20].

Cyber risk assessments are crucial processes that help businesses and organisations identify, analyse, evaluate, and prioritise potential risks in their information systems and assets [15]. The risk analysis phase involves the assessment of likelihood and impact of the identified risks to determine the level of risk faced by the organisations. By conducting cyber risk assessments, organisations can make informed decisions about resource allocation, control prioritisation, and risk mitigation strategy to protect their information systems and assets [27].

In many cases, organisations therefore rely on external consultants to provide a more independent perspective during risk assessment. Ensuring objectivity in cyber risk assessment is crucial, as subjective judgements or organisational biases may influence how risks are evaluated. Research in IT risk management highlights the importance of structured approaches that enhance objectivity in asset valuation and risk evaluation processes [21]. In practice, independent perspectives, such as those provided by external experts or consultants, can help mitigate internal biases and support more transparent and defensible risk assessments.

In consulting practice, however, cyber risk assessments are inherently judgement-intensive [12, 22]. These assessments are often perceived as subjective, making them difficult to justify, repeat, and compare across consultants and projects. Although widely recognised standards provide structured phases and terminology [14, 13, 15], the operationalisation of impact and likelihood scoring remains mainly dependent on contextual interpretation and reasoning. In practice, consultants must translate interview findings,

documentation, and control assessments into defensible impact and likelihood scores under time and resource constraints.

While the overall assessment structure is often similar (i.e., resulting in a risk level obtained from combining impact and likelihood), the approaches used to obtain, interpret, and justify these scores may differ across consultants, engagements, and organisational contexts [12, 22]. Judgements are influenced not only by organisational factors, but also by individual consultant experience, reasoning styles, and implicit knowledge, which may not always be explicitly documented or standardised across consultants and engagements.

As a result, risk assessment outcomes may be difficult to reproduce, to compare across projects, to consistently validate internally, or to clearly justify to external stakeholders [12, 3]. Justifications for impact and likelihood scores are often distributed across documents, interview notes, and the consultant's reasoning process, rather than being captured in a transparent and structured manner. This limits verifiability, explainability, and consistency, even when assessments are conducted under the same overarching framework.

Most importantly, this variability is not only organisational but also reflects a structural characteristic of non-prescriptive, goal-based cybersecurity standards, which intentionally avoid prescribing detailed scoring methodologies [24, 22]. While such flexibility enables adaptation to diverse contexts, it also creates interpretative space that may result in implicit reasoning and non-reproducible outcomes.

At the same time, fully standardising or automating impact and likelihood determination is neither feasible nor desirable, as effective risk assessment requires contextual understanding and professional accountability. This creates a structural gap between flexible standards and the need for consistent and explainable operationalisation in practice. This research addresses this gap by designing and evaluating the Cyber Risk Estimation Support Tool (CREST), an AI-assisted decision-support approach for cyber risk assessment in consulting practice. CREST is designed to structure evidence use and judgement logic while preserving flexibility, human judgement, and decision authority.

1.2 Research questions

The objective of this research is to investigate how AI-assisted decision support can support judgement-intensive cyber risk assessment in consulting practice. Specifically, the study focuses on improving the consistency, transparency, and reproducibility of the impact and likelihood scoring process conducted under non-prescriptive cybersecurity standards. To achieve this, the research examines how cyber risk assessments are currently operationalised in practice, how consultants interpret and translate evidence into scoring decisions, and how decision-support mechanisms can help structure reasoning and documentation without replacing human judgement. Based on this objective, the following research question is formulated:

RQ *How can AI-assisted decision support be designed and evaluated to improve consistency, transparency, and reproducibility of the judgement-intensive risk scoring process under non-prescriptive cybersecurity standards, while preserving human judgement and accountability?*

To answer the main research question, we have divided it into several sub-questions:

RQ.1 *How are cyber risk assessments currently operationalised in practice under flexible standards, and at which points do judgement-intensive scoring decisions occur?*

This question aims to understand how cyber risk assessments are currently performed in consulting practice and identify where judgement-intensive scoring decisions occur within the assessment workflow.

RQ.2 *How are different types of data and assessment inputs collected, interpreted, and translated into impact and likelihood scoring?*

This question investigates how consultants gather and interpret various forms of evidence and translate them into structured risk scoring decisions.

RQ.3 *How is judgement logic currently applied in practice, and to what extent does it remain implicit or insufficiently documented?*

This question examines how consultants currently apply reasoning when mapping assessment evidence and controls to risk scoring outcomes and to what extent this reasoning remains implicit.

RQ.4 *What requirements are needed to improve consistency, transparency, and explainability of risk scoring outcomes in consulting practice?*

This question identifies the functional and practical requirements needed to support more consistent, transparent and explainable risk scoring processes.

RQ.5 *How can AI-assisted decision support structure and document reasoning processes, interpretation consistency, and justification of scoring decisions without automating final judgement?*

This question addresses the design of the proposed decision-support approach and explores how AI can assist consultants in structuring and documenting judgement processes while preserving human decision authority.

1.3 Overview of the thesis

This thesis is structured as follows.

In Chapter 2, we discuss the background and related work relevant to this research. It introduces key concepts in cyber risk assessment, risk scoring, judgement-intensive assessment practices, decision support, and AI-assisted analysis, and concludes by identifying the research gap addressed in this research.

Chapter 3 describes the research methodology adopted in this thesis. It explains the Design Science Research Methodology (DSRM) approach, the research setting, the procedures used for requirement analysis, and the design of the demonstration and evaluation activities.

Chapter 4 presents the design of the Cyber Risk Estimation Support Tool (CREST), the AI-assisted decision-support artifact developed in this research. It introduces the solution objectives, describes the decision-support workflow, and explains CREST's architecture.

Chapter 5 evaluates CREST's technical performance and practical applicability. It presents the configuration comparison across different LLM models and retrieval-depth combinations, followed by the controlled case evaluation results, and consultant evaluation results.

Chapter 6 discusses the evaluation findings, reflecting on the effectiveness of the approach, its implications for cyber risk assessment practice, and the limitations of the research.

Finally, Chapter 7 concludes the thesis by summarising the main findings and contributions of the study and outlining directions for future work.

Chapter 2

Background and related work

In this chapter, we review the background and related work relevant to this research. This chapter discusses the broader literature on cyber risk assessment, with particular focus on risk scoring, judgement-intensive assessment practices, decision support, and AI-assisted analysis. This review is necessary because the present study addresses a practical problem that sits at the intersection of cyber risk assessment, human judgement, and emerging AI-based support.

The chapter is divided into five main sections. First, it introduces cyber risk assessment and the main standards and guidance used in the assessment process. Second, it discusses likelihood, impact, and risk scoring, including the role of judgement and the challenges of applying scoring in practice. Third, it introduces decision support as a response to these challenges. Fourth, it reviews AI-assisted approaches and LLM applications relevant to cybersecurity assessment. Finally, the chapter summarises these discussions to identify the research gap that motivates the design of the proposed artifact.

2.1 Cyber risk assessment

Risk assessment is a critical process for identifying, analysing, and evaluating risks that may affect an organisation's objectives. As digital systems become increasingly complex and interconnected, new issues arise regarding their safe and secure operation [2]. In this context, risk assessment has become a fundamental part of cybersecurity, as it supports the identification of relevant threats and vulnerabilities and helps organisations incorporate appropriate security controls into their system design and operation, therefore contributing to secure-by-design and cyber resilience [2].

In the cybersecurity domain, cyber risk assessment can be understood as part of broader risk management, where risks affecting digital information, systems, and related assets are identified, analysed, and evaluated in order to support decision-making [14, 15]. Overall, these perspectives show that cyber risk assessment is not only a technical analysis activity, but also a structured decision-support process within organisational risk management. The following subsections first outline the general cyber risk assessment process and then discuss the main standards that structure this process in practice.

2.1.1 Cyber risk assessment process

Although cyber risk assessment is discussed in different standards and frameworks, the overall process mainly follows a similar structure. Across standards, risk assessment is usually a part of a wider risk management cycle and includes several recurring activities, such as establishing the context of the assessment, identifying risks, analysing their significance, evaluating their acceptability, and determining appropriate treatment actions [14, 15, 13]. In the cybersecurity domain, these activities are applied to risks affecting digital information, systems, networks, and related organisational assets [15].

A first common step is context establishment, in which the scope, objectives, stakeholders, decision criteria, and risk acceptance thresholds are defined [14, 15]. This is followed by risk identification, where relevant risk scenarios are identified and described in terms of assets, threats, vulnerabilities, and possible adverse events [15, 13]. Next, in the risk analysis phase, the identified risks are examined in more detail by assessing their likelihood and impact, often while taking existing controls or countermeasures into

account [15, 14, 13]. The results are then used in the risk evaluation phase, in which analysed risks are compared against risk appetite and predefined criteria in order to determine whether they are acceptable, prioritised, or require treatment [14, 15]. Finally, in risk treatment, organisations decide how to respond to unacceptable risks, for example by reducing, avoiding, transferring, or accepting them [14, 15].

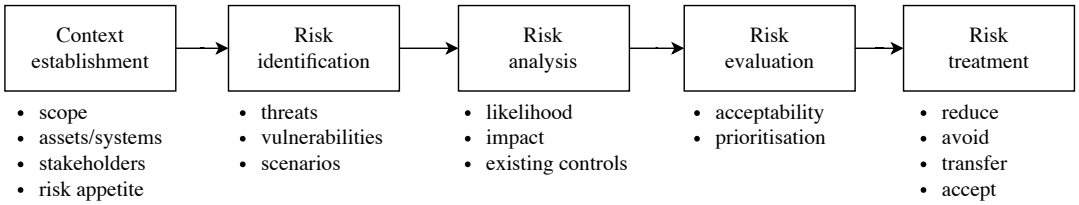


Figure 2.1: General cyber risk assessment process as reflected across major standards and guidance.

This general structure shown in Figure 2.1 can be observed across the main standards relevant to this thesis, although the terminology, scope, and level of operational detail differ across frameworks [14, 15, 13]. The following subsection discusses these standards in more detail.

For the purpose of this thesis, the risk analysis phase is particularly important. It is in this phase that available evidence must be interpreted and translated into likelihood and impact assessments, which then form the basis for the risk evaluation and treatment decisions [15, 13]. As a result, although cyber risk assessment is embedded in a broader management process, the analysis step becomes especially relevant for understanding where judgement-intensive scoring decisions happen in practice.

2.1.2 Cyber risk assessment standards and guidance

Several standards, regulatory frameworks, and guidance documents structure how cyber risk assessments are conducted in practice. Cybersecurity standards also differ in scope, focus area, and intended application domain, which can make standard selection and alignment challenging in practice [7]. Although they share a broadly similar process logic, they differ in scope, terminology, domain specificity, and the extent to which they prescribe operational assessment activities. These differences are important because they influence how organisations and consultants interpret risks, define assessment criteria, and translate evidence into risk-related judgements [14, 15, 13, 16, 9, 10].

ISO 31000. ISO 31000 provides the broadest framing of risk management. It offers general principles and process guidance that can be applied across sectors and risk domains, but it does not focus specifically on cybersecurity or prescribe detailed methods for determining likelihood and impact [14]. As a result, ISO 31000 is useful for establishing the overall logic of risk management, but it does not provide much detail on how cyber risks should be assessed in practice.

ISO/IEC 27005. ISO/IEC 27005 builds on this broader foundation by adapting risk management more explicitly to the information security domain. Compared with ISO 31000, it focuses more on threats, vulnerabilities, consequences, likelihood, and the role of existing controls when assessing information security risks [15]. At the same time, it remains relatively flexible in how organisations define scoring criteria and assessment methods. This makes ISO/IEC 27005 particularly relevant to this thesis, as it provides a structured cyber risk assessment process while still leaving room for interpretative judgement in operational scoring.

IEC 62443-3-2. IEC 62443-3-2 follows a similar flow, but provides a more detailed and domain-specific operational process for industrial automation and control systems. It introduces concepts such as zones and conduits, distinguishes between unmitigated and residual risk, and links risk assessment outcomes to target security levels and tolerable risk [13]. Compared with ISO 31000 and ISO/IEC 27005, IEC 62443-3-2 therefore provides a more prescriptive and domain-specific approach, especially for operational technology environments.

NIST SP 800-30. In addition to these standards, NIST Special Publication 800-30 provides guidance for conducting information security risk assessments. NIST positions risk assessment as a key component of an organisation-wide risk management process and describes it as supporting the identification of threats, vulnerabilities, impact, and likelihood in order to determine risk [16]. It also highlights that

organisations may use different risk models, assessment approaches, and analysis approaches depending on their context, and explicitly allows flexibility in the methodology, formality, and level of detail applied in practice [16]. This makes NIST SP 800-30 particularly useful as a complementary guidance source, especially because it provides a clearer conceptual discussion of likelihood, impact, risk models, and assessment approaches than many high-level standards.

NIS2 and ENISA guidance. In European contexts, cyber risk assessment practices are also increasingly influenced by the NIS2 Directive and supporting guidance published by ENISA. While the NIS2 Directive establishes broad cybersecurity risk-management obligations for essential and important entities, it does not prescribe a detailed operational risk assessment methodology [9]. More recently, ENISA has published technical implementation guidance that further elaborates cybersecurity risk-management practices, governance expectations, and operational security measures in more practical terms [10]. These documents are relevant because they increasingly shape regulatory expectations surrounding cyber risk assessment and governance in practice.

Overall, these standards, regulatory frameworks, and guidance documents provide a structured foundation for cyber risk assessment, but they differ in the level to which they operationalise the scoring process in practice. ISO 31000 offers a generic risk management backbone, ISO/IEC 27005 adapts this backbone to information security, IEC 62443-3-2 provides a more prescriptive domain-specific approach for industrial systems, and NIST SP 800-30 offers complementary guidance on how risk assessment can be framed and applied. Meanwhile, NIS2 and the related ENISA guidance increasingly shape regulatory expectations surrounding cybersecurity risk management and governance within European contexts. This comparison is important for the present study because it shows that, while the overall assessment process is well established, the practical application of likelihood and impact scoring often remains flexible and open to interpretation.

2.2 Risk scoring in cyber risk assessment

While the standards and guidance discussed in the previous section provide a structured process for cyber risk assessment, they leave a lot of room for interpretation in how identified risks are scored in practice. This means that consultants and organisations still need to determine how likelihood and impact should be assessed, combined, and translated into risk levels. Risk scoring therefore forms a crucial link between assessment result and decision-making, as it determines how risks are prioritised and whether they are considered acceptable or require treatment.

2.2.1 Likelihood, impact, and risk scoring

A central element of cyber risk assessment is translating identified risks into structured judgements about how important they are. This is commonly done through the assessment of likelihood and impact, which together provide the basis for determining the overall level of risk. In general terms, likelihood refers to the probability or plausibility that a threat event will occur, while impact refers to the magnitude of its consequences if it does occur [15, 16]. By combining these two dimensions, organisations can compare risks more systematically and prioritise which ones require treatment, monitoring, or acceptance [14, 15, 16].

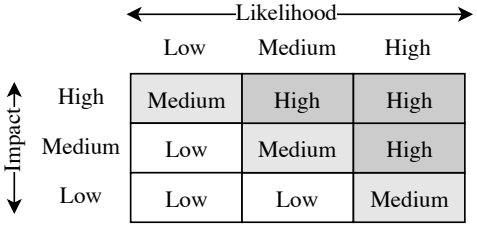


Figure 2.2: Example of a simple risk matrix combining likelihood and impact.

In practice, likelihood and impact can be expressed in qualitative, semi-quantitative, or quantitative

ways. Many organisations use qualitative scales such as low, medium, and high, while others apply numerical ranges, descriptive thresholds, or more detailed matrices depending on their context and maturity [16, 13]. Risk matrices are a common mechanism for combining the assessed likelihood and impact of a risk into an overall risk level. Figure 2.2 illustrates a simple risk matrix combining likelihood and impact. Although such matrices provide a structured way to compare risks, their outputs still depend heavily on how likelihood and impact are defined, interpreted, and scored in the first place.

This is precisely where flexibility in existing standards becomes important. Standards such as ISO/IEC 27005 and NIST SP 800-30 provide guidance on analysing risk in terms of likelihood and impact, but they do not prescribe a single universally applicable scoring model [15, 16]. Instead, organisations are expected to define scoring criteria that fit their context, objectives, and risk appetite. As a result, the application of likelihood and impact scoring in practice often varies across organisations and consulting engagements.

Figure 2.3 illustrates the general logic of risk scoring in cyber risk assessment, where assessment evidence is translated into likelihood and impact judgements and subsequently combined into an overall risk level.

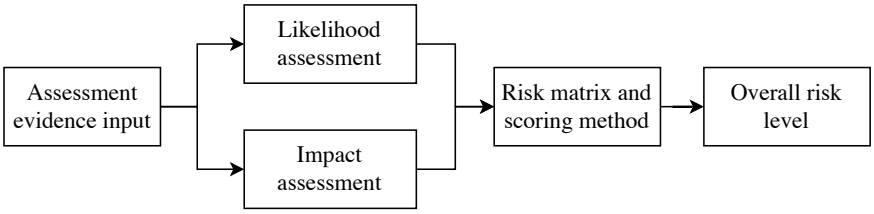


Figure 2.3: General logic of risk scoring in cyber risk assessment.

To support more explicit operational scoring, several practical methodologies have been developed. One example is the OWASP Risk Rating Methodology, which offers a more explicit and practice-oriented structure for assessing likelihood and impact [26]. In OWASP, likelihood is broken down into threat agent and vulnerability-related factors, while impact is broken down into technical impact and business impact. Each factor is linked to descriptors and scores that help translate qualitative judgement into a more systematic risk rating [26]. This is particularly relevant to this research because it provides an operational scoring structure that can be adapted to consulting practice.

Altogether, likelihood, impact, and their combination into overall risk levels form the core of cyber risk scoring. At the same time, the literature and standards show that while the need for such scoring is widely recognised, the current and concrete way to apply it in practice often remains flexible. This makes risk scoring not only a technical calculation step, but a judgement-intensive activity in which evidence interpretation, scoring logic, and methodological choice all play an important role.

2.2.2 Judgement in risk scoring and associated challenges

Although likelihood and impact scoring are often presented as structured elements of cyber risk assessment, their practical application remains highly judgement-intensive. Existing standards and guidance documents provide direction on what should be assessed, but they often leave a lot of flexibility in how likelihood, consequences, and overall risk levels are determined in practice [15, 14, 16]. As a result, scoring is rarely a purely mechanical exercise. Instead, assessors must interpret available evidence, determine which factors are most relevant, and translate incomplete or ambiguous information into scoring outcomes.

This reliance on judgement is closely related to the broader shift from prescriptive or "hard" regulatory regimes towards more goal-based or "soft" regimes [24]. In such settings, standards define objectives and expected outcomes, but do not prescribe one fixed technical method for applying them in practice. While this flexibility allows adaptation to different organisational contexts, it also increases interpretative burden. Different organisations or assessors may therefore follow the same broad framework while still applying different scoring criteria, evidence thresholds, and assumptions [24, 22].

The literature identifies several consequences of this flexibility. Scoring criteria may be insufficiently defined, risk matrices and calculation methods may vary or incorrectly developed, and the reasoning

behind assigned scores may only be partially documented [12, 22]. This can reduce comparability and reproducibility across assessments, even when the same backbone process is used. In addition, where justification remains implicit, it becomes more difficult to validate results internally or defend them clearly to external stakeholders [12, 3].

At the same time, the literature does not suggest that judgement should be removed from cyber risk scoring altogether. On the contrary, expert interpretation remains important because risk assessment often involves incomplete evidence, context-dependent information, and uncertain future events. The challenge is therefore not the existence of judgement itself, but the lack of structure and transparency in how that judgement is applied. Prior work has argued that there is a demand for more comparable and reproducible approaches to risk assessment [12], yet practical support for structuring such reasoning is still limited.

In summary, the literature suggests that the difficulty of cyber risk scoring is not only about defining likelihood and impact as abstract concepts, but in applying them in practice consistently under flexible, non-prescriptive standards. This makes judgement in risk scoring a central issue for this research, as it highlights the need for decision-support approaches that help structure reasoning, improve transparency, and preserve traceability without replacing professional judgement.

2.3 Decision support for judgement-intensive cyber risk assessment

The challenges discussed in the previous section suggest that the main difficulty in cyber risk scoring is not the presence of judgement itself, but the difficulty of applying and documenting that judgement in a consistent and transparent way [12, 24, 22]. This creates a need for decision support that does not replace expert interpretation, but helps structure it. In this context, decision support can be understood as mechanisms, tools, or representations that assist assessors in organising evidence, describing reasoning, and comparing scoring outcomes more systematically.

A first requirement is support for consistency. Because cyber risk scoring often involves diverse evidence and context-dependent interpretation, different assessors may arrive at different conclusions even when following the same framework [12, 22]. Decision support can help reduce such variation by providing a more structured basis for translating evidence into scoring inputs. A second requirement is support for structured and clear reasoning. When the reasoning behind a score remains implicit or only partially documented, the resulting assessment becomes more difficult to review, compare, and justify [12, 24]. Decision support is therefore valuable when it helps make judgement logic more explicit rather than leaving it fully implicit.

A third requirement is traceability and transparent justification. In judgement-intensive settings, outputs become more trustworthy when they are linked to the supporting evidence on which they are based. This is particularly important where evidence is distributed across multiple sources, such as policies, regulations, interview transcripts, and prior audit or gap assessment results. Prior work on automated assessment support also highlights the value of linking outputs back to supporting evidence in order to improve transparency and credibility [3]. A fourth requirement is the preservation of human judgement. Because risk scoring involves uncertainty, incomplete evidence, and context-sensitive trade-offs, professional judgement cannot simply be replaced by a fixed computational procedure [24]. Decision support is therefore most appropriate when it functions in a human-in-the-loop manner, where the system helps structure evidence and reasoning, but the final interpretation and decision remain under assessor control.

Seen as a whole, the literature suggests that useful decision support for judgement-intensive cyber risk assessment should support consistency, help structure reasoning, preserve traceability, and maintain human judgement. These requirements directly inform the design requirements defined in Chapter 3 and are also translated into the solution objectives presented in Chapter 4.

2.4 AI-assisted analysis and LLM applications

The need for more structured, transparent, and traceable decision support in cyber risk assessment has led to an increasing interest in the use of AI-assisted and LLM-based analysis. However, while AI-assisted and LLM-based approaches may improve efficiency, consistency, and evidence handling, they also

raise questions regarding reliability, explainability, and the appropriate role of automation in judgement-intensive work. This section therefore reviews existing AI-assisted approaches and LLM applications that are relevant to cyber risk assessment, before discussing their current limitations and the implications for the present study.

2.4.1 Existing approaches

Recent studies suggest that AI-assisted approaches are increasingly being explored to support cybersecurity assurance and risk-related tasks, although their focus areas are mainly different. Existing work can be broadly grouped into three relevant approaches: automated assessment review, LLM- and RAG-based threat modelling, and AI-driven automation of certification or assurance workflows.

Automated assessment review. Angelini et al. propose a methodology to support the automatic review of cyber risk assessments by mapping an existing control-based assessment to infrastructural aspects that can be assessed more objectively, such as ICT devices, policies, and governance-related elements. Their approach aims to reduce subjectivity and assessor bias by reviewing the trustworthiness of an already produced assessment rather than generating a new one from scratch. It combines governance-oriented information with a multi-layer attack graph and shows that automation can help detect possible underestimation and improve fit between assessed controls and actual infrastructure exposure [3].

LLM-and-RAG-based threat modelling. Erdogan et al. present the CORAS Threat Modeler, an open-source tool that uses LLMs, retrieval-augmented generation (RAG), and multi-agent orchestration to generate structured risk information and threat models from natural-language system descriptions. The tool integrates external knowledge bases such as CAPEC and CWE, retrieves contextually relevant attack patterns and weaknesses, and generates correct and interpretable threat models. Their evaluation shows that such an approach can broaden practitioners' risk perspectives and lower the barrier for non-experts, therefore demonstrating the potential of LLM-and-RAG-based support for risk-related analysis tasks [8].

AI-driven certification and assurance automation. Abba et al. investigate how AI can be applied to automate cybersecurity certification processes in order to improve efficiency, transparency, and governance accountability. Their work combines machine learning, process mining, and risk modelling to support control verification, evidence analysis, and continuous monitoring in certification workflows. The study reports improvements in efficiency and transparency, but also emphasises that opacity, bias, and accountability remain significant governance concerns, implying that explainability and human oversight are still necessary when automation is introduced into assurance settings [1].

All of these approaches show that AI can support cybersecurity assessment work in different ways. Some support the review of assessment outputs and the identification of gaps, while others demonstrate the potential of LLM-and-RAG-based support for risk-related analysis tasks. They also show that explainability and human oversight remain necessary when automation is introduced. However, these approaches focus on different parts of the broader assessment problem and therefore provide only partial support for the judgement-intensive task of translating diverse evidence into consistent likelihood and impact scoring outcomes.

2.4.2 Current Limitations

Despite these innovative approaches, the existing literature still leaves several limitations that are directly relevant to this research. First, many current approaches support related tasks rather than the operational scoring problem itself. Angelini et al. focus on the automatic review of an already produced risk assessment rather than on structuring how assessors translate diverse evidence into likelihood and impact judgements in the first place [3]. The CORAS Threat Modeler presented by Erdogan et al. is mainly directed towards threat modelling and structured risk information generation from system descriptions, rather than towards consultant-oriented evidence interpretation and score justification within flexible scoring frameworks [8].

Second, the literature indicates that current AI-assisted approaches still face limitations in coverage, treatment flexibility, and usability. In the evaluation of the CORAS Threat Modeler, participants considered the generated outputs useful and interpretable, but the study also reports the need for broader risk coverage, more flexible treatment support, and improved usability and interface design [8]. Similarly, Abba et al. highlight that while AI-driven automation can improve transparency and efficiency in

certification workflows, governance concerns such as opacity, bias, and accountability are still unresolved and require hybrid human–AI verification mechanisms and explainable control [1].

Third, relatively little work addresses how AI support can be specifically embedded into judgement-intensive scoring under flexible standards such as ISO/IEC 27005 [3, 8, 1]. Existing approaches demonstrate the potential of AI for automation, extraction, modelling, and review, but they do not systematically structure how evidence from documents, interviews, and past assessments is translated into explicit likelihood and impact reasoning that remains traceable, reviewable, and under human control. This limitation is particularly important for this research, because it suggests that the unresolved challenge is not only automation itself, but the design of decision support that can be applied for risk scoring while preserving transparency, traceability, and professional judgement.

2.5 Summary and research gap

The literature reviewed in this chapter shows that cyber risk assessment is supported by a well-established process backbone and a range of standards and guidance documents, including ISO 31000, ISO/IEC 27005, IEC 62443-3-2, and NIST SP 800-30 [14, 15, 13, 16]. These sources provide structured phases, shared terminology, and general guidance for identifying, analysing, evaluating, and treating risk. However, they also leave a lot of flexibility in how likelihood and impact are applied in practice, particularly in contexts where evidence is diverse and judgement is required [24, 22].

The literature further shows that risk scoring is not only a technical calculation step, but a judgement-intensive activity in which assessors must interpret available evidence, apply scoring logic, and justify their conclusions [12, 24, 22]. Although practical scoring methodologies provide a more explicit operational structure for likelihood and impact assessment [26], the way such scoring is applied in practice still depends heavily on contextual interpretation, evidentiary assumptions, and human reasoning. This creates challenges for consistency, transparency, traceability, and reproducibility of the risk scoring process across assessments [12, 3].

At the same time, recent work on AI-assisted cybersecurity analysis demonstrates the potential of automation, LLMs, and RAG to support assessment-related tasks, including automated review, threat modelling, and certification workflows [3, 8, 1]. These approaches show that such technologies can improve efficiency, evidence handling, and structured analysis. However, the reviewed studies do not directly address the central challenge identified in this thesis: how to support judgement-intensive likelihood and impact scoring in practice in a way that remains transparent, traceable, and under human control within flexible assessment frameworks [3, 8, 1].

Figure 2.4 summarises how these approaches lead into the research gap addressed in this research.

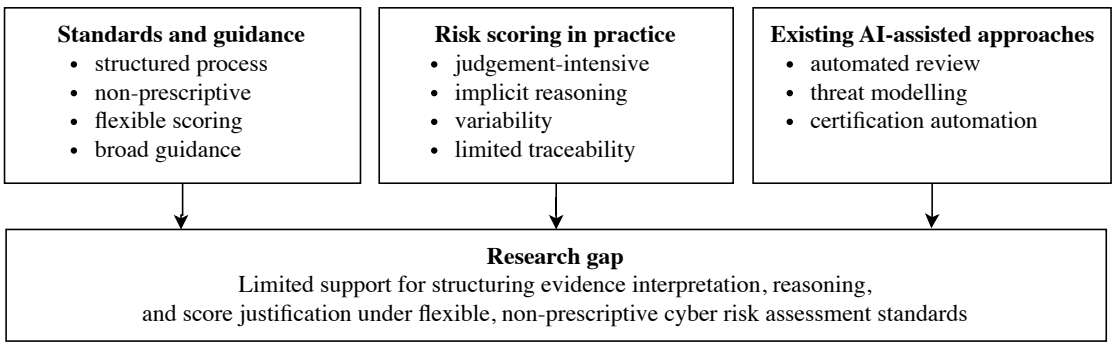


Figure 2.4: Summary of the literature reviewed in this chapter and the resulting research gap.

This reveals a structural gap in the literature. While existing standards preserve flexibility and existing AI-assisted approaches support related assessment tasks, limited work addresses how decision support can systematically structure evidence interpretation, reasoning, and score justification for cyber risk assessment in consulting practice. In particular, there is insufficient focus to how AI-assisted support can help translate diverse assessment inputs into more consistent and explainable scoring outcomes without

replacing professional judgement. This gap motivates this research, which designs and evaluates an AI-assisted decision-support approach intended to structure reasoning, improve traceability, and support consistency in cyber risk scoring while preserving consultant control and accountability.

Chapter 3

Method

In this chapter, we outlined the methodological approach adopted in this research. The research follows the Design Science Research Methodology (DSRM) to design, develop, and evaluate an AI-assisted decision-support artifact for cyber risk assessment. This approach is appropriate as the research aims to address a practical challenge in judgement-intensive risk scoring while generating structured and reproducible insights.

The research is structured according to the key phases of DSRM, including problem identification and objective definition, artifact design and development, demonstration, and evaluation. These phases are instantiated through a combination of literature review, stakeholder interviews, artifact development, controlled case construction, and multi-stage evaluation. For each phase, this chapter describes the applied methods and the rationale behind the methodological choices.

3.1 Research approach

This research adopts the Design Science Research Methodology (DSRM) as proposed by Peffers et al. [28], which is suitable for researches aiming to design, develop, and evaluate practical artifacts that address identified organisational problems. DSRM is particularly appropriate for this research, as the objective is not to automate decision-making, but to design an AI-assisted decision-support artifact that improves consistency, transparency, and verifiability in judgement-intensive cyber risk assessment practice.

The overall DSRM process applied in this research is illustrated in Figure 3.1.

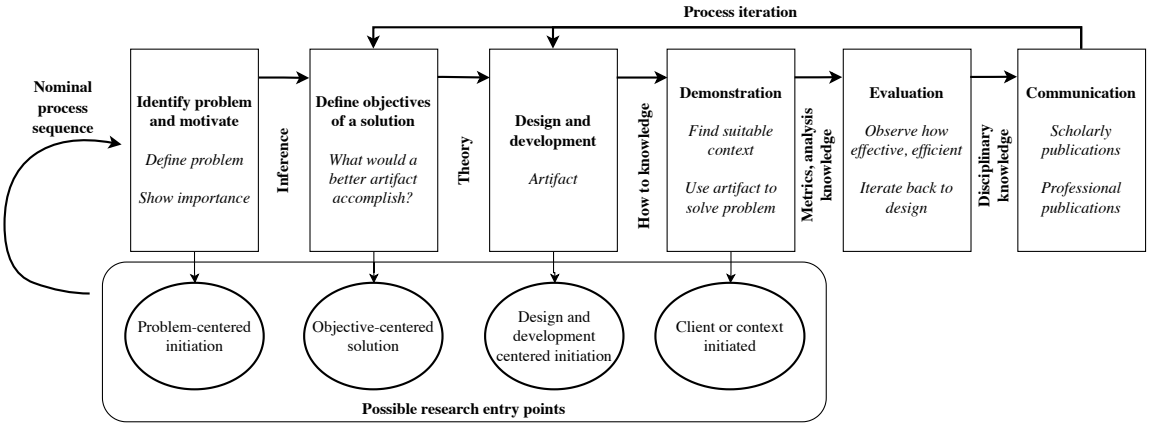


Figure 3.1: DSRM Process Model, adapted from Peffers et al. [28].

DSRM consists of six phases: problem identification, definition of objectives, design and development, demonstration, evaluation, and communication. In this research, these phases are instantiated as follows.

Problem identification and objective definition are conducted through a literature review, resulting in a preliminary problem definition that is then validated through interviews with senior cyber risk consultants. This is followed by the design and development phase, in which the Cyber Risk Estimation Support Tool (CREST) is developed as an AI-assisted decision-support artifact. CREST is then demonstrated to participating consultants and evaluated through a structured evaluation process. Finally, the results of the study are communicated through this thesis.

3.2 Problem identification and objective definition

This section outlines the problem identification and objective definition phases, which are done through a targeted literature review and stakeholder interviews.

3.2.1 Literature review

This research adopts a targeted literature review to understand existing approaches in cyber risk assessment and to identify key challenges. The review started with established risk assessment standards such as ISO 31000, IEC 62443-3-2, ISO/IEC 27005, and NIST SP 800-30 [14, 13, 15, 16]. These standards define high-level risk assessment phases and emphasise alignment with organisational and risk context, while intentionally allowing flexibility in how impact and likelihood are measured and calculated.

Based on these observations, relevant literature and prior studies on risk management frameworks and risk assessment practices were reviewed to understand how such flexibility is addressed in practice (see also Section 2.1.2). The literature shows that this flexibility increases reliance on expert judgement and interpretative reasoning, often leaving supporting reasoning implicit [24]. Consequently, although standards provide structure, they do not eliminate interpretation, resulting in diverse approaches, inconsistent scoring practices, and limited process reproducibility [22, 12].

Although prior research has explored structured scoring approaches and automation in cybersecurity assessment [3, 8, 1], limited work addresses how judgement-intensive risk scoring can be consistently structured and made transparent in practice. These insights help shape the problem definition and motivate the development of the AI-assisted decision-support artifact in this research.

3.2.2 Stakeholder interviews

Stakeholder interviews were conducted to validate preliminary findings from the literature review and to identify additional challenges encountered in practice. The interviews involved four senior cyber risk consultants with experience in IT and OT risk assessment engagements. The interview questions covered several themes, including current risk assessment processes and flexibility, data sources and data quality, judgement and decision-making practices, reporting and justification of results, and usability and improvement opportunities. The interviews were conducted online and lasted for one hour per consultant. The participants were selected because of their practical involvement in cyber risk assessment work.

A semi-structured interview format was used to allow flexibility in exploring relevant topics while still covering the main themes of the study. After the interviews, the collected notes were reviewed and grouped by recurring themes. The responses were compared across interviews to identify recurring issues related to flexibility in assessment practice, use of evidence, judgement in scoring, and reasoning documentation. The insights from these interviews were used to validate the identified challenges and to guide the design requirements of CREST. The interviews confirmed that risk scoring heavily relies on interpretation and experience of the consultants, particularly in situations where data is incomplete or ambiguous. Variability in scoring decisions across consultants was also identified as a common challenge. In addition, participants highlighted the need for better support in structuring and documenting reasoning during assessments.

3.3 Artifact design and development approach

This section outlines the artifact design and development phase of the research, describing the research setting, the type of artifact developed, and the design requirements derived from the literature review and stakeholder interviews.

3.3.1 Research setting

This research is conducted in the context of cybersecurity consulting, specifically within cyber risk assessment engagements involving information systems of organisations across various industries. In such engagements, consultants perform risk assessments based on a combination of data sources, including interviews with stakeholders and the analysis of system and policy documentation.

CREST is designed to support consultants in this context, particularly in structuring and interpreting assessment evidence. It aims to facilitate more consistent interpretation of evidence and reasoning across different consultants as users.

3.3.2 Artifact type

The artifact developed in this research is the Cyber Risk Estimation Support Tool (CREST), an AI-assisted decision-support prototype designed to support judgement-intensive risk scoring and to structure reasoning processes. In this thesis, CREST refers to the designed tool and workflow, including the Excel-centred assessment workbook, the Python-based control panel, and the backend processes used for evidence retrieval, LLM-assisted reasoning, and risk scoring. Rather than automating decision-making, CREST is intended to assist consultants by providing structured support while preserving human judgement. Hence, a human-in-the-loop approach is a core part of the design.

3.3.3 Design requirements

The design requirements are derived from the findings of the literature review and stakeholder interviews. These requirements reflect the need to address challenges related to judgement-intensive risk scoring and the lack of structured reasoning in current practices.

CREST is designed to meet the following requirements:

- **Improve consistency:** CREST should support more consistent scoring decisions across different consultants and engagements.
- **Structure reasoning processes:** CREST should assist in structuring the interpretation of evidence and the justification of scoring decisions.
- **Support traceability:** CREST should enable transparent reasoning by explicitly linking assessment outcomes to supporting evidence and data sources.
- **Maintain human judgement:** CREST should preserve human-in-the-loop decision-making, ensuring that final assessments remain under consultant control.

These requirements provide the basis for the solution objectives presented in Section 4.1. While these requirements describe what CREST needs to address from a methodological and practical perspective, the solution objectives translate them into more concrete goals that guide the structure of the proposed solution.

3.4 Demonstration

This section outlines the demonstration phase of the study, describing the construction and validation of the testing case used to demonstrate CREST, as well as the data used for the demonstrator.

3.4.1 Case construction and validation

To demonstrate CREST, a controlled assessment case was constructed based on a combination of real-world assessment data obtained from a multinational technology organisation (hereafter referred to as "the client") and structured scenario development. The case construction process aims to ensure that the generated risk scenarios are both realistic and aligned with available assessment evidence.

This research focuses on the risk analysis phase of the risk assessment process, where likelihood and impact are evaluated. Therefore, risk identification is considered out of scope. Instead of identifying risks from scratch, predefined risk scenarios are constructed to enable controlled and consistent evaluation of CREST. Figure 3.2 shows how the controlled assessment case was constructed in this study.

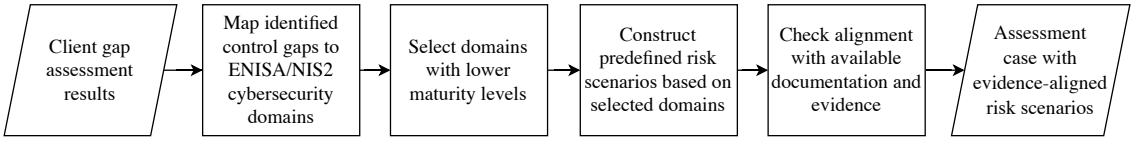


Figure 3.2: Case construction process for the controlled assessment case.

First, the client's existing gap assessment results based on ISO/IEC 27001 controls were used as the primary basis. These results reflect the maturity level of implemented controls and provide an indication of areas with potential weaknesses. To structure the assessment context, these control gaps were mapped to corresponding cybersecurity domains based on ENISA's guide on risk management measures in cybersecurity under the NIS2 Directive [11].

Based on this mapping, domains with lower maturity levels were selected as the basis for constructing risk scenarios. This approach ensures that the generated scenarios are grounded in identified control deficiencies rather than being randomly defined. For each selected domain, synthetic risk scenarios were developed by combining relevant threat events with vulnerabilities derived from the identified control gaps.

This alignment ensures that the constructed scenarios are consistent with the available documentation and assessment evidence. As a result, the retrieved evidence remains relevant to the scenarios, supporting a meaningful evaluation of CREST.

3.4.2 Case validation

To ensure the validity of the constructed assessment case, the generated risk scenarios were reviewed and refined in consultation with a senior cyber risk consultant. The validation process focused on assessing whether the scenarios are realistic, internally consistent, and representative of real-world situations.

Feedback was used to adjust scenario definitions, ensuring that they reflect realistic and representative threat and vulnerability combinations and are suitable for evaluation purposes.

3.4.3 Data used for the demonstrator

The data used for the demonstrator consists of assessment materials obtained from the client organisation, including policy documents, gap assessment results, assessment reports, and interview transcripts. These data sources reflect the types of information typically used by consultants during cyber risk assessment engagements.

The dataset consists of a total of 68 PDF documents and 4 Excel files. The documents vary in structure and format, including formal policy documents, tabular assessment results, semi-structured reports, and unstructured interview transcripts. As a result, the dataset represents a diverse collection of both structured and unstructured data.

This diversity is important to ensure that CREST is evaluated under realistic conditions, where consultants must interpret and integrate information from multiple sources with different levels of structure and completeness. The data is used as input for CREST's retrieval-augmented generation (RAG) component, where relevant information is extracted, structured, and used to support risk scoring and reasoning generation.

3.5 Evaluation method

This section outlines the evaluation phase of the study, describing the strategy, participants, metrics, as well as socio-technical considerations.

3.5.1 Evaluation strategy

CREST is evaluated using a multi-stage evaluation strategy designed to assess both its technical performance and practical applicability in cyber risk assessment practice. The evaluation consists of three stages.

First, configuration comparison is conducted to identify the most suitable combination of the LLM model and retrieval depth (k-value) for the retrieval-augmented generation (RAG) component. This stage focuses on optimizing system performance prior to further evaluation.

Second, a controlled case evaluation is performed to assess the accuracy and traceability of CREST when applied to predefined risk scenarios. This enables a systematic comparison between generated outputs and expected results.

Third, a consultant evaluation is conducted to assess CREST from a user perspective, including acceptability, usability, and overall user experience. This stage incorporates both structured instruments and qualitative feedback from practitioners.

The evaluation strategy combines controlled assessment scenarios and consultant feedback to assess CREST from both technical and practical perspectives.

3.5.2 Participants

The evaluation involves different participants across stages. The technical parameter testing and initial evaluation are conducted by the researcher. The consultant evaluation involves four senior cyber risk consultants.

3.5.3 Evaluation procedure

Figure 3.3 presents the multi-stage evaluation design applied in this research. The evaluation consists of four sequential stages, starting from technical parameter testing and controlled case evaluation, followed by practitioner-based evaluation. Each stage produces specific outputs that contribute to assessing both the technical performance and practical applicability of CREST.

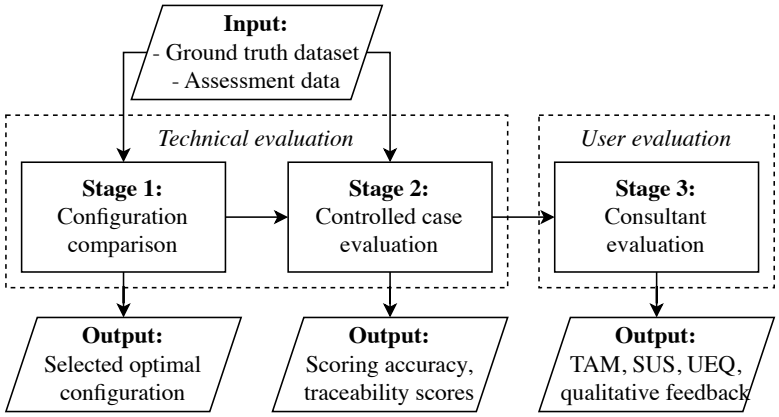


Figure 3.3: Evaluation design

Stage 1: Configuration comparison. To enable objective evaluation, a ground truth dataset is constructed for the selected risk scenarios. The ground truth defines the expected impact and likelihood scores, along with the corresponding reasoning. The initial ground truth is developed by the researcher and subsequently validated with a senior cyber risk consultant to ensure validity and alignment with real-world assessment practices.

A targeted subset of four predefined risk scenarios was selected for Stage 1. The selection was designed to cover different risk scoring levels based on the ground truth, namely critical, high, medium, and low risk, while also spreading the selected cases across different positions in the overall scenario set (Scenarios 1, 4, 8, and 12). This targeted sampling approach was used to keep the comparison manageable while still ensuring variation in expected scoring outcomes. These scenarios are used to test different combinations of LLM models and retrieval depth (k-value). Each configuration is evaluated based on scoring accuracy and reasoning traceability. Using a subset of representative scenarios enables efficient comparison of configurations while maintaining coverage across different risk levels.

Stage 2: Controlled case evaluation. The selected configuration is then applied to the full set of predefined risk scenarios. This stage aims to assess the overall performance of CREST across a broader range of cases. The generated outputs are compared against the ground truth to evaluate scoring accuracy, as well as the extent to which the generated reasoning can be traced back to and verified against the actual supporting evidence.

Stage 3: Consultant evaluation. CREST is demonstrated to senior cyber risk consultants, who are introduced to the prototype beforehand to understand its functionality. Following the demonstration, participants are asked to evaluate CREST through a combination of structured questionnaires and guided discussion. The evaluation focuses on perceived usefulness, usability, and overall user experience.

3.5.4 Experimental setup for configuration comparison

This section describes the experimental setup used for the configuration comparison in Stage 1 of the evaluation. While the evaluation procedure outlines the overall sequence of stages, this section specifies the key variables and configurations compared in this stage.

The configuration comparison focuses on two key variables within the RAG pipeline: (1) the choice of LLM model, and (2) the retrieval depth (k-value). These variables were selected because they directly affect both the quality of the generated reasoning and the relevance of the retrieved evidence. Other parameters, such as temperature and similarity threshold, were kept constant to limit the scope of the comparison and to make the effect of the selected variables easier to observe.

The comparison follows a controlled setup, in which different combinations of LLM models and k-values are applied to the same subset of predefined risk scenarios. This ensures that differences in performance can be linked to the tested configurations rather than variations in input data.

To ensure consistency, all configurations are evaluated using the same ground truth and evaluation metrics, allowing for systematic comparison of scoring accuracy and reasoning traceability.

3.5.4.1 LLM model selection

The selection of LLM models is constrained to the list of models available within the internal AI infrastructure of the host organisation. This infrastructure was used to ensure that client-related assessment data remained within the organisational environment and was not exposed to external model providers. Within these practical and data-governance constraints, three models were selected to represent different reasoning capabilities and context capacities. Table 3.1 summarises the selected models.

Table 3.1: Selected LLM models for technical parameter testing

Model	**Context window**	**Expected reasoning capability**
gpt-oss:20B	120K	Moderate
qwen3:30B	256K	Higher than baseline
deepseek-r1:32B	128K	Highest

First, a smaller open-weight model (gpt-oss:20B) is selected as a baseline, representing a conservative configuration with moderate reasoning capability and stable performance. Second, a mid-sized model (qwen3:30B) is included to represent improved reasoning ability and larger context capacity. Third, a more advanced reasoning-focused model (deepseek-r1:32B) is selected for its stronger performance on complex reasoning tasks.

This selection enables comparison across models with different reasoning strengths, while remaining within the constraints of practical deployment in the organisational setting.

3.5.4.2 Retrieval depth (k-value)

The retrieval depth (k-value) determines the number of document chunks provided as input to the LLM. In this research, k-values of 5 and 10 are selected.

This selection is based on prior empirical studies on Retrieval-Augmented Generation (RAG), which demonstrate that for question answering tasks, optimal performance is typically achieved when retrieving

between 5 and 10 document chunks [33]. Beyond this range, performance improvements show diminishing returns while additional documents introduce noise and increase computational overhead.

Therefore, k = 5 and k = 10 are chosen to represent a balanced and empirically grounded range for evaluating retrieval configurations.

3.5.5 Evaluation metrics

The evaluation of CREST covers both technical performance and user perception. To assess these aspects, the study uses a combination of quantitative and qualitative measures. Technical performance is evaluated through scoring accuracy and traceability, while user perception is assessed through structured questionnaire instruments and qualitative feedback. Together, these measures reflect both outcome correctness and reasoning quality, in line with the research objective of improving consistency, transparency, and reproducibility in the judgement-intensive risk scoring process.

3.5.5.1 Technical performance metrics

The technical performance of CREST is evaluated based on scoring accuracy and traceability.

Scoring accuracy refers to the extent to which the generated likelihood and impact scores align with the predefined ground truth. Accuracy is measured as the proportion of correctly predicted scores across all evaluated cases. A prediction is considered correct when the generated likelihood and impact levels exactly match the predefined ground truth.

Formally, accuracy is calculated as:

$$Accuracy = \frac{N_{correct}}{N_{total}} \tag{3.1}$$

where $N_{correct}$ is the number of correctly predicted scores, and N_{total} is the total number of evaluated cases.

Traceability refers to the extent to which the generated reasoning can be traced back to and verified against the actual supporting evidence. This includes whether the referenced evidence exists in the data, is relevant to the scenario, and adequately supports the assigned score.

Traceability is assessed using a structured rubric that evaluates both the verifiability of the reasoning and the correctness of cited sources. Each output is assigned a traceability level, ranging from fully traceable to untraceable. These levels are then mapped to numerical scores to enable quantitative analysis, and the overall traceability score is calculated as the average across all evaluated cases.

The rubric was developed iteratively to capture different levels of reasoning verifiability observed during evaluation. Initial versions with fewer categories were found to be insufficient to capture important variations in traceability. Through iterative refinement and application to sample outputs, additional levels were introduced to better reflect common patterns, such as vague sourcing, partially verifiable reasoning, and unsupported or incorrect references. This design allows for a more detailed assessment of reasoning quality beyond binary correct/incorrect evaluation.

Table 3.2 presents the final rubric used in this research.

Table 3.2: Traceability assessment rubric

Category	**Reasoning and evidence traceability**	**Evidence and citation accuracy**	**Weight**
Fully traceable	Aligned with risk context; specific source names provided	Direct quote correct, regardless of section ID or name	100%
Substantially traceable	Aligned with risk context; specific source names provided	No direct quote, but section ID or section name is correct	80%

Continued on next page

Category	Reasoning and evidence traceability	Evidence and citation accuracy	Weight
Partially traceable	Aligned with risk context; specific source names provided	No direct quote; section ID and section name incorrect, but the claim remains verifiable	60%
Generically traceable	Aligned with risk context; sources mentioned only in general terms	Broadly correct, but the claim cannot be verified precisely	40%
Minimally traceable	Aligned with risk context; sources absent or incorrectly mentioned	Sources are contradictory, irrelevant, or incorrect	20%
Untraceable	Not aligned with risk context; no identifiable source provided	Sources are nonexistent	0%

The assigned traceability levels are mapped to numerical scores (weight) to support quantitative comparison across evaluated cases. The weighting of the traceability levels reflects how easy or difficult it is for a reader to trace a generated claim back to its supporting evidence. Higher weights are assigned when the claim can be verified directly and with little effort, for example when a correct direct quote or clearly identifiable source is provided. Lower weights reflect cases where verification requires more effort because the citation is incomplete, vague, incorrect, or missing. At the lowest level, the claim cannot be traced back to any identifiable source and is therefore scored as untraceable.

The overall traceability score is calculated as the average of the assigned numerical scores across all evaluated cases:

$$Traceability = \frac{1}{N} \sum_{i=1}^{N} score_i \tag{3.2}$$

where $score_i$ is the traceability score assigned to the i-th evaluated case, and N is the total number of evaluated cases.

3.5.5.2 User perception metrics

To evaluate CREST from a user perspective, three established instruments are used: the Technology Acceptance Model (TAM), the System Usability Scale (SUS), and the User Experience Questionnaire (UEQ) [6, 4, 19]. These instruments were selected to capture complementary aspects of user perception, such as acceptance, usability, and broader user experience. Because the consultant evaluation involves a small group of expert users, these results are analysed descriptively rather than through inferential hypothesis testing.

Technology Acceptance Model (TAM)

TAM is used to assess perceived usefulness and perceived ease of use, indicating the extent to which consultants may be willing to adopt CREST in practice. Following Davis [6], TAM is applied through 12 questions divided into two constructs: perceived usefulness and perceived ease of use. TAM has also been applied in formative evaluations of early-stage technology prototypes with a small number of participants to obtain initial feedback on perceived usefulness, ease of use, and attitude toward technology use [30]. In this study, TAM is used as an instrument to assess whether consultants perceive CREST as useful and easy to use within the controlled evaluation setting.

An additional item is used to capture willingness to use the prototype in practice. Responses are coded on a five-point Likert scale, where 1 represents strongly disagree and 5 represents strongly agree. For each TAM construct, the score is calculated by averaging the responses to all items belonging to that construct. For example, the perceived usefulness score is calculated from Q11–Q16, while the perceived ease of use score is calculated from Q17–Q22. Since the Likert scale ranges from 1 to 5, the construct mean is also converted into a percentage of the maximum possible score by dividing it by 5.

System Usability Scale (SUS)

SUS is used to evaluate the overall usability of CREST through a standardised 10-item questionnaire [4]. It was selected because it provides a solid and widely used measure of perceived usability, which is suitable for evaluating whether the users experience CREST as usable within a short and limited evaluation session. SUS has also been used in usability testing of digital prototype tools and interactive platforms during iterative development [17]. Additionally, SUS has been used in formative usability studies where small groups of participants were used to identify usability issues before broader validation [23].

After obtaining the participant responses, each item is rated on a five-point Likert scale ranging from strongly disagree to strongly agree. SUS scores are calculated using the standard scoring method: for odd-numbered items, 1 is subtracted from the response value, while for even-numbered items, the response value is subtracted from 5. The adjusted values are summed and multiplied by 2.5, resulting in a total score ranging from 0 to 100 [5].

User Experience Questionnaire (UEQ)

UEQ is used to measure user experience aspects more than basic usability. It was selected because it provides a multidimensional assessment of user experience rather than a single overall score. The standard UEQ measures six dimensions: attractiveness, perspicuity, efficiency, dependability, stimulation, and novelty [19, 18]. These dimensions enable the evaluation to analyse between pragmatic qualities, such as efficiency and dependability, and hedonic qualities, such as stimulation and novelty.

The questionnaire consists of 26 semantic differential items rated on a seven-point scale between contrasting adjective pairs [19, 31]. The results are analysed by calculating the mean score for each UEQ dimension. This supports a more specific interpretation of how consultants experience CREST, for example whether weaknesses are related to ease of understanding, efficiency, control, or novelty. In addition, UEQ provides benchmark-based interpretation guidance, which helps position the results compared to reference values from previously evaluated products [31, 32, 18]. Therefore, in this study, the UEQ results are used to complement SUS by showing which aspects of the user experience are stronger or weaker, instead of only reporting a single usability score.

Combined interpretation of perception metrics

TAM, SUS, and UEQ are used in combination because they address related but distinct aspects of user perception. TAM is included to assess user perceptions on acceptance, especially whether consultants consider CREST useful and easy enough to support their work in practice. This is important because CREST is not only evaluated as an interface, but also as a decision-support artifact that may influence professional judgement and assessment workflow.

SUS is used as the main overall usability indicator. It provides a solid and widely recognised score that summarises whether CREST is perceived as usable as a whole. This is useful because the evaluation needs a clear overall judgement of usability, instead of only separate scores for individual experience dimensions. UEQ complements SUS by providing a more detailed profile of user experience in multiple dimensions, such as attractiveness, perspicuity, efficiency, dependability, stimulation, and novelty. This makes it easier to see which aspects of the interaction are stronger or weaker and where future improvements may be needed.

Although some conceptual overlap exists, particularly between usability, ease of use, and pragmatic user experience, the instruments are not seen as redundant. Instead, they provide different layers of interpretation: TAM captures acceptance potential, SUS captures overall usability, and UEQ captures a more detailed profile of the user experience. Together, these instruments provide a more comprehensive basis for interpreting how consultants perceive CREST during the controlled evaluation.

Qualitative feedback and questionnaire administration

In addition to these structured instruments, qualitative feedback is collected during the evaluation sessions to capture reflections on practical applicability, perceived strengths and limitations, socio-technical concerns, and possible future improvements. This qualitative input is used to complement and help interpret the questionnaire results rather than as a separate quantitative metric.

The questionnaires were administered via Qualtrics after the evaluation session. The form was configured not to collect participants' email addresses. Before filling in the questionnaire, participants were intro-

duced to CREST and guided through its main functions, including the scoring method, the configurable matrix elements, and the overall workflow of the tool. The questionnaire consisted of four parts: TAM, SUS, UEQ, and a socio-technical feedback section. This setup was intended to ensure that participants evaluated CREST based on an informed understanding of how the prototype operates, while qualitative reflections were collected to complement the structured perception measures.

3.5.6 Socio-technical considerations

In addition to technical and usability evaluation, this research considers several socio-technical aspects related to the use of AI-assisted decision-support in cyber risk assessment.

One key consideration is the risk of over-reliance on CREST. While CREST is designed to support decision-making, there is a potential risk that users may rely too heavily on the generated outputs without critically evaluating the underlying reasoning. To mitigate this, CREST is designed with a human-in-the-loop approach, ensuring that final decisions remain under the control of the consultant.

Another consideration is the potential shift in professional judgement practices. The introduction of structured reasoning support may influence how consultants interpret evidence and justify scoring decisions. While this can improve consistency and transparency, it may also affect individual judgement styles and flexibility in complex scenarios.

In addition, the use of AI introduces the risk of biased or incomplete outputs, particularly when the supporting data is limited or inconsistent in quality. This highlights the importance of transparency and traceability in the generated reasoning, allowing users to verify and critically assess the outputs.

Although these considerations are not included as direct evaluation metrics, they are still considered in the interpretation of the results and in the discussion of CREST's practical implications.

Chapter 4

Design

This chapter presents the design of CREST, the AI-assisted decision-support artifact developed in this research. Building on the problem analysis, research questions, and design requirements presented in the previous chapters, it shows how these inputs are translated into a structured solution for supporting judgement-intensive risk scoring in consulting practice.

The chapter is organised into three main parts. First, it introduces the solution objectives that guide CREST. Second, it describes the decision-support workflow, including how CREST processes assessment data, structures evidence, and supports the calculation of likelihood, impact, and final risk scores. Third, it presents CREST's architecture, including the Excel-centred workbook structure, Python-based control panel, evidence retrieval component, and LLM-assisted reasoning component.

4.1 Solution objectives

The solution objectives of CREST are based on the research questions presented in Chapter 1 and the design requirements defined in Chapter 3. They translate these inputs into concrete goals that guide its structure. In this way, they connect the identified challenges in judgement-intensive risk scoring to the design choices presented in the following sections.

- **Objective 1: Standardise the structuring of evidence for scoring.** CREST should reduce variation in how consultants translate diverse assessment evidence into inputs for impact and likelihood scoring. To support this, the design introduces a consistent workflow for extracting, organising, and preparing evidence for scoring, rather than leaving these steps fully implicit.
- **Objective 2: Make judgement logic more explicit and structured.** CREST should support consultants in articulating how evidence, control observations, and contextual factors contribute to scoring decisions. This addresses the problem that judgement logic often remains implicit and insufficiently documented in current practice.
- **Objective 3: Enable transparent and traceable justification.** CREST should link generated reasoning and scoring outputs to identifiable supporting evidence, allowing consultants to review, verify, and explain the basis of the assessment outcome.
- **Objective 4: Preserve consultant control and decision authority.** CREST should function as decision support rather than automated decision-making. Consultants remain responsible for reviewing extracted evidence, interpreting outputs, and making the final assessment decision.

4.2 Decision-support workflow

Although cyber risk assessment includes context establishment, risk identification, risk evaluation, and risk treatment, CREST focuses only on the risk analysis phase. This focus is justified by two main reasons. First, stakeholder interviews identified risk analysis as a key pain point in current consulting practice. Second, prior studies also highlight this phase as a point where consistency, transparency, and reproducibility in likelihood and impact scoring process are often lacking. In addition, risk analysis is

the stage where available evidence must be interpreted and translated into structured scoring outcomes, making it the most relevant stage for CREST.

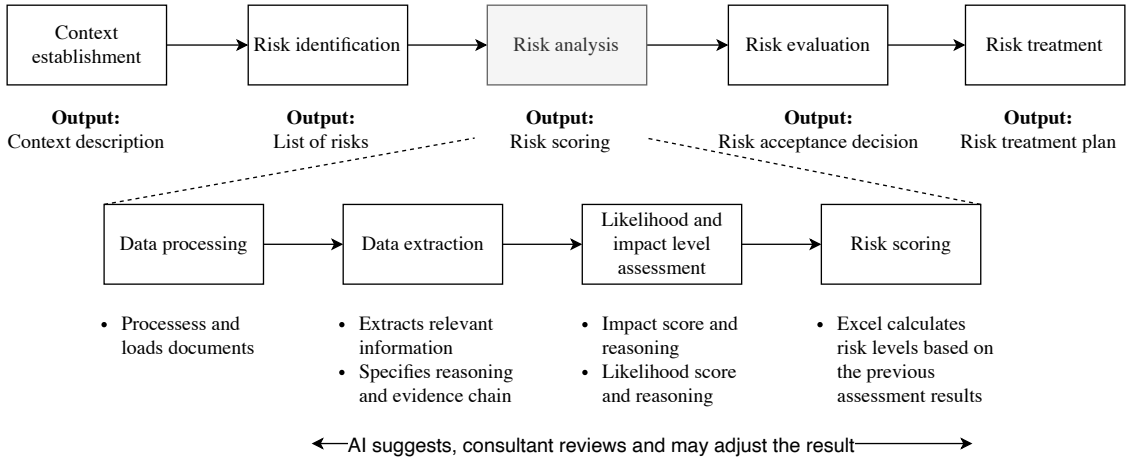


Figure 4.1: Scope of CREST within the broader cyber risk assessment process.

Figure 4.1 shows how CREST is positioned within the broader cyber risk assessment process. The artifact supports practical activities within the risk analysis phase, including data gathering, data processing, risk component assessment, and final risk scoring. Activities outside this phase, such as context establishment, risk identification, risk evaluation, and risk treatment, remain outside the current scope.

4.2.1 Assessment model and scoring dimensions

CREST uses the OWASP Risk Rating Methodology as the basis for its scoring model [26]. In this methodology, risk is determined by combining likelihood and impact, while the assessment itself is structured through predefined dimensions and scoring criteria. OWASP distinguishes two dimensions for likelihood, threat agent and vulnerability, and two dimensions for impact, technical impact and business impact. Each dimension consists of several factors that can be assessed through selectable descriptors linked to numerical scores. This structure provides a systematic basis for translating diverse assessment evidence into more comparable scoring inputs. Figure 4.2 shows the assessment model used in CREST.

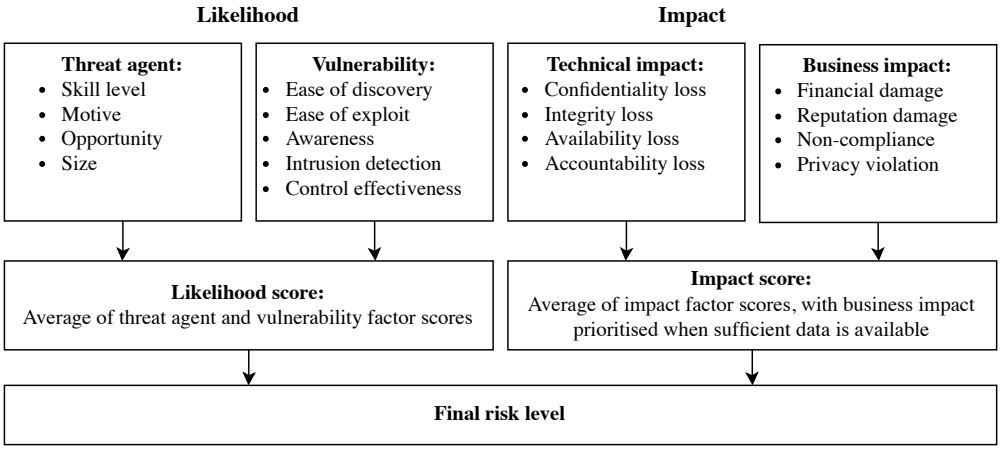


Figure 4.2: Assessment model used in CREST, adapted from the OWASP Risk Rating Methodology, and extended with *control effectiveness*.

Likelihood is assessed through the threat agent and vulnerability dimensions. Threat agent captures characteristics of the potential attacker, such as skill level, motive, opportunity, and size. Vulnerability

captures characteristics related to the weakness itself, including ease of discovery, ease of exploit, awareness, and intrusion detection. For impact, CREST follows the OWASP distinction between technical impact and business impact. Technical impact reflects consequences for systems and information assets, including loss of confidentiality, integrity, availability, and accountability, whereas business impact reflects organisational consequences such as financial damage, reputation damage, non-compliance, and privacy violation.

In addition to the original OWASP factors, CREST includes *control effectiveness* as an additional factor within the vulnerability dimension. This reflects the consulting context of the study, where much of the available evidence relates to control design, implementation, and identified gaps. Since ineffective or insufficiently implemented controls can increase the likelihood that a threat successfully exploits a vulnerability, adding control effectiveness makes the link between control assessment results and risk scoring more explicit. Rather than replacing the OWASP structure, this factor extends it in a way that is relevant to consulting practice.

A broader control assessment component was initially considered during design, but it was not included as a separate evaluated part of CREST. More explicit control assessment is technically possible and was partially explored during development. However, to keep the study manageable, this aspect was reduced in scope. In the current version, control-related information is included in a simplified form through the control effectiveness factor under the vulnerability dimension. A fuller control assessment component is therefore left for future work.

Although CREST uses OWASP as its conceptual basis, its implementation is designed to be semi-configurable in Excel. The main assessment dimensions remain fixed as threat agent, vulnerability, technical impact, and business impact in order to preserve methodological consistency. However, the factors, descriptors, and associated numerical scores within these dimensions can be adjusted to reflect client-specific terminology, assessment practices, or risk criteria. This allows the scoring framework to remain structured while still being adaptable to the organisational context.

4.2.2 Data processing and extraction

The data processing and extraction step prepares assessment evidence for later scoring. This step does not yet determine the final risk score. Instead, it converts unstructured and semi-structured assessment materials into structured factor-level inputs that can be reviewed and used in the subsequent likelihood and impact assessment stage. In this way, retrieval and extraction support the later scoring stage by organising relevant evidence into a more structured form.

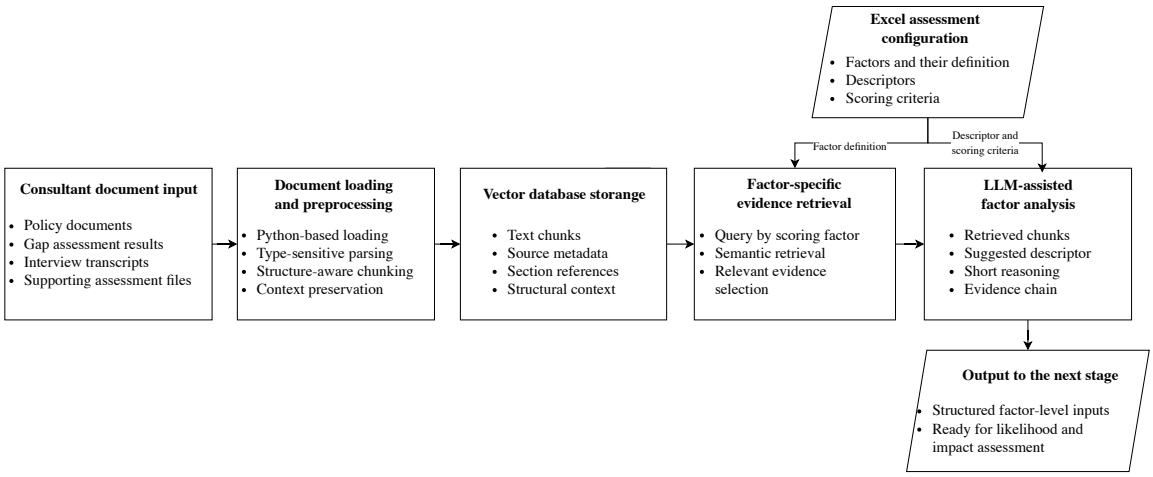


Figure 4.3: Data processing and extraction workflow in CREST.

To reduce manual preparation effort, CREST is designed so that consultants only need to place assessment data in a predefined folder structure before analysis. In practice, input documents are organised into designated folders by type, such as policy documents, gap assessment results, interview transcripts, and other supporting materials. This reflects the existing consulting workflow, where evidence is already

distributed across multiple files and formats, and avoids unnecessary preparation before the assessment begins.

Once the documents are placed in the appropriate folders, the Python-based processing component loads and preprocesses them for retrieval. Because the available data differ in structure and format, preprocessing applies document-type-sensitive parsing and chunking. More structured documents, such as policy documents or gap assessment outputs, are chunked in a way that preserves headings, sections, and tabular context where possible, while less structured materials, such as interview transcripts, are chunked more linearly. This helps preserve relevant context during retrieval without forcing all documents into the same structure.

After preprocessing, the resulting chunks are stored in a vector database together with metadata describing their source and structure. This metadata includes source document names, relevant section references, and other contextual identifiers that help maintain traceability during retrieval. The use of a vector database enables CREST to retrieve semantically relevant chunks across diverse documents rather than relying only on exact keyword matches.

For each scoring factor in the assessment model, the retrieval component queries the vector database to identify chunks that are relevant to the specific factor under consideration, such as a threat agent characteristic, a vulnerability-related factor, or an impact-related factor. The retrieved chunks are then passed to the LLM together with the relevant scoring criteria taken from the Excel-based assessment matrix. Based on this input, the model generates a structured analysis block for each factor, consisting of a suggested descriptor, a short reasoning statement, and an evidence chain indicating the basis of the suggestion.

The resulting factor-level outputs provide the structured inputs used in the likelihood and impact assessment described in the next subsection.

4.2.3 Likelihood and impact level assessment

After the factor-level analysis blocks have been generated, CREST translates these outputs into likelihood and impact assessments. At this stage, the proposed descriptor for each factor is matched against the scoring matrix defined in Excel, where each descriptor is linked to a numerical value. Because the factor structure is loaded dynamically from the configurable matrix sheet, the assessment logic does not rely on hardcoded factors or fixed scoring options. This allows the same workflow to remain applicable across different client-specific factor definitions, descriptors, and score ranges.

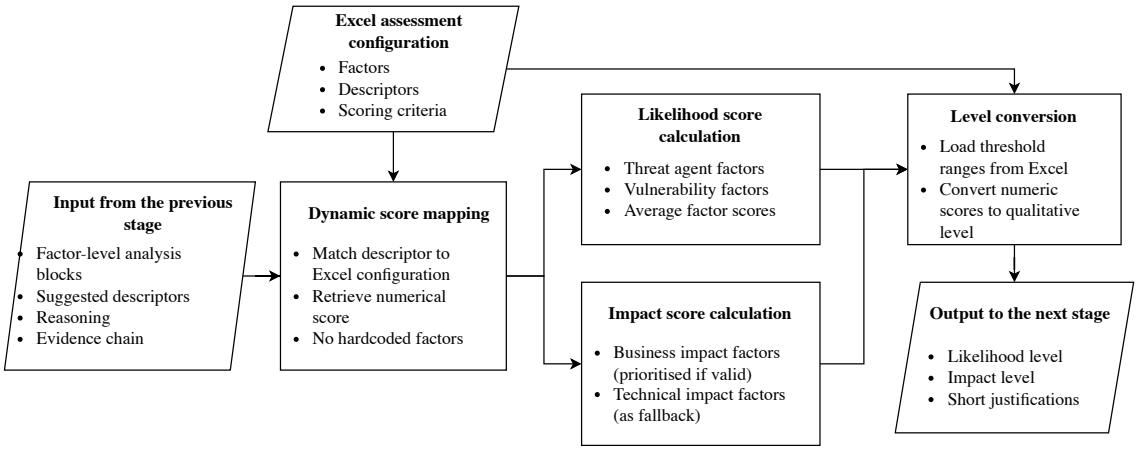


Figure 4.4: Likelihood and impact level assessment workflow in CREST.

For likelihood, CREST combines the scores of the threat agent and vulnerability factors. Once a descriptor has been identified for each factor, the corresponding numerical values are retrieved from the matrix and aggregated into an overall likelihood score. In the current implementation, this score is calculated as the average of the threat agent and vulnerability factor scores. This preserves the OWASP-based distinction between attacker-related characteristics and vulnerability-related conditions, while still allow-

ing implementation-specific factors such as control effectiveness to be included within the vulnerability dimension.

For impact assessment, the same procedure is applied to the business impact and technical impact dimensions. However, consistent with OWASP Risk Rating Methodology, business impact is prioritised when sufficient and valid information is available, as it better reflects organisational severity from a decision-making perspective. When business impact data is unavailable or flagged as insufficient, CREST uses technical impact as the fallback mechanism for estimating impact severity. In this way, the assessment remains operational even when the available evidence is inconsistent across impact dimensions.

After the numerical likelihood and impact scores have been calculated, they are translated into qualitative levels using configurable thresholds defined in Excel. These thresholds are read dynamically from the assessment matrix sheet rather than fixed in code, allowing the interpretation of numerical scores to be adapted to different client contexts. The output of this stage is therefore a structured likelihood level and impact level, which are used as inputs for the final risk scoring step.

4.2.4 Risk scoring

After the likelihood level and impact level have been determined, CREST calculates the final risk level by combining both levels through a configurable risk matrix implemented in Excel. This reflects the final stage of the OWASP-based scoring logic, in which risk severity is derived from the combination of likelihood and impact.

In the prototype, the final risk level is retrieved through an Excel-based lookup mechanism using `INDEX`/`MATCH`. Based on the likelihood and impact levels produced in the previous step, CREST retrieves the corresponding risk level from a predefined matrix stored in a configuration sheet. For example, a combination such as low likelihood and high impact may result in different final risk levels depending on the matrix used in the assessment context. This makes the final scoring logic explicit and directly visible within the working file.

A key design choice is that the risk matrix is maintained as configurable Excel content rather than fixed logic in Python. In practice, clients often use different risk matrices, ranging from simple three-by-three structures to more granular five-by-five or seven-by-seven formats. The prototype therefore uses an OWASP-style matrix as a default configuration, while still allowing the matrix values, labels, and dimensions to be adjusted in Excel. Using Excel for this step also allows consultants to inspect and adapt the applied scoring logic through a familiar interface without requiring changes to the Python code.

Although the previous workflow descriptions focus mainly on CREST's processing steps, the scoring process remains human-in-the-loop. Consultants review the factor-level scores, reasoning and calculated likelihood, impact, and risk scores in Excel before accepting or adjusting the assessment outcome. Figure 4.5 makes this review and adjustment role explicit.

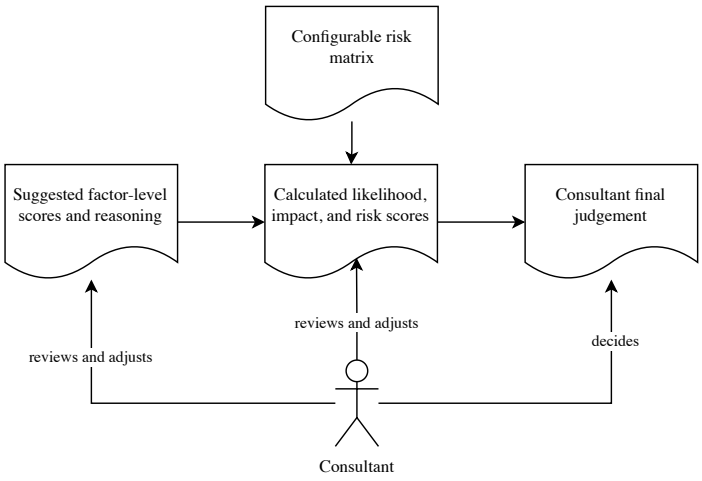


Figure 4.5: Human-in-the-loop review and adjustment in CREST's scoring process.

In this flow, CREST provides suggested reasoning and scoring support, while consultants remain responsible for reviewing the evidence, adapting the scoring configuration where needed, and confirming the final assessment result.

4.3 CREST architecture

This section presents the architecture of CREST. While the previous section described the decision-support workflow from input evidence to scoring output, this section focuses on the main system components that support that workflow. For clarity, the architecture is described through the Excel-centred workbook, the Python-based control panel, the evidence retrieval component, and the LLM-assisted reasoning component that transforms retrieved evidence into structured outputs.

4.3.1 User interface

CREST is designed with an Excel-centred user interface supported by a lightweight Python-based control panel. This reflects its the intended use in the host organisation, where Excel is already used as a familiar working-paper environment for documenting assessments, reviewing evidence, and recording results. Rather than requiring consultants to work directly with scripts or command-line tools, CREST allows them to trigger the main processing steps through a simple graphical interface while keeping assessment settings and outputs accessible in the Excel workbook.

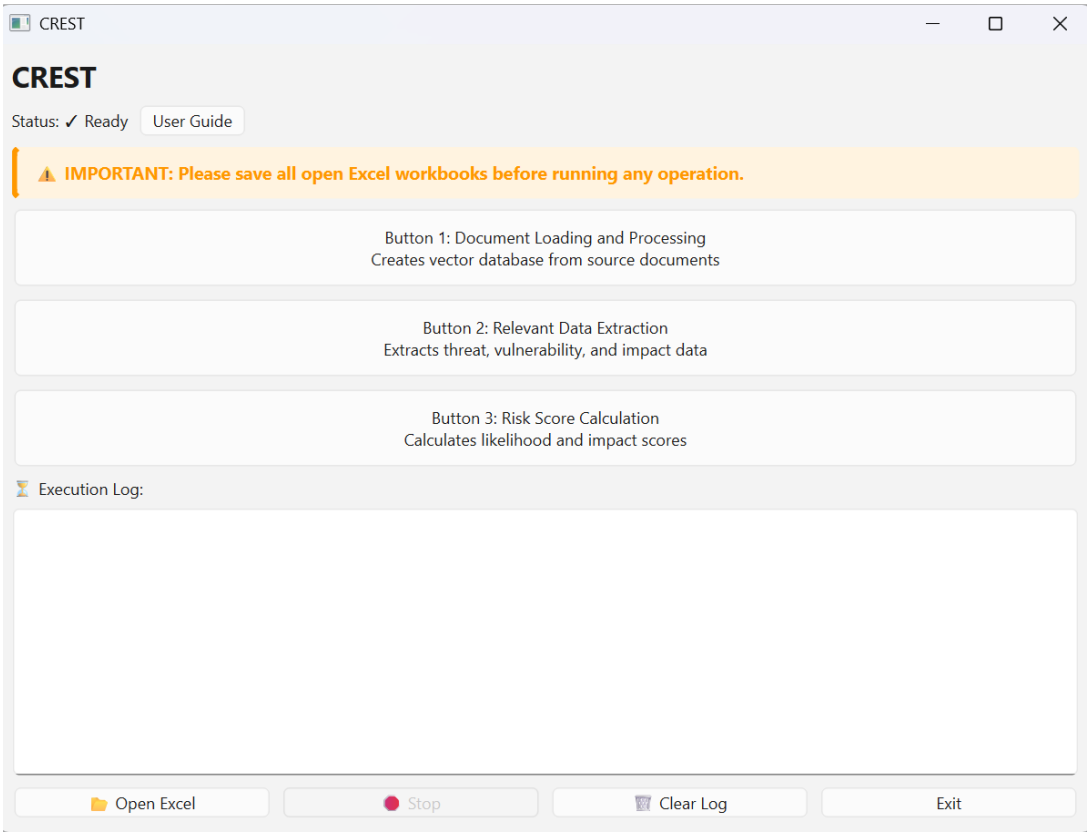


Figure 4.6: Control panel for running the workflow of CREST.

Figure 4.6 shows the control panel used to execute the main workflow of CREST. The interface is organised around three primary action buttons, each connected to one of the main processing stages in the workflow. The first button initiates document loading and preprocessing. The second button performs retrieval-based extraction and generates structured outputs for the scoring factors. The third button calculates likelihood and impact scores based on the generated outputs and the Excel-based assessment configuration.

In addition to the action buttons, the interface provides a process log that displays execution progress, summary messages, and error information. This is intended to improve usability during operation,

especially because some steps, such as document loading and retrieval-based analysis, may take time or depend on external model calls. The interface also includes a tool status indicator, a user guide button, and controls for opening the Excel workbook, stopping the current process, clearing the log, and exiting the control panel. Together, these elements provide a simple interaction model that reduces technical complexity for end users.

The Excel workbook serves as the main assessment workspace in which consultants provide context, review outputs, and adjust relevant settings. The Python-based control panel is integrated with the workbook by reading scenario inputs and scoring configuration from predefined sheets, and by writing generated reasoning, evidence references, and calculated scores back into predefined output cells. The workbook is not recreated during every run or execution. Consultant-provided context, scenario inputs, and configuration settings are retained, while generated output cells may be overwritten when the relevant processing step is run again. This keeps the processing logic separate from the assessment workbook while still allowing consultants to work in a familiar Excel environment.

4.3.2 Workbook structure

The Excel workbook is organised into a small number of core sheets, each with a distinct role in the assessment workflow. Table 4.1 summarises the main sheets used in the prototype.

Table 4.1: Core workbook structure of CREST

Sheet name	**Purpose**
Frontsheet	Stores consultant-provided contextual information such as organisation level, process, and system scope. This information is manually entered or adjusted and used for scope-aware justifications.
Risk Assessment	Contains scenario-related inputs and the structured outputs generated for likelihood, vulnerability, technical impact, and business impact assessment
Conf-RiskMatrix	Stores configurable factor definitions, descriptors, score mappings, and qualitative threshold ranges used in the scoring process

Together, these sheets provide the main workbook structure required to support CREST. The workbook functions not only as a place for entering and reviewing assessment data, but also as the main environment in which consultants inspect generated outputs and adjust scoring settings where needed.

4.3.3 Evidence retrieval architecture

This subsection presents the technical architecture of CREST's evidence retrieval component. While Section 4.2.2 described the role of data processing and extraction within the overall workflow, the focus here is on the main technical elements that prepare assessment documents for retrieval and store them in a searchable form.

In this prototype, FAISS is used as a local vector database for storing the document chunks. The index is generated by the Python processing component and stored as files in the prototype's working directory, together with metadata used for source tracing. It is not an external database server. Since the index can be regenerated from the original documents, retaining the source documents and metadata is more important for archival purposes, although keeping the FAISS index may support reproducibility of a specific run during evaluation.

The retrieval component is designed to process different types of assessment materials, including policy documents, gap assessment outputs, interview transcripts, and supporting Excel-based inputs. As summarised in Table 4.2, the component combines document parsing, document-type-sensitive chunking, metadata retention, embedding generation, and vector storage. Together, these elements transform diverse assessment materials into retrievable chunks while preserving contextual information needed for later review and source tracing.

Table 4.2: Technical architecture of the evidence retrieval component

Component	Details
PDF parsing	PDFMinerLoader as primary parser, with PyPDFLoader as fallback.
Excel parsing	Row-based parsing of tabular assessment files from client documents, where each row is treated as one chunk.
Chunking strategy	Hybrid chunking for structured documents such as policies and gap assessment outputs, preserving tables and paragraph boundaries where possible; paragraph-based chunking for less structured materials such as interview transcripts.
Metadata retention	Storage of source file names, section references, table headers, folder classification, and chunk length to support contextual retrieval and later traceability.
Embedding model	Embedding generation using a sentence-transformer model for vector representation of document chunks.
Vector storage output	Processed chunks stored as vectors in the FAISS database for later semantic retrieval.

Instead of retrieving information for the assessment case as a whole, this retrieval component selects document chunks relevant to specific scoring factors and risk context. Its role is not to interpret the evidence, but to identify and supply relevant and traceable document chunks to the reasoning component described in the next subsection.

4.3.4 LLM-assisted reasoning and explanation generation

This subsection describes CREST's LLM-assisted reasoning component, which transforms retrieved document chunks into structured, reviewable outputs for the scoring process. Building on the evidence retrieval architecture, the LLM-assisted reasoning component receives grounded input in the form of retrieved document chunks, relevant scoring criteria, and factor-specific prompts, and uses this input to generate factor-level analysis outputs.

Table 4.3 summarises the main elements of the LLM-assisted reasoning component. The component operates on factor-specific queries derived from the Excel-based assessment matrix. Separate queries are used for factors related to threat agent, vulnerability, technical impact, and business impact. The retrieved document chunks are then combined with the relevant scoring matrix and passed to the LLM, which generates structured output blocks consisting of a suggested descriptor, a short reasoning statement, and an evidence chain.

Table 4.3: Technical architecture of the LLM-assisted reasoning component

Component	Details
Factor-specific queries	Queries derived from the Excel-based assessment structure for threat agent, vulnerability, technical impact, and business impact factors.
LLM model interface	Internal LLM accessed through the host organisation's AI infrastructure.
Prompt mode	Factor-scoring prompts combining retrieved document evidence with the relevant scoring matrix and assessment context.
Evidence chain handling	Explicit reporting of document-grounded evidence and any separately flagged general knowledge used in the generated reasoning.

Continued on the next page

Component	Details
Structured output	Factor-level output block consisting of a suggested descriptor, a short reasoning statement, and an evidence chain indicating the basis of the suggestion.

Architecturally, this component is separated from the retrieval layer because its role is not to find relevant evidence, but to interpret the retrieved evidence in relation to the scoring model. This distinction helps preserve a clear separation between evidence selection and evidence interpretation. The resulting outputs remain reviewable by consultants and serve as structured input for the later likelihood and impact assessment stages.

Chapter 5

Evaluation

This chapter presents the evaluation results of CREST. The evaluation focuses on both technical performance and user perception of CREST in cyber risk assessment practice.

The chapter is organised into three main parts. First, it presents the configuration comparison results used to select the final CREST configuration. Second, it reports the evaluation results for scoring accuracy and reasoning traceability using the selected configuration. Third, it presents the consultant evaluation results, including questionnaire outcomes and qualitative feedback.

5.1 Configuration comparison results

This section presents the results of the configuration comparison step, in which six combinations of three LLM models and two retrieval depth settings were compared. The aim of this stage was to identify a suitable CREST configuration for the subsequent controlled case evaluation. Each configuration was applied to the same subset of four predefined risk scenarios and evaluated using scoring accuracy and reasoning traceability.

5.1.1 Comparison of tested combinations

Table 5.1 presents the scoring correctness of the six tested combinations across the four sampled scenarios. A value of *Correct* indicates that the generated final risk score matched the predefined ground truth, while *Incorrect* indicates a mismatch. For readability, Tables 5.1–5.3 use the short model labels gpt-oss, qwen3, and deepseek to refer to gpt-oss:20B, qwen3:30B, and deepseek-r1:32B, respectively.

Table 5.1: Scoring accuracy across the four sampled scenarios

Scenario	**gpt-oss** $k = 5$	**gpt-oss** $k = 10$	**qwen3** $k = 5$	**qwen3** $k = 10$	**deepseek** $k = 5$	**deepseek** $k = 10$
Scenario 1	Correct	Incorrect	Incorrect	Correct	Correct	Incorrect
Scenario 4	Incorrect	Incorrect	Correct	Correct	Incorrect	Correct
Scenario 8	Correct	Correct	Correct	Correct	Correct	Correct
Scenario 12	Correct	Correct	Correct	Correct	Correct	Correct
Accuracy	75%	50%	75%	100%	75%	75%

The results in Table 5.1 show that the tested combinations did not perform equally well across the four sampled scenarios. qwen3:30B with $k = 10$ was the only configuration that produced correct final risk scores for all four scenarios. By contrast, gpt-oss:20B with $k = 10$ achieved the lowest accuracy, with two incorrect predictions out of the four sampled cases.

Table 5.2 presents the average traceability scores obtained for each sampled scenario across the six tested combinations. These values show how traceable the generated reasoning was to supporting evidence in each case.

Table 5.2: Traceability scores across the four sampled scenarios

Scenario	**gpt-oss** $k = 5$	**gpt-oss** $k = 10$	**qwen3** $k = 5$	**qwen3** $k = 10$	**deepseek** $k = 5$	**deepseek** $k = 10$
Scenario 1	40.50%	42.50%	32.25%	44.75%	40.00%	38.50%
Scenario 4	63.25%	60.25%	73.50%	81.25%	46.75%	40.75%
Scenario 8	44.25%	46.75%	59.25%	52.00%	40.50%	39.75%
Scenario 12	67.75%	76.00%	79.50%	56.75%	44.25%	44.50%
Average	54%	56%	61%	59%	43%	41%

Table 5.2 shows that traceability also varied across both scenarios and configurations. The highest average traceability score was obtained by qwen3:30B with $k = 5$ at 61%, followed by qwen3:30B with $k = 10$ at 59%. The lowest average traceability score was observed for deepseek-r1:32B with $k = 10$ at 41%.

Table 5.3 summarises the average traceability, accuracy, and combined average score for all tested combinations.

Table 5.3: Summary of the tested model and retrieval depth combinations

No.	**Configuration**	**Traceability**	**Accuracy**	**Average**
1	qwen3 with $k = 10$	59%	100%	79%
2	qwen3 with $k = 5$	61%	75%	68%
3	gpt-oss with $k = 5$	54%	75%	64%
4	deepseek with $k = 5$	43%	75%	59%
5	deepseek with $k = 10$	41%	75%	58%
6	gpt-oss with $k = 10$	56%	50%	53%

In addition to the sample-based comparison above, Table 5.4 presents a supplementary full-set accuracy comparison across all 13 predefined scenarios. Unlike the sample-based comparison above, this comparison was used as a supplementary check to see whether the selected CREST configuration also remained strong when accuracy was assessed on the full scenario set. The supplementary check focuses only on accuracy, as scoring correctness was treated as the primary selection criterion and assessing traceability for all configurations across the full scenario set would require more manual evidence verification.

Table 5.4: Supplementary full-set accuracy comparison across the 13 predefined scenarios

Rank	**Configuration**	**Accuracy**
1	qwen3 with $k = 10$	77%
2	gpt-oss with $k = 5$	62%
3	deepseek with $k = 10$	54%
4	gpt-oss with $k = 10$	46%
5	qwen3 with $k = 5$	46%

Continued on the next page

Rank	Configuration	Accuracy
6	deepseek with $k = 5$	38%

The supplementary full-set accuracy comparison in Table 5.4 shows that qwen3:30B with $k = 10$ also achieved the highest accuracy when evaluated across all 13 predefined scenarios. This does not replace the sample-based comparison used for configuration selection, since the primary comparison needed to assess both accuracy and traceability on the same number of sampled cases. However, it provides additional support that the selected configuration remained the strongest performer in terms of scoring accuracy on the broader scenario set.

Unlike accuracy, traceability was not assessed across the full 13-scenario set for all configurations, as it required detailed manual review and verification of evidence links for each output. Therefore, the primary basis for configuration selection remained the sample-based comparison, where both metrics were evaluated on the same four sample scenarios.

Taken together, the sample-based comparison shows that the tested combinations did not perform equally well across the sampled scenarios. qwen3:30B with $k = 10$ achieved the highest overall average score at 79%, followed by qwen3:30B with $k = 5$ at 68%. The lowest overall score was observed for gpt-oss:20B with $k = 10$, which achieved an average of 53%.

The results also show that the configuration with the highest accuracy was not the one with the highest traceability. qwen3:30B with $k = 10$ achieved the highest accuracy in the sample-based comparison, whereas qwen3:30B with $k = 5$ obtained the highest traceability score. The additional full-set accuracy comparison further supports the selection of qwen3:30B with $k = 10$, as it also ranked highest when accuracy was assessed across all 13 predefined scenarios.

5.1.2 Selection of the final configuration

Based on the comparison results, qwen3:30B with $k = 10$ was selected as the final CREST configuration for the subsequent controlled case evaluation. This selection reflects the need for balanced configuration that performs well in risk score calculation while still providing sufficient and reviewable reasoning.

Although qwen3:30B with $k = 5$ achieved a slightly higher traceability score, qwen3:30B with $k = 10$ provided the strongest overall balance across the two comparison criteria used in this stage. In selecting the final configuration, scoring accuracy was given priority because traceable reasoning has less practical value if it supports an incorrect risk score. Traceability remains important for reviewability and justification, but it is most useful when the supported scoring outcome is sufficiently accurate. This selection was further supported by the supplementary full-set accuracy comparison, where qwen3:30B with $k = 10$ also achieved the highest accuracy across the 13 predefined scenarios. For this reason, it was selected as the final CREST configuration for the next evaluation step.

5.2 Controlled case evaluation results

This section presents the results of the controlled case evaluation. In this stage, CREST was used to assess the full set of predefined risk scenarios using the selected configuration, qwen3:30B with $k = 10$. The purpose of this stage was to assess the performance of the selected configuration beyond the initial four-scenario comparison subset by evaluating its scoring accuracy and reasoning traceability across the broader controlled case.

5.2.1 Accuracy findings

Table 5.5 presents the scoring results of CREST using the selected configuration across the 13 predefined risk scenarios. A value of *Correct* indicates that the generated final risk score matched the predefined ground truth, while *Incorrect* indicates a mismatch.

Table 5.5: Controlled case accuracy results for CREST using qwen3:30B with $k = 10$

Scenario	Expected risk	Generated risk	Check
1	Critical	Critical	Correct
2	Medium	Medium	Correct
3	High	Medium	Incorrect
4	Medium	Medium	Correct
5	High	High	Correct
6	Medium	Medium	Correct
7	Medium	Medium	Correct
8	High	High	Correct
9	Critical	Critical	Correct
10	Medium	Low	Incorrect
11	Low	Low	Correct
12	Low	Low	Correct
13	High	Medium	Incorrect
		Overall accuracy	**77%**

Overall, CREST achieved an accuracy of 77%, corresponding to 10 correct predictions out of 13 evaluated scenarios. This indicates that CREST was able to reproduce the predefined ground truth in most of the controlled cases. The three incorrect predictions all underestimated the predefined risk level, meaning that CREST generated a lower final risk score than expected in those scenarios.

As shown in Table 5.5, incorrect predictions were observed in Scenarios 3, 10, and 13. In all three cases, the generated risk level was lower than the predefined ground truth. Scenario 3 and Scenario 13 were expected to be high risk but were generated as medium risk, while Scenario 10 was expected to be medium risk but was generated as low risk. In the remaining ten scenarios, CREST produced the expected final risk score using the selected configuration.

5.2.2 Traceability findings

Table 5.6 presents the traceability scores of CREST using the selected configuration across the 13 predefined risk scenarios. For each scenario, the traceability score is calculated as the average of the assigned scores across the four assessed dimensions.

Table 5.6: Controlled case traceability results for CREST using qwen3:30B with $k = 10$

Scenario	Threat agent	Vulnerability	Technical impact	Business impact	Average
1	30.00%	44.00%	50.00%	55.00%	44.75%
2	60.00%	60.00%	55.00%	80.00%	63.75%
3	45.00%	76.00%	60.00%	85.00%	66.50%
4	85.00%	80.00%	85.00%	75.00%	81.25%
5	65.00%	76.00%	85.00%	85.00%	77.75%
6	70.00%	88.00%	50.00%	75.00%	70.75%
7	35.00%	36.00%	70.00%	55.00%	49.00%

Continued on the next page

Scenario	Threat agent	Vulnerability	Technical impact	Business impact	Average
8	45.00%	68.00%	55.00%	40.00%	52.00%
9	75.00%	72.00%	55.00%	70.00%	68.00%
10	20.00%	76.00%	90.00%	70.00%	64.00%
11	50.00%	28.00%	40.00%	25.00%	35.75%
12	50.00%	92.00%	50.00%	35.00%	56.75%
13	55.00%	40.00%	55.00%	70.00%	55.00%
				Overall traceability	**60%**

Using the selected configuration, CREST achieved an overall traceability score of 60% across the 13 evaluated scenarios. This indicates that the generated reasoning was, on average, partially traceable to the supporting evidence, although the level of traceability varied across cases.

As shown in Table 5.6, the highest traceability score was observed in Scenario 4 at 81.25%, followed by Scenario 5 at 77.75%. By contrast, the lowest traceability score was observed in Scenario 11 at 35.75%. These results suggest that the traceability of the generated reasoning was not uniform across the controlled case, but depended on the specific scenario and the supporting evidence available for it.

The traceability review showed several recurring failure modes. First, some retrieved document chunks were semantically relevant to the scoring factor but contextually misaligned. For example, the chunk text appeared relevant by topic, but the document source or title showed that it referred to a different organisational scope. Second, when the model relied on general cybersecurity knowledge, some sources were too broad and not specified. Third, some outputs contained inaccurate external reference identifiers, such as incorrect technique or framework reference numbers. These failure modes explain why some reasoning outputs were only partially or generically traceable, even when they appeared relevant at a high level.

5.3 Consultant evaluation results

This section presents the results of the consultant evaluation. The evaluation was conducted after the participating consultants had been introduced to CREST and its workflow. The purpose of this stage was to assess how the prototype was perceived from the consultant perspective, focusing on technology acceptance, usability, user experience, and socio-technical concerns. The questionnaire included structured items based on TAM, SUS, and UEQ, followed by open-ended questions on over-reliance, professional judgement, safeguards, and future improvements.

The results are interpreted descriptively because the consultant evaluation involved a small group of expert users. Therefore, the aim is not to make statistically generalisable claims, but to obtain structured practitioner feedback on whether CREST appears useful, usable, and acceptable as an early-stage decision-support prototype in cyber risk assessment practice.

5.3.1 Technology acceptance model

The Technology Acceptance Model (TAM) was used to obtain an initial indication of how the participating consultants perceived the usefulness and ease of use of CREST. As described in Chapter 3, the TAM responses were coded on a five-point Likert scale, where 1 represents 'strongly disagree' and 5 represents 'strongly agree'. For each TAM construct, the score was calculated by averaging the responses to all items belonging to that construct. The construct mean was also expressed as a percentage of the maximum possible score by dividing the mean by 5.

Table 5.7 summarises the TAM items used in the consultant evaluation. Perceived usefulness was measured using six items related to speed, quality, productivity, assessment objectives, ease of performing the assessment, and overall usefulness. Perceived ease of use was measured using six items related to ease of learning, ease of control, clarity of interaction, interaction fit, skill acquisition, and overall ease of

use. One additional item was used to capture willingness to use CREST in practice if it were available in practice.

Table 5.7: TAM constructs and questionnaire items

Construct	Item	Questionnaire statement
Perceived usefulness	Q11	Using this prototype would enable me to complete cyber risk assessment tasks more quickly.
	Q12	Using this prototype would improve the quality of my cyber risk assessment output.
	Q13	Using this prototype would help me complete more cyber risk assessment work within the same amount of time.
	Q14	Using this prototype would help me better achieve the objectives of cyber risk assessment.
	Q15	Using this prototype would make cyber risk assessment easier to perform.
	Q16	Overall, I would find this prototype useful in cyber risk assessment work.
Perceived ease of use	Q17	Learning to operate this prototype would be easy for me.
	Q18	I would find it easy to get this prototype to do what I want it to do.
	Q19	My interaction with this prototype would be clear and understandable.
	Q20	I would find it easy to interact with this prototype in the way I need.
	Q21	It would be easy for me to become skilful at using this prototype.
	Q22	Overall, I would find this prototype easy to use.
Willingness to use	Q23	If this prototype were available in practice, I would be willing to use it in my work.

Table 5.8 presents the TAM results at construct level for each participating consultant. To maintain anonymity, the respondents are shown as C1 to C4. The table reports construct-level scores, while the item-level descriptive results are provided in Appendix A.1. This aggregation is used because TAM is applied descriptively in this study to assess overall perceived usefulness, perceived ease of use, and willingness to use CREST in practice.

Table 5.8: TAM results by consultant

Construct	Items	C1	C2	C3	C4	Mean	Percentage
Perceived usefulness	Q11–Q16	4.00	3.67	4.67	4.00	4.08 / 5	81.7%
Perceived ease of use	Q17–Q22	3.67	4.67	4.00	3.67	4.00 / 5	80.0%
Willingness to use	Q23	4.00	4.00	4.00	4.00	4.00 / 5	80.0%

The perceived usefulness construct obtained a mean score of 4.08 out of 5, corresponding to 81.7% of the maximum possible score. This indicates that the participating consultants generally perceived CREST as useful for cyber risk assessment work. The perceived ease of use construct obtained a mean score of 4.00 out of 5, corresponding to 80.0% of the maximum possible score. This suggests that the prototype was also perceived as relatively easy to understand and operate after the participants had been introduced to the workflow. At consultant level, the scores show some variation, particularly for perceived ease of use, where C1 and C4 scored 3.67 while C2 scored 4.67. This suggests that ease of use may depend

on individual interaction experience or familiarity with the workflow. However, all consultants gave the same score of 4.00 for willingness to use, indicating a consistently positive but not unreserved willingness to use CREST in practice.

These findings should be interpreted as indicative practitioner feedback rather than statistically generalisable evidence of technology adoption, since the number of respondents is quite small. However, the results are still useful for the early evaluation of CREST because they provide structured feedback from experienced consultants who represent the intended user group of the prototype.

5.3.2 System usability scale

The System Usability Scale (SUS) was used to evaluate the perceived usability of CREST after the participating consultants had been introduced to the prototype and its workflow. The SUS score was calculated for each respondent using the standard scoring procedure described in Chapter 3. Each respondent therefore received one overall SUS score between 0 and 100.

Table 5.9 summarises the SUS items used in the questionnaire. The scale contains alternating positive and negative statements. This is why the scoring procedure adjusts odd-numbered and even-numbered items differently before calculating the final SUS score [4].

Table 5.9: SUS questionnaire items

Item	Questionnaire statement	Direction
Q1	I think that I would like to use this system frequently.	Positive
Q2	I found the system unnecessarily complex.	Negative
Q3	I thought the system was easy to use.	Positive
Q4	I think that I would need the support of a technical person to be able to use this system.	Negative
Q5	I found the various functions in this system were well integrated.	Positive
Q6	I thought there was too much inconsistency in this system.	Negative
Q7	I would imagine that most people would learn to use this system very quickly.	Positive
Q8	I found the system very cumbersome to use.	Negative
Q9	I felt very confident using the system.	Positive
Q10	I needed to learn a lot of things before I could get going with this system.	Negative

Table 5.10 shows the SUS scores for the four participating consultants. To maintain anonymity, the respondents are shown as C1 to C4. The table reports the final SUS score per consultant, while the item-level descriptive results are provided in Appendix A.2.

Table 5.10: SUS scores per consultant

Respondent	SUS score
C1	77.5
C2	100.0
C3	75.0
C4	72.5
Average	**81.25**

The average SUS score was 81.25 out of 100. This indicates a positive usability perception among the participating consultants. Individual scores ranged from 72.5 to 100.0, meaning that all respondents rated the prototype positively from a usability perspective. The score of C2 was higher than the others,

suggesting some variation in perceived usability across participants. However, since the remaining scores were also above 70, the overall pattern still shows a positive usability evaluation.

The results suggest that the Excel-based workbook and Python-based control panel did not create major usability issues during the guided evaluation. This is relevant because CREST is intended to support consultants within an existing assessment workflow rather than requiring them to learn a completely new assessment environment. At the same time, the SUS result should be interpreted as indicative feedback rather than a generalisable usability benchmark, as the evaluation involved a small group of expert users.

5.3.3 User experience questionnaire

The User Experience Questionnaire (UEQ) was used to assess broader user experience aspects of CREST beyond usability and acceptance. As described in Chapter 3, UEQ responses were converted to a scale from -3 to $+3$, where higher values indicate a more positive evaluation. The converted item scores were then averaged per UEQ scale. Based on the UEQ interpretation guidance, mean values between -0.8 and 0.8 indicate a neutral evaluation, values above 0.8 indicate a positive evaluation, and values below -0.8 indicate a negative evaluation [31].

The UEQ has 26 semantic-differential items grouped into six scales. These scales capture different aspects of user experience: attractiveness, perspicuity, efficiency, dependability, stimulation, and novelty. Table 5.11 summarises the scale structure and questionnaire items used in the consultant evaluation [31].

Table 5.11: UEQ scales and questionnaire items

Scale	**Item**	**Semantic differential pair**
Attractiveness	1	annoying / enjoyable
	12	good / bad
	14	unlikable / pleasing
	16	unpleasant / pleasant
	24	attractive / unattractive
	25	friendly / unfriendly
Perspicuity	2	not understandable / understandable
	4	easy to learn / difficult to learn
	13	complicated / easy
	21	clear / confusing
Efficiency	9	fast / slow
	20	inefficient / efficient
	22	impractical / practical
	23	organised / cluttered
Dependability	8	unpredictable / predictable
	11	obstructive / supportive
	17	secure / not secure
	19	meets expectations / does not meet expectations
Stimulation	5	valuable / inferior
	6	boring / exciting
	7	not interesting / interesting
	18	motivating / demotivating
Novelty	3	creative / dull
	10	inventive / conventional
	15	usual / leading edge

Continued on next page

Table 5.11: UEQ scales and questionnaire items (continued)

Scale	Item	Semantic differential pair
	26	conservative / innovative

Table 5.12 presents the UEQ results at scale level for each participating consultant. To maintain anonymity, the respondents are shown as C1 to C4. The table reports the average score per UEQ scale, while the item-level descriptive results are provided in Appendix A.3. All six scales obtained mean scores above 0.8, indicating a positive evaluation across the measured user experience dimensions.

Table 5.12: UEQ results by consultant

UEQ scale	C1	C2	C3	C4	Mean
Attractiveness	1.33	1.17	1.50	1.67	1.42
Perspicuity	0.50	2.75	1.50	1.00	1.44
Efficiency	1.25	2.00	1.25	2.00	1.63
Dependability	1.00	1.50	1.25	1.00	1.19
Stimulation	1.50	1.75	2.25	2.25	1.94
Novelty	0.25	1.00	1.50	1.50	1.06

The highest score was obtained for stimulation, with a mean score of 1.94. This suggests that the participating consultants perceived CREST as interesting and motivating to use. Efficiency also received a relatively high score of 1.63, indicating that the prototype was perceived as supporting task completion without unnecessary effort [31].

Novelty and dependability obtained the lowest scores, with mean scores of 1.06 and 1.19 respectively. Although these values are still within the positive range, they suggest that consultants were more cautious in evaluating how innovative the prototype felt and how predictable and dependable the interaction with CREST was. This is consistent with the prototype nature of CREST, where the workflow was demonstrated in a guided setting and further improvement would be needed before deployment [31].

At consultant level, the largest variation appears in perspicuity and novelty. C1 gave lower scores for these scales than the other consultants, suggesting that perceptions of how easy CREST was to understand and how innovative it felt may depend more strongly on individual interaction experience.

To further contextualise these results, the UEQ handbook also provides benchmark values for business software based on 158 product evaluations. Table 5.13 shows the benchmark values used for interpretation.

Table 5.13: UEQ benchmark values for business software, based on 158 product evaluations [31]

Category	Attract.	Persp.	Efficiency	Depend.	Stim.	Novelty
Excellent	1.72	1.85	1.70	1.61	1.53	1.49
Good	1.50	1.48	1.38	1.37	1.34	1.03
Above average	1.11	1.03	0.89	1.04	1.04	0.64
Below average	0.70	0.58	0.51	0.71	0.71	0.24

Compared with this benchmark, CREST was evaluated positively across all UEQ scales. Attractiveness, perspicuity, and dependability fall into the 'above average' category. Efficiency and novelty fall into the 'good' category, while stimulation falls into the 'excellent' category. This shows that the UEQ results are not only positive in general, but also compare quite well with benchmarked business software products.

Overall, the UEQ results show a positive user experience across all measured dimensions. Similar to the TAM and SUS results, these findings should be interpreted as descriptive feedback from a small group of expert users rather than as generalisable evidence of user experience quality.

5.3.4 Consultant feedback results

The structured questionnaire results were complemented with qualitative feedback from the participating consultants. This feedback helps explain how the consultants interpreted the prototype apart from the numerical TAM, SUS, and UEQ scores. In particular, it provides insight into what they found useful, where they saw limitations, and what would need to improve before CREST could be used in practice.

The feedback is presented in three parts. The first part focuses on usability and acceptance, including comments on the practical workflow and perceived usefulness of the prototype. The second part focuses on socio-technical issues, such as over-reliance, trust, accountability, and the preservation of professional judgement. The third part summarises broader expert interview findings on how CREST could fit into cyber risk assessment work.

5.3.4.1 Usability and acceptance feedback

In general, the qualitative feedback supported the structured questionnaire results by indicating that consultants saw practical value in CREST, especially as a way to make risk analysis more structured, transparent, and easier to review. The feedback also showed that acceptance depends on whether the tool remains integrated in the consultant workflow and produces outputs that are meaningful for both consultants and their clients.

Table 5.14: Summary of usability and acceptance feedback

Theme	**Consultant feedback**	**Implication for CREST**
Positive first impression	Consultants described CREST as interesting, promising, practical, and valuable.	Consistent with the questionnaire results, which indicated positive perceptions of usefulness, usability, and user experience.
Structured starting point	CREST was seen as useful for organising evidence and generating an initial view for discussion.	Confirms that CREST should be seen as decision support rather than automated decision-making tool.
Workflow familiarity	The Excel workbook, user guide, video demonstration, and simple control panel helped consultants understand the workflow.	Suggests that the Excel-centred design reduced adoption complexity during the guided evaluation.
Need for review and validation	Consultants emphasised that generated outputs should be reviewed, validated, and corrected rather than accepted automatically.	Supports the design of CREST that uses human-in-the-loop principle and clear review points in the workflow.
Business relevance	Consultants noted that outputs should be meaningful for business stakeholders, not only technical consultants.	Shows a need for improvement to include clearer business-facing risk statements and stronger links to business consequences in the output.

Continued on next page

Table 5.14: Summary of usability and acceptance feedback (continued)

Theme	Consultant feedback	Implication for CREST
Future workflow integration	Suggested improvements included a possible web-based version to reduce local system dependencies, clearer input preparation guidance, and stronger links to other risk assessment phases from risk identification to risk treatment to CREST workflow.	Shows that adoption would depend on integration with broader risk management phases and reporting workflows.

Overall, the feedback indicates that CREST was accepted as a potentially useful decision-support prototype, especially when considered as a tool that structures judgement and initial analysis instead of replacing consultant judgement. The overall comments on practicality and structured output indicated positive perceptions of usefulness, usability, and user experience. At the same time, the feedback shows that practical adoption would depend on further improvement, especially around business-facing risk framing, web-based version, and integration with broader risk management phases such as risk identification, risk evaluation, and risk treatment.

5.3.4.2 Socio-technical feedback

The consultant feedback also mentioned several socio-technical concerns related to the use of AI-assisted decision support tool like CREST in risk assessment. These concerns mainly relate to how consultants may interact with generated outputs, how professional judgement may be affected, and what safeguards are needed to prevent the tool from being treated as an automated decision maker by the users. Table 5.15 summarises the main socio-technical themes identified from the consultant feedback.

Table 5.15: Summary of socio-technical feedback

Theme	Consultant feedback	Implication for CREST
Risk of over-reliance	Consultants noted that users, especially less experienced consultants, may treat AI-generated outputs as correct by default if the tool appears to work well.	CREST should make it explicit that review and validation are mandatory, so generated outputs are treated as suggestions rather than final assessment decisions.
Shift in professional judgement	Feedback suggested that consultants may start from the AI-generated view and then validate it, instead of forming an initial judgement independently.	This aligns with the intended role of CREST as decision support rather than decision automation. Consultants still need sufficient training to critically review the output and recognise when reasoning or scoring needs correction.
Contextual interpretation risk	Consultants emphasised that risk assessment depends heavily on business context, system scope, sector, regulatory requirements, and organisational priorities. Generic outputs may then miss important insights.	CREST needs to support context-aware interpretation and clearly highlight assumptions, limitations, and evidence gaps in generated outputs.

Continued on next page

Table 5.15: Summary of socio-technical feedback (continued)

Theme	**Consultant feedback**	**Implication for CREST**
Incomplete or misleading outputs	Consultants mentioned that weak or incomplete outputs could result in incorrect or unactionable information, which may reduce trust in the assessment as a whole by misleading the stakeholders.	Generated reasoning should be traceable to evidence and should indicate where supporting evidence is weak, missing, or based on general knowledge.
Human validation as safeguard	Several consultants emphasised the need for consultants to review, confirm, or adjust outputs before they are finalised.	Human-in-the-loop validation should remain part of the workflow, with explicit review points before any result is used in reporting or decision-making.
Implementation and governance	One consultant noted that safeguards are not only a design issue, but also an implementation issue. Proper methodology, validation guidelines, and checklists are needed.	Adoption of CREST would require clear usage procedures, review responsibilities, and governance rules around how outputs may be relied on.

Overall, the socio-technical feedback supports the need to position CREST as a decision-support tool rather than an automated risk assessment system. Consultants saw value in AI-assisted structuring of evidence and reasoning, but also emphasised that generated outputs must remain reviewable, traceable, and based on professional judgement. This supports the human-in-the-loop design objective of CREST and shows that practical adoption would require not only technical improvement, but also clear review procedures and governance safeguards.

Chapter 6

Discussion

This chapter discusses the findings of the evaluation in relation to the research objective of this study. While Chapter 5 presented the evaluation results, the purpose of this chapter is to interpret what these results suggest about CREST, its practical relevance for cyber risk assessment, and the broader implications of introducing AI-assisted decision support into judgement-intensive consulting work.

The chapter is structured into three main parts. First, it interprets the findings from the different evaluation stages, including the configuration comparison, controlled case evaluation, and consultant evaluation. Second, it discusses the implications of these findings for cyber risk assessment practice, with particular attention to consistency, reasoning structure, traceability, and the role of CREST in supporting rather than replacing consultant judgement. Third, it reflects on the socio-technical implications of the proposed approach, including risks of over-reliance, changes in professional judgement practices, and issues of trust, explainability, and accountability. Finally, it outlines the main limitations of the study.

6.1 Interpretation of evaluation findings

This section discusses what the results presented in Chapter 5 mean for CREST. The aim is not to repeat the reported results, but to interpret what they suggest about CREST's technical performance, its practical value, and its ability to support more consistent, transparent, and traceable cyber risk scoring. The discussion is structured around the three evaluation stages: configuration comparison, controlled case evaluation, and consultant evaluation.

6.1.1 Configuration comparison and technical performance

The configuration comparison was an important first step in the evaluation, because it allowed the study to identify a suitable technical setup before using CREST to assess the full controlled case. This stage was not intended mainly to technically optimise the prototype, but to examine which combination of model and retrieval depth was most suitable for a decision-support task in which both scoring accuracy and reasoning traceability matter. In this respect, the results are important not only as a basis for configuration selection, but also as an early indication of how different technical choices affect CREST's ability to support judgement-intensive risk scoring.

The results showed clear performance differences across the six tested combinations. The two qwen3:30B configurations achieved the strongest overall results, with qwen3:30B with $k = 10$ producing the highest overall average score and qwen3:30B with $k = 5$ producing the highest average traceability score. The gpt-oss:20B configurations showed more mixed performance, while the deepseek-r1:32B configurations remained lower in traceability across both retrieval settings. This suggests that technical performance in this study was shaped by the interaction between LLM model and retrieval setting, rather than by the choice of the model alone. It also indicates that configurations that appear similar at a general level can still behave differently when applied to the same task and evidence base. This interpretation is also supported by the supplementary full-set accuracy comparison in Chapter 5, which showed that qwen3:30B with $k = 10$ remained the highest-performing configuration when accuracy was assessed across all 13 predefined scenarios.

Table 6.1 summarises the main strengths and limitations observed across the tested configurations, based on the comparison results presented in Chapter 5.

Table 6.1: Interpretive summary of the tested configurations

Configuration	**Main strength**	**Main limitation**
qwen3:30B with $k = 10$	Highest overall accuracy and strongest overall balance across the comparison criteria	Traceability was slightly lower than qwen3:30B with $k = 5$, partly because some outputs contained inaccurate external references such as incorrect attack technique or framework reference numbers.
qwen3:30B with $k = 5$	Highest average traceability and generally strong reasoning quality	Lower overall accuracy than qwen3:30B with $k = 10$
gpt-oss:20B with $k = 5$	Moderate overall performance and relatively stable traceability	Reasoning was generally less detailed than qwen3:30B
gpt-oss:20B with $k = 10$	Slightly higher traceability than the $k = 5$ setting	Lowest overall score among the tested combinations
deepseek-r1:32B with $k = 5$	Comparable accuracy to several other non-selected configurations	Lower traceability and weaker evidence linking than qwen3:30B
deepseek-r1:32B with $k = 10$	Larger retrieval setting allowed more contextual input	Lowest traceability performance, suggesting that additional retrieved chunks did not improve reasoning quality in this setup

A first important pattern is that scoring accuracy and traceability did not improve in exactly the same way. The configuration with the highest accuracy was not the same as the configuration with the highest traceability. This also explains why the final configuration selection prioritised scoring accuracy while still treating traceability as necessary for reviewability and justification. This matters because CREST is not intended only to generate a valid score, but also to support a reviewable reasoning process. In other words, a technically stronger configuration for this study was not simply the one that matched the ground truth most often, but the one that provided the strongest balance between correct scoring and verifiable justification. This is closely related to the solution objective of supporting transparent and structured reasoning rather than reliable prediction alone.

The comparison also suggests that the tested models are different not only in score outputs, but also in the way they constructed the reasoning. Based on the reviewed outputs, qwen3:30B generally produced the most detailed reasoning among the tested models, which contributed to its stronger overall performance. The qwen3:30B model used in this study corresponds to the Qwen3-30B-A3B released as part of the Qwen3 series in April 2025 [29], while gpt-oss:20B was released in August 2025 [25]. However, release date alone does not determine the reliability or recency of external references generated by a model. Some qwen3:30B outputs contained incorrect or inconsistent external references, such as inaccurate MITRE ATT&CK reference numbers. These issues may therefore indicate model-specific reference reliability issues, potentially related to training data, post-training, or hallucinated citation details, rather than simply because the model is older. This seems to have reduced its traceability score in some cases, even when the reasoning itself appeared relatively strong. By contrast, gpt-oss:20B produced less detailed reasoning overall, but its external references appeared more consistent and more recent in the reviewed outputs. This may explain why its traceability performance remained comparatively stable even though its overall reasoning depth was lower.

The results further suggest that deepseek-r1:32B was not very suitable to do the task in the tested setup. It showed the weakest traceability performance among the tested models, and increasing the retrieval depth from $k = 5$ to $k = 10$ did not improve its overall results. In some reviewed outputs, the larger retrieval setting also appeared to result in incomplete reasoning, including cases where the

evidence list was missing. A possible interpretation is that providing more retrieved chunks did not help this model connect its reasoning more clearly to the available evidence. While this study does not allow a conclusive technical explanation for this behaviour, it does show that adding more context is not automatically beneficial. For this prototype, the usefulness of retrieved evidence depended not only on how much information was provided, but also on how effectively the model could use it in a structured and traceable way.

This finding is relevant beyond the ranking of the six tested combinations. In many AI-assisted workflows, retrieval depth may be treated as a parameter that simply needs to be increased to provide more context. The present results shows that this assumption is not always applicable in judgement-intensive assessment tasks. Because CREST relies on factor-level evidence interpretation and reasoning generation, more retrieved chunks may also create additional noise or make it harder for the model to maintain clear connections between claims and sources. As a result, retrieval depth should be considered as part of a broader configuration setup rather than as a setting that is improved by simply increasing it.

Taken together, the configuration comparison supports the broader research objective of this study in two ways. First, it shows that CREST's technical setup has a direct effect on whether the resulting outputs are not only accurate, but also reviewable and traceable. Second, it shows that technical capability alone is not sufficient, what matters is whether the selected configuration supports the specific design goals of CREST, such as more consistent scoring, more explicit reasoning, and better traceability while still leaving room for consultant review. From this perspective, qwen3:30B with $k = 10$ was selected as the most suitable configuration, not because it was strongest on every aspect individually, but because it provided the best overall balance for the next evaluation stage.

6.1.2 Controlled case evaluation

The controlled case evaluation provides the main explanation of how CREST performs when it was applied to the full set of predefined risk scenarios. The results show that CREST reproduced the predefined ground truth in 10 out of 13 scenarios, which shows an overall accuracy of 77%. This suggests that the selected configuration was able to support risk estimation in a controlled case and provides initial evidence that the proposed decision-support approach is feasible for structuring cyber risk scoring.

However, the error pattern is important to further analyse. The three incorrect predictions all underestimated the risk level in the predefined ground truth. This is a relevant finding because underestimation is more problematic than general disagreement. It may cause a risk to appear less severe than it really is, which can affect prioritisation, treatment decisions, and risk communication. Thus, the controlled case evaluation does not only show how CREST performed, but also that its outputs require careful interpretation.

The traceability results provide another perspective on the controlled case evaluation result. CREST achieved an overall traceability score of 60%, showing that the generated reasoning was in general partially traceable to its supporting evidence. This matters because the aim of CREST is not only to produce a risk score, but also to make the basis of that score more explicit and reviewable. A high accuracy score without sufficient traceability would be less useful in consulting practice, because consultants still need to understand why a score was suggested and whether the supporting evidence is appropriate.

At the same time, the fact that accuracy was not perfect also helps to support the intended role of the artifact. CREST should not be positioned as a tool that produces final risk scores with full certainty. Instead, the controlled case evaluation results show that it can provide a structured starting point for risk estimation, while still requiring critical assessment of the generated reasoning and evidence chain. This is especially important because a perfectly accurate tool may also increase user expectations and the risk of over-reliance if its limitations are not made visible.

The variation in traceability across different scenarios shows that CREST's performance is sensitive not only to the availability of evidence, but also to how well retrieved evidence matches the specific risk context. In some cases, retrieved chunks were related to the assessed factor in terms of topic, but misaligned with the scenario in terms of context. For example, evidence about access control may be retrieved because it matches the general topic of access risk, even when the specific policy section concerns a different context, such as supplier access rather than internal organisational access. This reduces traceability because the generated reasoning may appear evidence-based while relying on evidence that is only partially relevant.

This suggests that the RAG pipeline and prompting strategy require further refinement. Prompts should more explicitly require the model to differentiate between evidence contexts before using retrieved chunks as support, while retrieval could be improved through richer metadata or context-aware filtering. The controlled case was based mainly on PDF and Excel-based assessment materials, which provided a realistic but still limited evaluation setting. In less controlled assessment environments, evidence may be more fragmented, incomplete, or represented in different formats. Therefore, the results should be interpreted as evidence of feasibility under controlled conditions, instead of as proof that the tool would perform equally well across all real client contexts.

Overall, the controlled case evaluation shows that CREST can support structured risk estimation and produce useful scoring suggestions in most cases. Its main value lies in making suggested scores, reasoning, and supporting evidence more explicit. However, the underestimation errors and partial traceability also show that further improvement is needed, particularly in strengthening evidence retrieval, improving traceability, and making uncertainty or weak evidence more visible in the output.

6.1.3 Consultant evaluation findings

The consultant evaluation findings provide an explanation of how CREST was perceived by its intended users. Overall, the results were positive across the structured questionnaire instruments. The TAM results suggest that consultants perceived CREST as useful and quite easy to use. The SUS result shows that the prototype did not create major usability issues during the guided evaluation. The UEQ results further show a positive user experience across all six measured dimensions. Taken together, these findings suggest that CREST was not only technically feasible, but also understandable and acceptable to the consultants who evaluated it.

The use of TAM, SUS, and UEQ also provided complementary perspectives even though the instruments partly overlap. TAM was useful for understanding whether consultants saw CREST as useful and easy to adopt in their work. SUS focused more specifically on perceived usability of the system interaction. UEQ added a broader view of user experience, including dimensions such as stimulation, dependability, and novelty. Therefore, the combined results provide a more rounded and comprehensive interpretation than any single instrument would provide. Importantly, the results suggest that consultants did not only see the tool as useful, but also found the interaction manageable and the overall experience positive.

This is important because CREST is designed as an interactive decision-support tool rather than a purely technical scoring model. Its practical value depends not only on whether it can generate risk scores, but also on whether consultants can understand, review, and integrate its outputs into their assessment workflow. The positive user perception results support the design choice to use a familiar Excel-based workspace. However, the evaluation also showed that the Python-based control panel may create deployment dependencies when run on different user environments. This suggests that while the Excel-centred design was suitable for prototyping and quick demonstration, a future deployment should reduce environment-specific dependencies, for example through web-based interface or a containerised setup using docker.

At the same time, the consultant feedback shows that acceptance is conditional. Consultants did not interpret CREST as a tool that could be directly deployed without further improvement. Instead, they saw value in the prototype as a structured starting point for risk analysis, especially where it helps to organise evidence, make reasoning more explicit, and provide an initial scoring suggestion for further review. This aligns with the intended role of CREST as decision support rather than decision automation. Acceptance also depends on the consultant's role and daily work. One consultant who participated in the evaluation was less directly involved in the type of risk assessment workflow targeted by CREST, which may explain some more neutral responses regarding the willingness to use the tool in practice. This shows that perceived usefulness may vary depending on how closely the tool matches the user's actual responsibilities.

The qualitative feedback also highlights several practical requirements for future adoption. Consultants emphasised that outputs should be meaningful not only for technical users, but also for business stakeholders. This suggests that future versions of CREST should provide clearer business-facing risk statements, stronger connections between risk scenarios and business consequences, and more concise summaries that can be used in reporting or risk register. In addition, feedback on simplified outputs, clearer input preparation, and links to other risk assessment phases such as risk identification, risk evaluation, and risk treatment shows that adoption would also depend on how well the tool fits into the broader

risk management workflow, not only on whether it can produce factor-level reasoning and scores.

The socio-technical feedback adds an additional layer to this interpretation. Consultants recognised the value of AI-assisted structuring, but also had concerns about over-reliance and the changing role of professional judgement. A key challenge is therefore not simply to increase trust in CREST, but to balance that trust appropriately. If consultants' trust towards the tool is too low, they may ignore its outputs and the tool will not be adopted or used at all. If they trust is too high, they may accept generated reasoning and scores without sufficient review. The ideal position is between these extremes, CREST should be trusted enough to be useful as a structured starting point, but not treated as something that removes the need for consultant assessment.

This also has implications for consultant training. The use of AI-assisted tools may shift some junior-level responsibilities from manually producing basic analysis towards reviewing, validating, and improving AI-generated drafts. This does not necessarily remove the need for junior consultant roles, but it changes what they need to learn. Training may need to focus more explicitly on risk assessment principles, evidence evaluation, contextual judgement, and the ability to recognise when AI-generated reasoning is appropriate or requires correction. Thus, CREST may change not only the assessment workflow, but also the way consultants develop judgement over time.

Overall, the consultant evaluation strengthens the interpretation of CREST as a promising but still early-stage decision-support tool. The positive questionnaire results show promising perceptions of usefulness, usability, and user experience, while the qualitative feedback clarifies the conditions needed by CREST to be more useful in practice.

6.2 Implications for cyber risk assessment practice

This section discusses the practical implications of the findings for cyber risk assessment work in consulting settings. Building on the interpretation of the evaluation results, the focus here is on what CREST may contribute to practice, particularly in relation to consistency, reasoning structure, traceability, and the appropriate role of AI-assisted support.

6.2.1 Support for consistency and structured reasoning

One of the main implications of this study for cyber risk assessment practice is that AI-assisted decision support may help reduce some of the inconsistency that appears when consultants translate diverse evidence into likelihood and impact scores. As discussed in Chapters 2 and 3, judgement-intensive scoring under non-prescriptive standards often relies on implicit reasoning, individual experience, and interpretation of available evidence. CREST does not remove these sources of variability completely, but it introduces a more structured workflow for how evidence is retrieved, interpreted, and translated into scoring inputs.

While a more structured workflow could also be achieved through templates or manual checklists, the use of LLMs is important because it reduces the effort required to create and document an explicit chain of evidence. Without AI assistance, consultants could still document such reasoning manually, but doing so consistently across multiple scenarios, factors, and evidence sources would increase assessment effort significantly. In practice, these manual and time-consuming tasks may lead consultants to keep parts of the reasoning implicit. CREST therefore does not only add structure to the workflow, but also lowers the practical burden of making evidence interpretation and scoring justification more explicit.

This is particularly relevant in consulting practice, where the same assessment framework may still be applied differently across consultants or engagements. By using a fixed assessment structure, predefined factors, and explicit descriptor-based reasoning, CREST provides a more consistent basis for moving from assessment data to scoring suggestions. The evaluation results show that this structure can support more stable scoring behaviour, especially when compared with a fully implicit process where reasoning is still mostly undocumented.

At the same time, the findings also suggest that consistency in this study is better understood as reducing unnecessary variation in reasoning and score justification, rather than eliminating variation altogether. CREST still depends on model behaviour, retrieval configuration, and the quality of the available evidence. In this sense, the contribution of the prototype is not that it eliminates judgement completely, but that it helps structure it. This is an important difference, because cyber risk assessment often requires

interpretation under incomplete, ambiguous, or context-dependent conditions. What CREST offers is therefore a more explicit and repeatable reasoning process, rather than a fixed or universally correct scoring mechanism.

Another related implication is the structuring of reasoning itself. In many consulting settings, the logic behind a score may remain partly internal to the consultant and may only be reflected indirectly in the final workbooks. CREST makes this reasoning process more visible and clearer by generating factor-level outputs that include a suggested descriptor, short reasoning, and supporting evidence. This can help consultants better articulate and explain why a particular score is being suggested, which may in turn support internal review, peer discussion, and more defensible assessment outcomes.

6.2.2 Traceability and reviewability in consulting practice

The second implication of the findings is that traceability can be improved through AI-assisted support, but that it remains a challenging requirement in practice. In cyber risk assessment, the ability to review how a score was calculated is important because scoring outcomes may later need to be discussed with colleagues, validated internally, or explained to clients. When reasoning is scattered across documents, interview notes, control findings, and consultant interpretation, this review process can become difficult and time-consuming.

CREST addresses this challenge by linking the LLM-generated reasoning to retrieved evidence and by presenting this linkage in a more explicit form. The traceability results suggest that this approach has practical value, because it makes the reasoning behind a score more reviewable than in a process where evidence usage remains mostly implicit. In this sense, CREST supports not only score generation, but also the documentation and verification of how a proposed score was formed.

However, the results also show that traceability is not guaranteed simply because evidence retrieval is included in the workflow. Even in the selected configuration, the traceability was still imperfect and varied across scenarios. This shows that evidence-grounded reasoning is still sensitive to model behaviour, prompt interpretation, use of external knowledge, and the quality of the retrieved chunks. In some cases, the generated reasoning was relatively strong while the supporting references were incomplete, inconsistent, or not precise. This means that reviewability is improved, but not fully reliable.

A further issue is related to the difference between document-based reasoning and reasoning supported by the model's general knowledge. In the design of CREST, the prompt first encourages the model to base its reasoning on retrieved document chunks. However, when relevant evidence is not sufficiently available in the retrieved materials, the model may still rely partly on general knowledge. This creates a clear difference in traceability. When reasoning is based on retrieved document chunks, the supporting statements can in principle be checked against the source text more directly. However, when reasoning is based on general knowledge, the resulting references are often broader, less specific, and more difficult to verify in the same way. Even when the prompt asks the model to indicate the basis of such knowledge, the outputs may still contain generic or insufficiently traceable references. This means that, in practice, strong traceability often depends on a combination of document-based evidence and model-supported interpretation, but the latter is inherently more difficult to verify in a precise way.

This difference also helps explain some of the traceability issues observed in the selected configuration. For example, qwen3:30B often produced detailed and credible reasoning, but some reviewed outputs referred to outdated or inconsistent external sources, such as older reports or incorrect reference numbers. This suggests that reasoning quality alone is not sufficient for strong traceability. A model may generate output that appears convincing, while still relying on references that are difficult to verify or that do not reflect the most current source base.

This has two important implications. First, traceability should be treated as a core evaluation aspect in AI-assisted cyber risk assessment, rather than as a secondary feature. A model that produces the most accurate scores but cannot reliably connect them to identifiable evidence may be of limited use in consulting work, where defensibility and explainability are often as important as the score itself. Second, traceability in this case should be understood as something that still requires human checking. CREST can make the reasoning path more explicit, but consultants still need to verify whether the cited evidence is relevant, correctly interpreted, and sufficiently supportive of the assigned score.

Seen in this way, the value of CREST lies not in producing perfectly traceable outputs in every case, but in shifting cyber risk assessment practice toward a more reviewable form of reasoning. This aligns

with the broader need identified in Chapters 2 and 3 for more transparent and explicit justification in judgement-intensive scoring work.

6.2.3 Role of CREST as decision support rather than decision maker

The findings of this study also emphasize the importance of positioning CREST as a decision-support tool rather than a decision maker. This distinction is fundamental to the design of CREST and crucial in interpreting its practical value. Although the selected configuration achieved reasonably good results in terms of accuracy and traceability, the evaluation also showed that the outputs were not consistently correct or fully traceable across all scenarios. For this reason, CREST should not be treated as a system that can automatically and reliably determine final risk scores without human oversight.

More importantly, cyber risk scoring is not only a technical classification problem. As discussed in Chapter 2, it is a judgement-intensive activity that depends on contextual understanding, interpretation of incomplete evidence, and professional experience. In many practical cases, consultants have to balance client data with contextual considerations that are difficult to be fully translated into a fixed computational process. CREST therefore provides structured support for this work, but does not remove the need for human interpretation.

From a practical perspective, this means that the prototype is best understood as supporting three tasks. First, it helps organise and retrieve relevant evidence from diverse assessment data. Second, it helps structure reasoning by translating retrieved evidence into factor-level scoring justification. Third, it helps make the basis of a proposed score more visible and reviewable. However, the final judgement still needs to remain with the consultant, who must decide whether the generated output is appropriate considering the overall assessment context.

This positioning analysis is important not only for academic reasons, but also for professional and organisational reasons. In consulting practice, risk assessments are often used to support control recommendations, treatment decisions, and client communication. These are activities where accountability cannot easily be delegated to a machine or AI component. Positioning CREST as decision support therefore preserves consultant responsibility and accountability while still allowing the benefits of AI-assisted structuring, retrieval, and reasoning support to be used in practice.

The findings of this study suggest that this human-in-the-loop positioning is not merely a cautious design choice, but a necessary one. CREST can improve consistency, structure, and traceability, but its outputs still require interpretation and verification. In this sense, the practical value of CREST is not about replacing consultant judgement, but about strengthening the process where that judgement is applied and documented.

6.3 Limitations

This research should also be interpreted considering several limitations. While the study provides useful insights into the design and evaluation of AI-assisted decision support for cyber risk assessment, the findings are influenced by the scale of the evaluation, the interpretative nature of the scoring task, and the specific prototype and data context used in this research. These limitations are discussed in the following subsections.

6.3.1 Evaluation scope and sample size

The first limitation of this study is related to the scope of the evaluation and the relatively small number of scenarios being tested. The configuration comparison was conducted on a targeted subset of four predefined risk scenarios, while the controlled case evaluation in the next step covered a total of 13 scenarios. Although this design was appropriate for a thesis project and allowed a manageable comparison across configurations, it still limits how much the evaluation can capture the full diversity of cyber risk assessment situations encountered in practice.

Another related limitation is about the number of practitioners involved in the evaluation process. The consultant evaluation was designed to provide initial insights into perceived usefulness, usability, and user experience of the prototype, but it was not intended to provide statistically extensive evidence about

adoption across the consulting profession. As a result, the findings should be seen as indicative rather than representative of all consultants or organisations.

A further limitation is that CREST was not evaluated in a live real-case engagement. The evaluation therefore shows performance under controlled and practitioner-reviewed conditions, but does not yet show how the tool performs during repeated use in an ongoing consulting project with real client decision-making consequences.

More generally, the evaluation design prioritised depth and feasibility over scale. This is understandable in the context of an initial design-oriented study, but it also means that the findings should be interpreted carefully. A larger evaluation involving more scenarios, more consultants, and more diverse organisational contexts would be needed to assess the robustness of CREST more extensively.

6.3.2 Ground truth and scoring subjectivity

The second limitation is related to the use of ground truth in a domain that is inherently judgement-intensive. In this study, the expected scores and reasoning for the predefined risk scenarios were first developed by the researcher and then validated with a senior cyber risk consultant. This was required to provide a necessary reference point or basis for evaluating scoring accuracy, but it does not remove the interpretative and subjective nature of cyber risk scoring itself.

In practice, cyber risk assessment rarely produces one objectively correct answer in the same way as a classification task. Likelihood and impact scores often depend on contextual assumptions, interpretation of incomplete evidence, and professional judgement. As a result, the ground truth used in this study should not be seen as an absolute or universally valid benchmark, but rather as a structured reference point developed for the purpose of controlled evaluation.

This has implications for how the reported accuracy results should be interpreted. A mismatch between the generated output and the predefined ground truth does not always mean that the output is unreasonable in practice. In some cases, it may instead reflect a different but still valid interpretation of the available evidence. For this reason, the accuracy metric is useful for structured comparison, but it should not be treated as a complete measure of quality on its own. This is also why the study includes traceability and qualitative interpretation beside score agreement.

6.3.3 Prototype and data constraints

The third limitation concerns the prototype itself and the data environment in which it was developed and evaluated. CREST was implemented as an Excel-centred prototype with a Python-based backend, designed to fit the current practical setting of the host organisation. This was an intentional design choice, as Excel provided a familiar workspace for consultants and made the prototype easier to demonstrate and evaluate in context. However, the limitation lies not in the use of Excel itself, but in the maturity of the current implementation. The prototype still depends on local execution, file-based workflows, and environment-specific configuration. Therefore, while an Excel-based interface could remain part of a future production version, the current implementation would still need further development, deployment hardening, and integration before it could be used as a stable operational tool.

The technical setup was also constrained by the internal AI infrastructure of the host organisation. Model selection was limited to the LLMs available within this environment, and the evaluation therefore does not cover the wider range of models that might be available in other settings. Stronger models could lead to better results compared to the results observed in this study. At the same time, this constraint reflects practical development limitations in many organisations, where model choice is shaped by data governance, confidentiality, and infrastructure requirements.

This model availability constraint also makes the evaluation more informative from a design perspective: because performance is not dominated by the use of the most capable LLM models, the results more clearly reveal the limitations of the other CREST elements, including retrieval quality, chunking, prompting, evidence selection, and workflow design. In this sense, the evaluation helps identify artifact-level limitations rather than allowing model capability to compensate for them. The results therefore depend on the specific interaction between the selected models, the prompting strategy, the retrieval depth, and the available assessment materials.

The data used in this study also reflects a specific organisational context. Although the dataset included multiple document types and a mix of structured and unstructured materials, it still originated from one client environment and one overall assessment setting. This means that CREST was not evaluated across a broad range of industries, documentation practices, or organisational maturity levels. However, cyber risk assessments are standardised at a high level across consulting practice through shared standards, assessment phases, reporting formats, and evidence types. Therefore, some transferability to similar assessment contexts can be expected, although the findings should still be validated across other organisational settings.

Finally, some aspects of the prototype were intentionally simplified to keep the study feasible. For example, control-related information was incorporated through the control effectiveness factor under the vulnerability domain rather than through a fully separate and comprehensive control assessment step. This was a practical design choice, but it also means that the current version of CREST does not yet capture the full complexity of control effectiveness as a standalone aspect.

Chapter 7

Conclusions

This chapter concludes the thesis by summarising the main findings and contributions of the research. The study investigated how AI-assisted decision support can be designed and evaluated to improve consistency, transparency, and reproducibility in the judgement-intensive cyber risk scoring process, while preserving human judgement and accountability. To address this objective, the research designed, developed, and evaluated CREST as an AI-assisted decision-support prototype for cyber risk estimation in consulting practice.

The chapter is structured into two main parts. First, it answers the main research question and sub-questions based on the findings from the literature review, stakeholder interviews, artifact design, and evaluation. Second, it outlines directions for future work, focusing on technical refinement, broader validation, and integration into consulting workflows.

7.1 Answers to the research questions

This section answers the main research question and the five sub-questions formulated in Chapter 1. The answers are based on the findings from the literature review, stakeholder interviews, artifact design, and evaluation of CREST.

Main research question

The main research question was:

> How can AI-assisted decision support be designed and evaluated to improve consistency, transparency, and reproducibility of the judgement-intensive risk scoring process under non-prescriptive cybersecurity standards, while preserving human judgement and accountability?

This research shows that AI-assisted decision support can be designed as a human-in-the-loop system that structures evidence use, scoring logic, and reasoning documentation without automating the final assessment decision. CREST implements this approach by combining an Excel-centred assessment workbook, a Python backend, retrieval-augmented generation, and LLM-assisted reasoning. The tool supports consultants by retrieving relevant evidence, generating factor-level scoring suggestions, and documenting reasoning and evidence chains for review.

The evaluation shows that this approach is feasible and positively perceived by consultants. In the controlled case evaluation, CREST achieved 77% scoring accuracy and an average traceability score of 60%. The consultant evaluation also showed positive perceptions across TAM, SUS, and UEQ. These findings suggest that CREST can support more structured and reviewable cyber risk scoring. In this sense, reviewability is part of process reproducibility, meaning that the goal is not to reproduce identical LLM outputs or final risk levels in every run, but to make the scoring process more explicit, traceable, and repeatable across assessments. However, the results also show that the tool should remain a decision-support rather than a decision-making tool. Underestimation errors, partial traceability, and consultant feedback on the risk of over-reliance confirm the need for human validation, balanced trust, and further improvement of retrieval, prompting, and output design.

RQ1: Current process of cyber risk assessment

RQ1 asked how cyber risk assessments are currently performed in practice under flexible standards, and at which points judgement-intensive scoring decisions occur.

The literature review and stakeholder interviews show that cyber risk assessment generally follows a structured process, including context establishment, risk identification, risk analysis, risk evaluation, and risk treatment. However, the implementation of likelihood and impact scoring remains flexible. Standards such as ISO/IEC 27005 and NIST SP 800-30 provide high-level guidance, but they do not prescribe a specific scoring model or evidence interpretation procedure. As a result, judgement-intensive decisions happen especially during the risk analysis phase, where consultants interpret available evidence into likelihood and impact scores.

RQ2: Data and assessment inputs

RQ2 asked how different types of data and assessment inputs are collected, interpreted, and translated into impact and likelihood scoring.

The research shows that consultants rely on various assessment inputs, including policy documents, gap assessment results, interview transcripts, assessment reports, and other supporting materials. These inputs differ in structure, completeness, and relevance. In practice, consultants interpret these materials by identifying evidence related to threats, vulnerabilities, control maturity level, and business or technical consequences. CREST supports this process by processing assessment materials, storing document chunks in a vector database, and retrieving evidence relevant to specific scoring factors. This makes the translation from assessment inputs to scoring suggestions more explicit and reviewable.

RQ3: Judgement logic in practice

RQ3 asked how judgement logic is currently applied in practice, and to what extent it remains implicit or insufficiently documented.

The findings show that judgement logic is often applied through consultant experience, contextual interpretation, and informal reasoning. While consultants may use predefined risk matrix, template, or assessment frameworks, the justification behind specific likelihood and impact scores is not always documented in a structured way. This creates challenges for process consistency, traceability, and reproducibility. CREST addresses this by generating structured factor-level reasoning, selected descriptors, and evidence chains. This does not remove judgement, but it makes parts of the reasoning process more explicit and easier to review.

RQ4: Requirements for improvement

RQ4 asked what requirements are needed to improve consistency, transparency, and explainability of risk scoring outcomes in consulting practice.

The research identified four main design requirements. First, the system should support more consistent scoring across consultants and engagements. Second, it should structure reasoning by linking evidence to scoring factors and descriptors. Third, it should support traceability by making evidence sources visible and reviewable. Fourth, it should preserve human judgement by keeping the consultant responsible for reviewing and finalising the assessment. These requirements guided the design of CREST and were translated into its solution objectives, workflow, and architecture.

RQ5: AI-assisted decision support

RQ5 asked how AI-assisted decision support can structure and document reasoning processes, interpretation consistency, and justification of scoring decisions without automating final judgement.

CREST validates one possible approach. The prototype uses RAG to retrieve relevant evidence and LLM-assisted reasoning to generate structured suggestions for each scoring factor. The output includes a suggested descriptor, short reasoning, and an evidence chain. The scoring logic is implemented using a configurable Excel-based matrix, allowing consultants to inspect and adjust the scoring model. In this way, CREST supports reasoning and justification while keeping the consultant in control. The evaluation confirms that this positioning is important, where consultants saw CREST as a tool to generate structured

starting point, but also emphasised the need for human validation, reviewability, and safeguards against over-reliance.

Overall, the research questions are answered by showing that AI-assisted decision support can improve the structure and reviewability of cyber risk scoring when it is designed surrounding evidence traceability, configurable scoring logic, and human-in-the-loop principle. At the same time, the findings show that such support remains dependent on retrieval quality, contextual relevance, user training, and appropriate governance of AI-assisted judgement.

7.2 Future work

While this research demonstrates the feasibility of AI-assisted decision support for cyber risk scoring, CREST remains as an early-stage prototype. The evaluation showed promising results in terms of scoring performance, traceability, and consultant perception, but also identified areas that require further improvement before the approach can be applied in practice. Future work should therefore focus on improving the technical robustness of the artifact and strengthening its integration into consulting workflows.

The following subsections outline two main directions for future work: performance optimisation and adoption and integration.

7.2.1 Performance optimisation

Future work should first focus on improving the technical performance and robustness of CREST. The controlled case evaluation showed that CREST can already reproduce the predefined ground truth of most scenarios, but the traceability results also showed that generated reasoning is not yet consistently verifiable. Performance optimisation should then address the model selection and configuration, retrieval pipeline, prompting strategy, output stability, and the robustness of CREST across different assessment contexts.

Extending LLM model comparison. One area for future work is to extend the comparison of LLM models. The current configuration comparison was limited to three models available in the internal AI infrastructure of the host organisation. This was appropriate for the prototype because it reflected practical deployment and data confidentiality constraints. However, future work could evaluate other additional models with stronger reasoning capabilities, more up-to-date knowledge, larger context windows, or better behaviour when following instructions. This would help to assess whether the observed limitations are partly related to the capabilities of the selected model. To distinguish model-related limitations from retrieval or prompting limitations more clearly, future work would also need to test these components separately.

Improving retrieval quality. A second area is further optimisation of the RAG pipeline. In the current prototype, retrieved evidence may be semantically related to a scoring factor but it is not always aligned with the specific risk context. For example, a retrieved chunk may discuss about access control policy in general, but refer to supplier access rather than the internal organisational access context of the assessed scenario. Future versions should then improve chunk retrieval by using richer metadata, context-aware filtering, document folder location, and document-type-specific retrieval strategies. This would help ensure that retrieved evidence is not only similar in topics, but also contextually relevant to the specific risk scenario, system scope, and scoring factor being assessed.

Refining prompts and evidence interpretation. A third direction is prompt refinement. The prompts used in the current prototype could be made more explicit in requiring the model to analyse different evidence contexts before using retrieved chunks as support. Future prompts should be able to instruct the model to check whether the evidence is related to the correct asset, threat, vulnerability, risk scenario, and overall assessment context before treating it as relevant. Prompts should also require the model to explicitly state when evidence is weak, indirect, missing, or only generally related to the scenario instead of making up non-existent evidence. This would improve the transparency of the reasoning process and reduce the risk that partially relevant evidence is presented as stronger support than it actually provides.

Reducing risk underestimation. Another important area for future work is reducing the risk of underestimating risk levels. In the controlled case evaluation, all incorrect predictions resulted in a lower

risk level than the predefined ground truth. This is an important finding to note because underestimation may be more problematic than overestimation in cyber risk assessment, as it can lead to wrong risk evaluation, insufficient prioritisation, or weaker treatment decisions. Future work should therefore analyse where this underestimation occurs in the workflow, for example whether it is caused by missing or weak retrieved evidence, insufficient attention to business impact, prompt interpretation, misleading risk matrix values, or the misleading interpretation from evidence to factor-level scores. Addressing this issue may require more explicit prompts for identifying severe impact factors, stronger checks for missing or incomplete evidence, and additional review flags when the generated score is lower than expected based on the available evidence.

Making uncertainty and output variation visible. Future work could also explore making uncertainty and output variation visible to the users. The current prototype presents a single suggested score and reasoning chain for each factor. However, risk estimation is inherently uncertain, and just displaying a single output may create a false impression of certainty. Future versions could use several model runs to show how stable or unstable a suggested score is across multiple generations. If the same scenario produces different scores or reasoning variations across runs, this could be shown to users as an indication of uncertainty. Such feature could help to manage user trust and expectation by showing when an output is relatively stable and when it should be treated with more caution.

Testing in less controlled settings. Future work should also evaluate CREST in less controlled assessment settings. The present evaluation was based on predefined scenarios and selected assessment materials, mainly PDF and Excel-based inputs. This has provided a useful first evaluation setting, but it does not fully capture the complexity of real consulting engagements, where data may be more inconsistent and risks may be different depending on client context. Further validation in real assessment projects would therefore be needed to test whether CREST remains robust when applied to less structured evidence and changing assessment contexts.

Using richer technical inputs. Lastly, performance optimisation should also consider the use of richer technical inputs. Consultant feedback suggested that CREST could benefit from additional sources such as threat models, technical drawings, network designs, or other architecture-related materials to produce more context-aware list of risk scenarios. Incorporating these inputs could improve contextual understanding, especially for system-specific technical impact and vulnerability assessment. However, this would also require further work on parsing, metadata extraction, and evidence linking to ensure that richer input sources improve traceability rather than introduce additional noise.

7.2.2 Adoption and integration

Future work should also focus on how CREST can be adopted and integrated into consulting practice. The current prototype was suitable for evaluation because it used an Excel-centred workflow that is familiar to consultants. However, broader adoption would require additional considerations not only on technical performance, but also ease of deployment, supporting methodology, business relevance, workflow integration, and practical validation in real consulting settings.

Improving deployment and accessibility. An area for future work to support adoption is improving the deployment model of CREST. The current prototype relies on local execution, which can create practical dependencies on individual consultant workspaces, software setup, and local configuration. A more robust deployment model could reduce these setup issues, make the tool easier to use across different engagements, and provide a more consistent user experience. This could be implemented through a web-based version, where consultants access CREST through a controlled environment, or through a containerised setup such as Docker, where the required Python environment and dependencies are packaged together. This form of deployment model could allow consultants to upload or connect assessment materials, run the analysis in a more consistent environment, and download structured outputs for review.

Developing operating procedures and review criteria. Future work should also define clearer procedures for using CREST in practice. Consultant feedback pointed out that the tool should not only generate outputs, but also be supported by clear procedures for preparing inputs, validating outputs, and deciding when results are sufficiently reliable. Future versions could therefore include input preparation templates, interview or context-gathering templates, validation criteria, and review checklists. These materials would help ensure that CREST is used consistently and that generated outputs remain subject to explicit consultant review.

Improving business-focused justification for final risk levels. Another important area for future work is improving the business relevance of CREST's outputs. The current prototype already includes business impact factors such as financial damage, reputational damage, non-compliance, and privacy violation as part of the factor-level scoring. However, consultant feedback highlighted the need for a clearer business-facing explanation of why a generated final risk level matters from a business perspective. Future versions could therefore add an output justification layer that summarises the main business consequences behind the final risk level, such as potential operational disruption, financial loss, regulatory exposure, or impact on customer trust. This would make the output more useful for risk registers, client reporting, and decision-making discussions.

Extending support across the risk assessment process. Future work should also extend CREST beyond the risk analysis phase. The current prototype focuses on likelihood, impact, and risk scoring, while other phases of the risk assessment process remain outside the main scope. Future versions could support automated risk identification by generating a list of risks from gap assessment results, threat analysis reports, or vulnerability assessment. They could also support risk evaluation by comparing generated risk levels against client-specific risk appetite, and support risk treatment by suggesting relevant control measures or treatment options. This way, CREST can be positioned as an end-to-end cyber risk assessment tool.

Integrating control maturity assessment more explicitly. A related improvement is stronger integration with control maturity assessment. During development, a more automated control assessment component was explored, but it was kept outside the evaluated scope of this thesis to keep the artifact evaluation manageable. In the current evaluated version of CREST, control-related information is simplified through a control effectiveness factor within the vulnerability dimension. Future work could further develop and validate this component, linking risks and vulnerability scoring more explicitly to the implementation status or maturity level of relevant controls in the client environment. This would allow consultants to assess not only the current risk level, but also how existing or potential future controls influence prioritisation and treatment decisions.

Piloting CREST in a real consulting engagement. Finally, future work should explore adoption through a limited pilot engagement in a real consulting setting. This pilot engagement should make it clear that CREST is experimental and that all outputs remain subject to manual consultant verification. This would allow the host organisation to evaluate whether CREST reduces assessment effort, improves reviewability, and produces outputs that are sufficiently useful and reliable for client-facing work. A pilot engagement would also help assess practical questions that cannot be fully answered in a controlled evaluation, such as how much correction effort is required, how consultants interact with the tool under time pressure, and whether CREST can become a sustainable service offering.

Taken together, these future work directions emphasise that CREST should be developed further as both a technical prototype and a practical consulting support mechanism. This thesis shows that AI-assisted decision support can help structure and document judgement-intensive cyber risk scoring, but its value depends on traceability, human validation, and careful integration into existing assessment practice.

Bibliography

[1] Suleiman S. Abba, Oluwadayo Mafolasere Olaniyi, Oluseun Babatunde Oladoyinbo, Olalekan Jamiu Okunleye, and Valerie Ojinika Ejiofor. AI-driven automation of cybersecurity certification processes: Evaluating efficiency, transparency and risk mitigation in digital governance systems. *Journal of Engineering Research and Reports*, 27(12):70–91, 2025. doi: 10.9734/jerr/2025/v27i121728.

[2] Ali Aghazadeh Ardebili, Marianna Lezzi, and Mahdad Pourmadadkar. Risk assessment for cyber resilience of critical infrastructures: Methods, governance, and standards. *Applied Sciences*, 14(24): 11807, 2024. doi: 10.3390/app142411807.

[3] Marco Angelini, Silvia Bonomi, and Alessandro Palma. A methodology to support automatic cyber risk assessment review. arXiv preprint arXiv:2207.03269, 2022.

[4] John Brooke. SUS: A "quick and dirty" usability scale. In Patrick W. Jordan, Bruce Thomas, Bernard A. Weerdmeester, and Ian L. McClelland, editors, *Usability Evaluation in Industry*, pages 189–194. Taylor & Francis, London, 1996.

[5] John Brooke. SUS: A retrospective. *Journal of Usability Studies*, 8(2):29–40, 2013.

[6] Fred D. Davis. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3):319–340, 1989. doi: 10.2307/249008.

[7] Fatiha Djebbar and Kim Nordström. A comparative analysis of industrial cybersecurity standards. *IEEE Access*, 11:85315–85334, 2023. doi: 10.1109/ACCESS.2023.3303205.

[8] Gencer Erdogan, Morgan Gillette, Simeon Tverdal, and Ragnhild Halvorsrud. From manual to automated cyber risk assessment: LLM- and RAG-driven multi-agent threat modeling with CORAS in healthcare case studies. In *Proceedings of the 2025 12th International Conference on Dependable Systems and Their Applications (DSA)*, pages 55–64, Sharjah, United Arab Emirates, 2025. IEEE. doi: 10.1109/DSA66321.2025.00018.

[9] European Parliament and Council of the European Union. Directive (EU) 2022/2555 of the european parliament and of the council on measures for a high common level of cybersecurity across the union (NIS2 Directive), 2022. URL `https://eur-lex.europa.eu/eli/dir/2022/2555`. Accessed: May 2026.

[10] European Union Agency for Cybersecurity. Technical implementation guidance on commission implementing regulation (EU) 2024/2690 regarding cybersecurity risk-management measures, 2025. URL `https://www.enisa.europa.eu/publications/nis2-technical-implementation-guidance`. Version 1.0, accessed: May 2026.

[11] France Cyber Maritime. Summary of the ENISA guide on risk management measures in cybersecurity (NIS2). URL `https://www.france-cyber-maritime.eu/wp-content/uploads/2025/09/Synthese_guide_ENISA_EN-2.pdf`. Summary document, accessed: May 2026.

[12] David C. Hall. Making risk assessments more comparable and repeatable. *Systems Engineering*, 14 (2):173–179, 2011. doi: 10.1002/sys.20169.

[13] International Electrotechnical Commission. IEC 62443-3-2:2020 – security for industrial automation and control systems – part 3-2: Security risk assessment for system design, 2020. International Standard.

[14] International Organization for Standardization. ISO 31000:2018 – risk management – guidelines, 2018. International Standard.

[15] International Organization for Standardization. ISO/IEC 27005:2022 – information security, cybersecurity and privacy protection – guidance on managing information security risks, 2022. International Standard.

[16] Joint Task Force Transformation Initiative. Guide for conducting risk assessments. Technical Report NIST Special Publication 800-30 Revision 1, National Institute of Standards and Technology, 2012.

[17] Maura M. Kepper, Allison J. L'Hotta, Thembekile Shato, Bethany M. Kwan, Russell E. Glasgow, Douglas Luke, Andrea K. Graham, Ana A. Baumann, Ross C. Brownson, and Brad Morse. Supporting teams with designing for dissemination and sustainability: The design, development, and usability of a digital interactive platform. *Implementation Science*, 19:82, 2024. doi: 10.1186/s13012-024-01410-7.

[18] Jessica Kollmorgen, Andreas Hinderks, and Jörg Thomaschewski. Selecting the appropriate user experience questionnaire and guidance for interpretation: The UEQ family. *International Journal of Interactive Multimedia and Artificial Intelligence*, 9(4):126–139, 2025. doi: 10.9781/ijimai.2024.08.005.

[19] Bettina Laugwitz, Martin Schrepp, and Theo Held. Construction and evaluation of a user experience questionnaire. In Andreas Holzinger, editor, *HCI and Usability for Education and Work*, volume 5298 of *Lecture Notes in Computer Science*, pages 63–76, Berlin, Heidelberg, 2008. Springer. doi: 10.1007/978-3-540-89350-9_6.

[20] In Lee. Cybersecurity: Risk management framework and investment cost analysis. *Business Horizons*, 64(5):659–671, 2021. doi: 10.1016/j.bushor.2021.02.022.

[21] Bilgin Metin, Sefa Duran, Eda Telli, Meltem Mutlutürk, and Martin Wynn. IT risk management: Towards a system for enhancing objectivity in asset valuation that engenders a security culture. *Information*, 15(1):55, 2024. doi: 10.3390/info15010055.

[22] Michael Matthias Naumann, Doru Alexandru Plesea, Marieta Olaru, and Fabian Pitz. A comparative study on different approaches to risk management in information security management systems. *International Journal of Advanced Engineering and Management Research*, 10(1):101–115, 2025. doi: 10.51505/ijaemr.2025.1007.

[23] Lyndsay A. Nelson, Carrie Reale, Shilo Anders, Russ Beebe, S. Trent Rosenbloom, Amber Hackstadt, Kryseana J. Harper, Lindsay S. Mayberry, Jared G. Cobb, Neeraja Peterson, Tom Elasy, Zhihong Yu, and William Martinez. Empowering patients to address diabetes care gaps: Formative usability testing of a novel patient portal intervention. *JAMIA Open*, 6(2):ooad030, 2023. doi: 10.1093/jamiaopen/ooad030.

[24] Christine L. Berner Nyvik, Roger Flage, and Seth Guikema. On the use of standards and guidelines as a tool to fulfil regulatory requirements. *Risk Analysis*, 41(10):1744–1750, 2021. doi: 10.1111/risa.13672.

[25] OpenAI. Introducing gpt-oss, 2025. URL `https://openai.com/index/introducing-gpt-oss/`. Accessed: May 2026.

[26] OWASP Foundation. OWASP risk rating methodology. URL `https://owasp.org/www-community/OWASP_Risk_Rating_Methodology`. Official OWASP documentation, accessed: 19 February 2026.

[27] Jyoti Parsola. Cybersecurity risk assessment and management for organizational security. *Neuro-Quantology*, 20(5):5330–5337, 2022. doi: 10.48047/nq.2022.20.5.nq22815.

[28] Ken Peffers, Tuure Tuunanen, Marcus A. Rothenberger, and Samir Chatterjee. A design science research methodology for information systems research. *Journal of Management Information Systems*, 24(3):45–77, 2007. doi: 10.2753/MIS0742-1222240302.

[29] Qwen Team. Qwen3: Think deeper, act faster, 2025. URL `https://qwenlm.github.io/blog/qwen3/`. Accessed: May 2026.

[30] Denaldi Rananda Saputra, Tranggono, and Mega Cattleya Prameswari Annisa Islami. Evaluation of a virtual reality-based introduction to hazardous and toxic waste management using the technology acceptance model. *Journal of Applied Informatics and Computing*, 9(4):1722–1729, 2025. doi: 10.30871/jaic.v9i4.10040.

[31] Martin Schrepp. User experience questionnaire handbook: All you need to know to apply the UEQ successfully in your projects. UEQ Handbook, 2023. URL `https://www.ueq-online.org/Material/Handbook.pdf`. Version 11, 12 September 2023.

[32] Martin Schrepp, Andreas Hinderks, and Jörg Thomaschewski. Applying the user experience questionnaire (UEQ) in different evaluation scenarios. In Aaron Marcus, editor, *Design, User Experience, and Usability: Theories, Methods, and Tools for Designing the User Experience*, volume 8517 of *Lecture Notes in Computer Science*, pages 383–392, Cham, 2014. Springer. doi: 10.1007/978-3-319-07668-3_37.

[33] Shengming Zhao, Yuchen Shao, Yuheng Huang, Jiayang Song, Zhijie Wang, Chengcheng Wan, and Lei Ma. Understanding the design decisions of retrieval-augmented generation systems. arXiv preprint arXiv:2411.19463, 2025.

Appendix A

Questionnaire item-level results

A.1 TAM item-level results

Table A.1 presents the item-level results for the TAM questionnaire. Responses were coded on a five-point Likert scale, where 1 represents strongly disagree, 2 disagree, 3 neutral, 4 agree, and 5 strongly agree. To maintain anonymity, the respondents are shown as C1 to C4.

Table A.1: TAM item-level results by consultant

Item	Construct	C1	C2	C3	C4	Mean
Q11	Perceived usefulness	4	4	5	4	4.25
Q12	Perceived usefulness	4	2	4	4	3.50
Q13	Perceived usefulness	4	5	5	3	4.25
Q14	Perceived usefulness	4	3	4	4	3.75
Q15	Perceived usefulness	4	4	5	4	4.25
Q16	Perceived usefulness	4	4	5	5	4.50
Q17	Perceived ease of use	4	5	5	4	4.50
Q18	Perceived ease of use	3	4	3	4	3.50
Q19	Perceived ease of use	4	5	4	3	4.00
Q20	Perceived ease of use	4	4	4	4	4.00
Q21	Perceived ease of use	3	5	4	3	3.75
Q22	Perceived ease of use	4	5	4	4	4.25
Q23	Willingness to use	4	4	4	4	4.00

Table A.2 provides the wording of the TAM items used in the questionnaire.

Table A.2: TAM item wording

Item	Question wording
Q11	Using this prototype would enable me to complete cyber risk assessment tasks more quickly.
Q12	Using this prototype would improve the quality of my cyber risk assessment output.

Continued on the next page

Item	Question wording
Q13	Using this prototype would help me complete more cyber risk assessment work within the same amount of time.
Q14	Using this prototype would help me better achieve the objectives of cyber risk assessment.
Q15	Using this prototype would make cyber risk assessment easier to perform.
Q16	Overall, I would find this prototype useful in cyber risk assessment work.
Q17	Learning to operate this prototype would be easy for me.
Q18	I would find it easy to get this prototype to do what I want it to do.
Q19	My interaction with this prototype would be clear and understandable.
Q20	I would find it easy to interact with this prototype in the way I need.
Q21	It would be easy for me to become skilful at using this prototype.
Q22	Overall, I would find this prototype easy to use.
Q23	If this prototype were available in practice, I would be willing to use it in my work.

A.2 SUS item-level results

Table A.3 presents the item-level results for the SUS questionnaire. Responses were coded on a five-point Likert scale, where 1 represents strongly disagree, 2 disagree, 3 neutral, 4 agree, and 5 strongly agree. To maintain anonymity, the respondents are shown as C1 to C4. The table reports the original coded responses before applying the SUS scoring transformation.

Table A.3: SUS item-level results by consultant

Item	Item type	C1	C2	C3	C4	Mean
Q1	Positive	4	5	3	4	4.00
Q2	Negative	2	1	2	2	1.75
Q3	Positive	4	5	4	4	4.25
Q4	Negative	2	1	2	3	2.00
Q5	Positive	4	5	3	5	4.25
Q6	Negative	1	1	2	1	1.25
Q7	Positive	3	5	4	4	4.00
Q8	Negative	2	1	1	2	1.50
Q9	Positive	4	5	4	3	4.00
Q10	Negative	1	1	1	3	1.50

Table A.4 provides the wording of the SUS items used in the questionnaire.

Table A.4: SUS item wording

Item	Question wording
Q1	I think that I would like to use this system frequently.
Q2	I found the system unnecessarily complex.
Q3	I thought the system was easy to use.

Continued on the next page

Item	Question wording
Q4	I think that I would need the support of a technical person to be able to use this system.
Q5	I found the various functions in this system were well integrated.
Q6	I thought there was too much inconsistency in this system.
Q7	I would imagine that most people would learn to use this system very quickly.
Q8	I found the system very cumbersome to use.
Q9	I felt very confident using the system.
Q10	I needed to learn a lot of things before I could get going with this system.

A.3 UEQ item-level results

Table A.5 presents the item-level results for the UEQ questionnaire. The original seven-point semantic differential responses were transformed to the standard UEQ scale from -3 to $+3$, where higher values indicate a more positive evaluation. To maintain anonymity, the respondents are shown as C1 to C4.

Table A.5: UEQ item-level results by consultant

Item	UEQ scale	C1	C2	C3	C4	Mean
Q24.1	Attractiveness	2	2	2	1	1.75
Q24.2	Perspicuity	1	2	2	1	1.50
Q24.3	Novelty	-2	-1	2	3	0.50
Q24.4	Perspicuity	-1	3	2	1	1.25
Q24.5	Stimulation	2	2	2	3	2.25
Q24.6	Stimulation	1	2	2	2	1.75
Q24.7	Stimulation	2	1	3	2	2.00
Q24.8	Dependability	0	0	1	0	0.25
Q24.9	Efficiency	0	0	2	2	1.00
Q24.10	Novelty	1	2	1	1	1.25
Q24.11	Dependability	2	2	2	2	2.00
Q24.12	Attractiveness	2	3	3	3	2.75
Q24.13	Perspicuity	1	3	1	0	1.25
Q24.14	Attractiveness	1	1	2	2	1.50
Q24.15	Novelty	1	1	1	1	1.00
Q24.16	Attractiveness	1	2	2	2	1.75
Q24.17	Dependability	1	2	0	0	0.75
Q24.18	Stimulation	1	2	2	2	1.75
Q24.19	Dependability	1	2	2	2	1.75
Q24.20	Efficiency	1	2	2	2	1.75
Q24.21	Perspicuity	1	3	1	2	1.75
Q24.22	Efficiency	2	3	1	2	2.00
Q24.23	Efficiency	2	3	0	2	1.75
Q24.24	Attractiveness	1	-1	-1	0	-0.25
Q24.25	Attractiveness	1	0	1	2	1.00
Q24.26	Novelty	1	2	2	1	1.50

Table A.6 provides the wording of the UEQ item pairs used in the questionnaire.

Table A.6: UEQ item wording

Item	**UEQ scale**	**Item pair**
Q24.1	Attractiveness	annoying / enjoyable
Q24.2	Perspicuity	not understandable / understandable
Q24.3	Novelty	creative / dull
Q24.4	Perspicuity	easy to learn / difficult to learn
Q24.5	Stimulation	valuable / inferior
Q24.6	Stimulation	boring / exciting
Q24.7	Stimulation	not interesting / interesting
Q24.8	Dependability	unpredictable / predictable
Q24.9	Efficiency	fast / slow
Q24.10	Novelty	inventive / conventional
Q24.11	Dependability	obstructive / supportive
Q24.12	Attractiveness	good / bad
Q24.13	Perspicuity	complicated / easy
Q24.14	Attractiveness	unlikeable / pleasing
Q24.15	Novelty	usual / leading edge
Q24.16	Attractiveness	unpleasant / pleasant
Q24.17	Dependability	secure / not secure
Q24.18	Stimulation	motivating / demotivating
Q24.19	Dependability	meets expectations / does not meet expectations
Q24.20	Efficiency	inefficient / efficient
Q24.21	Perspicuity	clear / confusing
Q24.22	Efficiency	impractical / practical
Q24.23	Efficiency	organized / cluttered
Q24.24	Attractiveness	attractive / unattractive
Q24.25	Attractiveness	friendly / unfriendly
Q24.26	Novelty	conservative / innovative