CHALMERS
UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG

Effects of Cognitive Load in Human-AI Requirements Engineering

Master's Thesis in Software Engineering and Technology

Niharika Nandi Shivamurthy Praveen
Laxmi Prashantraddi sasvihalli

Department of Computer Science and Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden 2025

Master's Thesis 2025

Effects of Cognitive Load in Human-AI Requirements Engineering

Niharika Nandi Shivamurthy Praveen
Laxmi Prashantraddi Savihalli

UNIVERSITY OF
GOTHENBURG

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Computer Science and Engineering
Chalmers University of Technology
University of Gothenburg
Gothenburg, Sweden 2025

Effects of Cognitive Load in Human-AI Requirements Engineering
Niharika Nandi Shivamurthy Praveen
Laxmi Prashantraddi Savihalli

© Niharika Nandi Shivamurthy Praveen and Laxmi Prashantraddi Sasviahlli, 2025.

Supervisor: Richard Berntsson Sevensson , Department of Computer Science and Engineering
Supervisor:Lekshmi Rani, Department of Computer Science and Engineering
Examiner: Gregory Gay, Department of Computer Science and Engineering

Master's Thesis 2025
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Typeset in LaTeX
Gothenburg, Sweden 2025

Effects of Coginitive Load in Human-AI Requirements Engineering
Niharika Nandi Shivamurthy Praveen and Laxmi Prashnatraddi Sasvihalli
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg

Abstract

As Artificial Intelligence becomes more integrated into software engineering, its role in decision-support systems within Requirements Engineering has grown. However, the cognitive demands placed on users interacting with these AI tools remain underexplored. This thesis investigates how explanation formats offered by Explainable AI affect mental effort, task difficulty, confidence, and correctness during requirements engineering inspired prioritization tasks. Through a controlled experiment with 61 participants, three XAI formats of bar charts, textual explanations, and confidence scores were evaluated across two task pairs of differing complexity. The study examined the influence of task complexity and explanation format, the impact of explanation type on decision-making quality, and whether participant preferences for certain formats aligned with improved performance and lower cognitive strain. Statistical analyses, including Spearman correlation and independent t-tests, revealed that task complexity consistently influenced cognitive load, while explanation format had no clear effect. Additionally, although preferred formats did not universally enhance task performance, participants who favored confidence scores showed marginally higher correctness and confidence levels. These findings suggest that cognitive effort in AI-assisted requirements engineering tasks is shaped more by task characteristics than explanation format alone, and that tailoring explanations to individual user preferences may offer subtle benefits.

Keywords: Requirement Engineering(RE), Cognitive Load(CL), Artificial Intelligence (AI), Explainable Artificial Intelligence (XAI), Weighted Shortest Job First (WSJF), Research Question (RQ), User Experience (UX).

Acknowledgements

We would like to sincerely thank our supervisors, Richard Svensson and Lekshmi Rani, for their valuable guidance, feedback, and encouragement throughout the course of this thesis. Their support has been instrumental in shaping our research. We would also like to thank our examiner, Gregory Gay, for his input and constructive advice. Additionally, we are grateful to all the participants who contributed their time and insights to our study. Finally, we would like to extend our appreciation to our families and friends for their continued support and motivation during this journey.

Niharika Nandi Shivamurthy Praveen and Laxmi Prashantraddi Sasvihalli, Gothenburg, September 2025

Contents

List of Figures

List of Tables

1

Introduction

Artificial Intelligence (AI) is rapidly reshaping software engineering, changing the way core development tasks are carried out. In particular, recent studies show that AI is becoming increasingly embedded in Requirements Engineering (RE), where it is used to support activities such as eliciting requirements, prioritizing features, and analyzing trade-offs [24, 5]. These activities are central to project success because they require stakeholders to weigh feasibility, manage risks, and maximize value [62, 65]. As AI systems take on a greater role in these decisions, the challenge is no longer only whether their outputs are accurate, but also whether practitioners can understand and reason with them [6, 27].

A central concern in this interaction is cognitive load, the mental effort required to process and integrate information during task execution [60, 48]. In RE, practitioners already operate under high cognitive demands due to the complexity of requirements, the diversity of stakeholders, and the presence of competing constraints [30, 2]. When AI-generated recommendations are opaque, vague, or misaligned with user expectations, they increase this mental effort and can quickly lead to cognitive overload [46, 45]. Such overload does not simply make tasks harder, it reduces the quality of decisions and undermines trust in AI systems [15, 27].

Explainable AI (XAI) has emerged as a promising way to address the challenges posed by opaque AI outputs. Techniques such as confidence scores, bar chart visualizations, and plain-language text explanations are designed to improve transparency and build user trust by clarifying how AI systems generate their results [6, 17]. Evidence from domains such as healthcare and other safety-critical settings suggests that well-designed explanations can enhance decision-making by making AI predictions more interpretable and actionable [27, 32]. Despite these advances, the influence of explanation format on cognitive load and decision-making performance within requirements engineering (RE) tasks remains insufficiently explored [23, 5]

The broader literature also highlights the cognitive demands of RE tasks themselves. Studies show that multitasking, task complexity, and ambiguous criteria can substantially increase the mental effort required for requirements prioritization and analysis [30, 2]. Research in behavioral software engineering further underscores the need to understand both individual and team cognition when engaging with decision-support tools [21, 51]. At the same time, findings from XAI research confirm that explanation design directly shapes users' performance, trust, and overall satisfaction [6, 45]. Yet, few studies bring these perspectives together, leaving an important gap in how different forms of explanation

influence cognitive load during RE prioritization tasks.

This thesis addresses that gap by empirically examining how three common explanation formats, text, bar charts, and confidence scores, shape cognitive load and decision-making performance in requirements prioritization tasks. Using a controlled survey experiment that varies task complexity (two criteria versus four criteria), the study provides systematic evidence on whether particular explanation designs can reduce mental effort and enhance decision quality. [54, 63, 45, 15].

The significance of this research lies in bridging explainability studies with cognitive load theory within the specific context of requirements engineering. While much prior work has assessed explanations primarily in terms of technical accuracy or model interpretability, this thesis shifts attention to the human perspective, focusing on how individuals experience and manage cognitive demands when making critical project decisions [60, 5, 23, 27, 6]. In doing so, the study offers practical insights for designing AI tools that better align with human cognitive capacities, enabling practitioners and organizations to adopt AI in ways that actively support rather than complicate prioritization and collaboration in software projects.

1.1 Thesis Outline

This thesis report is organized into several key sections to provide a clear and structured overview of the study. It begins with an introduction to the topic, outlining the problem space and explaining why the study matters. The background section then sets the foundation by discussing Cognitive Load Theory and its relevance to requirements engineering, along with ideas around how humans and AI can work together in this space. The next part covers related work, summarizing what past research has found, and pointing out the gaps this study aims to address. The methodology section walks through how the study was carried out, from the survey design and tasks to how the data was collected and analyzed. This is followed by the results chapter, which shares what was found in the responses and highlights the main patterns. The discussion then reflects on these findings, connecting them back to the research questions and existing studies, and considering what they mean for the use of AI in requirements engineering. After that, the thesis looks at potential limitations and factors that could have influenced the results. It ends with a conclusion that wraps everything up, highlights the study's contributions, and suggests where future research could go.

2

Background

2.1 Background

This section presents background information on Cognitive Load Theory (CLT) and the Requirements Engineering (RE) process, focusing on the activity of requirements prioritization. It also explores the relevance of CLT in RE contexts, especially as human engineers increasingly collaborate with Artificial Intelligence (AI) tools in decision-making processes.

This section also provides foundational context for the research, introducing Cognitive Load Theory and its theoretical underpinnings. It also explains the nature of Requirements Engineering in software development, emphasizing the cognitively intensive task of prioritizing requirements. The connection between CLT and RE is then elaborated, establishing the rationale for applying cognitive principles to the challenges of AI-assisted RE.

2.1.1 Cognitive Load Theory

Cognitive Load Theory (CLT), originally developed by John Sweller in the late 1980s, is a psychological theory concerned with how people process and retain information while learning or performing tasks [59]. The theory is based on the premise that working memory, the mental space in which we process information, is limited in both capacity and duration. When individuals are asked to perform complex tasks, especially those involving new or unstructured information, they may experience cognitive overload, impairing learning, problem-solving, or decision-making.

According to CLT, cognitive processing is divided into three kinds of load. Intrinsic cognitive load depends on the built-in complexity of the task itself. For example, analyzing interdependent software requirements involves holding multiple interacting elements in mind, which inherently increases the mental effort required [61][48]. Extraneous cognitive load results from the way information is presented to the learner. Poorly structured documentation or confusing user interfaces can add unnecessary load without supporting learning or task completion [47]. Germane cognitive load is the beneficial mental effort used to build knowledge structures or "schemas" that improve problem-solving and understanding. For instance, when requirements engineers reflect on prioritization strategies and gradually develop heuristics for evaluating trade-offs, they are investing cognitive effort that strengthens their long-term expertise.[61].

The key goal of CLT is to design information and tasks that minimize unnecessary load, manage complexity, and encourage productive learning. These principles are increasingly relevant in software development contexts where high cognitive demands can affect decision-making and productivity.

2.1.2 Requirements Engineering and Prioritization

Requirements Engineering (RE) is a structured process in software development focused on identifying, documenting, analyzing, and managing system requirements. The goal is to ensure that the final software product aligns with user needs, stakeholder goals, and system constraints. The RE process generally consists of several stages: elicitation, where requirements are gathered; specification, where they are documented; validation, where correctness is confirmed; and management, where changes are tracked throughout the lifecycle.[49].

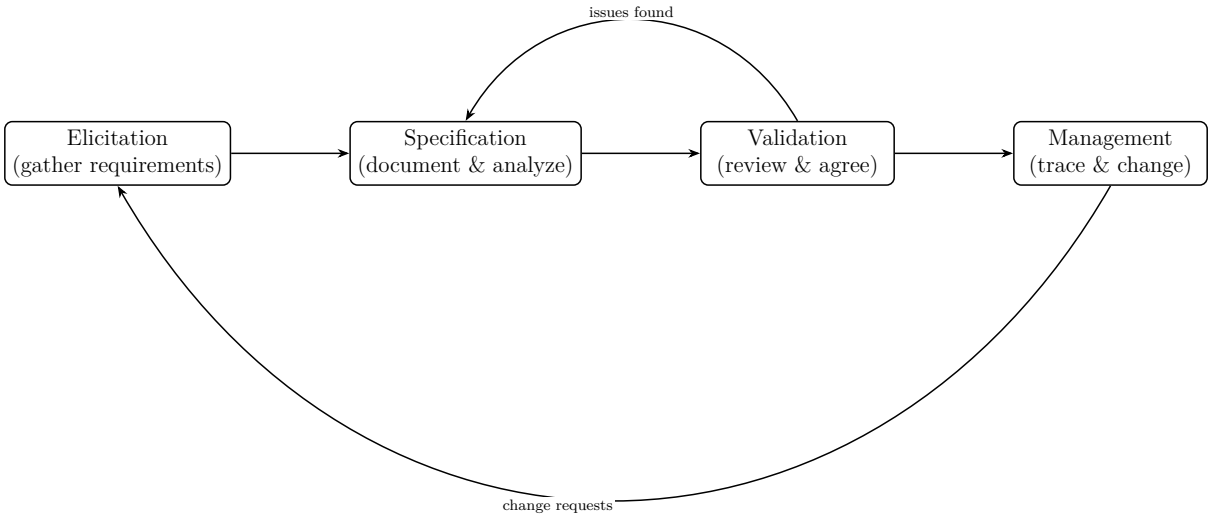


Figure 2.1: Requirements Engineering (RE) process with feedback loops ([49]).

One of the most critical and cognitively demanding steps in RE is requirements prioritization. This is the process of determining the relative importance of various requirements to guide decision-making and resource allocation. Engineers must often prioritize based on multiple, and sometimes conflicting, criteria such as stakeholder value, technical feasibility, cost, and implementation risk [1]. In multi-stakeholder environments, prioritization becomes even more complex due to differing opinions and business objectives.

Prioritization becomes increasingly complex in large-scale or multi-stakeholder projects, where competing interests must be balanced. Traditional methods such as the Analytic Hierarchy Process (AHP) and Cost-Value Approaches are commonly used, but they often require engineers to process large volumes of information and make difficult trade-offs [36]. This leads to significant cognitive effort, especially when requirements are ambiguous or when there are many dependencies between them.

The emergence of Artificial Intelligence tools, including Large Language Models (LLMs),

has added new capabilities to the prioritization process. These tools can analyze historical data, detect patterns, and propose ranked lists of requirements based on weighted factors. While AI can help reduce manual effort, it also introduces new challenges in managing cognitive load, particularly when AI outputs are poorly explained or misaligned with human expectations [56].

2.1.3 CLT and Its Relevance in Requirements Engineering

Cognitive Load Theory is highly relevant to Requirements Engineering, especially in activities such as elicitation, analysis, and prioritization, where engineers must process complex information and make judgments under uncertainty. As AI tools become more integrated into RE tasks, it is essential to ensure that these tools support human cognition rather than overwhelm it [3].

Studies have shown that engineers experience high levels of intrinsic and extraneous cognitive load when working with complex, unstructured requirements or when interpreting unclear AI-generated suggestions [5][30]. Poor management of these cognitive demands can result in decision fatigue, errors, and reduced stakeholder alignment. On the other hand, tools designed with CLT principles, such as those that present visual models, modularize information, or offer clear feedback, can reduce unnecessary load and improve task performance [44].

In particular, requirements prioritization benefits from CLT-informed AI tool design. Breaking down complex prioritization decisions into smaller, more manageable parts can help engineers focus better and reason more clearly. Similarly, AI tools that offer transparent, explainable recommendations rather than opaque outputs can reduce extraneous load and increase trust in the system. As such, CLT offers a theoretical lens through which the effectiveness of AI-assisted RE tools can be evaluated.

2.1.4 Explainable AI (XAI) and Its Role in Requirements Engineering

AI systems integrated into requirements engineering require humans to understand their outputs effectively [13]. The set of techniques known as Explainable AI (XAI) provides transparency into AI system behavior and decision-making processes, which human users can understand. The interpretability requirement in RE contexts becomes essential because engineers need to evaluate and validate AI-generated suggestions and potentially make changes to them [28].

The implementation of XAI techniques enhances trust and usability and cognitive efficiency in human-AI collaboration by minimizing the unclear aspects of AI outputs. Engineers face difficulties in understanding AI recommendation rationales because of a lack of explainability, which results in cognitive overload and misuse [15]. Well-designed XAI methods enable engineers to verify AI outputs efficiently, which strengthens user confidence and facilitates better decision-making processes [45].

In the context of software and requirements engineering, several prominent XAI techniques have gained traction. Model-agnostic approaches like LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (SHapley Additive exPlanations) are widely

adopted to interpret complex machine learning models by highlighting feature contributions for individual predictions [7]. Visual tools such as saliency maps and attention heatmaps are often used in domains involving image or text data, offering intuitive cues about the system's focus during decision-making. In requirements engineering specifically, more structured explanations such as decision trees, rule-based outputs, and ranked lists of features or criteria are frequently integrated to support traceability and justify prioritization decisions [33]. These methods aim to make AI outputs not only transparent but also actionable for engineers and stakeholders who rely on such insights for validating requirements, allocating resources, or managing trade-offs.

This research implements three XAI methods, which include confidence scores and bar charts, and text-based explanations. The selection of these modalities represents different ways to achieve interpretability. The AI provides quantitative certainty through confidence scores, and bar charts display feature importance for fast comparison, while text explanations deliver natural language explanations for AI-driven prioritization. Research by [7] demonstrates that these XAI methods both work effectively to enhance understanding and minimizing the mental work needed to understand AI recommendations.

3

Related Work

This section presents an overview of research on cognitive load, focusing on various domains relevant to this study. First, we summarize findings from a broad range of domains where cognitive load has been studied, such as education, healthcare, navigation, and marketing. Second, we review research specific to software engineering, where cognitive demands are prominent due to task complexity and system interdependencies. Finally, we discuss emerging literature on human-AI collaboration in SE and Requirements Engineering, particularly the role of Large Language Models and Explainable AI in shaping cognitive experiences during prioritization tasks.

3.1 Cognitive Load in General Domains

Research on cognitive load extends far beyond software engineering, and insights from other domains provide useful analogies for understanding task complexity, measurement, and mitigation strategies. We include studies from education, healthcare, navigation, marketing, and teamwork because they illustrate three points that are directly relevant to this thesis: (1) cognitive load consistently emerges as a barrier to effective performance across domains, (2) researchers have used diverse measurement techniques that can inform methodological choices in this work, and (3) strategies to mitigate cognitive load remain underdeveloped, motivating further investigation in software engineering contexts. The studies summarized in Table 3.1 were selected because they are frequently cited, represent methodological diversity (self-reports, physiological monitoring, behavioral measures), and exemplify how different factors such as emotional arousal, task complexity, and collaboration shape cognitive effort. We describe them as "key" studies, not because they exhaustively cover the field, but because they are illustrative and transferable to the challenges of requirements engineering and human-AI collaboration.

Table 3.1 organizes prior work by the "Actor type" (individual, human–AI collaboration, or human–human teams). For each study, we report the "Cognitive load factor" being investigated (e.g., task complexity, distractions, working memory load), the "Task" participants performed, the "Main findings", and whether "Mitigation strategies" were proposed or tested. The "Domain" column specifies the application area, while the "Measurement columns" indicate how cognitive load (CL Measure) and task performance (TP Measure) were assessed. The final column provides references. This organization allows comparison across domains and highlights recurring themes: task complexity and distractions consistently elevate cognitive load, measurement techniques vary widely, and mitigation remains more often theoretical than empirically validated.

To illustrate, Fraser [21] found that high emotional excitement in simulated clinical settings increased mental effort and hindered accurate task execution. Similarly, Skulmowski [58] showed that learners under high extraneous and intrinsic load experienced reduced focus and slower learning across complex online courses.

Another notable study by Žagar [68] examined navigation tasks and demonstrated how auditory distractions could increase error rates by elevating mental load. In marketing, Kakaria [35] observed that consumers who shopped without a plan showed higher EEG-based cognitive load, indicating more impulsive and less accurate decision-making. Furthermore, Whitney [64] showed how framing effects, combined with increased working memory load, shaped risky decisions in high-stress conditions.

These studies employed various cognitive load measurement techniques, including self-reports (e.g., 7-point rating scales), physiological monitoring (e.g., heart rate, EEG), and behavioral metrics (e.g., error rates, decision time). Despite the diversity of applications, task complexity consistently emerged as a primary cognitive load driver across domains [58][21]. However, few of these studies evaluated strategies to reduce load, but most mitigation approaches, such as training or adaptive system design, remained theoretical [68].

Actor Type	**CL Factor**	**Task**	**Findings**	**Mitigation**	**Domain**	**CL Measure**	**TP Measure**	**Ref**
Individual	Emotional state	Simulation training	More excitement increased CL; calmness reduced CL	Not addressed	Medical Education	7-pt scale	Heart sound ID	[21]
Individual	Extra/ intrinsic/ germane	Online learning	Extra load reduces focus, right load increases understanding	Theory: Constructive alignment . Empirical: No empirical evaluation	Education Psychology	Eye tracking, 7-pt scale, Pupilometry	Correct answers, Response rates	[58]
Individual	Distraction	Ship steering w/ alarm sounds	More errors,increased CL	Theory: Distraction training	Navigation	Heart rate and stress levels	Reaction time, errors	[68]
Individual	Purchase Planning	Virtual shopping	Unplanned increased cognitive load	Theory: Planning Strategies	E-commerce	EEG (gamma band)	Time, number of planned/ unplanned count, total expenditure	[35]
Individual	Working Memory Load	Risky decisions	High load decreased risky choices	Theory: Increased load limits WM.Empirical: Higher load reduces risky decisions.	Psychology	Dual-task	Decision bias	[64]
Individual	Memory Load	Decision-making with uncertainty information	High load reduced optimal decisions	Theory: Cognitive load hinders decisions. Empirical: Cognitive Load impairs optimal decisions.	Cognitive Psychology	Dual-task	Accuracy, decisions	[3]

Actor Type	CL Factor	Task	Findings	Mitigation	Domain	CL Measure	TP Measure	Ref
Individual	Task Complexity, Visual Distraction	Surgical decision,	High TC increased CL, worse decisions.	Theory: TC increases CL; Empirical: TC had no impact.	Medical	Measured using NASA-TLX, SURG-TLX, eye-tracking, EEG, EDA.	Time, errors, mental effort, and correct tasks.	[32]
Individual	Morphol -ogical Clarity	Image classification	Low MC increased CL; high MC reduced CL	Empirical: Adjacent visualizations and high MC CL	XAI and its impact on human-AI collaboration.	Pupil dilation,7-pt scale	Accuracy, confidence, time	[33]
Human-AI	TC, Cognitive Resources	AI chatbot learning	Reduced CL,increased learning	Theory: AI-assisted learning (iLearnTech chatbot)	Education	7-pt scale	Correct answers,time, accuracy, error	[46]
Human-AI	Task Complexity	Robot-assisted gait.	Real-time task adjustment maintained optimal CL	Theory: Adaptive difficulty .Empirical: System maintained cognitive load with 88% accuracy.	Medical	HR,EEG	Time, correct answers	[38]
Human - AI	Time scarcity, technology availability	Creative problem-solving	Time scarcity increased AI use, increased CL	Empirical: Time management strategies	Creative Teamwork	Self-report,task ratings	Success, creativity	[57]
Human - AI	TC, decision flexibility	Dynamic team tasks	Flexible AI reduced CL, improving adaptability.	Empirical: Adaptive AI	Workplace AI Integration	TC metrics, team adaptability measures	Success rate, goal achievement	[28]
Human-AI	AI explainability	COVID-19 decisions	Different XAI explanation types affected cognitive load and task performance; explanations focused on specific decisions led to reduced cognitive load and better performance.	Theory: Clear local explanations improve cognitive efficiency. Empirical: Local XAI explanations reducing cognitive load and improving TP	AI/ Healthcare	7-pt scale	Accuracy, Time	[29]
HH - Teams	Physiologi -cal synchronization, TC	Cardiac surgery	Increased synchronization increased performance.	Empirical: Feedback on synchronization.	Medical	HRV, entropy measures	Surgical errors, time	[16]
HH - Teams	Collabora -tive CL, transactive activities	Collaborative learning	Theoretical: Collaboration can reduce CL if guided	Theory: Structured guidance, role distribution.	Educational Psychology	Theoretical discussion - no direct empirical measurement	Theoretical	[37]

Actor Type	CL Factor	Task	Findings	Mitigation	Domain	CL Measure	TP Measure	Ref
HH-Teams	Cognitive effort, fatigue, TC	Team sport	High CL impaired physical and tactical performance in sports.	Empirical: Structured training.	Sports	NASA-TLX, PANAS, HR	Physical performance, tactical decision-making	[23]
HH - Teams	Team efficiency, TC	Military decisions	Improved decision-making and team performance reduced cognitive load.	Empirical: Decision-support systems improved team efficiency and reduced cognitive load.	Military	TCE Score	Task performance assessed through the Air Defense Warfare Team Performance Index (ATPI)	[34]
HH - Teams	Cognitive processing load, collaboration technology	Simulated - command and control	High CL increased errors, time.	Empirical: Task simplification and real-time feedback improved performance.	Military/ emergency	NASA-TLX,TC metrics	Error rates, time	[22]

Table 3.1: Cognitive Load in General Domains

3.2 Cognitive Load in Software Engineering

Cognitive load has also been studied in the context of software development, where complex problem-solving and information-intensive tasks are the norm. We include this body of work because it directly informs the challenges of requirements engineering and prioritization tasks addressed in this thesis. The studies summarized in Table 3.2 were selected because they represent diverse methodologies (EEG, eye-tracking, self-reports), focus on typical SE activities (e.g., coding, debugging, information sharing), and highlight both drivers of cognitive load and early attempts at mitigation. We describe them as “key” studies not because they exhaustively cover the field, but because they illustrate recurring patterns and gaps that are transferable to our problem space.

Table 3.2 organizes prior work by the "Actor type" (individual, human–AI collaboration, or human–human teams). Each row describes a study, reporting the "Cognitive load factor" under investigation (e.g., task complexity, distraction, trust), the "Task" participants performed, the "Main findings", and whether any "Mitigation strategies" were proposed or tested. The "Domain" column specifies the application area (e.g., software development, VR tasks, human–robot teaming). The final two measurement columns indicate how cognitive load (CL Measure) and task performance (TP Measure) were assessed, followed by the reference.

To illustrate, Goncales [26] used EEG sensors to show that higher task complexity increased cognitive load and reduced code accuracy. In human–computer interaction research, Ghulaxe [25] proposed AI-driven distraction reduction in development environments. While theoretically promising, these strategies were not empirically validated, reflecting a broader issue in SE research: a lack of rigorous evaluation of cognitive load interventions.

Across these studies, tools and methods to assess cognitive load vary: some use physiological measures (EEG, heart rate, pupil dilation), while others rely on behavioral performance or subjective ratings. While tasks such as prioritization, elicitation, and debugging are widely acknowledged as cognitively intense, there is still insufficient empirical work on effective interventions to support engineers in these phases [44].

Actor Type	**CL Factor**	**Task**	**Findings**	**Mitigation**	**Domain**	**CL Measure**	**TP Measure**	**Ref**
Individual	Task Complexity	Software dev(coding)	Increased TC led to higher CL, affecting code quality, speed	Theory: TC leads to higher CL, but no empirical mitigation strategies	Software Engineering	EEG	Code accuracy,time	[26]
Individual	Task Complexity	Cognitive tasks(varied)	Higher TC increased CL,via physiological signals	Theory: Accurate measurement of CL can help adaptive systems to reduce CL.Empirical: No specific mitigations	Human-Computer Interaction	Pupil, blinking rate,HR	None	[2]
Individual	Attention, Distraction	Driving Task	No empirical results; theoretical: AI reduces CL.	Theory: AI gaze tracking.	Automotive	The evaluation remained theoretical, based on proposed AI solutions like gaze tracking and blinking pattern detection	Theoretical only.	[25]
Human-AI	Task difficulty, trust	VR search task	Higher CL, reducing trust, performance.	Theory: Biosignal assessment.Empirical: No significant correlation found between biosignals, trust, and cognitive load.	AI	EEG, HR,7-pt scale	Time, correct answers	[27]
Human-AI	Task complexity, cognitive teaming, mental modeling	Rescue / exploration	Increased CL led to poor teaming	Theory: Adaptive mental modeling.	Human-Robot Teaming	Theoretical discussion; no empirical measurement.	Theoretical discussion; no empirical measurement.	[13]
Human-AI	Task complexity, packing difficulty	Collaborative packing task (virtual environment)	Higher CL, reduced task efficiency.	Empirical: AI-assisted packing guidance.	Human-AI Collaboration	NASA-TLX	Time, efficiency, errors	[39]
Human-AI	Cognitive capacity limitations, task complexity	Info sharing	Increased CL decreased sharing.	Empirical: HMM((Hidden Markov Model)-based cognitive load model improved information sharing and teamwork.	Human-Agent Collaboration	Secondary task performance, information recall (HMM-based)	Information recall,accuracy	[19]

Actor Type	CL Factor	Task	Findings	Mitigation	Domain	CL Measure	TP Measure	Ref
Human-AI	Task difficulty, agent reliability	N-back, shape selection tasks	Lower cognitive load improved task performance; agent reliability reduced cognitive strain.	Empirical: Reliable agent guidance reduced cognitive load and improved task efficiency.	VR-based Human-AI Interaction	EEG, GSR, HRV, self-reported cognitive load ratings	time, accuracy	[27]
Human - AI	Decision style, AI identity	Word-guessing game	Autocratic decision-making increased CL and reduced team efficacy; democratic style improved collaboration and lowered CL	Empirical: Democratic decision-making improved team efficacy and user satisfaction, reducing cognitive load	Human-AI Collaboration	NASA-TLX	Game win rate, accuracy	[43]
HH-Teams	Task Complexity	Emergency game	Higher TC increased CL,	Theory:Eye-tracking interfaces	Gaming	Eye-tracking metrics (e.g., pupil diameter)	Accuracy	[4]

Table 3.2: Cognitive Load in Software Engineering

3.3 Human-AI Collaboration and LLMs in Requirements Engineering

With the increasing adoption of Large Language Models (LLMs) in software engineering, researchers have begun to explore their potential across the requirements engineering (RE) lifecycle. Beyond traditional automation techniques, LLMs such as GPT-3.5 and GPT-4 are now being used to support elicitation, analysis, refinement, and prioritization tasks.

For elicitation, conversational LLMs have been studied as proxies for stakeholders during interviews. Lojo et al. [42] showed that students preferred LLM-based simulations over static transcripts when practicing elicitation, describing them as more realistic and engaging, though sometimes inconsistent. Similarly, Franch et al. [20] investigated how LLMs can generate stakeholder questions from software requirement patterns. While effective for broadening coverage, their approach sometimes produced redundant or out-of-scope requirements, requiring additional filtering effort from engineers.

Expanding on this idea, Ataei et al. [8] proposed "Elicitron", a multi-agent framework where LLMs simulate users, generate observations, and derive latent needs. This approach demonstrated improved coverage of design requirements but introduced interpretability challenges, underscoring the cognitive demands placed on engineers when reconciling multiple AI outputs. Quattrocchi et al. [50] benchmarked several LLMs for generating and evaluating user stories. They found that while LLMs matched humans in terms of coverage and style, they performed less well in creativity and acceptance criteria, shifting the cognitive burden to human reviewers for quality assurance.

In the area of prioritization, Sami et al. [55, 56] introduced a multi-agent system employing LLMs to improve user story quality and ranking accuracy. While their approach

showed productivity gains, it also revealed new issues: when AI-generated suggestions were unclear or overly numerous, engineers experienced extraneous cognitive load, often leading to confusion and delays.

These findings align with broader concerns in Explainable AI (XAI). Arrieta et al. [7] emphasize that for AI systems to be cognitively beneficial, they must provide explanations aligned with human reasoning. In RE, where decisions must be justified and traceable, explainability is essential. Techniques such as confidence scores, visualized importance weights, and natural language justifications have been proposed to increase interpretability, reduce mental effort, and improve trust.

Despite these advances, most LLM-based strategies remain underexplored in RE, particularly regarding their cognitive implications. The cost of interacting with opaque or overwhelming AI suggestions in sensitive tasks such as requirements elicitation and prioritization remains a significant research gap, motivating this thesis to investigate how human AI collaboration can be designed to support, rather than hinder, engineers' cognitive processes.

3.4 Summary

From the reviewed literature, we observed three consistent themes:

First, task complexity is consistently identified as a major source of cognitive load across domains, including software engineering. For example, Goncales et al. [26] showed that higher task complexity increases cognitive load during programming, but they did not investigate how developers could be supported in managing this demand, particularly in decision-intensive activities such as requirements prioritization.

Second, while AI tools such as LLMs are increasingly applied across the requirements engineering (RE) lifecycle, research on their role in prioritization remains limited. Sami et al. [55, 56] demonstrated that multi-agent LLM systems can improve user story quality and ranking accuracy, but their work did not consider the cognitive implications of interacting with such systems. Other studies have shown promising applications in elicitation and user story generation, such as Lojo et al. [42], Franch et al. [20], Quattrocchi et al. [50], yet prioritization, despite being a cognitively demanding and decision-critical task, has received comparatively little attention.

Third, explainability has been widely discussed as a way to make AI more understandable and trustworthy, but its impact in RE tasks is still largely untested. Arrieta et al. [7] provide a broad taxonomy of XAI techniques, yet no study has empirically examined how explanation styles influence engineers' mental effort and decision quality in prioritization contexts. Poorly explained outputs or overwhelming recommendations are, therefore, likely sources of extraneous cognitive load, but they remain underexplored in RE research.

Taken together, these gaps highlight the need to study requirements prioritization as a cognitively demanding RE activity where AI can both support and burden engineers. While prior work has shown that LLMs can assist in prioritization, no study has systematically investigated how AI support with or without explainability shapes the cognitive experience of engineers. This thesis addresses that gap by empirically evaluating how

AI-assisted prioritization affects cognitive effort and decision outcomes, contributing new insights at the intersection of prioritization, human cognition, and explainable AI.

4

Methodology

This study investigates the influence of XAI on cognitive load and decision-making performance during software requirements prioritization tasks. We focus on prioritization because it is one of the most cognitively demanding and decision-critical activities in requirements engineering. Engineers must weigh competing stakeholder needs, balance limited resources, and make trade-offs under uncertainty. While prior work has applied LLMs to elicitation and user story generation, research on prioritization has been comparatively scarce and has not examined the cognitive implications of AI support. Addressing this gap, our methodology followed a sequential process involving literature review, research design, and survey implementation. This approach ensured the work was grounded in theoretical understanding, refined through empirical testing, and systematically evaluated.

4.1 Research Design

A within-subject experimental design was adopted, where each participant performed prioritization tasks both without and with AI support. This design was chosen because it allowed participants to serve as their own control, enabling systematic comparisons between unassisted and assisted conditions. In particular, it supported analysis of:

- differences in cognitive load (RQ1),
- quality of decision-making (RQ2), and
- user preferences across explanation formats (RQ3).

The experiment was structured around two domains: **banking loan management** and **doctor appointment scheduling**. These domains were selected because they are widely understandable and reflect realistic decision-making contexts without requiring specialized knowledge. To introduce variation in complexity, the banking tasks involved two prioritization criteria, while the healthcare tasks involved four. This staged setup enabled systematic analysis of how task complexity and explanation format interact to influence cognitive load and performance.

Details of the task flow, prioritization criteria, and measurement instruments are explained in the sections below.

The research questions guiding this study are :

RQ1: How do different styles of XAI impact cognitive load during decision-making in requirements prioritization?

RQ2: How do different styles of XAI impact the quality of decision-making in requirements prioritization tasks?

RQ3: How do users' preferences for different XAI formats relate to their task performance, perceived mental effort, and trust in AI-supported requirements prioritization?

4.2 Methodology Process Overview

An overview of the methodological process is presented in Figure 4.1, showing the sequential steps from literature review through to data interpretation.

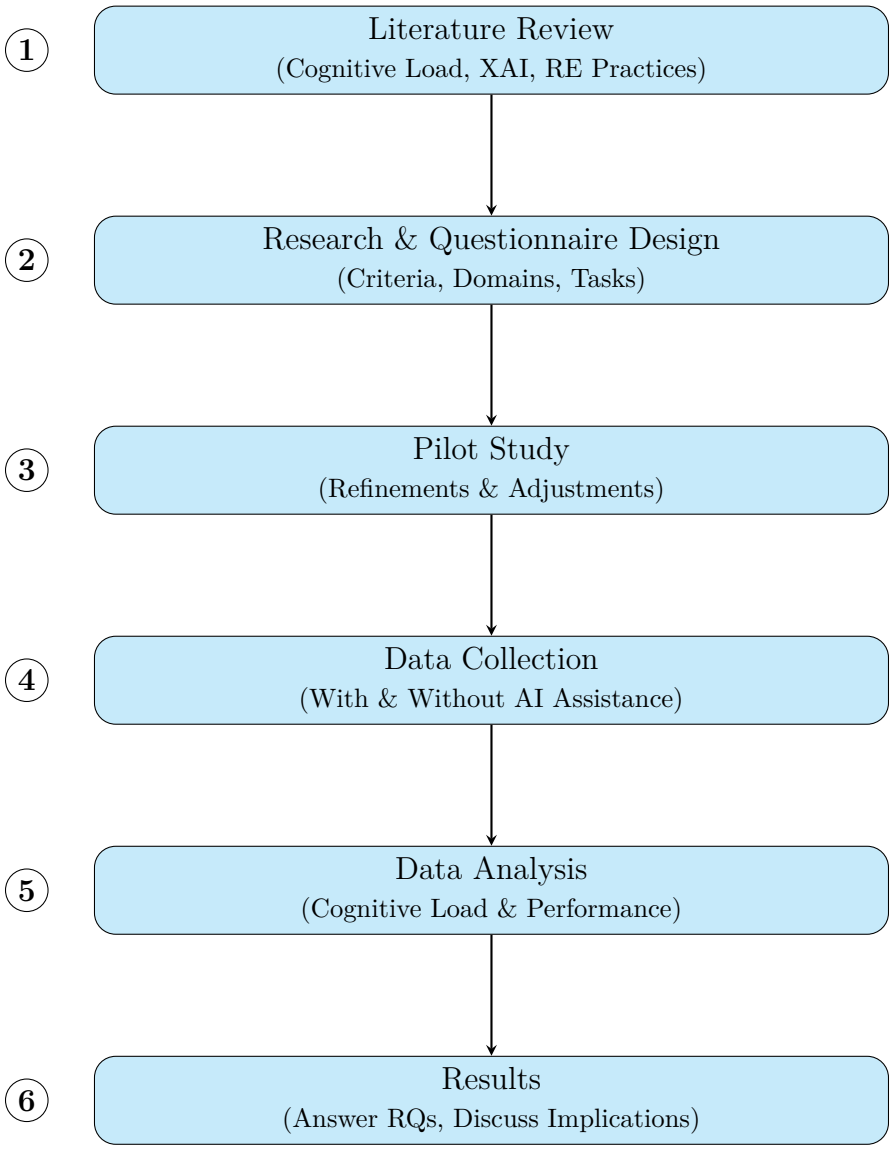


Figure 4.1: Methodology process flow (numbered steps).

4.3 Survey Design and Questionnaire

The survey instrument was developed using insights from the literature on cognitive load theory, requirements engineering, and XAI transparency, refined through supervisor feedback and a pilot study. The full instruments, including task descriptions, instructions, and XAI prompts of the AI-generated explanations, are provided in the link in the appendix B. Because Microsoft Forms does not support random assignment of alternative explanation types within a single survey, we created three separate versions of the survey. Each version contained a different combination of XAI explanation formats (e.g., text with confidence, bar chart with confidence, or text with bar chart). Participants were distributed across these versions to ensure that no participant was exposed to the same explanation format twice, while still allowing comparison between numeric, visual, and textual explanations. Participants were randomly assigned to one survey version, meaning that each individual was exposed to only one explanation format per domain, while across the full sample, all three formats were tested.

4.3.1 Survey Flow

The survey began with demographic questions, followed by requirements prioritization tasks in two domains. In each domain, participants first completed a baseline task without AI assistance, followed by a comparable task with AI-generated prioritization presented in one of three explanation formats. After each task, participants rated their perceived cognitive load. The survey concluded with questions on usability, trust, and preferences for the explanation format experienced, along with open-ended feedback.

4.3.2 Demographics

The demographics section collected participant details such as professional role, years of experience with requirements prioritization, and prior exposure to AI tools. This contextual information was important for interpreting variation in task performance and workload ratings. The target population was professionals and students with exposure to requirements engineering, as they regularly engage in prioritization decisions and are familiar with the challenges of balancing competing criteria. Their expertise provided both realism and validity to the evaluation.

4.3.3 Prioritization Tasks

The main body of the questionnaire contained two domains: a **Banking Loan Management System** and a **Doctor Appointment System**. Each task required participants to prioritize ten functional requirements. These requirements were created by the researchers, inspired by typical functionalities from publicly available system descriptions and prior RE literature, to ensure they were realistic yet domain-neutral. Using ten requirements, balanced realism and feasibility were achieved, resulting in a number that was large enough to represent the complexity of real-world decision-making but still manageable within the time constraints of an online survey.

To systematically vary complexity, the banking task (Task 1) required prioritization based on *two criteria*: development time and customer value. The healthcare task (Task 2)

required prioritization based on *four criteria*: development time, customer value, risk, and time sensitivity. This staged design allowed us to examine how cognitive load changes when task complexity increases while holding other parameters constant. These four criteria are widely recognized in the requirements prioritization literature [36, 65, 9, 10, 11, 14, 40], providing theoretical grounding for their selection.

More specifically, in the banking domain, Task 1.1 involved prioritizing requirements for a loan management system without AI assistance, while Task 1.2 involved prioritizing requirements for an online banking system with AI assistance. In the healthcare domain, Task 2.1 focused on prioritizing requirements for an emergency doctor appointment booking system without AI support, while Task 2.2 addressed a general doctor appointment booking system with AI support. These domains and task variations were selected because they are familiar to most participants, involve multi-criteria trade-offs similar to software requirements decisions, and help reduce the risk of bias. By ensuring that the with-AI tasks (1.2 and 2.2) were not identical to the without-AI tasks (1.1 and 2.1), participants were less likely to find the second task easier simply due to prior exposure. This design, combined with the use of three survey versions, minimized potential carryover effects while still allowing comparison of cognitive load and decision quality with and without AI support.

4.3.4 XAI Explanation Formats

Three explanation formats were tested:
1. **Confidence scores** (numerical probabilities, e.g., "Requirement A is recommended with 72% confidence") were selected for their ability to communicate model certainty in a compact form.
2. **Bar charts** (visual ranking of requirements by importance) were selected for their ability to present comparisons quickly and clearly.
3. **Textual explanations** (natural language reasoning, e.g., "Requirement B is prioritized because it reduces waiting time, which users rated highly") were selected because natural language is intuitive and widely used in LLM interfaces.

These formats were chosen because they represent common explanation styles in XAI research [7, 45, 15], and together they allow us to compare numeric, visual, and textual communication of AI reasoning.

4.3.5 Implementation of AI Support

Participants did not interact directly with a live AI tool. Instead, all requirements, prioritizations, and explanations were pre-generated for consistency across participants and survey versions. The explanations (confidence scores, bar charts, and textual justifications) were generated with ChatGPT-4 (OpenAI). To preserve authenticity, these outputs were presented as screenshots embedded directly into Microsoft Forms, so participants viewed them exactly as produced by ChatGPT-4. This ensured exposure to realistic AI-generated content without introducing variability from system interfaces or user interaction. In the appendix B, we provide the generated explanations as they appeared in Microsoft Forms, showing the different XAI formats.

4.3.6 Measurement Approach

Cognitive load was measured after each task using a 7-point Likert scale assessing mental demand, effort, complexity, and confidence [48, 41]. Decision-making performance was evaluated by accuracy, using the WSJF method. Usability, trust, and preference ratings for the explanation formats were also collected through Likert-scale items and open-ended responses. This mixed-methods approach provided both subjective (self-reported load, trust, satisfaction) and objective (accuracy, time) measures.

4.4 Pilot Study

A pilot study with a small participant group was conducted to evaluate the clarity, usability, and pacing of the questionnaire. The objectives were to test whether the task instructions were understandable, verify that the XAI explanations were interpretable, and measure the time needed to complete the survey. Findings showed that some participants misunderstood the meaning of the prioritization criteria, prompting revisions to the instructions and inclusion of illustrative examples. The demographic section was streamlined to reduce participant fatigue, and task ordering was adjusted so simpler tasks appeared first to improve engagement and minimize dropout. Descriptions of AI explanations were also refined to ensure consistency in how participants interpreted each format. These refinements increased the reliability and user-friendliness of the final instrument.

4.5 Data Collection

Participants were recruited using a convenience sampling approach [18], leveraging university mailing lists, LinkedIn, WhatsApp groups, Discord servers, and both personal and professional networks. Additional outreach was conducted via supervisors' industry contacts to enhance diversity. This non-probabilistic sampling method was chosen due to its practicality and ability to reach participants with relevant experience in software engineering and requirements prioritization. The survey was administered via Microsoft Forms and remained open for three weeks to allow sufficient time for responses. Participation was voluntary, and the survey was designed to take approximately 12 minutes to complete based on pilot testing. A total of 61 completed responses were collected, representing participants from diverse backgrounds, including software developers, testers, product owners, requirements engineers, and students in software engineering programs. This diversity helped ensure that the study captured a range of perspectives on cognitive load in AI-assisted requirements prioritization.

4.6 Data Analysis

Before analysis, all completed survey responses were consolidated into a single dataset. As described in Section 4.3, participants were distributed across three survey versions, each of which contained a different combination of XAI explanation formats (e.g., text with confidence, bar chart with confidence, or text with bar chart). This ensured that each participant was exposed to only one explanation type per domain, while still allowing comparison of all three formats across the full sample.

For analysis, responses were grouped according to the specific XAI technique presented (bar chart, confidence score, or text explanation), and then subdivided into tasks performed with and without AI assistance. Task performance and cognitive load responses were aligned to their corresponding task identifiers. Scores for prioritization accuracy were calculated using the WSJF-based gold standard described in Section 4.6.2.1. Cognitive load scores were computed as the average of Likert-scale responses across four dimensions: mental demand, effort, complexity, and confidence.

4.6.1 Data Cleaning

During data cleaning, only the responses collected during the pilot study were removed to ensure that the analysis was based solely on data from the final version of the survey. The remaining 61 valid responses included both task types completed by every participant: (1) baseline prioritization without AI assistance and (2) prioritization with AI-generated recommendations. For analysis, responses were sorted by these two task types, while also distinguishing between the three explanation formats used in the AI-assisted tasks.

4.6.2 Defining the Correct Prioritization Order

A reference prioritization was created for each task using the Weighted Shortest Job First (WSJF) method to evaluate participant performance objectively. The tasks included ten functional requirements, with WSJF scores calculated according to the approach detailed in Section 4.6.2.1. The resulting scores were used to organize requirements into three priority levels: High, Medium, and Low.

The number of requirements grouped within each group changed from one task to another because WSJF values were distributed differently across tasks. The grouping process used natural breaks in WSJF scores instead of fixed numbers (e.g., 2-3-5) to determine the relative [12][55]

Requirement	**Customer Value**	**Development Time**	**WSJF Score**	**Priority Group**
Loan Payment Reminder Notifications	5	2.5	2.00	High
Loan Interest Rate Calculator	4	2.0	2.00	High
Loan Application Form	5	3.5	1.43	Medium
Automated Loan Status Updates	4	3.0	1.33	Medium
Loan Summary & Statement Generation	4	3.5	1.14	Medium
Loan Repayment Schedule Generator	3	3.5	0.86	Low
Loan Eligibility Checker	3	4.0	0.75	Low
Document Upload & Verification	2	3.0	0.67	Low
Loan Approval & Verification Process	2	4.0	0.50	Low
Personalized Loan Offers	1	3.0	0.33	Low

Table 4.1: Example of WSJF Grouping for Task 1.1 – Loan Management Task

In this example 4.1, the WSJF score was calculated using only Customer Value and Development Time. The two highest scores formed the High Priority group, the next three requirements formed Medium Priority, and the remaining five were classified as

Low Priority. This method ensured that features delivering the highest value in the shortest time were addressed first.

In the remaining three tasks (Tasks 1.2, 2.1, and 2.2), the same WSJF-based grouping logic was applied, but the specific priority group sizes varied depending on the distribution of WSJF scores in each scenario. In Task 1.2 (AI-assisted banking), the grouping followed a similar pattern to Task 1.1 but with a different set of requirements and slightly different group sizes. In Task 2.1 (emergency doctor booking), the WSJF formula incorporated four criteria of Customer Value, Risk Reduction, Time Sensitivity, and Development Time, resulting in group sizes determined by natural score gaps. Task 2.2 (AI-assisted doctor booking) also used the four-criterion WSJF calculation, producing a distinct distribution of High, Medium, and Low priority requirements. This consistent yet adaptive grouping method ensured that each task reflected the most valuable and time-efficient features for its domain while allowing fair comparison between AI-assisted and non-assisted conditions.

4.6.2.1 WSJF Calculation Method

The Weighted Shortest Job First (WSJF) method [52] was used to determine the reference prioritization order for each task. WSJF helps identify which requirements deliver the most value in the shortest time and is widely used in agile prioritization. However, the calculation parameters varied between Task 1 and Task 2 due to differences in scenario complexity and available attribute data.

WSJF Formula for Task 1 (Loan Management and Online Banking System) For Task 1.1 (without AI) and Task 1.2 (with AI), WSJF was calculated using only two factors:

$$\text{WSJF} = \frac{\text{Customer Value}}{\text{Development Time}} \tag{WSJF-1}$$

- Customer Value: Rated between 1 and 5 based on the perceived importance of the requirement to users.
- Development Time: Estimated effort or time required to implement the requirement.

WSJF Formula for Task 2 (Emergency and Doctor Appointment Systems) For Task 2.1 (without AI) and Task 2.2 (with AI), a more detailed version of WSJF was used to reflect the higher complexity of healthcare-related decision-making:

$$\text{WSJF} = \frac{\text{Customer Value} + (5 - \text{Risk}) + \text{Time Sensitivity}}{\text{Development Time}} \tag{WSJF-2}$$

- Customer Value: Scored from 1 to 5.
- Risk Reduction / Opportunity Enablement: Scored from 1 to 5, capturing the potential to reduce failure or enable significant gains.
- Time Sensitivity: Reflected how urgent the requirement was in terms of delivery impact.

- Development Time: Estimated implementation time or effort.

This more comprehensive formula allowed for a richer prioritization context in tasks involving time-critical healthcare scenarios. By adapting the WSJF model to each task domain, the study ensured that prioritization benchmarks were realistic and contextually appropriate [55]. The calculated WSJF scores were used to rank the features and group them into High, Medium, and Low priority categories, as described in Section 6.2.

4.6.3 Prioritization Accuracy Scoring

Each participant's prioritization output was compared to the standard grouping (High, Medium, Low). The accuracy score reflected the number of requirements correctly classified into the same group as the reference. For example, if 7 out of 10 requirements were placed in the correct group, the accuracy score was 0.70. These scores were calculated for both:

- **Manual Tasks (1.1, 2.1) – no AI support**
- **AI-assisted Tasks (1.2, 2.2) – using different XAI formats**

4.6.4 Cognitive Load Analysis

Perceived cognitive load was measured using 7-point Likert scale-derived questions. Each participant rated the mental demand, task difficulty, effort, and confidence after each task. Scores were normalized and averaged for composite analysis. Separate mean load scores were computed for: Tasks without AI (baseline) and Tasks with XAI support, segmented by explanation type. This allowed direct comparisons of mental effort under varying AI support conditions.

4.6.5 Descriptive Statistics

Descriptive statistics were used to summarize the participants' demographics through frequency distributions and to analyze task performance and self-reported cognitive load ratings using means and standard deviations. Cognitive load was assessed on a 7-point Likert scale, where 1 indicated very low demand/effort and 7 indicated very high demand/effort. Average prioritization accuracy scores and cognitive load ratings were computed for both AI-assisted and non-assisted conditions, enabling comparison across XAI techniques and task complexity levels.

Independent- and paired-samples t-tests were used to compare (a) performance and load between AI-assisted and non-assisted conditions, and (b) across explanation formats. While we did not formally test for normality, t-tests are widely used in studies with Likert-scale measures and moderate sample sizes, and we acknowledge this assumption as a limitation.

The null hypotheses stated that there would be no significant differences in (H0a) prioritization accuracy between AI-assisted and non-assisted tasks, (H0b) self-reported cognitive load between AI-assisted and non-assisted tasks, and (H0c) either accuracy or cognitive load across the three explanation formats. The corresponding alternative hypotheses

proposed that AI assistance and explanation format would exert significant effects on accuracy, cognitive load, or both.

4.7 Ethics

The study was conducted in accordance with established ethical guidelines, specifically the Declaration of Helsinki [66]. Participation was entirely voluntary, and informed consent was obtained from all respondents before they began the survey. Participants were clearly informed about the study's purpose, what their participation would involve, and their right to withdraw at any time without consequence. To protect anonymity, no personally identifiable information was collected. All responses were stored securely and were accessible only to the research team. The survey instructions also highlighted the confidentiality of the data and confirmed that it would be used solely for academic research purposes.

4.8 Validity of the Study

Construct, internal, and external validity were addressed through the use of multiple strategies [67]. To support construct validity, the tasks and instructions were standardized, and established measurement tools were adapted, including Likert-scale items assessing perceived mental demand, task difficulty, effort, and confidence in performance. These items are widely used in cognitive load research and provide a reliable basis for capturing subjective workload. To enhance internal validity, task contexts were varied across related tasks to minimize learning effects. For example, in Task 1, participants prioritized features in a loan management system, whereas subsequent banking tasks involved general online banking activities. Similarly, in Task 2, the first task involved emergency doctor appointments, and later tasks involved routine bookings.

External validity was strengthened through participant diversity, ensuring findings were relevant to both academic and industry contexts. Random assignment of participants to one of three XAI technique conditions (text explanations, bar charts, confidence scores) reduced potential bias from prior exposure.[53]

Despite these controls, the study's online administration meant participants completed tasks in uncontrolled environments, potentially introducing distractions. The reliance on self-reported measures also means results may be subject to personal bias; however, validated scales and clear instructions were used to mitigate these risks.[31]

5

Results

5.1 Introduction

The research questions guiding this thesis focused broadly on identifying cognitive load drivers and their effect on decision-making. The focus is also aligned with the role of XAI in shaping participants' cognitive experiences and outcomes during requirements prioritization tasks.

The research questions reflect the nature of the data collected, which specifically evaluated how different forms of XAI, such as bar charts, textual explanations, and confidence scores, affect cognitive load and influence decision outcomes. The questions aim to capture these dynamics more precisely and are as follows:

RQ1: How do different styles of XAI impact cognitive load during decision-making in requirements prioritization?

RQ2: How do different styles of XAI impact the quality of decision-making in requirements prioritization tasks?

RQ3: How do users' preferences for different XAI formats relate to their task performance, perceived mental effort, and trust in AI-supported requirements prioritization?

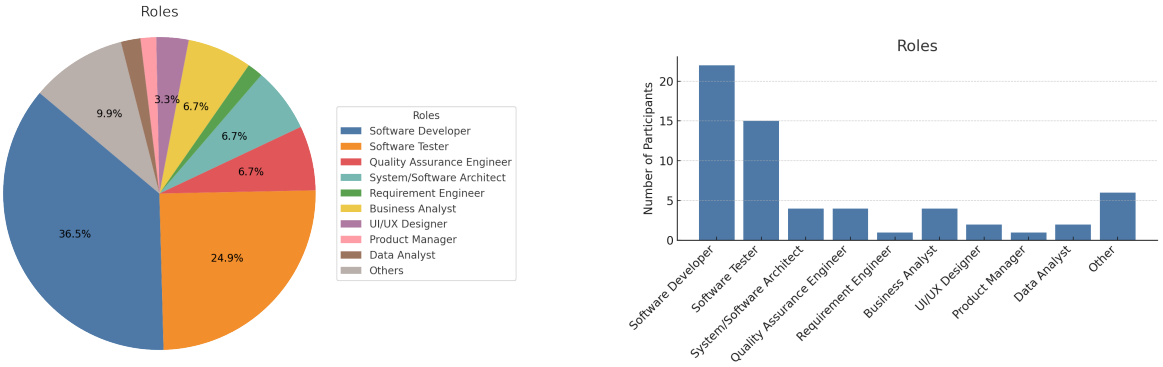
The remainder of this chapter presents the findings aligned with these research questions, beginning with participant demographics, followed by an analysis of XAI influence on cognitive load and decision outcomes.

5.2 Demographics of Survey Participants

To better understand the context in which participants engaged with the decision-making tasks, the survey included three demographic questions: (1) participants' primary professional role, (2) their years of experience in requirements engineering, and (3) how often they prioritize requirements in their current roles. In later sections, these data provide a foundation for interpreting participants' interactions with XAI.

Participants represented a diverse range of roles across the software development lifecycle.

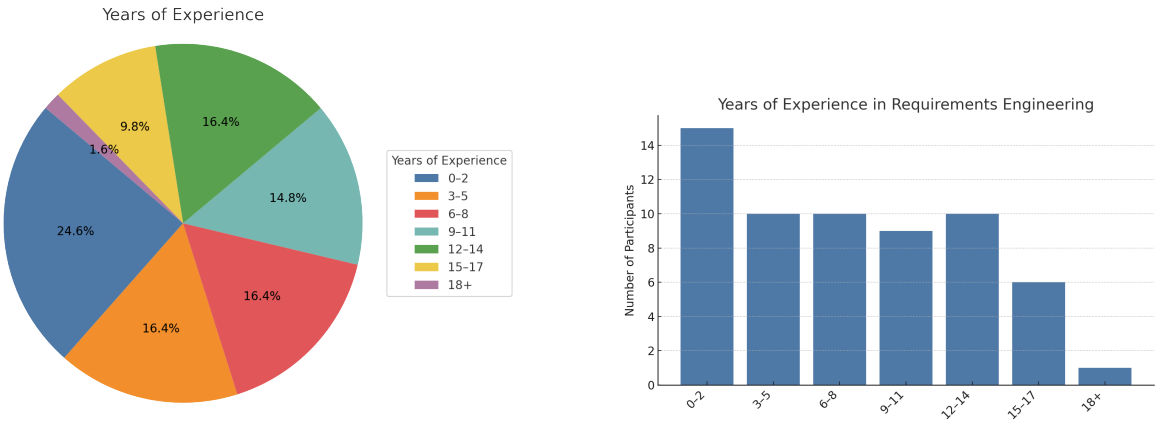
The largest group identified as Software Developers (36.1%), followed by Software Testers (24.6%). Other notable roles included System/Software Architects, Quality Assurance Engineers, Project Managers, Business Analysts, UI/UX Designers, and Requirements Engineers. A small portion also classified themselves under Other roles, including hybrid or interdisciplinary functions. This spread indicates a broad participation base, ensuring the results are informed by varying perspectives across industry roles.
See Figures 5.1 a and b for a visual breakdown of participants' roles.



(a) (b)

Figure 5.1: Distribution of participants' professional roles

The participants' experience in requirements engineering ranged from 0 to 18 years. The mean experience was approximately 7.4 years, with a median of 7 years, indicating a balanced representation of both early-career and seasoned professionals. A few participants reported no experience, while several others had more than a decade of involvement in RE tasks. This distribution reflects a suitable range for analyzing cognitive responses across experience levels.
See Figures 5.2 a and b for a visual breakdown of participants' experience.

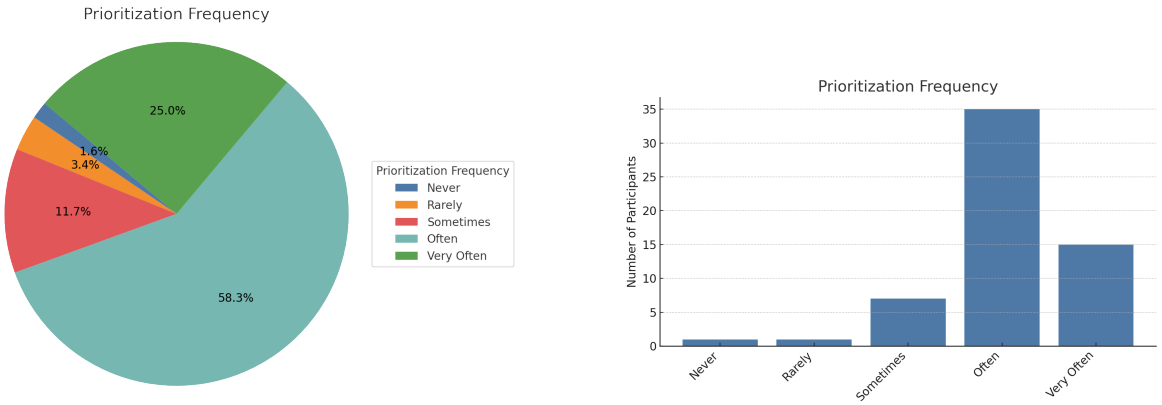


(a) (b)

Figure 5.2: Distribution of participants' experience

When asked how often they prioritize requirements as part of their role, a majority of participants indicated that they engage in this activity either "Often" (57.4%) or "Very

Often" (24.6%). A smaller subset reported doing it only "Sometimes" (11.5%), while very few chose "Rarely" or "Never". These findings confirm that the task of requirements prioritization is a common and regular part of participants' workflows, making them suitable evaluators of XAI support during such decision-making activities.
See Figures 5.3 a and b for visualization of prioritization frequency.



(a) (b)

Figure 5.3: Distribution of participants' prioritization frequency

5.3 Results Aligned with Research Questions

This section presents the results of the survey aligned with the research questions, focusing on how different forms of XAI influence cognitive load and decision quality in requirements prioritization tasks. The findings are organized by each research question.

5.3.1 Overview of Key Task Metrics

Task / XAI Type	Correct Answers	Correct SD	Mental Effort	Effort SD	Task Complexity	Complexity SD	Confidence in Answers	Confidence SD
Task 1.1	3.90	2.17	4.79	1.40	4.43	1.51	4.89	1.59
Task 1.2 – Bar Chart	8.47	3.03	4.53	1.43	4.68	1.63	4.58	1.64
Task 1.2 – Text Explanations	8.52	2.87	4.45	1.36	4.75	1.29	4.40	1.39
Task 1.2 – Confidence Scores	8.50	2.37	4.32	1.49	3.95	1.36	4.90	1.60
Task 2.1	3.47	1.72	4.77	1.56	4.77	1.44	4.88	1.63
Task 2.2 – Bar Chart	4.32	0.23	4.31	1.54	4.27	1.32	4.90	1.51
Task 2.2 – Text Explanations	4.10	0.45	4.37	1.45	4.21	1.39	4.94	1.46
Task 2.2 – Confidence Scores	4.25	0.85	4.00	1.55	4.60	1.61	4.65	1.54

Table 5.1: Summary of average scores across key metrics by task and XAI type.

Before addressing the research questions individually, a high-level overview of participants' performance and self-reported cognitive measures across all four tasks is presented in Table 5.1. This summary includes average scores for correctness, mental effort, task difficulty, and confidence, capturing the general effect of XAI on task experience. The correct answers are in the scale of 1-10 and all the other columns in the scale of 1-7.

The box plots in Figure 5.4 illustrates the distribution of participant responses across the four tasks for all measured metrics. These plots highlight not only the central tendency but also the spread and outliers in the data. Metrics such as effort and confidence display wider distributions across tasks, indicating greater variability in participant perceptions. The individual data points further show how responses are spread within each task.

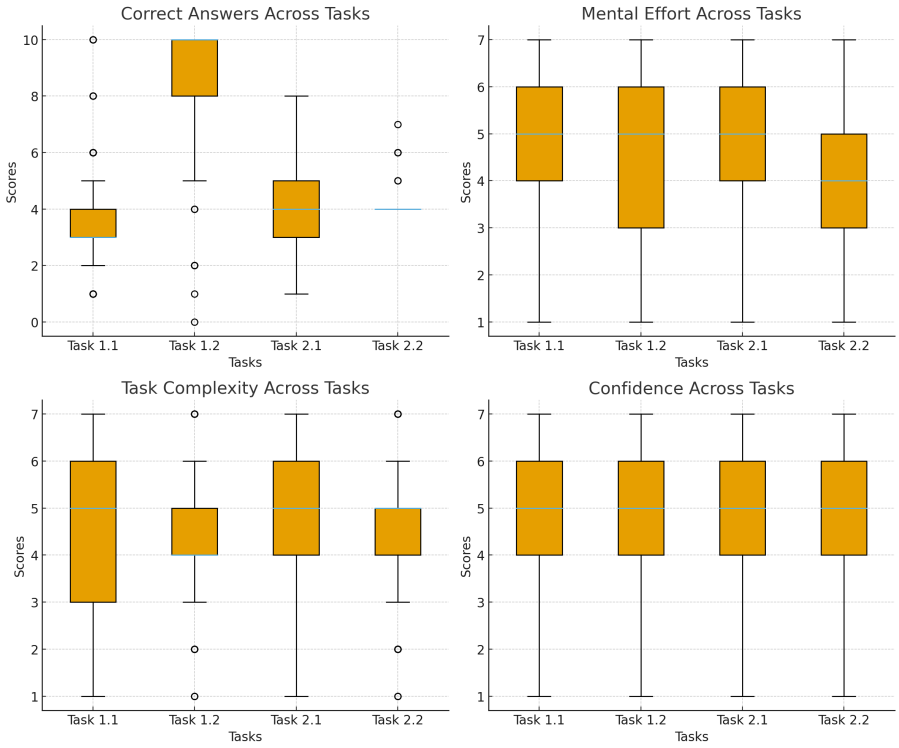


Figure 5.4: Box plots of all participant results

From the graph in Figure 5.5, it is evident that Task 1.2(the first task involving XAI) was associated with the highest correctness and slightly reduced effort and difficulty, suggesting a positive impact of XAI support. Confidence remained relatively stable, with minor variations across tasks. These trends provide context for the more detailed analyses that follow in the upcoming sections.

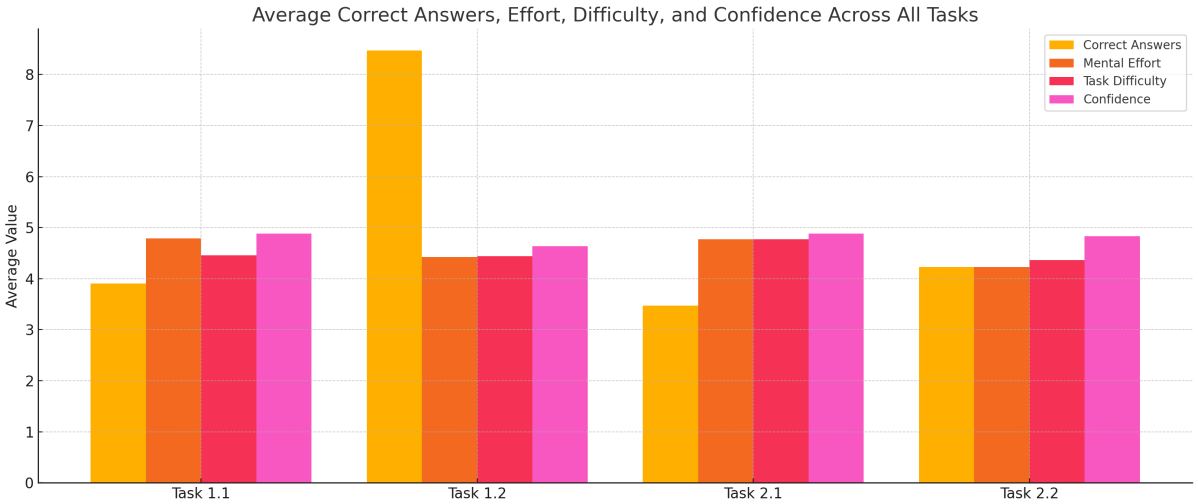


Figure 5.5: Key Task Metrics

5.3.2 RQ1: How do different styles of XAI impact cognitive load during decision-making in requirements prioritization?

5.3.2.1 Correlation Between Tasks: Evidence of XAI's Influence on Cognitive Load

To understand how different styles of XAI influence cognitive processing during decision-making in requirements prioritization, this section focuses on two key indicators of cognitive load: mental effort and task difficulty. These metrics reflect participants' perceived cognitive burden while working through tasks with and without XAI support.

To perceive how XAI influenced cognitive load, this study used both Spearman correlation and paired t-tests to see different types of relationships within the data. Spearman correlation was chosen because the variables of mental effort and task difficulty were measured on ordinal Likert scales, making this non-parametric test more appropriate than the Pearson correlation. It allowed the analysis to detect trends in how cognitive load changed between tasks with and without XAI, without assuming a linear relationship or normal distribution.

The results indicate moderate to strong correlations between the metrics across tasks. Full details of the correlation coefficients are provided in Appendix A.1.

Mental Effort: Participants' reported mental effort showed strong positive relationships between several tasks that included XAI. For instance, effort ratings between Task 1.2 and Task 2.2 were closely related ($r = 0.582$, $\rho < 0.001$), as were those between Task 1.2 and Task 2.1 ($r = 0.495$, $\rho < 0.001$), and between Task 2.1 and Task 2.2 ($r = 0.492$, $\rho < 0.001$). These findings suggest that participants tended to experience similar levels of mental workload when working with XAI, even though the tasks varied. This might reflect a consistent way of thinking or approaching the tasks when support was available.

Task Difficulty: A similar pattern appeared in the way participants rated task difficulty. There were strong positive correlations between Task 1.2 and Task 2.2 ($r = 0.571$, $\rho < 0.001$), Task 1.2 and Task 2.1 ($r = 0.473$, $\rho < 0.001$), and Task 2.1 and Task 2.2 ($r = 0.475$, $\rho < 0.001$). These findings suggest that XAI influenced how challenging the tasks felt, making perceptions of difficulty more consistent across the survey. However, when comparing Task 1.1, which did not include XAI, to later tasks that did, the correlations were negative. For example, the link between Task 1.1 and Task 2.1 was($r = 0.356$ and $\rho = 0.005$), and between Task 1.1 and Task 2.2 it was ($r = 0.294$ and $\rho = 0.020$). These results may reflect a shift in how participants judged complexity, depending on whether they had XAI support.

Overall, the correlation results suggest that the presence of XAI influences cognitive load factors, especially effort and difficulty across tasks. Participants reported more consistent cognitive responses in XAI-supported tasks, while performance remained task-dependent and did not consistently improve with XAI.

5.3.2.2 Statistical Differences in Cognitive Load Measures

To assess whether the presence of XAI meaningfully influenced cognitive load or decision performance, paired t-tests were conducted across the dimensions of effort and difficulty.

Comparisons were made between tasks with and without XAI (Tasks 1.1/2.1 vs. 1.2/2.2), as well as between the two XAI-supported tasks themselves (1.2 vs. 2.2). Since the same participants completed both tasks, the paired design helped control for individual differences, focusing the analysis on the effect of the XAI intervention itself. Together, these tests provided a more complete view of whether and how XAI influenced users' mental effort and perceived difficulty across varying task conditions.

Comparison	**t-statistic**	**ρ-value**	**Interpretation**
Effort_1_1 vs Effort_1_2	1.403	0.1658	This result is not statistically significant
Effort_2_1 vs Effort_2_2	2.928	0.0048	This result is statistically significant
Effort_1_1 vs Effort_2_1	0.058	0.9539	This result is not statistically significant
Effort_1_2 vs Effort_2_2	1.177	0.2437	This result is not statistically significant
Difficulty_1_1 vs Difficulty_1_2	0.060	0.9524	This result is not statistically significant
Difficulty_2_1 vs Difficulty_2_2	2.313	0.0241	This result is statistically significant
Difficulty_1_1 vs Difficulty_2_1	-1.040	0.3023	This result is not statistically significant
Difficulty_1_2 vs Difficulty_2_2	0.471	0.6394	This result is not statistically significant

Table 5.2: Paired t-test results for mental effort and task difficulty across tasks.

Mental Effort: The analysis of perceived mental effort revealed some variation across conditions. A significant difference was found between Task 2.2 and Task 2.1 ($t = 2.928$, $\rho = 0.0048$), where the task supported by XAI appeared to require lesser levels of cognitive engagement. This suggests that the presence of XAI may have affected how participants approached or processed the task.

Task Difficulty: In terms of task difficulty, the results also pointed to some variation linked to XAI. Task 2.2 was rated as significantly less difficult than Task 2.1 ($t = 2.313$, $\rho = 0.0241$). This suggests that the XAI intervention may have altered how challenging the task felt to participants.

By contrast, no notable differences in difficulty were reported between Task 1.1 and Task 1.2 ($\rho = 0.9524$), nor between the two XAI tasks, Task 1.2 and Task 2.2 ($\rho = 0.6394$). These results imply that while XAI had some influence on how difficulty was experienced, its effect was not consistent across all task pairs.

5.3.2.3 Impact of Different XAI Types on Cognitive Load

To answer RQ1, we begin by presenting the average values reported by participants for each XAI-supported task. Table 5.1 above shows the mean number of correct answers, self-reported mental effort, perceived task complexity, and confidence in answers across the three XAI types in Tasks 1.2 and 2.2.

The descriptive statistics reveal subtle differences in how participants experienced each XAI format. For example, confidence scores were associated with lower effort ratings, while bar charts and text explanations tended to result in higher task complexity ratings

depending on the task. These patterns are further investigated through correlation and significance testing below.

5.3.2.4 Correlation Test for different XAI

Following the descriptive results in Table 5.1, which provided insight into the average performance and perceptions for each XAI type, we now examine the consistency of participant experience across the two XAI-supported tasks (1.2 and 2.2). Spearman correlation tests were used to evaluate whether participants showed similar patterns of correctness, effort, difficulty, and confidence across different XAI types.

XAI Type	**Metric Pair**	**Spearman correlation**	**Interpretation**
Bar Chart	Effort_1.2 vs Effort_2.2	0.656	Positive
Bar Chart	Difficulty_1.2 vs Difficulty_2.2	0.615	Positive
Confidence Scores	Effort_1.2 vs Effort_2.2	0.585	Positive
Confidence Scores	Difficulty_1.2 vs Difficulty_2.2	0.694	Positive
Text Explanations	Effort_1.2 vs Effort_2.2	0.493	Positive
Text Explanations	Difficulty_1.2 vs Difficulty_2.2	0.499	Positive

Table 5.3: Spearman correlation between Tasks 1.2 and 2.2 across key metrics for each XAI type.

Bar Charts: Between Task 1.2 and Task 2.2, positive correlations were observed in both mental effort (r = 0.656) and task difficulty (r = 0.615). This suggests that participants experienced bar charts as similarly demanding in both decision-making scenarios.

Confidence Scores: Correlations between tasks were similarly positive when participants interacted with confidence scores. Effort and difficulty both yielded strong positive correlations (r = 0.585 and r = 0.694, respectively), pointing to a consistent cognitive experience across different contexts.

Text Explanations: Text-based explanations produced positive correlations in both effort (r = 0.493) and difficulty (r = 0.499), indicating that the perceived cognitive demand remained relatively stable across tasks.

5.3.2.5 Statistical Differences in Cognitive Load by XAI Type

To determine whether the same XAI type produced significantly different experiences across two tasks, we conducted paired t-tests comparing participants' ratings between Task 1.2 and Task 2.2 for each form of XAI. We analyzed their mental effort and perceived difficulty in answering.

XAI Type	**Metric**	**Task 1.2 Mean**	**Task 2.2 Mean**	**Mean Difference**	**t-statistic**	**p-value**	**Interpretation**
Bar Chart	Effort	4.53	4.37	-0.16	0.567	0.5778	Result is not statistically significant
Bar Chart	Difficulty	4.68	4.21	-0.47	1.531	0.1431	Result is not statistically significant
Confidence Scores	Effort	4.32	4.32	0.00	0.000	1	Result is not statistically significant
Confidence Scores	Difficulty	3.95	4.27	0.32	-1.322	0.2005	Result is not statistically significant
Text Explanations	Effort	4.45	4.00	-0.45	1.308	0.2063	Result is not statistically significant
Text Explanations	Difficulty	4.75	4.60	-0.15	0.420	0.6794	Result is not statistically significant

Table 5.4: Paired t-test comparison of Task 1.2 and Task 2.2 across XAI types

Bar Charts: No significant differences were seen in how much mental effort participants reported ($\rho = 0.578$), how difficult the tasks felt ($\rho = 0.143$). These stable scores suggest that bar charts offered a familiar and steady experience, even if the final outcomes didn't always match that comfort.

Confidence Scores: Participants rated their mental effort and perceived difficulty at almost the same level in both tasks (all ρ-values above 0.20). This points to a kind of cognitive consistency, participants seemed to process and trust the confidence scores similarly across tasks, even though the actual results were not as consistent.

Text Explanations: The scores for effort and difficulty stayed relatively stable (ρ-values all above 0.20). This suggests that while participants felt just as engaged and assured using text explanations, those explanations may not have always helped them make better decisions, depending on the task.

5.3.3 RQ2: How do different styles of XAI impact the quality of decision-making in requirements prioritization tasks?

5.3.3.1 Correlation Between Tasks: Evidence of XAI's Influence on Decision Quality

To understand how different styles of XAI influence the quality of decision-making in requirements prioritization, this section focuses on two key indicators of cognitive load: Correctness and Confidence. These metrics reflect participants' perceived cognitive burden while working through tasks with and without XAI support

The results indicate moderate to strong correlations between the metrics across tasks. Full details of the correlation coefficients are provided in Appendix A.1.

Correctness: The correlation between Task 1.2 and Task 2.2, both of which included XAI support was negative (r = 0.347, $\rho = 0.006$). This suggests that participants who performed well in one XAI-supported task did not necessarily do well in the other. In some cases, doing well in one task actually aligned with performing less accurately in the next. On the other hand, a positive correlation was observed between Task 1.2 and Task 2.1 (r = 0.263, $\rho = 0.039$). This may indicate that task performance was shaped not only by the presence of XAI, but also by the nature of the tasks themselves. Other comparisons, such as between Task 1.1 and Task 1.2 or between Task 2.1 and Task 2.2, were not statistically significant. These results suggest that correctness did not consistently carry over across tasks, regardless of whether XAI was used.

Confidence: Confidence ratings showed the strongest and most consistent correlations, especially across tasks that included XAI. The connection between Task 1.2 and Task 2.2 was strong (r = 0.718, $\rho < 0.001$), followed by the link between Task 1.2 and Task 2.1 (r = 0.639, $\rho < 0.001$), and between Task 2.1 and Task 2.2 (r = 0.653, $\rho < 0.001$). These patterns suggest that XAI may have helped participants feel more certain in their choices, even as the task structure changed. By contrast, confidence in Task 1.1, which had no XAI, did not significantly correlate with the other tasks. This might mean that the presence of XAI made the experience of decision-making feel more stable and reliable overall.

5.3.3.2 Statistical Differences in Cognitive Load Measures

To assess whether the presence of XAI meaningfully influenced decision quality, paired t-tests were conducted across dimensions of correctness and confidence. Comparisons were made between tasks with and without XAI (Tasks 1.1/2.1 vs. 1.2/2.2), as well as between the two XAI-supported tasks themselves (1.2 vs. 2.2).

Comparison	**t-statistic**	**ρ-value**	**Interpretation**
Correct_1_1 vs Correct_1_2	-10.624	0.0000	Result is Statistically significant
Correct_2_1 vs Correct_2_2	-2.595	0.0118	Result is Statistically significant
Correct_1_1 vs Correct_2_1	1.491	0.1410	Result is Not statistically significant
Correct_1_2 vs Correct_2_2	10.905	0.0000	Result is Statistically significant
Confidence_1_1 vs Confidence_1_2	0.823	0.4135	Result is Not statistically significant
Confidence_2_1 vs Confidence_2_2	0.314	0.7547	Result is Not statistically significant
Confidence_1_1 vs Confidence_2_1	0.000	1.0000	Result is Not statistically significant
Confidence_1_2 vs Confidence_2_2	-1.450	0.1523	Result is Not statistically significant

Table 5.5: Paired t-test results for correctness and confidence across tasks.

Correctness: When comparing correctness scores between tasks, several notable differences emerged. Task 1.2, which included support from an XAI system, showed a clear improvement over Task 1.1, which did not include any AI assistance ($t = -10.624$, $\rho < 0.001$). A similar result was observed in the second task pair, where Task 2.2, again supported by XAI, outperformed Task 2.1 ($t = -2.595$, $\rho = 0.0118$). These findings point to a consistent pattern in which tasks that incorporated XAI were associated with higher correctness scores than those without it.

Interestingly, even when comparing two tasks that both included XAI, Task 1.2 and Task 2.2, the difference in performance remained statistically significant ($t = 10.905$, $\rho < 0.001$). This suggests that factors beyond the mere presence of XAI, such as the way information was presented or the specific nature of each task, may have contributed to performance variation. On the other hand, when comparing the two tasks that lacked XAI, namely Task 1.1 and Task 2.1, there was no significant difference in correctness ($\rho = 0.1410$), which further highlights the impact of XAI within these scenarios.

Confidence: Confidence ratings remained relatively steady across all task conditions. None of the comparisons produced statistically significant results, including those between Task 1.1 and Task 1.2 ($\rho = 0.4135$), Task 2.1 and Task 2.2 ($\rho = 0.7547$), or Task 1.2 and Task 2.2 ($\rho = 0.1523$). This suggests that participants' level of self-assuredness in their responses was largely unaffected by whether XAI was present or not. Despite differences in correctness or effort, the introduction of AI support did not appear to influence how confident participants felt about their decisions.

5.3.3.3 Impact of Different XAI Types on Decision Quality

5.3.3.4 Correlation Test for different XAI

Bar Chart: Confidence ratings showed a strong positive correlation ($r = 0.820$), indicating that bar charts consistently contributed to a sense of confidence across tasks. Correctness scores showed a positive correlation ($r = 0.500$), which may reflect a connection between perceived clarity and actual performance.

XAI Type	Metric Pair	Spearman Correlation	Interpretation
Bar Chart	Correct_1.2 vs Correct_2.2	0.500	Positive
Bar Chart	Confidence_1.2 vs Confidence_2.2	0.820	Positive
Confidence Scores	Correct_1.2 vs Correct_2.2	-0.382	Negative
Confidence Scores	Confidence_1.2 vs Confidence_2.2	0.758	Positive
Text Explanations	Correct_1.2 vs Correct_2.2	-0.239	Negative
Text Explanations	Confidence_1.2 vs Confidence_2.2	0.507	Positive

Table 5.6: Spearman correlation between Tasks 1.2 and 2.2 across key metrics for each XAI type.

Confidence Scores: Confidence in answers remained positively correlated (r = 0.758), suggesting that participants trusted this form of explanation across tasks. However, the correctness scores showed a negative correlation (r = –0.382), implying that high self-assurance may not always align with task accuracy.

Text Explanations: Confidence levels were positively correlated (r = 0.507), suggesting participants felt equally sure in both instances. However, correctness showed a negative correlation (r = –0.239), which may indicate variability in how effectively these explanations supported accurate decisions.

5.3.3.5 Statistical Differences in Decision Quality by XAI Type

Bar Chart: Participants performed noticeably better in Task 1.2 compared to Task 2.2 when using bar charts, as shown by a significant difference in correctness scores (t = 5.678, $\rho < 0.001$). Even though the same type of explanation was used, something about the second task may have made it harder to apply the information as effectively. On the other hand, no significant differences were seen in how confident they were in their decisions ($\rho = 0.130$).

Confidence Scores: The drop in correctness was again significant between Task 1.2 and Task 2.2 when confidence scores were used (t = 7.213, $\rho < 0.001$). Despite this, participants rated confidence at almost exactly the same level in both tasks (all ρ-values above 0.20). This points to a kind of cognitive consistency, participants seemed to process and trust the confidence scores similarly across tasks, even though the actual results were not as consistent.

XAI Type	Metric	Task 1.2 Mean	Task 2.2 Mean	Mean Difference	t-statistic	ρ-value	Interpretation
Bar Chart	Correct	8.47	4.11	-4.37	5.678	0	Result is statistically significant
Bar Chart	Confidence	4.58	4.95	0.37	-1.587	0.1298	Result is not statistically significant
Confidence Scores	Correct	8.50	4.32	-4.18	7.213	0	Result is statistically significant
Confidence Scores	Confidence	4.91	4.91	0.00	0.000	1	Result is not statistically significant
Text Explanations	Correct	8.45	4.25	-4.20	5.581	0	Result is statistically significant
Text Explanations	Confidence	4.40	4.65	0.25	-0.839	0.4120	Result is not statistically significant

Table 5.7: Paired t-test comparison of Task 1.2 and Task 2.2 across XAI types

Text Explanations: For text-based explanations the correctness declined significantly from Task 1.2 to Task 2.2 (t=5.581, $\rho<0.001$). But again, the scores for confidence stayed relatively stable (ρ-values all above 0.20). This suggests that while participants felt just as engaged and assured using text explanations, those explanations may not have always helped them make better decisions, depending on the task.

Across all three XAI types, the only consistent significant change between tasks was in correctness scores, with participants performing better in Task 1.2. However, their self-reported effort, difficulty, and confidence remained statistically unchanged in most cases. This indicates that while participants perceived their cognitive load as stable, the effectiveness of the XAI types in supporting correct decisions varied with context, potentially due to differences in task structure or complexity.

5.3.4 RQ3: How do users' preferences for different XAI formats relate to their task performance, perceived mental effort, and trust in AI-supported requirements prioritization?

5.3.4.1 Participant Preferences for XAI Types

Before analyzing how participant preferences relate to decision quality, effort, and confidence, we summarize the distribution of preferences for each XAI type in Figure 5.6. Participants were asked to indicate which explanation they found the easiest to understand, the most useful for decision-making, and which they preferred overall.

As shown, bar charts were most often selected as the easiest to understand, while text explanations and confidence scores were more frequently rated as most useful and overall preferred. These patterns suggest that while participants found bar charts visually simple, they may have valued the depth or clarity offered by other explanation formats during actual decision-making. The following subsections investigate whether these subjective preferences influenced task performance or cognitive load.

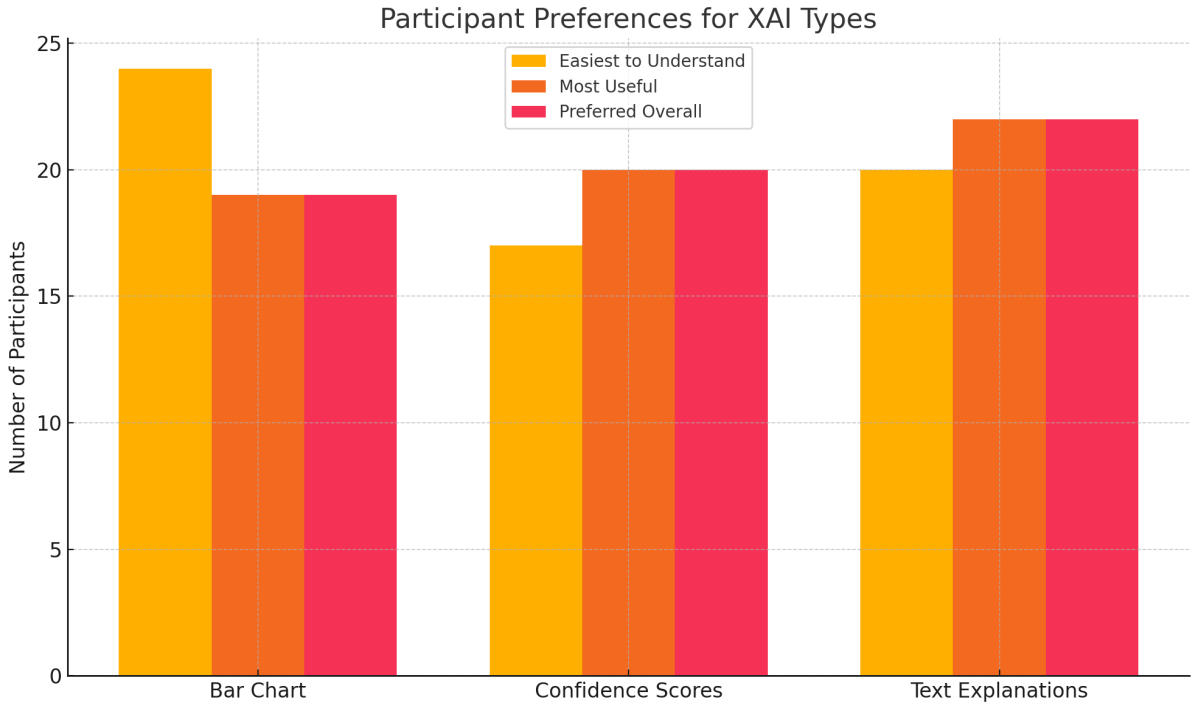


Figure 5.6: Distribution of Participant Preferences for Each XAI Type by Category

5.3.4.2 Correlation Between XAI Preferences and Decision Quality

To explore how participants' subjective preferences and perceptions of XAI types influenced the quality of their decision-making, a series of Spearman correlation tests were conducted. These correlations compare three preferences related survey variables: ease of understanding, perceived usefulness and overall preference against the correctness of participants, mental effort, perceived difficulty, and confidence in Tasks 1.2 and 2.2, where XAI was present.

No correlations were found between most preference variables and performance, with one low correlation between Confidence Scores and Correctness in Task 2.2. Details are in Appendix A.2

Ease of Understanding and Performance: Participants who rated a specific XAI type as easiest to understand showed notable relationships with performance in Task 2.2: Confidence scores were positively correlated with correctness in Task 2.2 ($r = 0.331$), suggesting that perceiving this format as easier to understand was linked with better outcomes in later tasks. In contrast, bar charts were negatively correlated with correctness in Task 2.2 ($r = –0.210$), possibly indicating that although they were preferred for clarity by some, they did not necessarily lead to improved decisions. Text explanations showed weak or no correlation with correctness, effort, or confidence.

Overall, these results suggest that ease of understanding alone does not guarantee better performance, though confidence scores appear to have offered some benefit.

Perceived Usefulness and Cognitive Load: When asked which XAI they found most useful for decision-making: Bar charts had a negative correlation with difficulty ($r = –0.241$) and correctness in Task 2.2 ($r = –0.156$), suggesting that despite their appeal, they may have contributed to cognitive strain or confusion in actual performance. Confidence scores and text explanations showed no correlations across all metrics, indicating that perceived usefulness was not a strong predictor of how participants experienced the task cognitively.

Overall Preference and Task Performance: General preference showed only weak trends. A positive correlation between confidence scores and correctness in Task 2.2 ($r = 0.277$) again reinforced the idea that some formats may have helped performance modestly. Other correlations, including those for effort, difficulty, and confidence, were negligible across all three formats.

In sum, the data shows that participants' preferences or perceived ease of use do not strongly correlate with actual decision performance. However, some patterns emerged, confidence scores were both positively perceived and modestly linked with better correctness in Task 2.2, suggesting they may have supported clearer judgment. In contrast, bar charts, despite being widely seen as easy to understand, did not lead to stronger decisions and were in some cases associated with lower correctness or higher perceived difficulty.

5.3.4.3 Significance Between XAI Preferences and Decision Quality

To complement the correlation analysis, we conducted independent samples t-tests comparing participants who preferred a specific XAI format (bar chart, text explanation, or

confidence scores) against those who did not. This analysis evaluated whether preference for an XAI type had a statistically significant impact on participants' correctness, perceived effort, task difficulty, and confidence across Tasks 1.2 and 2.2.

Correctness: For all three XAI types, the difference in correctness between those who preferred a particular format and others was not statistically significant. The closest to significance was observed for confidence scores in Task 2.2 ($\rho = 0.0719$), where participants who preferred this format scored higher on average (4.50) than others (4.10). While not conclusive, this result hints that confidence scores might have supported more accurate decisions.

Effort and Difficulty: Across all XAI types and both tasks, no significant differences were found in perceived mental effort or task difficulty between preferred and non-preferred groups. Interestingly, participants who preferred bar charts rated Task 2.2 as less difficult on average (mean = 3.89) compared to others (mean = 4.57), with the result approaching significance ($\rho = 0.0973$). A similar trend was seen for text explanations in Task 1.2 difficulty ($\rho = 0.0981$), though again this was not statistically significant.

Confidence: Participants' self-reported confidence levels did not differ significantly based on preferred XAI type across either task. All p-values remained well above 0.1, indicating no measurable effect of XAI preference on confidence in decision-making.

Although none of the observed differences reached statistical significance, the consistently higher correctness scores among those who preferred confidence scores, particularly in Task 2.2, suggest this format may have had a positive but subtle effect on decision quality. Meanwhile, bar charts and text explanations showed no reliable advantage in enhancing performance or reducing cognitive load, even when they were favored by participants. These findings reinforce the idea that subjective preference alone does not reliably predict objective performance or perceived task difficulty, though certain trends warrant further investigation in larger studies.

5.3.4.4 Significance between Correctness and perceived easiest to understand XAI

To deepen the analysis of XAI's influence on decision-making, we examined whether participants' perceived ease of understanding of a specific XAI type was linked to actual performance, measured by correctness in Tasks 1.2 and 2.2. Specifically, we ran independent samples t-tests comparing correctness scores between participants who selected a given XAI (bar chart, confidence scores, or text explanations) as the easiest to understand and those who did not.

Preference	Performance Metric	Mean (Preferred Group)	Mean (Others)	t-statistic	ρ-value	Significance
Bar Chart	Correct_1.2	8.96	8.17	1.163	0.2499	Not significant
Bar Chart	Correct_2.2	4.08	4.32	-1.584	0.1186	Not significant
Text Explanations	Correct_1.2	8.15	8.63	-0.643	0.5243	Not significant
Text Explanations	Correct_2.2	4.10	4.29	-1.160	0.2519	Not significant
Confidence Scores	Correct_1.2	8.18	8.59	-0.502	0.6198	Not significant
Confidence Scores	Correct_2.2	4.59	4.09	2.074	0.0518	Not significant

Table 5.8: Comparison of correctness scores based on participants' preferred XAI type.

Participants in Task 1.2 who identified bar charts as the easiest to understand had slightly higher mean correctness (M = 8.96) than others (M = 8.17), though this difference was not statistically significant ($\rho = 0.25$). Those who selected text explanations or confidence scores as easiest also showed no significant advantage in Task 1.2 correctness ($\rho = 0.52$ and $\rho = 0.62$, respectively).

In Task 2.2 the only near-significant result emerged for confidence scores. Participants who rated this format as the easiest to understand scored higher (M = 4.59) than others (M = 4.09), with a ρ-value of 0.052. While not statistically significant at the conventional 0.05 level, this suggests a possible link between perceived explainability and performance in more complex tasks. No significant differences were observed for bar charts ($\rho = 0.119$) or text explanations ($\rho = 0.252$).

While these results do not demonstrate statistically significant effects, they offer valuable insight. Participants who perceived confidence scores as more understandable may have performed better in challenging prioritization tasks. This trend echoes earlier findings and suggests that perceived explainability, particularly of confidence scores, might play a role in supporting better decision-making.

5.3.4.5 Significance between mental effort and perceived overall preferred XAI

This analysis aimed to uncover whether favoring a particular XAI correlated with feeling more cognitively at ease during the tasks.

Participants in Task 1.2 who preferred bar charts reported slightly lower mental effort (M = 4.32) than others (M = 4.48), but this difference was not statistically significant ($\rho = 0.71$). For those who favored text explanations, the effort difference was negligible ($\rho = 0.94$). Interestingly, those who preferred confidence scores reported slightly higher effort (M = 4.55) compared to others (M = 4.37), though again this difference was not significant ($\rho = 0.61$).

Similar trends continued in Task 2.2, where none of the differences in reported effort reached statistical significance: Bar chart preference: $\rho = 0.68$; Text explanations: $\rho = 0.99$; Confidence scores: $\rho = 0.64$.

These findings suggest that a participant's overall preference for a specific XAI format did not meaningfully influence their perceived cognitive effort during the prioritization tasks. While individual opinions may guide comfort or familiarity, they did not significantly reduce (or increase) the mental workload experienced across XAI conditions.

5.3.4.6 Significance between Trust in XAI and Reported Confidence in Decisions

Participants were grouped based on whether they rated the XAI as "Very Trustworthy," "Trustworthy," "Somewhat Trustworthy," or "Neither Trustworthy nor Untrustworthy."

Participants in Task 1.2 who rated the XAI as Very Trustworthy had the highest average confidence (M = 6.00), compared to others (M = 4.59), but the result was not statistically significant ($\rho = 0.39$). Similarly, those who found the XAI simply Trustworthy reported

higher confidence (M = 4.86 vs. 4.33), but again, this difference did not reach significance (ρ = 0.20). Ratings of Somewhat Trustworthy and Neither also showed no significant differences in Task 1.2 confidence.

Once again in Task 2.2, Very Trustworthy ratings were associated with the highest mean confidence (M = 6.50 vs. 4.78), yet this result was not statistically significant (p = 0.14). The only statistically significant difference emerged for participants who selected "Neither Trustworthy nor Untrustworthy." This group reported lower confidence (M = 3.73) compared to others (M = 5.08), with a significant ρ-value of 0.040.

Although most group differences were not statistically significant, the data suggest that higher trust in XAI may correspond to higher reported confidence during decision-making. Notably, those who expressed neutral trust (neither high nor low) had significantly lower confidence, indicating that ambivalence toward XAI trustworthiness might reduce a user's assurance in their choices.

5.3.5 Participant Perceptions of XAI Trust, Confidence, and Future Use

Following the completion of all prioritization tasks, participants were asked a series of general reflection questions to assess their perceptions of XAI support. These included items measuring perceived trustworthiness, the degree to which XAI boosted their confidence in decisions, and their comfort with using similar explanations in the future. The results are visualized in Figure 5.7.

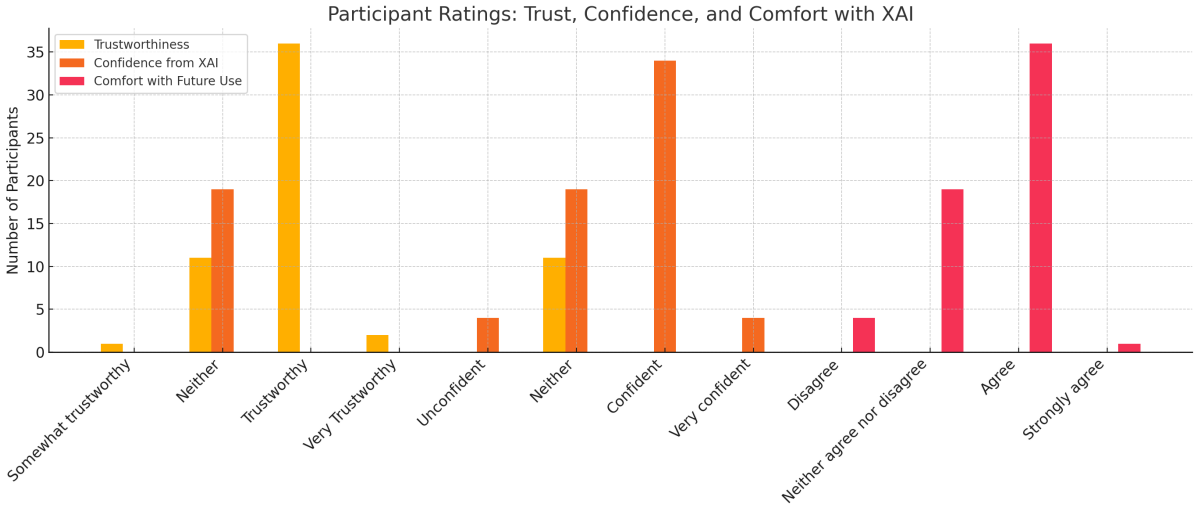


Figure 5.7: Participant Ratings of Trustworthiness, Confidence, and Comfort with Future Use of XAI

The graph reveals that most participants found the XAI either "trustworthy" or "very trustworthy", with only a small number expressing skepticism. A majority also reported feeling confident or very confident in their decisions when supported by XAI. Finally, comfort with using XAI in future decision-making scenarios was overwhelmingly positive, with most respondents selecting "agree" or "strongly agree".

These findings indicate a general openness toward integrating XAI into requirements

engineering, provided the explanations remain interpretable and supportive of user trust.

6

Discussion

6.1 RQ1: How do different styles of XAI impact cognitive load during decision-making in requirements prioritization?

The results indicated that while mental effort and task difficulty were moderately consistent across tasks with and without XAI, these measures did not reveal strong or systematic reductions in cognitive load due to explanation support. Although participants tended to report similar levels of effort and difficulty when using the same XAI type in different tasks, these trends did not always reach statistical significance and varied depending on task pairing. The observed correlations suggest that XAI support may contribute to shaping how users cognitively experience tasks, but this influence appears more subtle than expected.

The t-test results further support this interpretation. In particular, task difficulty and effort ratings remained relatively stable across task variations, suggesting that even though XAI assistance was available, it did not always lead to a reduction in perceived mental demand. This points to a possible gap between the intended clarity of explanations and their actual cognitive impact, especially when task complexity changes. While a few significant changes were observed in effort ratings between specific task pairs, overall differences in cognitive load indicators were minimal, reinforcing the view that XAI support alone does not consistently reduce mental effort or perceived difficulty.

These findings align with the foundational principles of Cognitive Load Theory, as proposed by Sweller et al. [60], which suggest that effective instructional design must manage intrinsic and extraneous load to optimize user processing. In the context of XAI, however, the present results highlight that merely providing explanations is not sufficient to ease cognitive demands, especially in tasks where users are required to integrate explanations with decision-making under uncertainty. Similar patterns were also discussed in the work of Ahmad et al. [2], where physiological measures revealed that explanation-rich environments did not always correlate with reduced mental load, suggesting that explanation usefulness is highly context-dependent. Furthermore, insights from Herm [30] also emphasize that the relationship between XAI interpretability and cognitive burden is nuanced, often requiring personalization or tailoring to the user's mental model to be truly effective.

Together, these findings underscore the importance of moving beyond one-size-fits-all explanation strategies in decision support systems. While XAI offers interpretability, its cognitive benefits may only manifest when explanations are closely aligned with users' prior knowledge, task expectations, and the complexity of information processing required. This highlights a broader implication for AI design in requirements engineering, where explanation design must consider both informational clarity and cognitive economy.

6.2 RQ2: How do different styles of XAI impact the quality of decision-making in requirements prioritization tasks?

The results revealed that XAI support had a measurable but inconsistent influence on participants' task performance. While participants performed slightly better on tasks with explanation support in some cases, these differences were not statistically significant across all task pairs. This suggests that although XAI may assist users in navigating complex decision-making scenarios, its overall impact on accuracy is modest and highly dependent on task-specific factors. Notably, even when performance improved with certain XAI formats, these gains were not sustained consistently across all combinations of tasks and explanation types.

The t-test comparisons showed that the presence of XAI explanations led to significant improvements in correctness in some task transitions but not in others. These mixed results point to the possibility that XAI's benefits are more likely to surface when users are already somewhat familiar with the decision context or explanation style. In other cases, even with explanation support, performance remained stable, indicating that users may not have fully integrated or trusted the guidance provided. The overall pattern suggests that while XAI has potential to guide user decisions, it does not guarantee performance improvements across varying task complexities.

These findings align with the work of Appel et al. [4], who noted that cognitive response to information is strongly influenced by the nature of the task and situational demands. Their study highlighted that explanation benefits often depend on how well users can adapt to the task environment, particularly under time constraints or shifting goals. Similar results were observed by Funke and Galster [22], who found that support systems designed to ease workload in high-stakes environments improved performance in certain scenarios, but not uniformly across all task types. Both studies reinforce the idea that external aids like XAI can be helpful, but only when contextually and cognitively appropriate.

However, these findings stand in contrast to the assumptions proposed by Kirschner et al. [37], who argue that scaffolding mechanisms such as instructional explanations should consistently enhance performance by reducing extraneous cognitive load. The lack of consistent gains in the current results challenges this notion and underscores that interpretability alone may not be enough. When explanation design fails to match user expectations or when cognitive resources are already taxed, additional information might not translate into better outcomes.

Overall, these results emphasize the need for more adaptive and user-sensitive approaches to designing explanation support. XAI should not only inform but also respond to the dynamic needs of users, taking into account their familiarity, task history, and preferred reasoning strategies. Without this alignment, explanation tools may fall short of enhancing decision quality in complex requirements engineering scenarios.

6.3 RQ3: How do users' preferences for different XAI formats relate to their task performance, perceived mental effort, and trust in AI-supported requirements prioritization?

This research question explored whether participants' preferences for different explanation formats had any notable effect on perceived cognitive load. The analysis revealed no significant correlation between the preferred explanation format and self-reported mental effort or task difficulty. Regardless of which format participants favored, their cognitive load indicators remained largely unchanged. This suggests that subjective preferences, while useful for gauging perceived usefulness or appeal, may not necessarily translate into measurable cognitive relief.

The results indicated that even when participants rated certain XAI types as easier to understand or more helpful, these preferences did not correspond with lower cognitive effort or reduced difficulty. For instance, some formats like bar charts were frequently chosen as the most useful, yet those choices did not align with decreased mental workload. This points to a possible disconnect between what users believe helps them and what actually reduces their cognitive strain during task execution. It may also reflect the influence of individual differences in reasoning style or previous exposure to visual aids.

The results align with the work of Teteris et al.[21], who reported that learners tend to maintain relatively stable levels of cognitive effort during complex training simulations, even when additional support materials are available. Their study emphasized that familiarity and task design often play a stronger role in cognitive experience than subjective preferences. Similarly, Sami et al. [55] noted in the context of AI-based elicitation tools that user feedback often favors certain explanation types, but these preferences do not always align with improvements in cognitive processing or output quality. Both studies reinforce the notion that user satisfaction does not always mirror actual reductions in workload.

On the other hand, the results challenge findings such as those by Whitney et al. [64], who argue that framing effects and preferred formats, especially those aligning with working memory strengths, can shape how effortful a task feels. According to their view, well-matched explanation formats should reduce cognitive friction. However, the absence of such an effect in the current study suggests that in requirements engineering contexts, perceived usefulness alone may not be enough to influence mental effort, particularly when users face unfamiliar or complex prioritization scenarios.

Altogether, the findings indicate that while preference data can offer valuable insights into user experience and perceived ease of use, it should not be solely relied upon when

designing explanation strategies. Tailoring XAI support requires a deeper understanding of how explanation types interact with cognitive demands, task framing, and individual differences in processing. Designing with this nuance in mind may offer a more accurate pathway toward reducing cognitive load in practical software engineering environments.

6.4 Summary of Discussion

The findings across all three research questions offer nuanced insights into how XAI impacts cognitive load in requirements prioritization tasks. While participants often preferred certain explanation formats like bar charts or confidence scores, these subjective choices did not reliably reduce perceived mental effort or task difficulty. This suggests that the perceived usefulness or clarity of a format may not always align with its actual cognitive impact, especially in decision-making settings that involve complex trade-offs.

Throughout the results, XAI formats demonstrated limited and inconsistent effects on cognitive load. Even though tasks supported by XAI showed improvements in correctness, they did not lead to significantly lower effort or difficulty ratings. These findings align with core ideas from Cognitive Load Theory [59], which emphasize that well-structured support materials are only effective when matched with a learner's cognitive needs and task complexity. Similarly, Barredo Arrieta et al. [7] have noted that explainability in AI is context-dependent and does not guarantee reduced mental strain on its own.

The broader patterns in this study reinforce prior work by Paas and Ayres [48], who observed that cognitive load is influenced more by task design and processing demands than by surface-level preferences. Additionally, while previous research such as Chakraborti et al. [13] emphasizes the potential of XAI in improving collaborative decision-making, the current results suggest that these benefits may not extend to reducing individual cognitive effort in software engineering tasks. Teteris et al. [21] similarly caution that effort ratings may remain stable even when support systems are present, particularly in high-pressure or simulation-based settings.

Overall, the study underscores the importance of tailoring XAI not just to user preferences but also to the specific cognitive challenges of the task at hand. Designing explainability features that align more closely with how users process and manage information, rather than solely relying on what they find visually or intuitively appealing, may hold greater promise for easing cognitive load in human–AI collaboration.

7

Conclusion

The research examined how cognitive load affects human-AI collaboration during software requirements prioritization and investigated XAI's role in supporting decision-making processes. The integration of AI tools into requirements engineering workflows requires understanding their effects on engineers' mental effort and task performance and their trust in AI-generated recommendations.

The study used a structured survey to assess how different XAI explanation types affected participants' cognitive load and prioritization accuracy when performing tasks at various complexity levels among software professionals and academic participants. The Weighted Shortest Job First (WSJF) technique was used as a benchmark prioritization framework, allowing for a quantitative comparison of decision-making outcomes between AI-assisted and non-AI-assisted settings.

The study found that XAI explanations had task-dependent effects. While some tasks with XAI support showed significant improvements in correctness, others did not. Measures of mental effort and difficulty remained largely unchanged, indicating that explanations did not consistently reduce perceived cognitive load. Confidence ratings were stable across most conditions, though trust in XAI sometimes aligned with higher confidence. Participants expressed preferences for visual and textual explanations, but these preferences did not reliably correspond to lower effort or improved outcomes.

This research combined performance metrics with Likert-scale cognitive load assessments and qualitative feedback to provide a nuanced view of XAI in requirements engineering. The findings suggest that while XAI can support decision-making, its effectiveness depends on explanation clarity, contextual fit, and task demands. Participants' varied preferences indicate that customizable or multi-modal XAI may be more suitable than one-size-fits-all solutions.

The successful adoption of AI in modern software engineering depends on research like this, which addresses both technical and human-centered aspects of AI adoption. The research should continue by measuring cognitive load through real-time physiological data and by studying group decision-making in RE collaborative settings.

7.1 Limitations

While this study is designed to investigate the impact of AI-generated explanations on cognitive load during requirements prioritization, it has some limitations that should be acknowledged.

Firstly, the study relies on self-reported measures of cognitive load through Likert scale questions. Self-assessment can introduce bias, as participants may interpret the questions differently or may not accurately recall their mental effort. Although we used well-established measurement methods to enhance consistency, the inherent subjectivity of self-reported data remains a limitation.

While the study primarily targets university students and working professionals from our network, this may limit the generalizability of the findings. Participants may have varied levels of experience in requirements engineering, which could affect how they perceive cognitive load and interact with AI-generated explanations.

Additionally, the use of hypothetical scenarios (online banking and doctor appointment systems) may not fully capture real-world complexity. While these scenarios were carefully chosen to be realistic and relevant, they might not reflect the specific challenges that professionals encounter in their actual work environments.

While this study primarily relied on paired t-tests to assess statistical significance across task conditions, a Wilcoxon signed-rank test was also conducted as a non-parametric alternative. This was done to account for the possibility that the data may not fully meet the assumptions of normality required by t-tests. The Wilcoxon test produced results that were largely consistent with the original t-test findings, reinforcing the validity of the reported trends. However, there were a few task comparisons, where the Wilcoxon test detected significant differences not identified in the t-tests. These subtle discrepancies highlight the importance of selecting appropriate statistical tests based on data characteristics. While the main results remain valid, future work may benefit from incorporating non-parametric methods like the Wilcoxon test from the outset, especially when the distribution of participant data is uncertain or potentially skewed.

Another limitation concerns the analysis of users' preferences for different XAI formats and how these relate to their task performance, perceived effort, and trust. Although some moderate patterns were observed, such as participants who preferred bar charts reporting slightly lower mental effort, these associations were not consistent across formats. Preferences are highly subjective and may reflect comfort or familiarity rather than functional effectiveness, meaning that the conclusions drawn from this analysis should be interpreted with caution.

Lastly, while this study focused on three commonly used XAI types, it is possible that alternative explanation formats might have better supported participants during prioritization tasks. For example, interactive explanations or counterfactual examples, which allow users to explore why a different recommendation was not made, may have reduced cognitive effort even further. Similarly, using a combination of modalities, such as layered explanations that adapt based on user feedback or task complexity, could have offered more tailored support. Exploring these richer or more adaptive explanation strategies

could be a fruitful direction for future studies in the context of requirements prioritization.

7.2 Future Work

While this study focuses on understanding how task complexity influences cognitive load during AI-assisted requirements prioritization, there are several directions for future research that could build upon our findings.

First, this study mainly considered task complexity as a cognitive load driver. Future research could explore other factors that influence cognitive load, such as emotional stress, time pressure, or task switching in human-AI collaboration. Examining how these additional factors interact with AI assistance could provide a broader understanding of the challenges engineers face.

Second, although this study used a survey-based approach to gather self-reported measures of cognitive load, future work could employ experimental or longitudinal studies. Observing participants over longer periods or in real-time project environments could offer deeper insights into how cognitive load evolves during actual requirements engineering workflows.

Third, the participant group in this study was relatively diverse but still limited in size. Larger and more industry-focused studies could be conducted to confirm and extend these findings across different organizational settings and engineering cultures.

Fourth, this study measured perceived cognitive load, future research could combine this with physiological measures such as eye tracking, pupil dilation, or EEG to obtain a more objective and comprehensive view of mental effort during AI-supported tasks.

Finally, although this study concentrated on the task of requirements prioritization, cognitive load in other phases of requirements engineering, such as elicitation, validation, and conflict resolution, remains an open area for further exploration. Understanding how cognitive load varies across different types of tasks could help design even more targeted AI tools to support engineers effectively.

By exploring these areas, future work can contribute to developing a more complete understanding of cognitive load in AI-supported requirements engineering and help create better, more human-centered AI tools for the field.

7.3 Use of generative AI in this thesis

The authors employed Generative AI tools ChatGPT by OpenAI for limited purposes during thesis preparation and writing, while maintaining ethical responsibility. The tools helped create draft explanatory content while improving English text clarity and grammar, and assisting with LaTeX formatting for tables and figures, and references.

During the survey design phase ChatGPT assisted in developing question structures and phrasing while generating XAI-based prioritization suggestions, which included text

explanations and bar chart summaries, and confidence scores for survey tasks. The authors thoroughly examined the generated outputs before validating them and making necessary adjustments to maintain research objectives and academic standards.

The research process did not use generative AI tools to create academic content from scratch or interpret data, or replace any evaluative or creative aspects. The authors performed all essential aspects of thesis work, including research question development and survey instrument completion and result interpretation, and theoretical framework construction.

The AI integration received strict supervision to boost efficiency and clarity without compromising scholarly standards. The responsible AI tool implementation follows current academic standards, which use AI instruments to assist human work instead of replacing it.

Bibliography

[1] Philip Achimugu, Ali Selamat, R. Ibrahim, and M. N. Mahrin. A systematic literature review of software requirements prioritization research. *Information and Software Technology*, 56(6):568–585, 2024.

[2] Muneeb Imtiaz Ahmad, Ingo Keller, David A. Robb, and Katrin S. Lohan. A framework to estimate cognitive load using physiological data. *Journal of Biomedical Informatics*, 125:103969, 2022.

[3] Pamela M. Allen, John A. Edwards, Frank J. Snyder, Kevin A. Makinson, and David M. Hamby. The effect of cognitive load on decision making with graphically displayed uncertainty information. *Risk Analysis*, 34(8):1495–1505, 2014.

[4] Timo Appel, Philipp Gerjets, Stefan Hoffmann, Klaus Moeller, Martin Ninaus, Christoph Scharinger, Nikita Sevcenko, Florian Wortha, and Ehsan Kasneci. Cross-task and cross-participant classification of cognitive load in an emergency simulation game. 2025.

[5] R. Arora, J. Smith, and M. Patel. Optimizing cognitive load in software requirements engineering. *Journal of Software Engineering*, 45(2):112–130, 2023.

[6] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Andrés Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020.

[7] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Andrés Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020.

[8] Yasaman Ataei, Wenzhuo Xu, and Frank Maurer. Elicitron: An llm agent-based simulation framework for design requirements elicitation. *arXiv preprint arXiv:2404.16045*, 2024.

[9] Patrik Berander and Anneliese Andrews. Requirements prioritization. In *Engineering and Managing Software Requirements*, pages 69–94. Springer, Berlin, Heidelberg,

2005.

[10] Barry W. Boehm. *Software Engineering Economics*. Prentice Hall, Englewood Cliffs, NJ, 1981.

[11] Barry W. Boehm and Li G. Huang. Value-based software engineering: A case study. *IEEE Computer*, 36(3):33–41, 2003.

[12] Noor Hazlini Borhan, Hazura Zulzalil, Sa'adah Hassan, and Norhayati Mohd Ali. Requirements prioritization techniques focusing on agile software development: A systematic literature review. *International Journal of Scientific & Technology Research*, 8(11):2118–2125, 2019.

[13] Tathagata Chakraborti, Sarath Sreedharan, Yu Zhang, and Subbarao Kambhampati. Ai challenges in human-robot cognitive teaming. 2017.

[14] Robert N. Charette. *Software Engineering Risk Analysis and Management*. McGraw-Hill, New York, NY, 1989.

[15] Yin Cheng, Bernard J. Jansen, Fan Zeng, Wei Gao, Xingyu Ma, and Tat-Seng Chua. Measuring explanation satisfaction of explainable ai: Development and validation of the explanation satisfaction scale. *Electronics*, 12(12):2594, 2023.

[16] R. D. Dias, M. A. Zenati, R. Stevens, J. M. Gabany, and S. J. Yule. Physiological synchronization and entropy as measures of team cognitive load.

[17] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[18] Ilker Etikan, Sulaiman Abubakar Musa, and Rukayya Sunusi Alkassim. Comparison of convenience sampling and purposive sampling. *American Journal of Theoretical and Applied Statistics*, 5(1):1–4, 2016.

[19] Xiaocong Fan and John Yen. Realistic cognitive load modeling for enhancing shared mental models in human-agent collaboration. 2007.

[20] Xavier Franch et al. Leveraging requirements elicitation through software requirement patterns and llms. In *Proceedings of the 31st International Working Conference on Requirements Engineering: Foundation for Software Quality (REFSQ 2025)*. Springer, 2025.

[21] Ma I. Teteris E. Baxter H. Wright B. McLaughlin K. Fraser, K. Emotion, cognitive load, and learning outcomes during simulation training. *Medical Education, 46(8), 785-792*, 2012.

[22] G. J. Funke and S. M. Galster. The effects of cognitive processing load and collaboration technology on team performance in a simulated command and control environment. *International Journal of Industrial Ergonomics*, 39(3):541–547, 2008.

[23] J. Fuster, T. Caparrós, and L. Capdevila. Evaluation of cognitive load in team sports: Literature review.

[24] Ana Fuster-Guilló, María Luisa Pertegal-Felices, Jorge Azorín-López, Ramón Rico, and Antonio Jimeno-Morenilla. A study on cognitive load in software requirements engineering supported by artificial intelligence tools. *IEEE Access*, 9:31286–31296, 2021.

[25] Vivek Ghulaxe. Neuro-adaptive ai for dynamic distraction mitigation in autonomous vehicle environments. *Journal Name*, 2024.

[26] Lucian José Gonçales, Kleinner Farias, and Bruno C. da Silva. Measuring the cognitive load of software developers: An extended systematic mapping study. In *Proceedings of the 35th Brazilian Symposium on Software Engineering (SBES)*, 2021.

[27] Kunal Gupta, Ryo Hajika, Yun Suen Pai, Andreas Duenser, Martin Lochner, and Mark Billinghurst. In ai we trust: Investigating the relationship between biosignals, trust, and cognitive load in vr. 2019.

[28] Verena Hagemann, Matthias Rieth, Anjana Suresh, and Frank Kirchner. Human-ai teams—challenges for a team-centered ai at work. *Frontiers in Artificial Intelligence*, 6, 2023.

[29] Lukas-Valentin Herm. Impact of explainable ai on cognitive load: Insights from an empirical study.

[30] P. Herm. Cognitive load implications in ai-based software engineering. *AI & Engineering Review*, 12(3):78–95, 2023.

[31] Heather Hickam et al. Validity evidence for an instrument for cognitive load in virtual health care education. *Advances in Health Sciences Education*, 27:1037–1056, 2022.

[32] Emma E. Howie et al. Cognitive load management: An invaluable tool for safe and effective surgical training.

[33] Audrey Hudon, Thomas Demazure, Andrew Karran, Pierre-Majorique Léger, and Stéphane Sénécal. Explainable artificial intelligence (xai): How the visualization of ai predictions affects user cognitive load and confidence. *Information Systems Frontiers*, 25(2):395–412, 2021.

[34] J. H. Johnston, S. M. Fiore, C. Paris, and C. A. P. Smith. Application of cognitive load theory to develop a measure of team cognitive efficiency. *Military Psychology*, 25(3):252–265, 2013.

[35] Saurabh Kakaria, Fatemeh Saffari, Thomas Zoëga Ramsøy, and Enrique Bigné. Cognitive load during planned and unplanned virtual shopping: Evidence from a neurophysiological perspective. *[Journal Name]*, 2025. To be updated with journal name, volume, pages, and DOI if available.

[36] Joakim Karlsson and Kevin Ryan. A cost–value approach for prioritizing requirements. *IEEE Software*, 14(5):67–74, 1997.

[37] Paul A. Kirschner, John Sweller, Fred Kirschner, and Fred Paas. From cognitive

load theory to collaborative cognitive load theory. *Educational Psychology Review*, 30.

[38] Andreas Koenig, Dario Novak, Xavier Omlin, Markus Pulfer, Eric Perreault, Lukas Zimmerli, Matjaž Mihelj, and Robert Riener. Real-time closed-loop control of cognitive load in neurological patients during robot-assisted gait training. *IEEE Transactions on Robotics*, 41(1):1–12, 2025.

[39] Grace Lee, Charalampos Mavrogiannis, and Siddhartha S. Srinivasa. Towards effective human-ai teams: The case of collaborative packing. 2019.

[40] Laura Lehtola, Marjo Kauppinen, and Sari Kujala. Requirements prioritization challenges in practice. In *Product-Focused Software Process Improvement (PROFES 2004)*, volume 3009 of *Lecture Notes in Computer Science*, pages 497–508, Berlin, Heidelberg, 2004. Springer.

[41] Rensis Likert. Likert scale. `https://en.wikipedia.org/wiki/Likert_scale`, 1932.

[42] Nelson Lojo, Rafael González, Rohan Philip, José Antonio Parejo, Amador Durán Toro, Armando Fox, and Pablo Fernández. Using large language models to develop requirements elicitation skills. *arXiv preprint arXiv:2503.07800*, 2025. CS.SE.

[43] Ivy Munyaka, Zahra Ashktorab, Casey Dugan, Jeff Johnson, and Qinlan Pan. Decision making strategies and team efficacy in human-ai teams. *Proceedings of the ACM on Human-Computer Interaction*.

[44] K. Neyigapula. Human cognition and ai-driven decision making in software engineering. *Cognitive Computing Journal*, 18(4):55–72, 2023.

[45] Mahsan Nourani, Sina Mohseni, and Eric D. Ragan. Assessing trust and explanation satisfaction in xai user studies. *Frontiers in Computer Science*, 5:1096257, 2023.

[46] Damilola Olugbade, Benjamin I. Edwards, and Oluwafunmilayo A. Ojo. Facilitating cognitive load management and improved learning outcomes and attitudes in middle school technology and vocational education through ai chatbot. *Journal Name*, 2024.

[47] Fred Paas and Paul Ayres. Cognitive load theory: A broader view on the role of memory in learning and education. *Educational Psychology Review*, 26(2):191–195, 2014.

[48] Fred G. W. C. Paas. Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology*, 84(4):429–434, 1992.

[49] Klaus Pohl. *Requirements Engineering: Fundamentals, Principles, and Techniques*. Springer, 2010.

[50] Giorgio Quattrocchi et al. Can llms generate user stories and assess their quality?

arXiv preprint arXiv:2507.15157, 2025.

[51] Paul Ralph. Toward a theory of behavioural software engineering. *ACM Transactions on Software Engineering and Methodology*, 29(2):1–47, 2020.

[52] Donald G. Reinertsen. *The Principles of Product Development Flow: Second Generation Lean Product Development*. Celeritas Publishing, Redondo Beach, CA, 2009.

[53] Lars Repke, Felix Birkenmaier, and Clemens Lechner. Validity in survey research: From research design to measurement instruments. Technical report, GESIS – Leibniz Institute for the Social Sciences, 2024.

[54] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, 2016.

[55] Malik Abdul Sami, Muhammad Waseem, Zheying Zhang, Zeeshan Rasheed, Kari Systä, and Pekka Abrahamsson. Ai based multiagent approach for requirements elicitation and analysis. *arXiv preprint arXiv:2409.00038*, 2024.

[56] Muhammad A. Sami et al. Prioritizing software requirements using large language models. *arXiv preprint arXiv:2405.01564*, 2024.

[57] S. J. Shaikh and I. F. Cruz. Ai in human teams: Effects on technology use, members' interactions, and creative performance under time scarcity. *AI & SOCIETY*.

[58] Alexander Skulmowski and Kai M. Xu. Understanding cognitive load in digital and online learning: A new perspective on extraneous cognitive load. *Educational Psychology Review*, 33(4):907–930, 2021.

[59] John Sweller. Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2):257–285, 1988.

[60] John Sweller. *Cognitive Load Theory*. Springer, New York, 2011.

[61] John Sweller, Jeroen J. G. van Merriënboer, and Fred Paas. Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review*, 31(2):261–292, 2019.

[62] Axel van Lamsweerde. Requirements engineering in the year 00: A research perspective. *Proceedings of the 22nd International Conference on Software Engineering (ICSE)*, pages 5–19, 2000.

[63] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. Designing theory-driven user-centric explainable ai. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–15. ACM, 2019.

[64] Paul Whitney, Christina A. Rinehart, and John M. Hinson. Framing effects under cognitive load: The role of working memory in risky decisions. *Psychonomic Bulletin & Review*, 15(6):1179–1184, 2008.

[65] Karl E. Wiegers. First things first: Prioritizing requirements. `https://www.processimpact.com/articles/prioritizing.html`, 1999.

[66] World Medical Association. World medical association declaration of helsinki: Ethical principles for medical research involving human subjects. *JAMA*, 310(20):2191–2194, 2013.

[67] Robert K. Yin. *Case Study Research and Applications: Design and Methods*. Sage Publications, Thousand Oaks, CA, 6 edition, 2018.

[68] Dušan Žagar, Miha Svetina, Aljaž Košir, and Franc Dimc. Human factor in navigation: Overview of cognitive load measurement during simulated navigational tasks. 2020.

A

Appendix

A.1 Supplementary Tables

Comparison	**Spearman Correlation**	**Spearman ρ**
Correct_1_1 vs Correct_1_2	-0.072	0.578
Correct_1_1 vs Correct_2_1	0.234	0.067
Correct_1_1 vs Correct_2_2	0.083	0.520
Correct_1_2 vs Correct_2_1	0.263	0.039
Correct_1_2 vs Correct_2_2	-0.347	0.006
Correct_2_1 vs Correct_2_2	-0.025	0.850
Effort_1_1 vs Effort_1_2	-0.045	0.727
Effort_1_1 vs Effort_2_1	-0.136	0.292
Effort_1_1 vs Effort_2_2	-0.213	0.097
Effort_1_2 vs Effort_2_1	0.495	0.000
Effort_1_2 vs Effort_2_2	0.582	0.000
Effort_2_1 vs Effort_2_2	0.492	0.000
Difficulty_1_1 vs Difficulty_1_2	-0.143	0.266
Difficulty_1_1 vs Difficulty_2_1	-0.356	0.005
Difficulty_1_1 vs Difficulty_2_2	-0.294	0.020
Difficulty_1_2 vs Difficulty_2_1	0.473	0.000
Difficulty_1_2 vs Difficulty_2_2	0.571	0.000
Difficulty_2_1 vs Difficulty_2_2	0.475	0.000
Confidence_1_1 vs Confidence_1_2	-0.150	0.245
Confidence_1_1 vs Confidence_2_1	0.027	0.838
Confidence_1_1 vs Confidence_2_2	-0.064	0.623
Confidence_1_2 vs Confidence_2_1	0.639	0.000
Confidence_1_2 vs Confidence_2_2	0.718	0.000
Confidence_2_1 vs Confidence_2_2	0.653	0.000

Table A.1: Spearman correlations between task pairs for correctness, effort, difficulty, and confidence.

Preference-Easiest to understand	Performance Metric	Spearman correlation	Interpretation
Bar Chart	Correct_1.2	0.182	Positive correlation
Bar Chart	Effort_1.2	-0.117	Negative correlation
Bar Chart	Difficulty_1.2	-0.189	Negative correlation
Bar Chart	Confidence_1.2	-0.062	Negative correlation
Bar Chart	Correct_2.2	-0.210	Negative correlation
Confidence Scores	Correct_1.2	-0.056	Negative correlation
Confidence Scores	Effort_1.2	0.124	Positive correlation
Confidence Scores	Difficulty_1.2	0.039	Negative correlation
Confidence Scores	Confidence_1.2	0.096	Negative correlation
Confidence Scores	Correct_2.2	0.331	Positive correlation
Text Explanations	Correct_1.2	-0.109	Negative correlation
Text Explanations	Effort_1.2	0.009	Negative correlation
Text Explanations	Difficulty_1.2	0.159	Positive correlation
Text Explanations	Confidence_1.2	-0.014	Negative correlation
Text Explanations	Correct_2.2	-0.159	Negative correlation

Table A.2: Spearman correlation between perceived understandability of XAI types and performance metrics.

B

Survey Instrument

This appendix includes the full survey instrument used in the study, including task descriptions, prioritization scenarios, and XAI explanation prompts. The survey used in this study was implemented in Microsoft Forms. All surveys were implemented using Microsoft Forms and can be accessed via the following links:

- : https://forms.office.com/Pages/DesignPageV2.aspx?subpage=design&FormId=ZXoUKW1T-UO4AuChtc-dv0qA6yxLP5xFkznU3lO-k7ZUMlhDUENQU1lVUlFLMlZIM1VMWEY0R1k2SS4u
- :https://forms.office.com/Pages/DesignPageV2.aspx?subpage=design&FormId=ZXoUKW1T-UO4AuChtc-dv0qA6yxLP5xFkznU3lO-k7ZUMTVKWTRMOEZHM0MyNlhLMEowMzRXRU01Vy4u
- :https://https://forms.office.com/Pages/DesignPageV2.aspx?subpage=design&FormId=ZXoUKW1T-UO4AuChtc-dv0qA6yxLP5xFkznU3lO-k7ZURUU0WDFQOUdKS1YxMElJSEEwQ1A1R