

MASTERARBEIT

# **Immersive Data Visualization and LLM-Supported Decision-Making in Digital Twins for Predictive Maintenance of Industrial Robots**

---

vorgelegt am 30. Juni 2026  
Manfred Moritz Treu

Erstprüferin: Prof. Dr. Eike Langbehn  
Zweitprüfer: Prof. Dr. Gerd Bruder

**HOCHSCHULE FÜR ANGEWANDTE  
WISSENSCHAFTEN HAMBURG**

Fakultät Informatik und Digitale Gesellschaft  
Finkenau 35  
22081 Hamburg

# Zusammenfassung

Diese Arbeit untersucht, inwieweit immersive Datenvisualisierung in Virtual Reality (VR) und Entscheidungsunterstützung durch Large Language Models (LLMs) die Fehlerdiagnose in digitalen Zwillingen von Industrierobotern verbessern können. Hierfür wurde ein Unity-basierter VR-Prototyp für die Meta Quest VR-Headsets entwickelt, der einen fünfachsigigen Industrieroboter als digitalen Zwilling simuliert und neun Fehlerszenarien abdeckt. Der Prototyp implementiert zwei Visualisierungsmodi (die „Labels“-Darstellung mit numerischen Feldern und eine multimodale Schicht mit Temperatur-Heatmap, Ghost-Roboter-Vibrationsoverlay und haptischem Feedback) sowie einen sprachbasierten LLM-Assistenten auf Basis einer RAG-Pipeline mit einem verkörperten Avatar.

In einer kontrollierten 2×2-within-subject-Studie mit  $N = 22$  Teilnehmenden wurden Diagnosegenauigkeit, Aufgabenabschlusszeit, kognitive Belastung (NASA-TLX), Vertrauen (TAS), Nutzererleben (UEQ-S) und Entscheidungssicherheit gemessen. Die LLM-Unterstützung verbesserte die Diagnosegenauigkeit signifikant um ca. 25 Prozentpunkte ( $F(1, 21) = 11.08$ ,  $p = 0.003$ ,  $\eta_p^2 = 0.345$ ), reduzierte die mentale Beanspruchung ( $p = 0.010$ ) und steigerte das Nutzererleben deutlich ( $F(1, 21) = 29.85$ ,  $p < 0.001$ ,  $\eta_p^2 = 0.59$ ), allerdings auf Kosten einer längeren Bearbeitungszeit. Da der Assistent über eine Ground-Truth-Injektion als Halluzinationsschutz abgesichert war, ist der Genauigkeitsgewinn als Obergrenze gut fundierter LLM-Unterstützung zu verstehen. Die multimodale VR-Visualisierung zeigte hingegen keinen signifikanten Effekt auf die Genauigkeit, senkte aber die Entscheidungssicherheit und erhöhte die Frustration bei Nicht-Experten. Interaktionseffekte beider Faktoren wurden nicht gefunden. Angesichts der moderaten Stichprobengröße ( $N = 22$ ) sind die teils sehr großen Effektstärken als Befunde einer explorativen Studie zu verstehen. Die Ergebnisse zeigen, dass LLM-Integration der wirksamste Hebel zur Verbesserung von Diagnoseleistung und Nutzererfahrung in VR-Digital-Twin-Systemen ist, während multimodale Visualisierung primär für erfahrene Domänenexperten von Nutzen sein könnte. Die Variante mit den klassischen numerischen Feldern und einem LLM-Assistenten erzielte die besten Ergebnisse in der vorliegenden Stichprobe.

## Abstract

This thesis investigates the extent to which immersive multimodal visualization in virtual reality (VR) and large language model (LLM) based decision support improve fault diagnosis in digital twins of industrial robots. A Unity-based VR prototype was developed for the Meta Quest standalone headset, simulating a five-axis industrial robot arm with nine fault types. The system implements two visualization modes (numerical text labels and a multimodal

layer comprising a temperature heatmap shader, a ghost-robot vibration overlay, and haptic controller feedback) alongside a voice-based LLM assistant powered by a retrieval-augmented generation pipeline and delivered through a photorealistic lip-synchronized avatar.

A controlled 2×2 within-subject laboratory study with  $N = 22$  participants measured diagnostic correctness, task completion time, cognitive load (NASA-TLX), system trust (TAS), user experience (UEQ-S), and decision confidence. LLM assistance significantly improved diagnostic correctness by approximately 25 percentage points ( $F(1, 21) = 11.08$ ,  $p = 0.003$ ,  $\eta_p^2 = 0.345$ ), reduced perceived mental demand ( $p = 0.010$ ), and produced a very large improvement in user experience ( $F(1, 21) = 29.85$ ,  $p < 0.001$ ,  $\eta_p^2 = 0.59$ ), at the cost of increased task completion time. Because the assistant was grounded through a ground-truth injection that served as a hallucination guardrail, the correctness gain should be read as an upper bound of well-grounded LLM assistance. Multimodal VR visualization produced no significant effect on diagnostic correctness and was associated with both lower decision confidence and higher frustration in the non-expert sample. No significant interaction between the two factors was found. Given the modest sample size ( $N = 22$ ), the partly very large effect sizes should be interpreted as findings of an exploratory study. These results indicate that LLM voice assistance is the dominant driver of diagnostic performance gains in VR digital twin systems for industrial maintenance, while the benefit of multimodal visualization appears conditional on domain expertise that was absent in the present convenience sample. The condition combining the classic numerical text labels with LLM assistance achieved the best overall results in the present sample.

## Acknowledgements

Part of this work was carried out during a research stay at the Synthetic Reality Lab (SREAL)<sup>1</sup> at the University of Central Florida (UCF) in Orlando, Florida. SREAL is an extended-reality and human–computer interaction research laboratory affiliated with UCF’s Institute for Simulation & Training (IST). I am grateful to the lab and its members for their hospitality, for providing access to their facilities and XR equipment, and for the valuable discussions that helped shape this thesis.

---

<sup>1</sup><https://sreal.ucf.edu>

# Contents

|                                                                    |             |
|--------------------------------------------------------------------|-------------|
| <b>List of Figures</b>                                             | <b>V</b>    |
| <b>List of Tables</b>                                              | <b>VII</b>  |
| <b>List of Abbreviations</b>                                       | <b>VIII</b> |
| <b>1. Introduction</b>                                             | <b>1</b>    |
| 1.1. Motivation & Problem Statement . . . . .                      | 1           |
| 1.2. Research Questions and Hypotheses . . . . .                   | 2           |
| 1.3. Scope and Delimitations . . . . .                             | 2           |
| <b>2. Theoretical Background &amp; Related Work</b>                | <b>4</b>    |
| 2.1. Industry 4.0 and Predictive Maintenance . . . . .             | 4           |
| 2.1.1. Concepts and Definitions . . . . .                          | 4           |
| 2.1.2. Sensor Data in Industrial Robotics . . . . .                | 5           |
| 2.1.3. Current Approaches to Condition Monitoring . . . . .        | 6           |
| 2.2. Digital Twins . . . . .                                       | 6           |
| 2.2.1. Definition and Taxonomy . . . . .                           | 6           |
| 2.2.2. Digital Twins in Industrial Maintenance . . . . .           | 7           |
| 2.2.3. Existing Digital Twin Systems for Robots . . . . .          | 8           |
| 2.3. Virtual Reality in Industrial Applications . . . . .          | 8           |
| 2.3.1. Immersive Visualization Technologies . . . . .              | 8           |
| 2.3.2. Multimodal Representation (Visual, Audio, Haptic) . . . . . | 9           |
| 2.3.3. VR in Maintenance and Training . . . . .                    | 10          |
| 2.4. Human Factors & Cognitive Load . . . . .                      | 10          |
| 2.4.1. Cognitive Load Theory . . . . .                             | 10          |
| 2.4.2. Natural Mapping in HCI . . . . .                            | 11          |
| 2.5. Large Language Models as Decision Support . . . . .           | 12          |
| 2.5.1. Overview of LLMs . . . . .                                  | 12          |
| 2.5.2. LLM-Assisted Decision Making . . . . .                      | 12          |
| 2.5.3. Trust in Automated and AI Systems . . . . .                 | 13          |
| 2.5.4. Embodied Agents and Virtual Assistants in VR . . . . .      | 13          |
| 2.6. Research Gap and Positioning of this Work . . . . .           | 14          |

|                                                                              |           |
|------------------------------------------------------------------------------|-----------|
| <b>3. System Design &amp; Implementation</b>                                 | <b>16</b> |
| 3.1. System Architecture Overview . . . . .                                  | 16        |
| 3.1.1. Component Overview (Unity Frontend, Python Backend) . . . . .         | 16        |
| 3.1.2. Data Flow and Communication . . . . .                                 | 17        |
| 3.1.3. Initial Concept and Iterative Development Process . . . . .           | 17        |
| 3.2. Digital Twin Simulation . . . . .                                       | 18        |
| 3.2.1. Robot Model and Fault Scenarios . . . . .                             | 18        |
| 3.2.2. Robot Kinematics and Inverse-Kinematics Solver . . . . .              | 19        |
| 3.2.3. Sensor Data Simulation (Temperature, Vibration, Joint Angles) . . . . | 21        |
| 3.2.4. Scenario Design (Nine Scenarios across Three Severity Categories) .   | 22        |
| 3.3. VR Application (Unity) . . . . .                                        | 23        |
| 3.3.1. Development Environment and Hardware . . . . .                        | 23        |
| 3.3.2. Multimodal Visualization Components . . . . .                         | 24        |
| 3.3.3. Text Data Panel (Baseline Condition) . . . . .                        | 26        |
| 3.3.4. Condition Switching and Scenario Management . . . . .                 | 26        |
| 3.4. LLM Integration . . . . .                                               | 27        |
| 3.4.1. LLM Selection and API Integration . . . . .                           | 27        |
| 3.4.2. Context and Prompt Engineering . . . . .                              | 27        |
| 3.4.3. Voice Input / Speech-to-Text . . . . .                                | 28        |
| 3.4.4. Embodied Virtual Agent (Avatar) . . . . .                             | 28        |
| 3.5. Study Infrastructure . . . . .                                          | 29        |
| 3.5.1. Logging System (Correctness, Task Completion Time) . . . . .          | 29        |
| 3.5.2. Questionnaire Integration . . . . .                                   | 30        |
| 3.5.3. Counterbalancing Implementation . . . . .                             | 30        |
| <b>4. Study Design &amp; Methodology</b>                                     | <b>32</b> |
| 4.1. Research Design . . . . .                                               | 32        |
| 4.1.1. Experimental Approach (Controlled Lab Study) . . . . .                | 32        |
| 4.1.2. Within-Subject Design and Justification . . . . .                     | 32        |
| 4.1.3. 2×2 Factorial Design . . . . .                                        | 32        |
| 4.2. Participants . . . . .                                                  | 33        |
| 4.2.1. Target Population and Recruitment . . . . .                           | 33        |
| 4.2.2. Inclusion/Exclusion Criteria . . . . .                                | 33        |
| 4.2.3. Sample Size and Power Considerations . . . . .                        | 34        |
| 4.2.4. Realized Sample . . . . .                                             | 34        |
| 4.3. Experimental Conditions . . . . .                                       | 34        |
| 4.3.1. Condition 1 – Baseline (Text Panel, No LLM) . . . . .                 | 34        |
| 4.3.2. Condition 2 – Immersive VR (Multimodal, No LLM) . . . . .             | 35        |
| 4.3.3. Condition 3 – LLM Only (Text Panel + LLM) . . . . .                   | 36        |
| 4.3.4. Condition 4 – Combined (Multimodal VR + LLM) . . . . .                | 36        |

|           |                                                                                            |           |
|-----------|--------------------------------------------------------------------------------------------|-----------|
| 4.4.      | Scenarios and Tasks . . . . .                                                              | 37        |
| 4.4.1.    | Scenario Design and Assignment (12 Trials per Participant) . . . . .                       | 37        |
| 4.4.2.    | Task Description (Maintenance Decision + Confidence Rating) . . . . .                      | 37        |
| 4.4.3.    | Fault Types and Diagnostic Categories . . . . .                                            | 37        |
| 4.5.      | Measures . . . . .                                                                         | 39        |
| 4.5.1.    | Subjective Measures (NASA-TLX, TAS, UEQ-S, Decision Confidence) . . . . .                  | 39        |
| 4.5.2.    | Objective Measures (Correctness, Task Completion Time) . . . . .                           | 39        |
| 4.6.      | Procedure . . . . .                                                                        | 40        |
| 4.6.1.    | Session Flow and Duration . . . . .                                                        | 40        |
| 4.6.2.    | Counterbalancing Strategy . . . . .                                                        | 41        |
| 4.6.3.    | Briefing and Informed Consent . . . . .                                                    | 41        |
| 4.7.      | Data Analysis Plan . . . . .                                                               | 41        |
| 4.7.1.    | Repeated Measures ANOVA (rmANOVA) . . . . .                                                | 41        |
| 4.7.2.    | Sphericity . . . . .                                                                       | 42        |
| 4.7.3.    | Effect Sizes . . . . .                                                                     | 42        |
| <b>5.</b> | <b>Results</b>                                                                             | <b>43</b> |
| 5.1.      | Diagnostic Correctness . . . . .                                                           | 43        |
| 5.2.      | Task Completion Time . . . . .                                                             | 43        |
| 5.3.      | Trust and Decision Confidence . . . . .                                                    | 45        |
| 5.4.      | Cognitive Load (NASA-TLX) . . . . .                                                        | 46        |
| 5.5.      | User Experience: UEQ-S . . . . .                                                           | 47        |
| 5.6.      | Condition Preferences, Perceived Helpfulness, and Qualitative Feedback . . . . .           | 47        |
| 5.6.1.    | Condition Preferences and Helpfulness Ratings . . . . .                                    | 47        |
| 5.6.2.    | Qualitative Participant Feedback . . . . .                                                 | 48        |
| 5.7.      | Summary of Results . . . . .                                                               | 49        |
| <b>6.</b> | <b>Discussion</b>                                                                          | <b>50</b> |
| 6.1.      | Interpretation of Results . . . . .                                                        | 50        |
| 6.1.1.    | H1: LLM Support (Supported for Correctness, Cognitive Load, and User Experience) . . . . . | 50        |
| 6.1.2.    | H2: Multimodal Visualization (Not Supported) . . . . .                                     | 52        |
| 6.1.3.    | Interaction Effects and Additivity . . . . .                                               | 54        |
| 6.2.      | Comparison with Related Work . . . . .                                                     | 54        |
| 6.3.      | Practical Implications . . . . .                                                           | 55        |
| 6.3.1.    | Design Recommendations for VR Maintenance Systems . . . . .                                | 55        |
| 6.3.2.    | Guidelines for LLM Integration in Industrial HCI . . . . .                                 | 56        |
| 6.4.      | Limitations . . . . .                                                                      | 57        |
| 6.4.1.    | Sample Size and Generalizability . . . . .                                                 | 57        |
| 6.4.2.    | No Familiarization Phase . . . . .                                                         | 58        |

|           |                                                     |           |
|-----------|-----------------------------------------------------|-----------|
| 6.4.3.    | Ecological Validity and Scenario Design . . . . .   | 58        |
| 6.4.4.    | LLM Variability and Design Choices . . . . .        | 59        |
| 6.4.5.    | Trust Measurement and a Shifting Referent . . . . . | 60        |
| 6.5.      | Future Work . . . . .                               | 60        |
| <b>7.</b> | <b>Conclusion</b>                                   | <b>62</b> |
| 7.1.      | Summary of Contributions . . . . .                  | 62        |
| 7.2.      | Answers to Research Questions . . . . .             | 63        |
| 7.3.      | Final Remarks . . . . .                             | 64        |
|           | <b>Bibliography</b>                                 | <b>66</b> |
| <b>A.</b> | <b>Questionnaire Materials</b>                      | <b>71</b> |
| <b>B.</b> | <b>Informed Consent Form</b>                        | <b>74</b> |

# List of Figures

|      |                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |    |
|------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 2.1. | Three stages of digital integration after Kritzinger et al. . . . .                                                                                                                                                                                                                                                                                                                                                                                                      | 7  |
| 2.2. | Reality–virtuality continuum after Milgram and Kishino . . . . .                                                                                                                                                                                                                                                                                                                                                                                                         | 9  |
| 3.1. | High-level system architecture. The standalone Unity/Quest application communicates with the Python FastAPI backend over HTTP REST on a local network; the backend brokers the LLM (Groq), the RAG knowledge base (ChromaDB/VoyageAI), and text-to-speech (Cartesia), and stores all data locally. . . . .                                                                                                                                                               | 17 |
| 3.2. | The three-stage iterative development process; dashed arrows indicate the iterative refinement between stages. . . . .                                                                                                                                                                                                                                                                                                                                                   | 18 |
| 3.3. | The KUKA arm under inverse-kinematics control. Green: planned end-effector path; red: actual path from the CCD solver. Their divergence reflects the positional error, which grows under injected joint faults. . . . .                                                                                                                                                                                                                                                  | 20 |
| 3.4. | Multimodal visualization across three fault-severity states. The temperature heatmap shades the robot mesh from turquoise (healthy) through yellow to red (critical); the semi-transparent ghost overlay trails the arm to make vibration spatially perceptible, becoming more pronounced as severity rises; and the green/red end-effector path lines indicate trajectory deviation. Together these encode the perceptual channels of the multimodal condition. . . . . | 25 |
| 3.5. | The embodied LLM assistant, positioned to the left of the robot within the participant’s field of view, providing a spatial anchor for the spoken responses. . . . .                                                                                                                                                                                                                                                                                                     | 29 |
| 4.1. | Participant’s view in Condition 1 (text labels, no LLM): each joint shows a floating label with temperature, vibration, and end-effector deviation. . . . .                                                                                                                                                                                                                                                                                                              | 35 |
| 4.2. | Participant’s view in Condition 3 (text labels + LLM): the Condition 1 labels with the embodied LLM assistant added to the left. . . . .                                                                                                                                                                                                                                                                                                                                 | 36 |
| 4.3. | The in-VR maintenance sheet for submitting a diagnosis: participants selected one of nine categories (each with a short signature description) and rated their confidence before confirming. . . . .                                                                                                                                                                                                                                                                     | 38 |
| 4.4. | Flow of a single experimental session (approximately 60–90 minutes, including short breaks). . . . .                                                                                                                                                                                                                                                                                                                                                                     | 40 |
| 5.1. | Diagnostic correctness by condition (higher is better). The bar marks a significant LLM main effect (** : $p < 0.01$ ). . . . .                                                                                                                                                                                                                                                                                                                                          | 44 |

|      |                                                                                                                                                                                          |    |
|------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 5.2. | Task completion time (seconds) by condition (higher is worse). The bar marks a significant LLM main effect (** : $p < 0.01$ ). . . . .                                                   | 44 |
| 5.3. | Trust ratings by condition (higher is better). . . . .                                                                                                                                   | 45 |
| 5.4. | Decision confidence by condition (higher is better). The bar marks a significant <i>Labels</i> effect (* : $p < 0.05$ , Labels > No-Labels); the LLM effect was not significant. . . . . | 45 |
| 5.5. | NASA-TLX subscales by condition (higher is worse). Bars mark significant main effects (* : $p < 0.05$ ): LLM on <i>Mental Demand</i> , Labels on <i>Frustration</i> . . . .              | 46 |
| 5.6. | UEQ-S scales by condition (higher is better). The bars mark a significant LLM main effect (***) : $p < 0.001$ . . . . .                                                                  | 47 |

# List of Tables

|                                                                                                                                                                                      |    |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 3.1. Fault signature patterns for the nine scenarios. Deviations are relative to the normal baseline; severity is the urgency category used for balanced scenario selection. . . . . | 23 |
| 4.1. The four experimental conditions resulting from the $2 \times 2$ factorial design. Each participant completed all four conditions. . . . .                                      | 33 |
| 5.1. Summary of two-way RM-ANOVA results. Significant effects are highlighted in bold. . . . .                                                                                       | 49 |

# List of Abbreviations

|                 |                                        |
|-----------------|----------------------------------------|
| <b>ANOVA</b>    | Analysis of Variance                   |
| <b>API</b>      | Application Programming Interface      |
| <b>AR</b>       | Augmented Reality                      |
| <b>CCD</b>      | Cyclic Coordinate Descent              |
| <b>CSV</b>      | Comma-Separated Values                 |
| <b>DT</b>       | Digital Twin                           |
| <b>HAW</b>      | Hamburg University of Applied Sciences |
| <b>HCI</b>      | Human–Computer Interaction             |
| <b>HMD</b>      | Head-Mounted Display                   |
| <b>HTTP</b>     | Hypertext Transfer Protocol            |
| <b>IK</b>       | Inverse Kinematics                     |
| <b>JSON</b>     | JavaScript Object Notation             |
| <b>LLM</b>      | Large Language Model                   |
| <b>ML</b>       | Machine Learning                       |
| <b>NASA-TLX</b> | NASA Task Load Index                   |
| <b>PdM</b>      | Predictive Maintenance                 |
| <b>RAG</b>      | Retrieval-Augmented Generation         |
| <b>REST</b>     | Representational State Transfer        |
| <b>RM-ANOVA</b> | Repeated-Measures Analysis of Variance |
| <b>RQ</b>       | Research Question                      |
| <b>TAS</b>      | Trust in Automated Systems             |
| <b>TCT</b>      | Task Completion Time                   |

|              |                                       |
|--------------|---------------------------------------|
| <b>TTS</b>   | Text-to-Speech                        |
| <b>UEQ-S</b> | User Experience Questionnaire (Short) |
| <b>UX</b>    | User Experience                       |
| <b>VR</b>    | Virtual Reality                       |
| <b>XR</b>    | Extended Reality                      |

# 1. Introduction

## 1.1. Motivation & Problem Statement

Predictive maintenance is a critical aspect of modern industrial operations, particularly in the context of Industry 4.0, where cyber-physical systems and data-driven analytics are fundamentally transforming manufacturing processes. As industrial robots become increasingly central to production workflows, ensuring their reliable operation through timely fault detection is essential to minimizing downtime and reducing operational costs.

Current condition monitoring systems predominantly rely on text-based reports and 2D dashboards, which can overwhelm operators with fragmented, decontextualized information. Under time pressure, this information density leads to increased cognitive load and suboptimal diagnostic decisions, a problem that intensifies as robot systems grow in complexity.

Digital twin technology offers a promising foundation for addressing these limitations by providing a continuously updated virtual representation of physical assets. However, most existing digital twin interfaces remain screen-based and lack the spatial depth and interactivity needed to fully leverage the technology's potential. Simultaneously, large language models (LLMs) have demonstrated strong capabilities in natural language understanding and context-aware reasoning, yet their integration into industrial maintenance workflows remains largely unexplored.

This thesis investigates how immersive visualization through virtual reality (VR) and LLM-supported decision assistance can be combined within a digital twin framework to improve diagnostic correctness and reduce cognitive load for operators managing industrial robots. To this end, a Unity-based VR prototype is developed, simulating sensor data for a variety of common fault types, and evaluated in a controlled lab study measuring diagnostic correctness, decision confidence, and perceived usability. The work aims to provide both a practical implementation and an empirical contribution to the emerging intersection of immersive interfaces and AI-driven decision support in industrial environments.

## 1.2. Research Questions and Hypotheses

The central research question guiding this thesis is:

*To what extent does multimodal data representation in combination with LLM-supported analysis improve efficiency, decision-making quality, and trust in the decision in fault diagnosis for digital twins of industrial robots, compared to isolated solutions?*

This question is addressed through a controlled 2×2 within-subject study, systematically varying the *representation type* (immersive multimodal VR vs. textual data labels) and *LLM assistance* (available vs. not available), leading to four experimental conditions. Dependent variables span both objective performance measures (diagnostic correctness and task completion time) and subjective assessments via NASA-TLX (Hart & Staveland, 1988), the Trust in Automated Systems scale (Jian et al., 2000), the short User Experience Questionnaire (UEQ-S) (Schrepp et al., 2017), and a single-item confidence scale.

The following hypotheses are derived accordingly:

**H1 – Effect of LLM Assistance:** The use of an LLM reduces task completion time, increases correctness, enhances trust and confidence, and reduces subjective cognitive workload compared to no LLM.

**H2 – Effect of Representation Type:** The multimodal representation reduces task completion time, increases correctness, enhances trust and confidence, and reduces subjective cognitive workload compared to the textual label representation.

Each hypothesis bundles several dependent variables; accordingly, H1 and H2 are not evaluated as a single all-or-nothing claim but are tested separately for each dependent variable (correctness, task completion time, trust, decision confidence, and cognitive workload), and judged “supported” only for those variables on which the predicted effect is observed.

## 1.3. Scope and Delimitations

This thesis focuses on the design, implementation, and empirical evaluation of a VR-based digital twin prototype for predictive maintenance of industrial robots. The system is developed in Unity and operates on simulated sensor data, covering nine diagnostic categories: Bearing Damage, Motor Overheating, Loose Mounting, Gearbox Play, Lubrication Needed, Overload, Coupling Defect, Cooling Failure, and No Error. Two visualization conditions are compared, immersive multimodal representation and textual data labels per joint, in combination with

and without LLM-based decision assistance, evaluated in a controlled within-subject lab study.

Several aspects are explicitly outside the scope of this work. The prototype does not interface with real robot hardware or live sensor streams from an operational factory environment. The study is conducted under controlled laboratory conditions, and findings may not directly generalize to real-world deployment scenarios or fault types beyond those simulated. Long-term usability, system robustness under production conditions, and economic viability are likewise not addressed. The LLM component is evaluated solely as a decision-support interface and not as a standalone diagnostic system.

The study draws on a convenience sample of university students and researchers rather than domain experts such as maintenance technicians or robot operators. Participants received a standardized introduction to the fault types and the system prior to the study. Accordingly, findings should be interpreted in the context of a non-expert population, and generalizability to experienced industrial operators remains to be validated in future work.

## 2. Theoretical Background & Related Work

This chapter establishes the conceptual foundations on which the present work builds and situates it within the existing body of research. It proceeds from the broad industrial context, Industry 4.0 and predictive maintenance (Section 2.1), to the enabling technology of digital twins (Section 2.2), the immersive visualization paradigm of virtual reality (Section 2.3), the human-factors principles that govern interface design (Section 2.4), and the use of large language models as conversational decision support (Section 2.5). Section 2.6 synthesizes the identified gaps and positions the contribution of this thesis accordingly.

### 2.1. Industry 4.0 and Predictive Maintenance

#### 2.1.1. Concepts and Definitions

Industry 4.0 describes the ongoing transformation of manufacturing through the integration of cyber-physical systems (CPS), the Industrial Internet of Things (IIoT), and advanced data analytics into production environments (Lasi et al., 2014; Schwab, 2016). At its core, Industry 4.0 enables physical assets to communicate, monitor themselves, and interact with digital systems in real time, creating a continuous feedback loop between the physical and the digital world (Lu, 2017). Within this paradigm, industrial robots have become central production resources, and their reliable operation is a prerequisite for the highly automated, tightly coupled processes that characterize modern factories. As collaboration between human operators and robotic systems intensifies, effective human–robot communication becomes a decisive factor for both safety and productivity (Mavridis, 2015), and the interfaces through which operators perceive and act upon machine state are central to it.

Maintenance strategies in industrial settings are commonly classified along a spectrum of reactivity. *Reactive maintenance* addresses failures only after they occur, resulting in unplanned downtime and potentially high repair costs. *Preventive maintenance* follows fixed schedules irrespective of the actual condition of the asset, which avoids some failures but can lead to unnecessary interventions and premature part replacement. *Predictive maintenance*

(PdM), by contrast, uses continuous condition monitoring and data-driven models to anticipate failures before they manifest, enabling timely and targeted interventions (Ismail et al., 2025). As Ismail et al. (2025) note in their systematic review, the adoption of digital twin technology has significantly accelerated PdM capabilities by providing a continuously updated virtual replica of physical assets, supporting both real-time monitoring and what-if simulation. Concurrently, the maturation of machine learning (ML) and artificial intelligence (AI) has opened new opportunities for predictive analytics (see Section 2.5).

The economic significance of effective maintenance strategies is substantial. Unplanned downtime is widely regarded as a major cost factor in manufacturing, and as robot fleets grow in complexity and deployment density, the demand for scalable, intelligent maintenance solutions intensifies (Ismail et al., 2025). However, the value of a prediction is only realized when it is communicated to a human operator in a form that supports a timely and correct decision, a requirement that motivates the interface-centered perspective of this thesis.

### **2.1.2. Sensor Data in Industrial Robotics**

Industrial robots are equipped with a range of sensors that continuously capture operational state information. The most diagnostically relevant signals include joint temperature, vibration (typically measured via accelerometers), joint torque and motor current, and joint angle encoders (Qin et al., 2026). Each of these data streams encodes distinct fault signatures. Elevated and rapidly rising joint temperature, for instance, may indicate motor overheating or bearing damage, whereas abnormal vibration patterns localized at a single joint often point to loose mounting or coupling defects. Mechanical wear in a joint’s gear transmission, such as backlash from gearbox play, expresses itself primarily as a deviation of the end-effector path and can therefore remain inconspicuous in the joint’s own temperature and vibration channels (Slamani et al., 2012). Because a single physical fault frequently expresses itself across several channels at once, the combined interpretation of multiple signals is often more informative than any individual reading, an observation that directly motivates the multimodal representation explored in this work.

Fault detection from raw sensor streams is non-trivial. Signals are subject to noise, sensor drift, and cross-joint dependencies, so preprocessing steps such as filtering, normalization, and feature extraction are typically required before meaningful fault patterns can be identified (Jardine et al., 2006). In this thesis, sensor data is simulated for nine diagnostic categories derived from fault taxonomies reported in recent reviews of industrial-robot condition monitoring (Lei et al., 2025; Zhang et al., 2025), enabling controlled experimental conditions while preserving the characteristic signal profiles of each fault type (see Section 3.2). To keep the diagnostic task tractable and focused on the representation question, the simulated data

sources are deliberately limited to joint temperature and vibration, which are among the most widely used signals for fault diagnosis in industrial robots (Qin et al., 2026).

### 2.1.3. Current Approaches to Condition Monitoring

State-of-the-art condition monitoring systems employ a variety of methods ranging from classical signal processing to machine learning. Classical signal-processing techniques such as vibration analysis remain common in practice, while data-driven approaches such as autoencoders, support vector machines (SVMs), and convolutional neural networks (CNNs) applied to time-series data have demonstrated strong performance in automated anomaly detection (Lei et al., 2020).

Despite these advances, a critical challenge remains: translating model outputs into actionable operator decisions. In practice, diagnostic information is still commonly surfaced through 2D dashboards or text-based reports, presenting operators with fragmented, spatially decoupled data. Cognitive Load Theory predicts that mentally integrating such disjointed information imposes a high extraneous load on a limited working memory (Paas et al., 2003; Sweller, 1988); a field study of control-room operators links excessive and poorly arranged information presentation to elevated operator cognitive load (Afzal et al., 2022). The decisive bottleneck is therefore frequently not the accuracy of the underlying detection model but the cognitive cost of the human interpretation step. This gap between automated detection and effective human decision-making is precisely what the design approach taken in this thesis seeks to narrow.

## 2.2. Digital Twins

### 2.2.1. Definition and Taxonomy

The concept of the digital twin (DT) was originally introduced by Grieves as a virtual counterpart to a physical product, enabling lifecycle management through continuous data exchange (Grieves, 2014); the term “digital twin” saw early formal use in NASA’s technology roadmap (Shafto et al., 2010) and is commonly attributed to John Vickers of NASA, with the concept’s origins later consolidated by Grieves and Vickers (2017). A widely adopted classification distinguishes three levels of digital representation according to the degree of automation in the data flow between the physical and the virtual entity (Kritzinger et al., 2018): a *Digital Model*, a static virtual representation without automated data flow; a *Digital Shadow*, which receives data unidirectionally from the physical asset; and a full *Digital Twin*, which maintains a bidirectional data connection so that the virtual model both reflects and

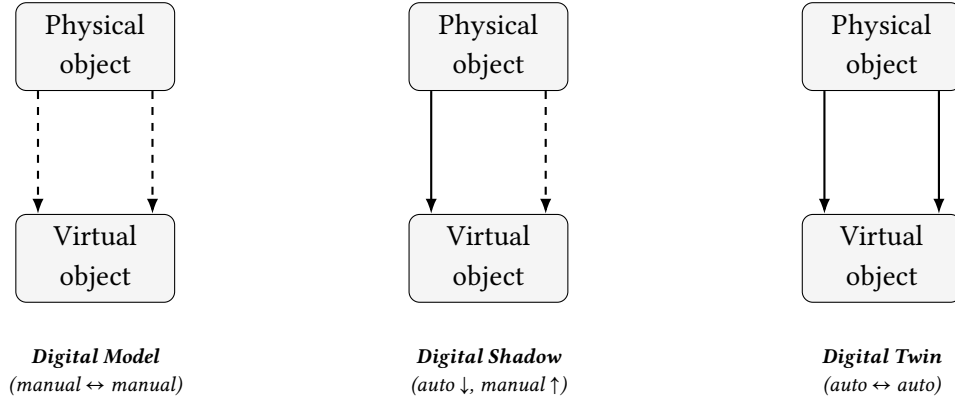


Figure 2.1.: Three levels of digital integration after Kritzing et al. (2018): solid arrows denote automatic data flow, dashed arrows manual transfer. This thesis implements a *Digital Model*.

influences the physical system (Figure 2.1). More recent surveys broaden this view, framing the digital twin as an integration of high-fidelity modeling, real-time data, and analytical services rather than a single artifact, and identifying data connectivity, model fidelity, and service integration as the principal enabling technologies and open challenges (Fuller et al., 2020).

For the purposes of this thesis, the implemented system relies exclusively on simulated sensor data and can therefore be classified as a *Digital Model*, since it does not receive real-time data from a physical asset. The design and evaluation of the immersive visualization and LLM-based decision-support components, however, are independent of the origin of the data and apply equally to more advanced DT implementations; the system architecture can be extended to support real-time data integration with limited effort. In this sense, the overall system and its design principles are applicable to a full digital twin. For readability, the term “digital twin” is used throughout the thesis to refer to the implemented VR-based prototype, with the understanding that it represents one specific instantiation within the broader DT taxonomy.

### 2.2.2. Digital Twins in Industrial Maintenance

Digital twins have emerged as a central enabler of predictive maintenance in industrial environments (Fuller et al., 2020; Ismail et al., 2025). By maintaining a continuously updated model of asset health, a DT supports fault prediction, remaining useful life (RUL) estimation, and scenario simulation without interrupting physical operations. Ismail et al. (2025) describe a four-layer architecture for DT-driven PdM systems: a *physical layer* (the monitored assets), a *data acquisition and transmission layer* (sensors and IIoT connectivity), a *digital-twin layer*

(the virtual model and its analytics), and a *PdM service layer*, which provides the interface through which operators and maintenance teams interact with the twin. The present work concentrates on this last, human-facing layer. All four layers are nonetheless represented in the developed prototype (simulated data standing in for the physical and acquisition layers and simplified analytical models for the digital-twin layer), so that the service and decision-support layer can be studied in isolation.

The transformative potential of digital twins for manufacturing has been highlighted consistently in recent academic reviews of the field, underscoring their growing adoption across sectors (Ismail et al., 2025). Representative case studies illustrate the breadth of the paradigm across the product lifecycle, including a digital-twin-driven prognostics and health management framework for wind turbines (Tao et al., 2018). What unites these cases is a reliance on screen-based dashboards for the final human-facing step, a design choice this thesis revisits.

### **2.2.3. Existing Digital Twin Systems for Robots**

Robot-specific digital twin implementations have been developed across a variety of platforms, including Unity, Unreal Engine, and ROS-based frameworks (Qin et al., 2026). Qin et al. (2026) provide a comprehensive review of robot DT systems in manufacturing and identify key application areas such as assembly, machining, material handling, human–robot interaction, and predictive maintenance. Their analysis further documents a marked recent rise in the integration of extended reality (XR) — both VR and AR — into robot DT systems. This immersive integration, however, has concentrated on assembly, robot programming, and human–robot collaboration rather than on fault diagnosis, leaving the diagnostic use of immersive digital twins comparatively underexplored.

Unity has established itself as a strong option for DT visualization owing to its real-time rendering capabilities, cross-platform support, and native integration with VR/AR hardware (Qin et al., 2026). In this thesis, Unity serves as the foundation for both the robot model and the immersive visualization environment, providing the bridge between the simulated sensor data and the operator’s perceptual experience.

## **2.3. Virtual Reality in Industrial Applications**

### **2.3.1. Immersive Visualization Technologies**

Virtual reality (VR) provides users with a computer-generated environment that replaces the physical surroundings and is experienced through head-mounted displays (HMDs). VR is placed at the far end of the reality-virtuality continuum introduced by Milgram and Kishino

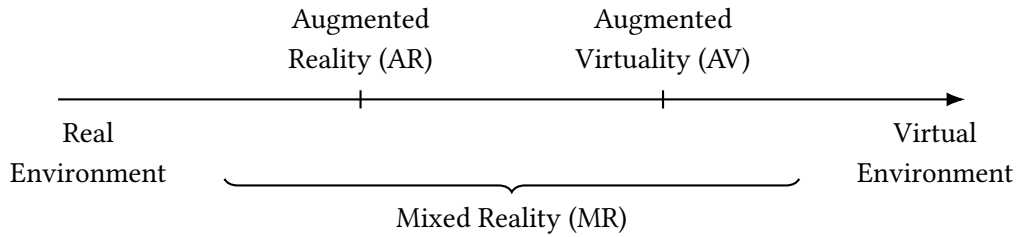


Figure 2.2.: The reality–virtuality continuum after Milgram and Kishino (1994); this thesis sits at the fully virtual end.

(1994), which spans real environments, augmented reality (AR), augmented virtuality, and fully virtual environments (Figure 2.2).

Contemporary HMDs range from standalone devices such as the Meta Quest series to tethered high-fidelity systems such as the HTC Vive Pro. Two constructs are central to the quality of a VR experience and are frequently conflated: *immersion*, the objective technical capability of a system to deliver sensory stimuli, and *presence*, the subjective sense of “being there” that the user experiences (Slater, 2003). Presence is commonly assessed with validated questionnaires such as that of Witmer and Singer (1998), and a higher sense of presence has been linked to stronger engagement and emotional involvement in immersive settings (Makransky & Lilleholt, 2018). Compared to desktop or AR interfaces, fully immersive VR offers superior spatial depth perception and uninterrupted focus, at the cost of physical isolation from the real environment. For diagnostic tasks that require a holistic assessment of a robot’s joint-level condition, the spatial affordances of VR are particularly well suited.

### 2.3.2. Multimodal Representation (Visual, Audio, Haptic)

Multimodal data representation leverages several perceptual channels simultaneously to communicate information more effectively than single-channel approaches (Turk, 2014). In the context of VR-based condition monitoring, the relevant modalities include:

- **Visual encoding:** color-mapped heatmaps for temperature gradients and exaggerated joint animations for vibration intensity provide intuitive, spatially grounded representations of fault states.
- **Spatial audio:** directional audio cues can reinforce visual signals and direct operator attention to critical joints.
- **Haptic feedback:** controller vibration can convey urgency or signal that a threshold has been exceeded.

The theoretical foundation for multimodal encoding lies in the principle of *natural mapping* (Norman, 2013): representations that correspond spatially and perceptually to their referents reduce the cognitive effort required for interpretation. Distributing information across complementary channels can additionally increase the salience of critical events and provide perceptual redundancy, so that a signal missed in one modality may still be caught in another. By mapping temperature directly onto color at the affected joint and vibration onto animated motion at its source, the VR prototype in this thesis aims to minimize extraneous cognitive load (see Section 2.4.1).

### 2.3.3. VR in Maintenance and Training

VR and AR have both been studied extensively as training media for maintenance procedures, particularly for assembly, disassembly, and fault-correction tasks. On the AR side, surveys document a broad range of applications across design, manufacturing, and maintenance (Nee et al., 2012; Palmarini et al., 2018). For immersive VR, a systematic review documents broad adoption across higher-education domains (Radianti et al., 2020), while meta-analytic evidence shows positive but moderate learning gains, with the largest effects in technical domains such as engineering (Coban et al., 2022). Gavish et al. (2015) found measurable error reductions for AR-based but not for VR-based training, underscoring that immersion alone does not guarantee transfer. A consistent finding across this literature is nonetheless that the benefits of immersion are greatest when the task has a strong spatial or procedural component that abstract media represent poorly.

A clear gap nonetheless persists. While VR has been examined intensively for *training*, its application to real-time fault *diagnosis* and decision-making within immersive digital twins remains, to the author's knowledge, largely unexplored. Qin et al. (2026) document growing XR integration in robot DT systems, but predominantly for assembly, programming, and human-robot collaboration rather than for interactive diagnostic decision support. This thesis addresses that gap by treating the immersive interface not as a training aid but as the primary medium through which an open-ended diagnostic decision is made.

## 2.4. Human Factors & Cognitive Load

### 2.4.1. Cognitive Load Theory

Cognitive Load Theory (CLT), developed by Sweller (1988) and subsequently extended to instructional design (Paas et al., 2003), distinguishes three types of cognitive load: *intrinsic load*, determined by the inherent complexity of the task; *extraneous load*, arising from the

way information is presented; and *germane load*, associated with the formation of mental schemas. Because working memory is limited, the three types compete for the same finite resource; minimizing extraneous load is therefore crucial, as it otherwise draws capacity away from task-relevant processing.

In fault diagnosis, complex 2D dashboards can impose high extraneous load by forcing operators to mentally integrate spatially disjointed sensor readings and to map abstract numbers back onto the physical machine. Immersive, spatially consistent visualization as offered by VR-based digital twins has the potential to reduce this extraneous load by providing a direct perceptual correspondence between the representation and the physical system. A well-designed immersive environment may additionally raise engagement and sustained attention, although the same richness carries the risk of introducing distracting detail if it is not carefully controlled.

Subjective cognitive workload in this thesis is measured using the NASA Task Load Index (NASA-TLX), a multidimensional instrument that assesses workload across six subscales: mental demand, physical demand, temporal demand, performance, effort, and frustration (Hart & Staveland, 1988). NASA-TLX has been validated across a wide range of domains and is among the most widely used subjective workload measures in human-factors research.

#### **2.4.2. Natural Mapping in HCI**

The principle of natural mapping, introduced by Norman (2013), holds that interfaces are most intuitive when the spatial and conceptual relationship between controls or representations and their referents mirrors real-world correspondences. Applied to information visualization, this implies that spatial representations of physical systems, where data is anchored to the location it describes, are processed more efficiently than abstract, spatially decoupled displays.

Empirical evidence from information-visualization research supports this claim: color-coded and spatially grounded representations exploit pre-attentive perceptual processing and have been shown to reduce response times and error rates compared to tabular or text-based formats (Ware, 2012). This principle directly motivates the design of the immersive visualization condition in this thesis, in which fault indicators are rendered at their physical location on the robot model rather than in a detached dashboard, so that the operator can read machine state from the machine itself.

## 2.5. Large Language Models as Decision Support

### 2.5.1. Overview of LLMs

Large language models (LLMs) are neural networks trained on large corpora of text to model the probability distribution of natural language. The transformer architecture introduced by Vaswani et al. (2017) established the foundation for modern LLMs by replacing recurrent computation with self-attention, enabling efficient parallel training and strong performance across a wide range of language tasks. Scaling this architecture to billions of parameters, as first demonstrated with GPT-3 (Brown et al., 2020), yielded emergent few-shot reasoning capabilities, and subsequent models such as GPT-4 (OpenAI, 2023) have shown remarkable competence in natural language understanding, contextual reasoning, and domain-specific question answering. Because such models can be adapted to many downstream tasks without task-specific retraining, they have been characterized as *foundation models* (Bommasani et al., 2021). Open-weight alternatives such as LLaMA (Touvron et al., 2023) have further broadened access to these capabilities for research and industry.

The properties most relevant to this thesis are the ability to generate natural-language explanations of sensor states, to respond to voice-based queries, and to synthesize diagnostic reasoning from structured input. These properties position LLMs as a natural complement to immersive visualization: where VR supplies spatial and perceptual context, the LLM assistant supplies verbal, on-demand reasoning support.

### 2.5.2. LLM-Assisted Decision Making

The integration of LLMs as decision-support co-pilots has been explored across domains including clinical reasoning, legal analysis, and education (Bommasani et al., 2021). Controlled studies have begun to quantify the resulting productivity gains; Noy and Zhang (2023), for example, report substantial improvements in both speed and output quality on professional writing tasks when participants are assisted by a generative model. In the context of robot digital twins, Qin et al. (2026) identify LLMs as an emerging enabling technology, pointing specifically to generative, LLM-based modeling as a future research direction. Building on this, the present work investigates a complementary role for LLMs at the operator interface: bridging the human-machine cognitive gap through natural interaction that lowers the expert-knowledge barrier for non-expert operators.

Effective deployment in a constrained domain requires careful prompt engineering to anchor the model's responses within the relevant knowledge domain and to suppress irrelevant output (White et al., 2023). A complementary technique is retrieval-augmented generation (RAG), in which the model is supplied at inference time with passages retrieved from an

external knowledge base, grounding its answers in verifiable source material (Lewis et al., 2020). The principal risks are model *hallucination* (the generation of plausible but factually incorrect information Ji et al., 2023) and *automation bias*, the tendency of users to over-rely on automated recommendations irrespective of their accuracy (Parasuraman & Riley, 1997). Design guidance for human–AI interaction stresses that systems should make their confidence and limitations legible precisely to support appropriately calibrated reliance (Amershi et al., 2019). In this thesis, the LLM component is constrained through system prompting and retrieval to the domain of industrial robot fault diagnosis, and its effect on operator trust and decision quality is evaluated empirically.

### 2.5.3. Trust in Automated and AI Systems

Trust in automation is a critical determinant of whether operators appropriately rely on automated decision support. The seminal framework of Lee and See (2004) conceptualizes trust as *calibrated reliance*: the degree to which an operator’s use of an automated system matches the system’s actual reliability. They ground trust in three bases of information about the automation: its *performance* (what it does), its *process* (how it operates), and its *purpose* (why it was designed), which together shape appropriate reliance. Both *under-trust*, which leads to disuse of beneficial automation, and *over-trust* (automation bias) represent failure modes that degrade decision quality. A meta-analysis of human–robot interaction confirms that system performance characteristics are among the strongest determinants of trust, while operator- and environment-related factors play a more moderate role (Hancock et al., 2011).

In the context of this thesis, trust is of particular relevance because participants interact with both an immersive VR interface and an LLM assistant, each of which may shape reliance independently. Trust is measured using the Trust in Automated Systems (TAS) scale (Jian et al., 2000), a validated 12-item instrument measuring trust and distrust in automated systems.

### 2.5.4. Embodied Agents and Virtual Assistants in VR

Within VR environments, AI-driven assistance can be delivered through *embodied agents*, anthropomorphic avatars that occupy the shared virtual space, or through disembodied interfaces such as floating text panels or voice-only systems. The “computers are social actors” paradigm holds that users apply social rules and expectations to computers largely automatically, even when fully aware of their artificial nature (Nass & Moon, 2000; Nass et al., 1994). Building on this, Bickmore and Picard (2005) show that relational agent behavior fosters trust and engagement over repeated interaction. Recent work on VR-based embodied conversational agents (ECAs) indicates that embodied presentation with synchronized lip movement (Peixoto et al., 2021) and non-verbal cues such as gaze and body sway (Garau

et al., 2003) increases users' sense of presence relative to text- or voice-only agents (Lim et al., 2025). Non-verbal channels of this kind reinforce the perception of the agent as a social actor, paralleling the role that non-verbal communication plays in human–robot collaboration more broadly (Mavridis, 2015).

For the VR prototype in this thesis, the LLM assistant is rendered as an embodied avatar collocated with the robot model. This choice is motivated by the expectation that embodied presentation will strengthen the perceived credibility of the assistant and facilitate more natural interaction. The trade-off between the development overhead of a fully animated avatar and the potential benefits of social presence was resolved in favor of a semi-animated agent with lip-sync and idle animations, following evidence that synchronized lip movement and subtle non-verbal behavior already contribute to perceived presence in VR even without full-body animation (Garau et al., 2003; Peixoto et al., 2021).

## 2.6. Research Gap and Positioning of this Work

The preceding review reveals three interlocking gaps in the existing literature that this thesis aims to address.

First, while both immersive VR and LLM-based decision support have been studied in isolation, no empirical work has investigated their *combined* effect on fault diagnosis in digital twin environments for industrial robots. Recent reviews of robot digital twins document the rising adoption of XR and identify LLM-based techniques as an emerging direction (Qin et al., 2026), while systematic reviews of digital-twin-driven predictive maintenance concentrate on architectures and analytics rather than the operator-facing interface (Ismail et al., 2025); the intersection of immersive visualization and LLM assistance for the diagnostic task itself remains unaddressed.

Second, existing studies on VR in industrial maintenance focus predominantly on training scenarios with predefined correct procedures, rather than on open-ended diagnostic decision-making under uncertainty. Evaluating the effect of visualization modality on diagnostic correctness within a controlled factorial design addresses this gap directly.

Third, trust measurement in VR-based digital twin contexts has received limited attention. Applying the validated TAS instrument in this setting contributes a methodologically grounded assessment of how immersive interfaces and AI assistance jointly shape operator trust in automated systems.

This thesis positions itself at the intersection of immersive visualization, AI-driven decision support, and human factors in industrial maintenance. It contributes a functional VR proto-

type (including the conceptual design and implementation of an LLM pipeline) together with an empirical dataset from a 2×2 within-subject study spanning four experimental conditions.

## 3. System Design & Implementation

### 3.1. System Architecture Overview

#### 3.1.1. Component Overview (Unity Frontend, Python Backend)

The system developed for this thesis consists of two main components: a standalone virtual reality application running on a head-mounted display (HMD) and a Python-based server backend hosted on a laptop computer. Both components communicate over a local wireless network (Figure 3.1).

The VR application, built with Unity 6 (version 6000.0.60f1) and the XR Interaction Toolkit 3.2.1, constitutes the user-facing frontend. It is responsible for rendering the three-dimensional digital twin of the robot, presenting sensor data in the respective visualization format for each experimental condition, managing the study flow, collecting participant decisions, and forwarding voice queries to the backend when the LLM assistant is active. It is thus the primary interface through which participants interact with the system, as well as the central component of the overall architecture.

The backend is implemented as a Python FastAPI application and fulfills three functions: (1) it serves as the interface to the large language model, assembling the context-enriched prompt from the current sensor data and the participant query before calling the Groq API<sup>1</sup>; (2) it maintains a local retrieval-augmented generation (RAG) knowledge base for domain-specific context retrieval, covering all fault types and their characteristic signatures; and (3) it receives and persists study data, writing participant decisions and study events to local CSV files and synchronizing them in real time to a dedicated folder in the Unity project. In summary, the backend coordinates study-data management and LLM integration, while the Unity application handles the simulated digital twin, user interaction, and real-time visualization.

---

<sup>1</sup>Groq, <https://groq.com> (accessed 2026-06-30).

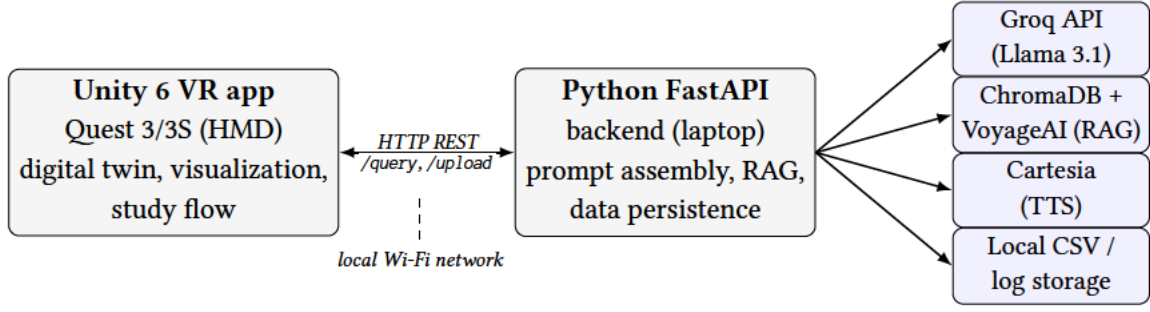


Figure 3.1.: High-level system architecture. The standalone Unity/Quest application communicates with the Python FastAPI backend over HTTP REST on a local network; the backend brokers the LLM (Groq), the RAG knowledge base (ChromaDB/VoyageAI), and text-to-speech (Cartesia), and stores all data locally.

### 3.1.2. Data Flow and Communication

Communication between the Unity application and the Python backend is realized via HTTP REST. The Quest HMD and the backend laptop are connected to the same local Wi-Fi network; the Unity client stores the backend IP address in player preferences and constructs endpoint URLs at runtime, allowing the IP to be configured per session without rebuilding the application.

Two primary communication channels are active during the study. First, when a participant submits a voice query in a condition with LLM support, the Unity application sends a POST /query request containing the transcribed question, the current sensor readings for all five joints (temperature in °C and vibration normalized to the range [0, 1]), and an internally injected field identifying the correct scenario diagnosis. The backend processes this request and returns the LLM-generated response as plain text, which Unity then routes to the text-to-speech pipeline and plays the audio. Second, upon each participant decision, the application sends a POST /upload-study-data request containing the full decision record, including outcome, confidence rating and response time.

The sensor simulation within Unity updates continuously; network requests are event-driven and do not occur on a fixed polling interval.

### 3.1.3. Initial Concept and Iterative Development Process

The system was initially conceived as a VR-based digital twin for the predictive maintenance of industrial robots, augmented with a range of tools intended to support operator decision-making. This concept evolved through an iterative process in which several interface approaches and candidate features were prototyped and tested before the final design was

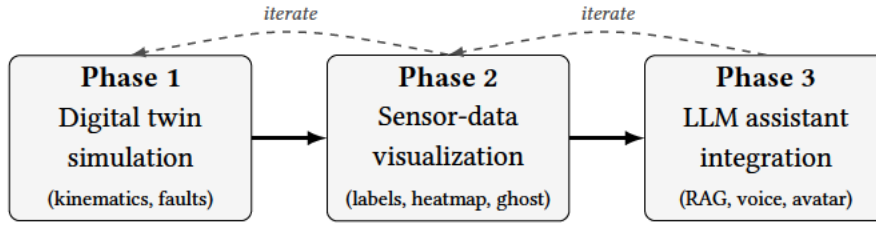


Figure 3.2.: The three-stage iterative development process; dashed arrows indicate the iterative refinement between stages.

fixed. The development proceeded in three main stages, each refined on the basis of testing and informal feedback: first, the digital twin simulation; second, the visualization methods for the sensor data; and third, the integration of the LLM-based assistant (Figure 3.2).

Following these exploratory prototypes, the design was consolidated to align with the study described in Chapter 4. The remainder of this chapter describes the final design and its implementation in detail.

## 3.2. Digital Twin Simulation

### 3.2.1. Robot Model and Fault Scenarios

The digital twin is modeled after a five-axis articulated KUKA robot provided by the robotics laboratory of Hamburg University of Applied Sciences (HAW Hamburg). The three-dimensional model was integrated into the Unity scene with all five rotational axes fully animated, allowing the robot to execute continuous pick-and-place trajectories during each study scenario. This continuous motion serves two purposes: it provides a realistic operational context and enables participants to observe kinematic deviations associated with certain fault types, most notably end-effector drift in the case of Gearbox Play.

The choice of robot platform was preceded by a deliberate evaluation of alternatives. Several commercially available collaborative arms were considered as potential physical study objects, including the UFACTORY Lite 6<sup>2</sup>, the DOBOT MG400<sup>3</sup>, and the Elephant Robotics myCobot 320<sup>4</sup>. These platforms would have offered direct access to real motion and sensor data, but each entailed substantial drawbacks for the scope of this work: procurement costs

<sup>2</sup>UFACTORY Lite 6, <https://www.ufactory.us/product/lite-6> (accessed 2026-06-29).

<sup>3</sup>DOBOT MG400 Robotic Arm, <https://top3dshop.com/product/dobot-mg400-robotic-arm> (accessed 2026-06-29).

<sup>4</sup>Elephant Robotics myCobot 320 M5 (2022), <https://shop.elephantrobotics.com/collections/mycobot-pro-320/products/commercial-and-economic-six-axis-collaborative-robot> (accessed 2026-06-29).

in the range of roughly \$2,400–\$4,000 and the need for additional instrumentation to capture the temperature and vibration signals central to the diagnostic task. Against these options, the existing KUKA model from the HAW Hamburg robotics laboratory was selected because it was available at no additional cost and was well suited as a study object. Its principal limitation, the absence of a live data connection, is precisely what motivated the simulation-based approach adopted here and the resulting classification of the system as a *Digital Model* (see Section 2.2.1).

The simulation covers nine diagnostic categories, each constituting a distinct ground truth for the experimental scenarios: *No Error*, *Bearing Damage*, *Motor Overheating*, *Loose Mounting*, *Gearbox Play*, *Overload*, *Coupling Defect*, *Cooling Failure*, and *Lubrication Needed*. Each scenario is defined as a Unity ScriptableObject (ScenarioConfig) and assigned to one of three severity categories (*Low Urgency*, *Medium Severity*, and *Critical*) to enable balanced scenario selection across experimental conditions (see Section 4.6.2).

The selection of these categories is not arbitrary but follows the fault taxonomies established in recent reviews on condition monitoring and fault diagnosis of industrial robots (Lei et al., 2025; Zhang et al., 2025). Such reviews consistently identify the drivetrain and joint assembly—bearings, gear and harmonic-drive transmissions, servo motors, and their mountings—as the dominant sources of degradation in articulated manipulators, together with the lubrication and thermal conditions that govern their wear. The eight fault categories used here map directly onto these components: *Bearing Damage* and *Coupling Defect* target the rolling and connecting elements of the drivetrain; *Gearbox Play* and *Overload* the transmission and its loading; *Motor Overheating* and *Cooling Failure* the servo and its thermal management; and *Loose Mounting* and *Lubrication Needed* the mechanical fixation and tribological state. *No Error* provides the healthy baseline required to assess false-positive behavior. The set is therefore intended not as an exhaustive failure catalogue but as a representative spread across the principal failure modes of an articulated arm, deliberately chosen so that each category expresses a distinguishable signature across the simulated diagnostic channels. These channels are the two sensor modalities of temperature and vibration, which are among the signals most widely used for robot fault diagnosis (Qin et al., 2026; Zhang et al., 2025), complemented by the end-effector path deviation through which a purely kinematic fault such as *Gearbox Play*—which leaves the thermal and vibration signals unaffected—becomes observable. The concrete signal profiles realized for each category are described in Section 3.2.3 and summarized in Table 3.1.

### 3.2.2. Robot Kinematics and Inverse-Kinematics Solver

Driving the five-axis arm along continuous, collision-free pick-and-place trajectories required solving the inverse-kinematics (IK) problem: given a target position for the gripper, the

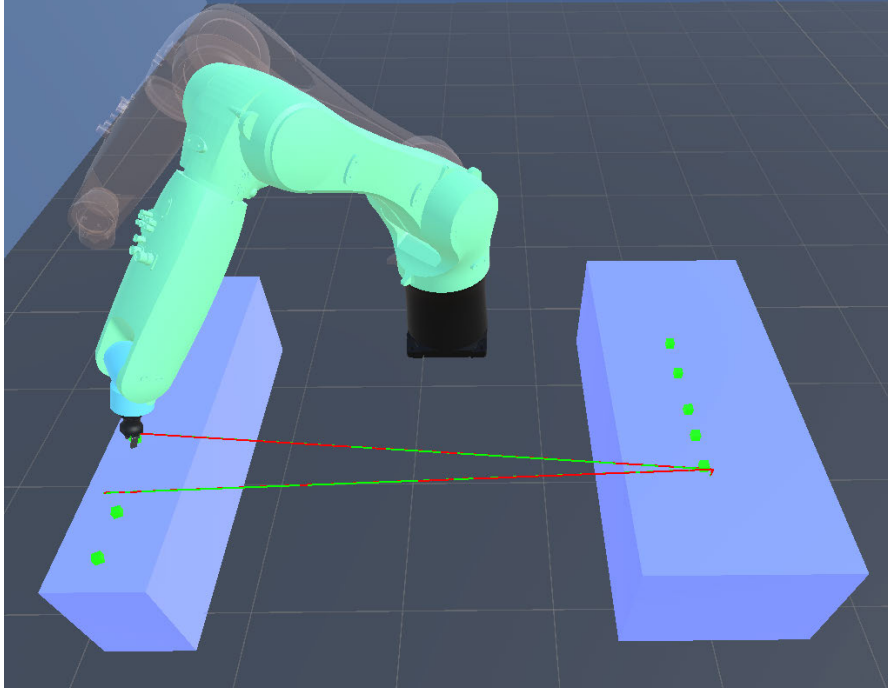


Figure 3.3.: The KUKA arm under inverse-kinematics control. Green: planned end-effector path; red: actual path from the CCD solver. Their divergence reflects the positional error, which grows under injected joint faults.

joint angles that place the end effector at that position must be determined every physics frame. Because no closed-form analytical solution was available for the laboratory robot’s specific link geometry, and because off-the-shelf IK packages did not integrate cleanly with the fault-injection and visualization requirements of the study, a custom solver was developed. This turned out to be one of the most demanding parts of the implementation and was arrived at only after an explicit comparison of alternative approaches.

**Exploration of candidate approaches** Two implementation strategies were prototyped before the final design was fixed. The first was a *physics-based* approach in which each joint is an articulated rigid body (*ArticulationBody*) actuated through Unity’s built-in reduced-coordinate drive system, with the gripper closing on physical colliders. While physically faithful, this approach proved difficult to control precisely: drive stiffness, damping, and solver-iteration settings interacted in ways that produced jitter and overshoot, and the physical grasp was unreliable for the small transported objects. It also made the deterministic, repeatable fault behavior the study required (a precisely controlled end-effector drift, for instance) hard to enforce, since the motion emerged from the physics solver rather than being directly prescribed. The second approach, adopted for the final system, is a *transform-based kinematic* solver that drives the joint rotations directly and attaches transported objects by

reparenting rather than by physical grasping. This yielded smooth, fully reproducible motion and, importantly, a single point at which fault-specific perturbations could be injected into the kinematic chain.

**Cyclic-coordinate-descent solver** The final solver implements a heuristic *cyclic coordinate descent* (CCD) scheme. On each physics step the chain is traversed from the joint closest to the end effector back towards the base. For each joint  $j$  with rotation axis  $\mathbf{a}_j$ , the solver computes the signed angle between the vector from the joint to the current end-effector position and the vector from the joint to the target, and rotates the joint about its axis to reduce that angle. The per-step rotation is clamped to a maximum step angle and scaled by a per-joint rotation weight and a position-error gain, so that large initial errors are corrected quickly while small residual errors are approached smoothly without oscillation. The update is additionally low-pass filtered (`smoothFactor`) to suppress frame-to-frame jitter, and each joint angle is constrained to a configured  $[\theta_{\min}, \theta_{\max}]$  range to keep the arm within mechanically plausible limits. Several CCD iterations are executed per frame (typically on the order of ten to twenty), followed by a short fine-correction loop that tightens the end-effector error below a fixed threshold before the gripper acts. Target poses are interpolated with smooth-step easing between hover, descent, and transport waypoints, producing the characteristic continuous pick-and-place cycle observed in every scenario.

**Coupling faults into the kinematic chain** A key advantage of the transform-based solver is that fault signatures can be expressed directly as disturbances of the IK target and of the joint update. End-effector path deviation for the Gearbox Play scenario, for example, is realized by adding a vibration-scaled random offset to the IK target each step, so that the gripper visibly departs from its planned trajectory; the system records both the planned and the actually realized end-effector path as separate world-space polylines, making the deviation inspectable (Figure 3.3). Vibration additionally modulates the solver itself: as the average vibration level rises, the number of IK iterations per frame is reduced, degrading tracking precision in a manner that mimics a mechanically compromised joint. The same kinematic layer drives a semi-transparent duplicate of the robot used for the ghost-overlay vibration visualization (Section 3.3.2), so that the perceptual cues remain spatially registered with the physical motion at all times.

### 3.2.3. Sensor Data Simulation (Temperature, Vibration, Joint Angles)

Rather than streaming live data from a physical robot, the system generates sensor values procedurally within Unity according to per-scenario parameters defined in the corresponding `ScenarioConfig` asset. Two sensor modalities are simulated for each of the five joints:

**Temperature** Initial temperature values are specified per joint in degrees Celsius (valid range: 25–90 °C) together with per-joint rise rates in °C per second. A healthy baseline distributes joint temperatures within normal operating range with near-zero rise rates. Fault scenarios elevate the initial value and/or the rise rate at the affected joint(s), producing a time-evolving thermal signature that reflects the characteristic pattern of each fault type. Motor Overheating, for example, manifests as a rapidly rising temperature at a single joint, while Cooling Failure produces simultaneous temperature increases across multiple joints.

**Vibration** Vibration levels are represented as normalized values in the range [0, 1], displayed to participants as a percentage (0 % = normal, 100 % = maximum). The same initial-value and rise-rate scheme applies. Bearing Damage presents elevated vibration alongside elevated temperature at the same joint, whereas Loose Mounting manifests as isolated vibration elevation with temperatures remaining within normal range.

Beyond temperature and vibration, the simulation also derives end-effector path deviation from the vibration level and a scenario-specific drift-influence parameter, making trajectory irregularities visible for faults such as Gearbox Play. Joint torque and an aggregate motor-current value are additionally computed internally, but they are a remnant of an earlier design iteration: they are neither part of any fault signature (Table 3.1) nor shown to participants, and were not used as a diagnostic channel in the study.

The complete fault-signature patterns used to parameterize each scenario are summarized in Table 3.1.

### 3.2.4. Scenario Design (Nine Scenarios across Three Severity Categories)

Nine individual scenarios were designed, one per diagnostic category, each assigned to one of three severity categories (Low Urgency, Medium Severity, Critical). Each experimental condition draws exactly three scenarios, one from each severity category, selected deterministically from a composite seed derived from the participant identifier and the condition index, which keeps the assignment balanced and reproducible. Because each participant completes three scenarios in each of the four conditions (twelve trials in total) drawn from the pool of nine distinct scenarios, individual scenarios necessarily recur across conditions for a given participant, although never more than once within the same condition. The order of the three scenarios within each condition is additionally randomized per participant to mitigate within-condition ordering effects.

Table 3.1.: Fault signature patterns for the nine scenarios. Deviations are relative to the normal baseline; severity is the urgency category used for balanced scenario selection.

| Fault Type         | Temperature                 | Vibration                    | End-Effector      | Severity    |
|--------------------|-----------------------------|------------------------------|-------------------|-------------|
| No Error           | Normal, stable              | Normal                       | Normal            | Low urgency |
| Motor Overheating  | High ↑↑, one joint          | Normal                       | Normal            | Critical    |
| Cooling Failure    | High ↑↑, multi-joint        | Normal                       | Normal            | Critical    |
| Bearing Damage     | Elevated ↑, one joint       | Elevated ↑, same joint       | Normal            | Critical    |
| Loose Mounting     | Normal                      | High, one joint              | Normal            | Medium      |
| Coupling Defect    | Normal                      | Elevated, joints 4–5         | Normal            | Medium      |
| Overload           | Elevated, joints 1–3        | Normal                       | Normal            | Medium      |
| Gearbox Play       | Normal                      | Normal                       | Deviation visible | Medium      |
| Lubrication Needed | Slightly elev. ↑, one joint | Slightly elev. ↑, same joint | Normal            | Low urgency |

### 3.3. VR Application (Unity)

#### 3.3.1. Development Environment and Hardware

The VR application was developed in Unity 6 (version 6000.0.60f1) using the XR Interaction Toolkit 3.2.1 and the XR Hands package 1.6.1. The application targets the Meta Quest platform and runs as a standalone Android build, requiring no tethered PC connection during operation. Development was conducted using a Meta Quest 3; the user study was subsequently conducted with a Meta Quest 3S, which shares the same software platform and interaction model while offering a reduced form factor. The OpenXR backend with the Meta XR feature set was used, providing access to controller input and haptic output via the HapticImpulsePlayer component from the XR Interaction Toolkit. The backend laptop and the HMD were connected to the same local Wi-Fi network throughout all sessions. Target rendering performance was 90 Hz to ensure comfortable immersion and minimize the risk of cybersickness.

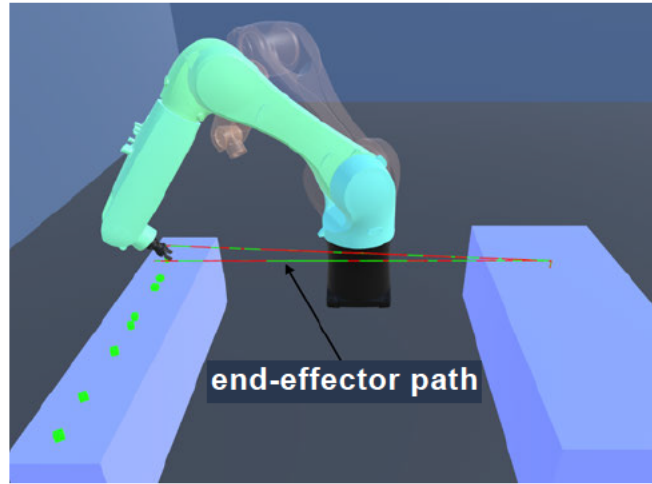
### 3.3.2. Multimodal Visualization Components

The multimodal visualization condition renders fault-relevant sensor states through three simultaneously active perceptual channels: a temperature heatmap applied to the robot mesh, a ghost-robot vibration overlay, and haptic controller feedback. All three channels are driven in real time by the sensor values produced by the scenario simulation.

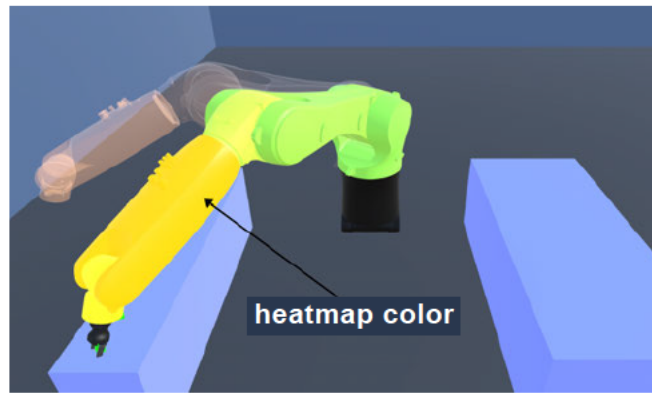
**Temperature Heatmap** Joint temperature is encoded via a shader applied directly to the robot mesh. The color scale transitions from a cool turquoise-blue, representing normal operating temperatures, through yellow to red at elevated fault-level temperatures, following the perceptual principle of natural mapping (Norman, 2013): heat is intuitively associated with the warm end of the color spectrum, requiring no legend for correct interpretation. At the critical red threshold, an additional emissive glow effect is applied, further increasing the perceptual salience of overheating conditions. The shader updates continuously as sensor values evolve over the course of a scenario (Figure 3.4).

**Ghost-Robot Vibration Overlay** Vibration is made spatially perceptible through a purpose-built ghost overlay: a complete duplicate of the robot model, rendered semi-transparently and offset slightly in time relative to the primary mesh, so that it appears to trail behind the physical robot's motion. The positional lag of the ghost is amplified by a factor of approximately 6–10× relative to the physically simulated vibration amplitude, ensuring that even moderate fault-level vibration produces a clearly visible tremor effect. The opacity of the ghost mesh at each joint is inversely coupled to its transparency: as vibration intensity at a joint increases, the ghost becomes more opaque at that location, providing an additional visual channel that spatially localizes the fault on the robot body. Joints operating within normal parameters remain nearly invisible in the ghost layer (Figure 3.4, where the ghost grows more pronounced from the healthy to the critical state).

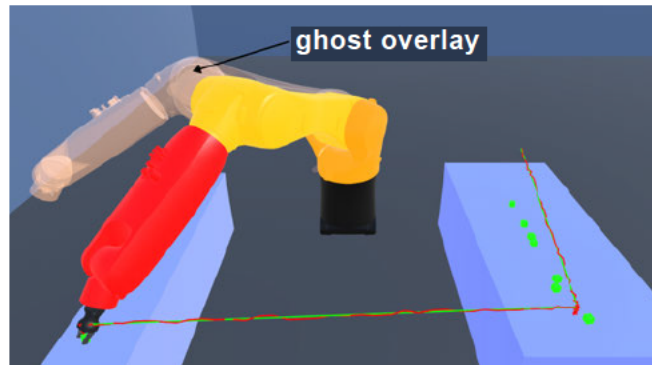
**Auditory Feedback** An audio source is attached to the robot model and emits ambient operational sound throughout all scenarios. This serves to increase environmental immersion rather than to encode fault-specific information, and is therefore present uniformly across fault and no-fault scenarios alike. Fault-specific auditory cues were deliberately not implemented as a distinct diagnostic channel. This decision was motivated by the goal of not overloading participants with too many simultaneous cues, which could increase cognitive load and reduce the overall interpretability of the multimodal display. The auditory channel thus serves primarily to enhance the realism of the simulation rather than to convey explicit diagnostic information.



(a) Healthy



(b) Elevated



(c) Critical

Figure 3.4.: Multimodal visualization across three fault-severity states. The temperature heatmap shades the robot mesh from turquoise (healthy) through yellow to red (critical); the semi-transparent ghost overlay trails the arm to make vibration spatially perceptible, becoming more pronounced as severity rises; and the green/red end-effector path lines indicate trajectory deviation. Together these encode the perceptual channels of the multimodal condition.

**Haptic Controller Feedback** Haptic feedback is delivered via the Meta Quest controllers using the `HapticImpulsePlayer` API and is active exclusively in the multimodal conditions. When a participant aims the controller ray at a robot joint, haptic vibration pulses are triggered in the controller, with the intensity reflecting the current vibration level of the targeted joint. This allows participants to probe individual joints for fault-related vibration through directed physical interaction, adding a tactile dimension to the spatial diagnostic task. Although implemented at an early stage of development, the feature was retained throughout the iterative process, as it proved a valuable addition to the multimodal experience without introducing undue complexity or cognitive load.

### 3.3.3. Text Data Panel (Baseline Condition)

In the two label-based conditions (*VR\_Label\_NoLLM* and *VR\_Label\_LLM*), sensor data is presented as floating billboard labels anchored to each of the five robot joints in world space. Each label displays the joint name, the current temperature formatted as `T: XX.X°C`, and the vibration level as a percentage (`vib: XX%`). An additional label near the end effector reports the estimated path deviation in millimeters, derived from the average vibration level weighted by the scenario-specific drift-influence parameter. All labels face the user camera at all times (billboard rendering) and update at 10 Hz. This design replicates the information content of a conventional numerical monitoring dashboard within the VR environment, serving as the text-only baseline against which the multimodal condition is compared.

### 3.3.4. Condition Switching and Scenario Management

Scene and condition management is handled by `StudyManager`, a singleton that persists across Unity scene loads via `DontDestroyOnLoad`. Each of the four experimental conditions corresponds to a dedicated Unity scene (*VR\_label\_noLLM*, *VR\_label\_LLM*, *VR\_multi\_noLLM*, *VR\_multi\_LLM*), ensuring clean separation of visualization components per condition. Upon condition start, `StudyManager` loads the appropriate scene, initializes the correct scenario sequence for the current participant, and executes a forced garbage collection pass to prevent memory accumulation across the four-condition session. Within each scene, `ScenarioLoader` applies the active `ScenarioConfig` parameters to the robot simulation at scenario start.

## 3.4. LLM Integration

### 3.4.1. LLM Selection and API Integration

The LLM component is powered by the Groq API using the llama-3.1-8b-instant model. This model was selected for its low inference latency on Groq’s hardware-accelerated inference infrastructure, which is critical for maintaining a fluid conversational experience within VR. Responses are constrained to a maximum of 120 tokens at a sampling temperature of 0.3 to prioritize concise, deterministic outputs over verbose explanations.

The integration follows a three-step pipeline executed on the Python backend. First, the Unity client sends the transcribed participant query together with the current sensor readings to the POST /query endpoint. Second, the backend optionally retrieves relevant document snippets from the local ChromaDB vector store (see Section 3.4.2) before assembling the full prompt. Third, the assembled prompt is forwarded to the Groq API, and the returned response text is sent back to Unity, where it is passed to the text-to-speech pipeline.

### 3.4.2. Context and Prompt Engineering

The backend maintains a retrieval-augmented generation (RAG) knowledge base (Lewis et al., 2020) using ChromaDB<sup>5</sup> as the vector store and the VoyageAI<sup>6</sup> voyage-3 model for text embeddings. The document corpus comprises four plain-text files: a purpose-written fault-signature reference sheet mapping each of the nine diagnostic categories to their characteristic sensor patterns, and three general-purpose texts on industrial robot maintenance sourced from publicly available documentation. At query time, the top-ranked document chunk is retrieved and appended as background context to the prompt.

The system prompt assigns the LLM the role of a concise technical maintenance advisor and enforces strict response constraints: answers are limited to one or two sentences, numerical values are formatted in domain-appropriate units (temperature in °C, vibration as a percentage), and disclaimers or off-topic responses are suppressed. The prompt additionally encodes the characteristic sensor signatures of all nine fault types, allowing the model to reason over the current sensor state even without explicit retrieval.

A key design choice is the injection of the scenario’s correct diagnosis as a hidden internal field in the prompt, formatted as [INTERNAL: the correct diagnosis is `X`]. This field is invisible to the participant but instructs the LLM to guide the user towards the correct diagnosis when directly asked for a recommendation. When the participant poses open-ended questions, the

---

<sup>5</sup>Chroma, <https://www.trychroma.com> (accessed 2026-06-30).

<sup>6</sup>Voyage AI, <https://www.voyageai.com> (accessed 2026-06-30).

model instead describes the relevant sensor patterns without revealing the answer directly. This design ensures that the LLM acts as a genuine decision-support tool rather than an answer oracle, while still being able to confirm or correct a participant’s hypothesis when prompted.

This is a deliberate design decision: by supplying the ground-truth diagnosis to the assistant, the study isolates the effect of *having* a reliable decision-support dialogue from the orthogonal question of whether a deployed LLM would diagnose correctly on its own. The trade-off is that the assistant is more reliable than a real-world system would be; the resulting implications for ecological validity are taken up explicitly in Sections 6.1.1 and 6.4.4.

Conversation history is maintained across turns within a condition: each prior exchange is prepended to the current request, allowing the model to provide contextually coherent multi-turn responses throughout a scenario.

### 3.4.3. Voice Input / Speech-to-Text

Voice interaction is the exclusive input modality for the LLM assistant, chosen to preserve the hands-free character of the VR experience. Participants activate the microphone via a push-to-talk button on the Meta Quest controller. The recorded audio is transcribed using the Groq Whisper API via the GroqTranscriber component in Unity; the resulting text is then forwarded to the backend query pipeline.

The assistant’s response is delivered exclusively as synthesized speech: no response text is displayed in the scene. Text-to-speech synthesis is performed by the Cartesia API<sup>7</sup>, and the resulting audio file is played back within the VR application. This design ensures that participants engage with the assistant through natural dialogue rather than reading, which is consistent with the hands-free VR context and reduces the risk of visual attention being diverted from the robot to a text panel.

The system supports both English and German. The study was conducted primarily in English; however, participants were explicitly offered the option to interact with the assistant in German, and a subset chose to do so. Questionnaires were available in both languages.

### 3.4.4. Embodied Virtual Agent (Avatar)

The LLM assistant is embodied as a realistic humanoid avatar positioned to the left of the robot, within the participant’s natural field of view when inspecting the digital twin. The avatar was created using a photogrammetric face-scanning tool applied to the researcher’s

---

<sup>7</sup>Cartesia, <https://cartesia.ai> (accessed 2026-06-30).

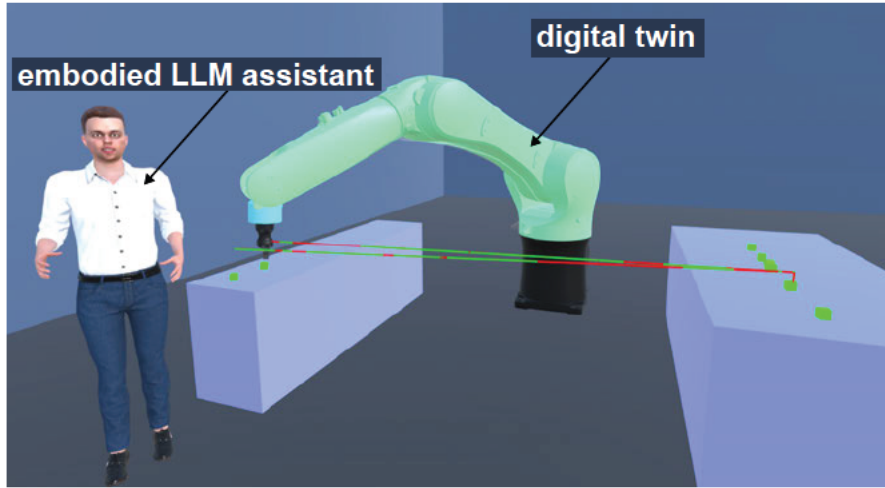


Figure 3.5.: The embodied LLM assistant, positioned to the left of the robot within the participant’s field of view, providing a spatial anchor for the spoken responses.

own likeness and dressed in casual professional attire (shirt and trousers) to convey a collegial, non-authoritative persona.

The avatar features a custom lip-synchronization system (AvatarJawLipSync, AvatarLipSync) that drives jaw and blendshape animations in real time from the Cartesia TTS audio stream, producing the visual impression of speech during response playback. Between responses, the avatar displays idle animations to maintain a sense of presence. This combination of photorealistic appearance, professional attire, and synchronized lip movement was intended to maximize social presence and perceived credibility of the assistant, consistent with evidence that embodied agents with synchronized lip movement (Peixoto et al., 2021) and a realistic appearance heighten social presence in VR (Garau et al., 2003; Lim et al., 2025). Figure 3.5 shows the spatial arrangement of the avatar next to the robot from the participant’s perspective.

## 3.5. Study Infrastructure

### 3.5.1. Logging System (Correctness, Task Completion Time)

All study-relevant events are logged automatically by the DecisionLogger component attached to StudyManager. Two separate CSV files are created per participant session: a decisions file and an events file. The decisions file records one row per submitted diagnosis and contains the participant identifier, session identifier, UTC timestamp, scenario identifier, experimental condition, participant diagnosis, correct diagnosis, correctness flag, confidence rating (1–7

Likert scale), task completion time in seconds, and the instantaneous sensor readings for all five joints (temperatures in °C and vibration levels as normalized values) at the moment of submission. The events file captures coarser session-level events including SESSION\_START, SCENARIO\_START, PHASE\_CHANGE, LLM\_QUERY, and SESSION\_END.

Task completion time is derived as the elapsed time between the SCENARIO\_START event and the moment the participant submits their diagnosis. Diagnostic correctness is computed post-hoc as a binary correctness flag per scenario and aggregated to a proportion per condition.

Log files are written locally to the Quest’s persistent data path and transmitted over the local network to the study backend, where they are stored as CSV and log files on the experimenter’s machine. No participant data are sent to external or cloud-based services; all records remain under local control in accordance with the data-protection terms of the informed-consent form (Appendix B).

### **3.5.2. Questionnaire Integration**

After completing the three scenarios of each experimental condition, participants removed the headset and completed a questionnaire block at a PC using a purpose-built HTML survey administered locally in the browser. The survey collected three standardized instruments: the NASA Task Load Index (NASA-TLX) measuring perceived cognitive workload across six subscales, the Trust in Automated Systems (TAS) scale measuring trust in the system, and the User Experience Questionnaire Short (UEQ-S) assessing overall user experience. Decision confidence, by contrast, was not part of this PC survey: it was captured inside the VR environment immediately after each individual diagnosis on the hand-held clipboard, as a single 7-point Likert item (see Section 4.4.2), and the three per-scenario ratings were subsequently averaged into one confidence score per condition. Questionnaire responses were stored locally as JSON and collected by the experimenter at the end of each session. The survey was available in both English and German.

### **3.5.3. Counterbalancing Implementation**

Condition order across the four experimental conditions was counterbalanced using a  $4 \times 4$  Latin Square, yielding four distinct orderings each of which exposes every condition at each position exactly once across the participant group. Participants were assigned to a Latin Square row based on the numeric component of their participant identifier modulo 4, ensuring automatic assignment without experimenter intervention.

Within each condition, the three scenarios are selected deterministically from the scenario pool using a seed computed from the participant identifier and the condition index, guaranteeing that each participant receives a unique but reproducible combination of scenarios across the twelve trials (three scenarios per condition, four conditions). The within-condition presentation order is additionally shuffled per participant to prevent scenario-position effects. This design ensures that potential learning effects and scenario familiarity are distributed evenly across conditions and participants.

## 4. Study Design & Methodology

### 4.1. Research Design

#### 4.1.1. Experimental Approach (Controlled Lab Study)

This study employs a controlled laboratory experiment to investigate the individual and combined effects of immersive multimodal visualization and LLM-based decision support on fault diagnosis in a virtual reality digital twin. The experiment was conducted at Hamburg University of Applied Sciences (HAW Hamburg). Two independent variables were manipulated: (1) the *visualization modality* (text label panel vs. multimodal VR representation) and (2) *LLM assistance* (absent vs. present). Dependent variables encompass both objective performance measures (diagnostic correctness and task completion time) and subjective measures collected via standardized questionnaires and a single confidence item.

#### 4.1.2. Within-Subject Design and Justification

All participants completed all four experimental conditions within a single session (within-subject design). This approach was chosen for two reasons: it maximizes statistical power by eliminating between-subject variance in baseline diagnostic ability, and it allows each participant to serve as their own control, which is particularly important given the diversity of prior VR and AI experience in the sample. The primary risk associated with within-subject designs, ordering and learning effects, was addressed through Latin Square counterbalancing of the condition order and deterministic randomization of the scenario assignment (see Section 4.6.2).

#### 4.1.3. 2×2 Factorial Design

The study follows a  $2 \times 2$  fully crossed factorial design. Factor A is the visualization modality with two levels: a text-only label condition (floating numerical readouts per joint) and a multimodal VR condition (heatmap shader, ghost-robot vibration overlay, haptic feedback).

Factor B is LLM assistance with two levels: absent and present. The full crossing of these two factors yields four experimental conditions, as shown in Table 4.1. For brevity, the statistical analysis in Chapter 5 abbreviates the two visualization levels as *Labels* (text panel) and *No-Labels* (multimodal).

Table 4.1.: The four experimental conditions resulting from the  $2 \times 2$  factorial design. Each participant completed all four conditions.

|             | No LLM      | LLM         |
|-------------|-------------|-------------|
| Text Labels | Condition 1 | Condition 3 |
| Multimodal  | Condition 2 | Condition 4 |

In Chapter 5 the visualization levels **Text Labels** and **Multimodal** are abbreviated *Labels* and *No-Labels*, respectively.

## 4.2. Participants

### 4.2.1. Target Population and Recruitment

A total of  $N = 22$  participants were recruited via a university-wide announcement at HAW Hamburg and through personal networks and word of mouth. The target population consisted of students and researchers with a general technical or scientific background; no prior experience with industrial robot maintenance was required or expected, as the study explicitly targets non-expert operators as the primary user group. Participants received no monetary compensation for their participation.

### 4.2.2. Inclusion/Exclusion Criteria

Participants were required to be of legal age (at least 18 years). For safety reasons, individuals with a known heart condition, anxiety disorders, or a susceptibility to seizures were not considered, and participation during pregnancy was advised against. Participants were also asked not to take part if they were acutely suffering from nausea, vomiting, headache, dizziness, or migraine, or if they were under the influence of alcohol, drugs, or perception-altering medication. These criteria follow the approved ethics protocol and were restated on the informed-consent form (Appendix B). Consistent with the study’s focus on non-expert operators, no prior experience in industrial robot maintenance was required.

### 4.2.3. Sample Size and Power Considerations

Because the achievable sample size was constrained by participant availability rather than fixed a priori, a sensitivity power analysis was conducted to determine the smallest effect the design could reliably detect with the realized sample. In the  $2 \times 2$  within-subject design, each of the two main effects and their interaction is tested with a single numerator degree of freedom ( $df_1 = 1$ ,  $df_2 = N - 1 = 21$ ) and is equivalent to a paired contrast over the  $N = 22$  participants. For  $\alpha = 0.05$  and a desired power of  $1 - \beta = 0.80$  (two-tailed), this configuration is sensitive to standardized within-subject mean differences of  $d_z \geq 0.63$ , corresponding to a medium-to-large effect (Cohen’s  $f \approx 0.31$ ,  $\eta_p^2 \approx 0.09$ ). The study is therefore adequately powered to detect medium-to-large main effects, whereas smaller effects may have remained undetected; this is taken into account when interpreting the non-significant results (Chapter 6). The sample size was ultimately bounded by participant availability; its consequences for the detectability of smaller effects are revisited in Section 6.4.

### 4.2.4. Realized Sample

The final sample comprised  $N = 22$  participants (17 male, 5 female) with a mean age of  $M = 30.9$  years ( $SD = 10.8$ ; range 21–61, median 26); the age distribution was right-skewed, with most participants in their twenties and a few older participants. The sample was predominantly academic in terms of formal education (11 held a Bachelor’s and 5 a Master’s degree, the remainder a vocational qualification or a secondary-school leaving certificate). Prior experience with virtual reality varied across the sample, ranging from no prior exposure ( $n = 4$ ) through little ( $n = 12$ ) and moderate ( $n = 3$ ) to regular use ( $n = 3$ ). Prior exposure to AI-based tools was, by contrast, generally high ( $n = 12$  “much”,  $n = 8$  “moderate”,  $n = 2$  “little”). Consistent with the inclusion criteria (Section 4.2.2) and the study’s focus on non-expert operators, none of the participants reported professional experience in industrial robot maintenance; prior exposure to robotics ranged from none ( $n = 18$ ) to at most limited or purely theoretical familiarity ( $n = 4$ ). All 22 participants completed all four experimental conditions; no data were excluded.

## 4.3. Experimental Conditions

### 4.3.1. Condition 1 – Baseline (Text Panel, No LLM)

In Condition 1, participants diagnosed robot faults relying exclusively on numerical sensor data presented as floating billboard labels attached to each of the five joints (temperature in °C, vibration as a percentage, and end-effector deviation in millimeters). No multimodal

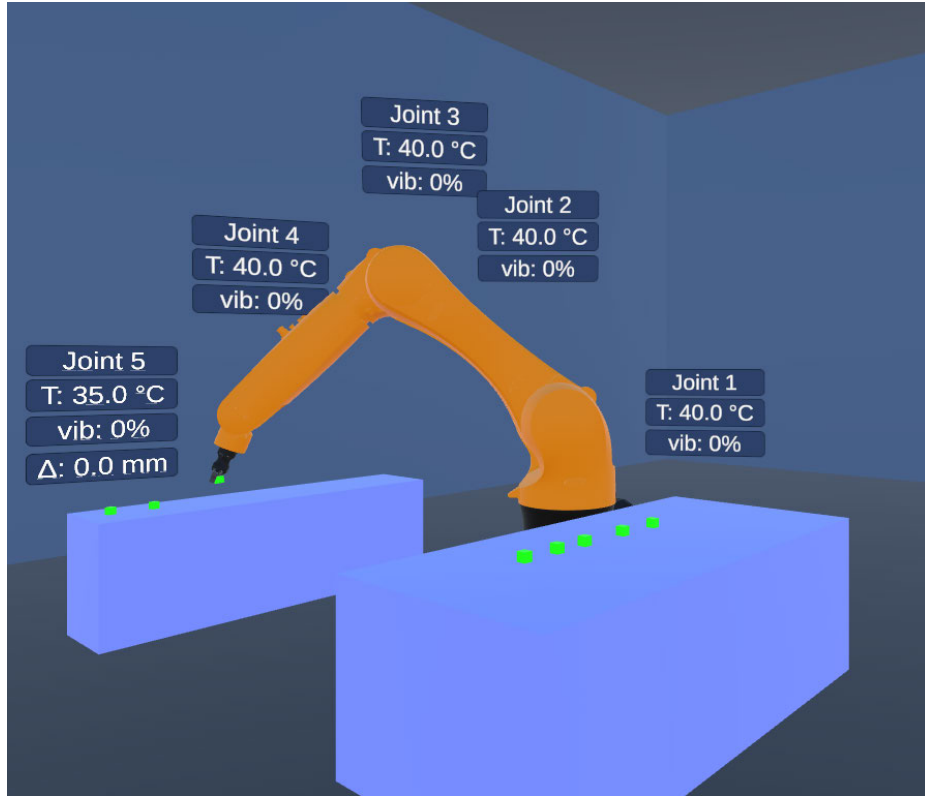


Figure 4.1.: Participant's view in Condition 1 (text labels, no LLM): each joint shows a floating label with temperature, vibration, and end-effector deviation.

visualization and no LLM assistant were available. This condition constitutes the closest approximation to a conventional text-based monitoring dashboard and serves as the baseline against which all other conditions are compared. Figure 4.1 shows the participant's view in this condition.

#### 4.3.2. Condition 2 – Immersive VR (Multimodal, No LLM)

Condition 2 activates the full multimodal visualization layer described in Section 3.3.2 and illustrated across severity states in Figure 3.4: the temperature heatmap shader, the ghost-robot vibration overlay, and haptic controller feedback. The numerical text labels are *not* shown in this condition; the sensor state is conveyed exclusively through the perceptual cues, so that the multimodal representation genuinely replaces, rather than supplements, the numerical readout. No LLM assistant is available. This condition isolates the effect of multimodal visualization on diagnostic performance.

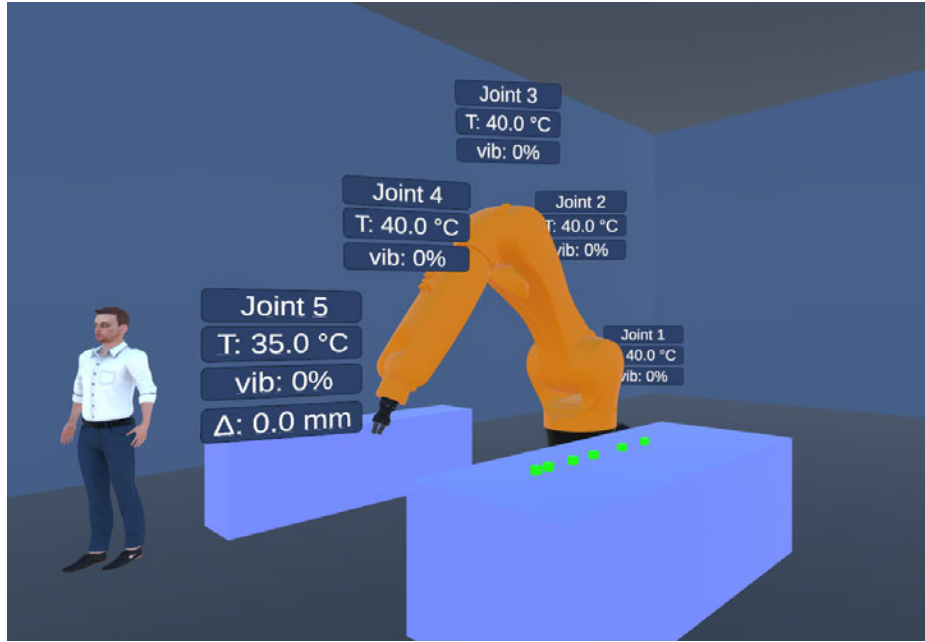


Figure 4.2.: Participant's view in Condition 3 (text labels + LLM): the Condition 1 labels with the embodied LLM assistant added to the left.

#### 4.3.3. Condition 3 – LLM Only (Text Panel + LLM)

Condition 3 presents the same text-only label interface as Condition 1, with the addition of the voice-based LLM assistant. Participants could query the embodied avatar agent at any time by pressing the push-to-talk button on the Meta Quest controller, speaking their question in English or German, and receiving a spoken response synthesized by the Cartesia TTS system. The avatar's lip-synchronization animations were active during playback. This condition isolates the effect of LLM assistance without any multimodal visual augmentation (Figure 4.2).

#### 4.3.4. Condition 4 – Combined (Multimodal VR + LLM)

Condition 4 combines the full multimodal visualization of Condition 2 with the LLM voice assistant of Condition 3, constituting the complete intended system. Participants had simultaneous access to all perceptual channels and could additionally consult the AI assistant via voice. This condition is expected to yield the highest diagnostic correctness and lowest cognitive workload if the two interventions act synergistically. The embodied assistant alongside the multimodal robot is shown in Figure 3.5.

## 4.4. Scenarios and Tasks

### 4.4.1. Scenario Design and Assignment (12 Trials per Participant)

Each participant completed twelve scenario trials in total, three per experimental condition, drawn from the pool of nine distinct scenarios introduced in Section 3.2.4. Each condition's three scenarios were drawn from the three severity categories (Low Urgency, Medium Severity, Critical), one per category, so that every condition presented a comparable spread of severities. Because twelve trials are drawn from nine scenarios, individual scenarios recurred across conditions for a given participant; the assignment was deterministic, seeded by the participant identifier, so that it remained balanced and reproducible across the sample. The within-condition presentation order was additionally randomized per participant.

### 4.4.2. Task Description (Maintenance Decision + Confidence Rating)

At the start of each scenario, the robot simulation was initialized with the corresponding fault parameters and began its continuous pick-and-place movement. Participants were free to observe the robot for as long as needed before submitting a diagnosis; no time limit was enforced.

To submit a diagnosis, participants activated a hand-mounted clipboard interface via a button on the left Meta Quest controller. The clipboard displayed all nine possible diagnostic categories (see Section 4.4.3) as a selectable list, from which participants chose one using the right controller's ray pointer. Following the diagnosis selection, participants rated their confidence in the decision on a 7-point Likert scale (1 = not at all confident, 7 = extremely confident). Diagnostic correctness was scored as a binary outcome (correct / incorrect) per scenario. The clipboard interface used for this combined diagnosis-and-confidence step is shown in Figure 4.3.

### 4.4.3. Fault Types and Diagnostic Categories

Participants were required to distinguish among nine diagnostic categories. Prior to the study, all fault types were explained in a standardized verbal briefing, and a reference overview of the diagnostic categories was available on the in-VR hand clipboard throughout each scenario. The nine categories and their characteristic sensor signatures are as follows:

- **No Error:** sensor values evenly distributed across joints and stable over time.
- **Bearing Damage:** elevated temperature *and* vibration at the same joint, both rising.

**Maintenance Sheet**  
Select your diagnoses

|                    |                                                                                                        |        |
|--------------------|--------------------------------------------------------------------------------------------------------|--------|
| Bearing Damage     | Elevated temperature AND elevated vibration at the same joint, both rising                             | Select |
| Motor Overheating  | Significantly elevated temperature at one joint, rapidly rising. Vibration normal                      | Select |
| Loose Mounting     | Significantly elevated vibration at one joint. Temperature normal                                      | Select |
| Gearbox Play       | Values normal, but visible path deviation at the end effector                                          | Select |
| Lubrication Needed | Slightly elevated temperature AND slightly elevated vibration at the same joint, slowly rising         | Select |
| No Error           | Values may be elevated, but evenly distributed across all joints and stable                            | Select |
| Overload           | Elevated temperature at multiple/all joints (lower joints more than upper). Vibration normal           | Select |
| Coupling Defect    | Elevated vibration at an upper joint. Temperature normal                                               | Select |
| Cooling Failure    | Significantly elevated temperature at multiple joints simultaneously, rapidly rising. Vibration normal | Select |

**Confidence**      How Confident are you?

1 7

Confirm

Figure 4.3.: The in-VR maintenance sheet for submitting a diagnosis: participants selected one of nine categories (each with a short signature description) and rated their confidence before confirming.

- **Motor Overheating:** significantly elevated and rapidly rising temperature at one joint; vibration within normal range.
- **Loose Mounting:** significantly elevated vibration at one joint; temperature within normal range.
- **Gearbox Play:** sensor values within normal range, but visible end-effector path deviation.
- **Overload:** elevated temperature concentrated at the lower joints (joints 1–3); vibration normal.
- **Coupling Defect:** elevated vibration at the upper joints (joints 4–5); temperature normal.

- **Cooling Failure:** significantly elevated temperature at multiple joints simultaneously, rising rapidly; vibration normal.
- **Lubrication Needed:** slightly elevated temperature *and* vibration at the same joint, rising slowly.

## 4.5. Measures

### 4.5.1. Subjective Measures (NASA-TLX, TAS, UEQ-S, Decision Confidence)

Three standardized instruments were administered via a PC-based HTML survey after the three scenarios of each experimental condition. Participants removed the headset between conditions and completed the questionnaire block before continuing with the next condition.

**NASA Task Load Index (NASA-TLX)** (Hart & Staveland, 1988) measures perceived workload across six subscales: mental demand, physical demand, temporal demand, performance, effort, and frustration. Each subscale was rated on a 21-point scale (1 = very low, 21 = very high). Raw (unweighted) TLX scores are reported, consistent with recent recommendations in the human factors literature.

**Trust in Automated Systems (TAS) scale** (Jian et al., 2000) assesses trust in automated systems and was administered after each of the four conditions, including the two without LLM support, so that trust could be compared across conditions.

**User Experience Questionnaire Short (UEQ-S)** (Schrepp et al., 2017) is an eight-item instrument assessing pragmatic quality (usability) and hedonic quality (stimulation, novelty) on 7-point bipolar scales. It was administered after each condition. The full item wording for all questionnaire instruments is reproduced in Appendix A.

In addition, participants rated their **decision confidence** after each individual scenario on a 7-point Likert scale (1 = not at all confident, 7 = extremely confident), yielding three ratings per condition that were subsequently averaged.

### 4.5.2. Objective Measures (Correctness, Task Completion Time)

**Diagnostic correctness** was computed per condition as the proportion of correctly identified fault types across the three scenarios (range: 0–1, equivalent to 0–3 correct diagnoses). A diagnosis was scored as correct if and only if the selected `ErrorType` matched the `correctDiagnosis` field of the active `ScenarioConfig`.

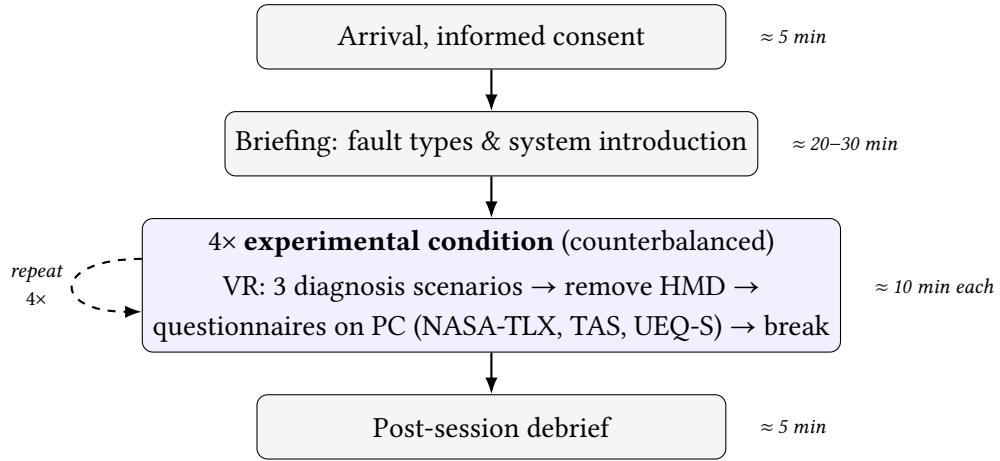


Figure 4.4.: Flow of a single experimental session (approximately 60–90 minutes, including short breaks).

**Task completion time** (TCT) was logged automatically as the elapsed time in seconds from the `SCENARIO_START` event to the moment the participant submitted their diagnosis. Three TCT values per condition were averaged to yield a single per-condition score. Cases where the time limit was not imposed and participants deliberated for extended periods are included without exclusion to preserve the naturalistic character of the diagnostic task.

## 4.6. Procedure

### 4.6.1. Session Flow and Duration

Each session followed a standardized procedure. Upon arrival, participants received a written and verbal explanation of the study purpose, task instructions, and data anonymization procedures, and provided written informed consent. The experimenter then introduced all nine fault types and their characteristic sensor signatures, using the in-VR clipboard as a visual reference. Participants were encouraged to ask clarifying questions at this stage. No formal VR familiarization or practice scenario was included; the briefing was intended to compensate for the absence of a warm-up trial. Post-hoc observation suggested that some participants in their first condition required additional time to orient themselves to the interface, which is discussed as a limitation in Section 6.4.

Following the briefing, participants completed the four experimental conditions in their assigned counterbalanced order. Each condition consisted of three scenarios conducted within the VR headset, after which the participant removed the headset and completed the three questionnaires (NASA-TLX, TAS, UEQ-S) and reviewed their confidence ratings for the

just-completed scenarios on the PC survey. A short break was offered between conditions. The full session lasted approximately 60 to 90 minutes. Figure 4.4 summarizes the overall session procedure.

#### **4.6.2. Counterbalancing Strategy**

Condition order was counterbalanced across participants using a  $4 \times 4$  Latin Square, yielding four unique orderings such that each condition appears at each serial position exactly once across the participant group. Participants were automatically assigned to a Latin Square row based on the numeric component of their participant identifier modulo 4. The scenario assignment within each condition was additionally randomized per participant, as described in Section 3.5.3.

#### **4.6.3. Briefing and Informed Consent**

An ethics application for the study was reviewed by the ethics review board of HAW Hamburg and received a favorable opinion on 5 May 2026. All participants provided written informed consent before any data were recorded, and participation was voluntary with the right to withdraw at any time without consequences. Data were collected anonymously under participant identifiers, with no names or other personal data stored alongside the study data.

### **4.7. Data Analysis Plan**

#### **4.7.1. Repeated Measures ANOVA (rmANOVA)**

The primary analysis is a  $2 \times 2$  repeated-measures ANOVA with Visualization (Labels vs. No-Labels) and LLM support (absent vs. present) as within-subject factors, computed separately for each dependent variable: diagnostic correctness, task completion time, the six NASA-TLX subscales, the TAS trust score, the UEQ-S pragmatic, hedonic, and overall scores, and mean decision confidence. The analysis was carried out in IBM SPSS Statistics. For each effect (the two main effects and their interaction) the analysis yields  $F(1, 21)$ , the corresponding  $p$ -value, and partial eta-squared ( $\eta_p^2$ ) as the effect-size measure.

### **4.7.2. Sphericity**

Because each within-subject factor has only two levels, the sphericity assumption is trivially satisfied, so that no Mauchly test or Greenhouse–Geisser correction is required.

### **4.7.3. Effect Sizes**

Effect sizes are reported as partial eta-squared ( $\eta_p^2$ ) for each main effect and the interaction, interpreted using the conventional benchmarks (small: 0.01, medium: 0.06, large: 0.14; Cohen, 1988). Because each factor has only two levels, the main effects are read directly from the omnibus tests and require no post-hoc pairwise decomposition.

## 5. Results

In this chapter, we present the descriptive statistics and inferential statistical tests for our experiment. We analyzed the responses with two-way repeated-measures analyses of variance (RM-ANOVAs) for the two factors LLM and Labels at the 5% significance level. Because each factor has only two levels, the main effects are read directly from the ANOVA and require no post-hoc pairwise comparisons.

For brevity, this chapter and all result figures refer to the two factors by their level names rather than by the condition descriptions used in Chapter 4. The *visualization* factor distinguishes *Labels* — the numerical text-label interface (Conditions 1 and 3) — from *No-Labels* — the immersive multimodal interface, in which the numerical labels are replaced by the heatmap, ghost-robot, and haptic cues (Conditions 2 and 4). The *LLM* factor distinguishes *LLM* (assistant available, Conditions 3 and 4) from *No-LLM* (no assistant, Conditions 1 and 2). In other words, *No-Labels* denotes the multimodal variant introduced in Section 4.3.

We found no significant interaction effects, hence we proceeded with analyses of the main effects.

### 5.1. Diagnostic Correctness

The descriptive statistics for the Correctness results are shown in Figure 5.1.

We found a significant effect of LLM on *Correctness*,  $F(1, 21) = 11.08$ ,  $p = \mathbf{0.003}$ ,  $\eta_p^2 = 0.345$ , indicating that the LLM conditions were significantly higher in correctness than the No-LLM conditions.

We found no significant effect of Labels on *Correctness*,  $F(1, 21) = 0.02$ ,  $p = 0.887$ ,  $\eta_p^2 = 0.001$ .

### 5.2. Task Completion Time

The descriptive statistics for the Time results are shown in Figure 5.2.

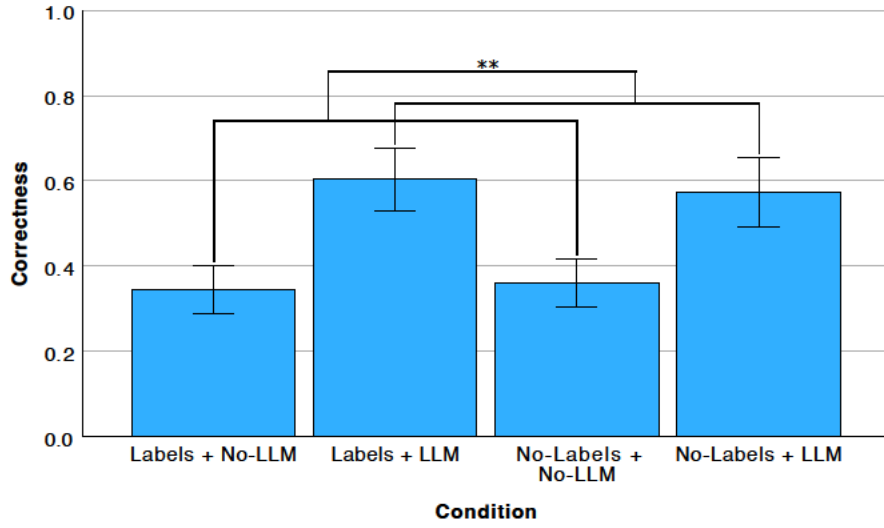


Figure 5.1.: Diagnostic correctness by condition (higher is better). The bar marks a significant LLM main effect (\*\* :  $p < 0.01$ ).

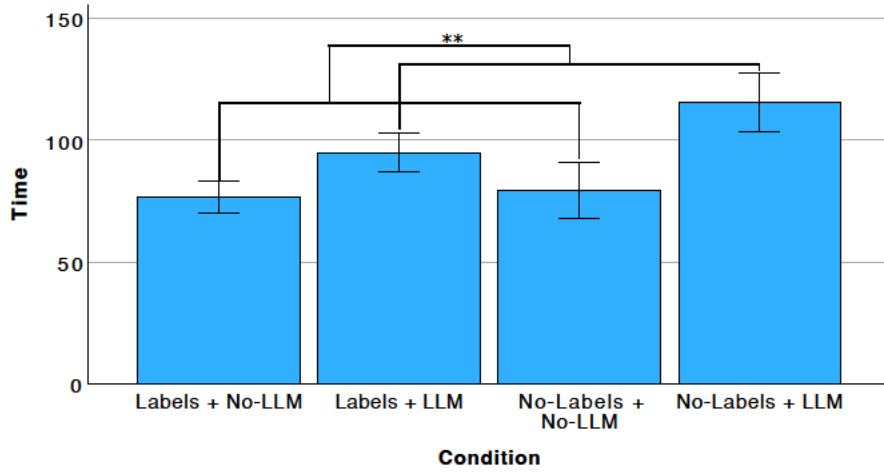


Figure 5.2.: Task completion time (seconds) by condition (higher is worse). The bar marks a significant LLM main effect (\*\* :  $p < 0.01$ ).

We found a significant effect of LLM on *Time*,  $F(1, 21) = 8.68$ ,  $p = \mathbf{0.008}$ ,  $\eta_p^2 = 0.292$ , indicating that LLM-assisted conditions required significantly more time to complete tasks.

We found no significant effect of Labels on *Time*,  $F(1, 21) = 2.21$ ,  $p = 0.152$ ,  $\eta_p^2 = 0.10$ .

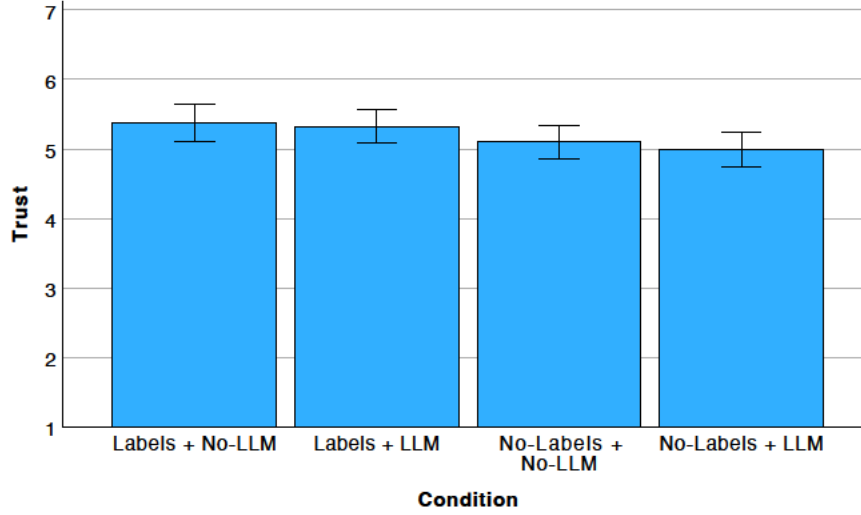


Figure 5.3.: Trust ratings by condition (higher is better).

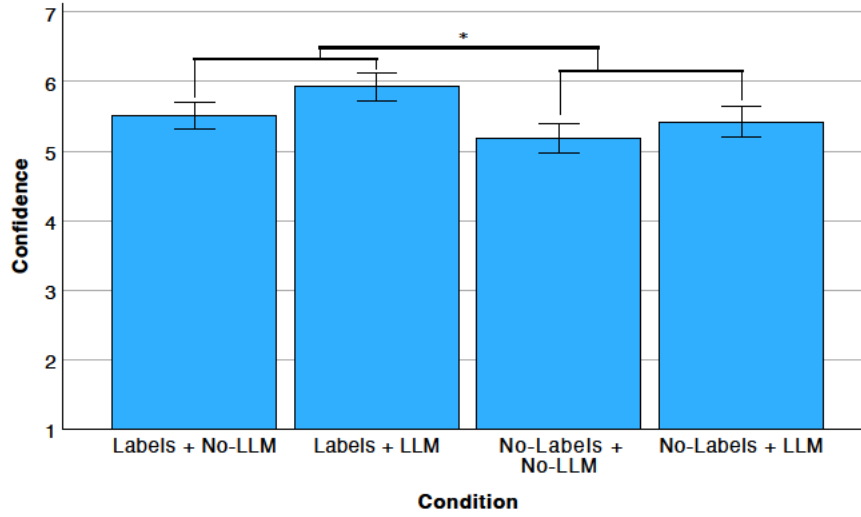


Figure 5.4.: Decision confidence by condition (higher is better). The bar marks a significant *Labels* effect (\* :  $p < 0.05$ , Labels > No-Labels); the LLM effect was not significant.

### 5.3. Trust and Decision Confidence

The descriptive statistics for the Trust and Confidence results are shown in Figure 5.3 and Figure 5.4.

We found no significant effect of LLM on *Trust*,  $F(1, 21) = 0.21$ ,  $p = 0.648$ ,  $\eta_p^2 = 0.010$ , on *Confidence*,  $F(1, 21) = 3.12$ ,  $p = 0.092$ ,  $\eta_p^2 = 0.129$ , and no significant effect of Labels on *Trust*,  $F(1, 21) = 2.67$ ,  $p = 0.118$ ,  $\eta_p^2 = 0.11$ .

We found a significant effect of Labels on *Confidence*,  $F(1, 21) = 7.26$ ,  $p = \mathbf{0.014}$ ,  $\eta_p^2 = 0.26$ ,

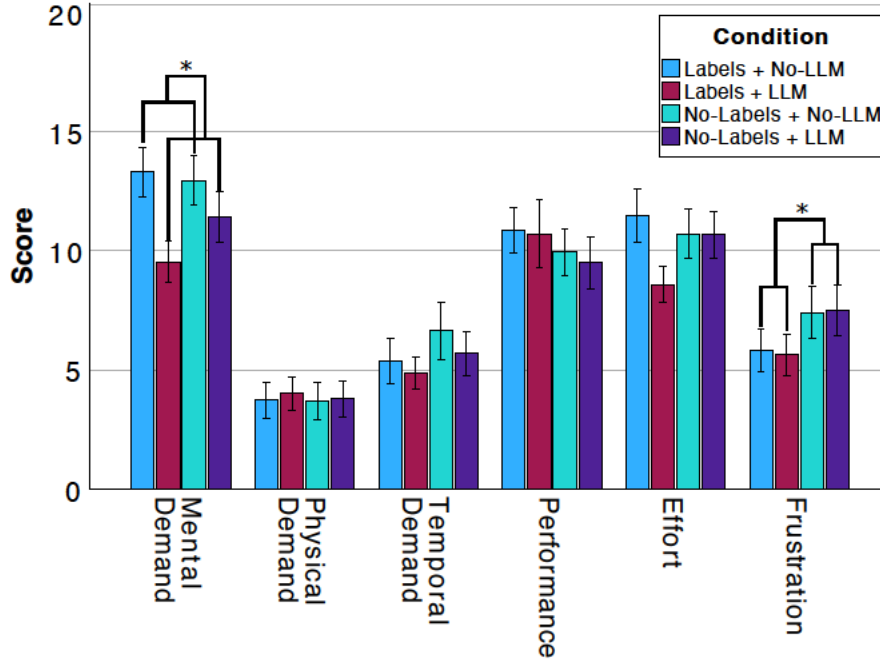


Figure 5.5.: NASA-TLX subscales by condition (higher is worse). Bars mark significant main effects (\* :  $p < 0.05$ ): LLM on *Mental Demand*, Labels on *Frustration*.

indicating that the Labels conditions were rated higher than the No-Labels conditions.

## 5.4. Cognitive Load (NASA-TLX)

The descriptive statistics for the NASA-TLX results are shown in Figure 5.5.

We found a significant effect of LLM on *Mental Demand*,  $F(1, 21) = 8.10$ ,  $p = \mathbf{0.010}$ ,  $\eta_p^2 = 0.28$ , indicating that the No-LLM conditions were rated higher than the LLM conditions.

We found no significant effect of LLM on *Physical Demand*,  $F(1, 21) = 0.31$ ,  $p = 0.582$ ,  $\eta_p^2 = 0.02$ , on *Temporal Demand*,  $F(1, 21) = 0.83$ ,  $p = 0.372$ ,  $\eta_p^2 = 0.038$ , on *Performance*,  $F(1, 21) = 0.09$ ,  $p = 0.773$ ,  $\eta_p^2 = 0.001$ , on *Effort*,  $F(1, 21) = 2.75$ ,  $p = 0.112$ ,  $\eta_p^2 = 0.12$ , on *Frustration*,  $F(1, 21) = 0.003$ ,  $p = 0.953$ ,  $\eta_p^2 = 0.001$ .

We found no significant effect of Labels on *Mental Demand*,  $F(1, 21) = 0.92$ ,  $p = 0.349$ ,  $\eta_p^2 = 0.04$ , on *Physical Demand*,  $F(1, 21) = 0.18$ ,  $p = 0.677$ ,  $\eta_p^2 = 0.01$ , on *Temporal Demand*,  $F(1, 21) = 1.61$ ,  $p = 0.219$ ,  $\eta_p^2 = 0.07$ , on *Performance*,  $F(1, 21) = 1.15$ ,  $p = 0.295$ ,  $\eta_p^2 = 0.05$ , on *Effort*,  $F(1, 21) = 0.80$ ,  $p = 0.382$ ,  $\eta_p^2 = 0.04$ .

We found a significant effect of Labels on *Frustration*,  $F(1, 21) = 4.58$ ,  $p = \mathbf{0.044}$ ,  $\eta_p^2 = 0.18$ , indicating that the Labels conditions were rated lower than the No-Labels conditions.

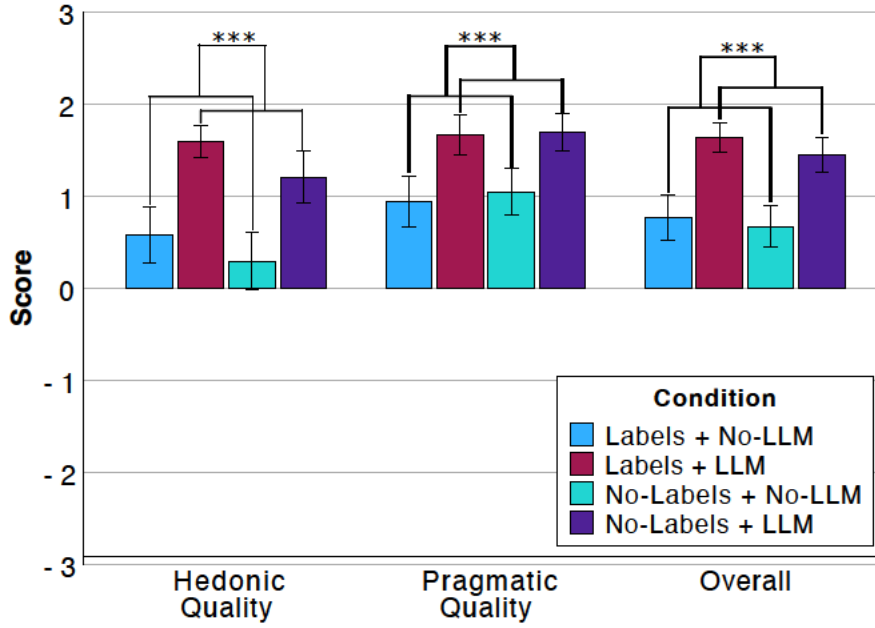


Figure 5.6.: UEQ-S scales by condition (higher is better). The bars mark a significant LLM main effect (\*\*\*:  $p < 0.001$ ).

## 5.5. User Experience: UEQ-S

The descriptive statistics for the UEQ-S results are shown in Figure 5.6.

We found a significant effect of LLM on *Hedonic Quality*,  $F(1, 21) = 16.05$ ,  $p < 0.001$ ,  $\eta_p^2 = 0.43$ , *Pragmatic Quality*,  $F(1, 21) = 18.53$ ,  $p < 0.001$ ,  $\eta_p^2 = 0.47$ , and *Overall Usability*,  $F(1, 21) = 29.85$ ,  $p < 0.001$ ,  $\eta_p^2 = 0.59$ . All three main effects indicate that the LLM conditions were rated significantly higher than the No-LLM conditions.

We found no significant effect of Labels on *Hedonic Quality*,  $F(1, 21) = 1.76$ ,  $p = 0.198$ ,  $\eta_p^2 = 0.08$ , *Pragmatic Quality*,  $F(1, 21) = 0.16$ ,  $p = 0.696$ ,  $\eta_p^2 = 0.01$ , nor *Overall Usability*,  $F(1, 21) = 0.67$ ,  $p = 0.423$ ,  $\eta_p^2 = 0.03$ .

## 5.6. Condition Preferences, Perceived Helpfulness, and Qualitative Feedback

### 5.6.1. Condition Preferences and Helpfulness Ratings

Seventeen of 22 participants preferred the Labels + LLM condition; 4 preferred No-Labels + LLM, and 1 preferred Labels + No-LLM. Similarly, 16 participants rated Labels + LLM as most

helpful.

### **5.6.2. Qualitative Participant Feedback**

Ten of 22 participants provided written comments. Feedback clustered around two themes: LLM utility and visualization clarity. Regarding LLM assistance, several participants noted its potential for industrial applications, with one stating it “should definitely be researched further.” However, critical observations were also raised. One participant reported instances where the LLM suggested non-existent fault diagnoses and remained unresponsive to corrections, highlighting concerns about over-reliance on potentially hallucinated outputs. Another participant noted an adaptation effect: the LLM felt confusing during the first encounter but became more helpful in following runs, suggesting that familiarity with the system may moderate its perceived utility. Regarding visualization, participants generally favored the Labels + LLM condition. One participant noted that color transitions in the multimodal condition were slow to perceive and proposed a gradient indicator to signal change direction. Another found the multimodal vibration cues difficult to interpret initially, with clarity improving only in the second run. One participant suggested a minimal labels-only variant that omits visual overlays entirely.

## 5.7. Summary of Results

Table 5.1.: Summary of two-way RM-ANOVA results. Significant effects are highlighted in bold.

| Outcome                                                                          | Factor | F(1,21) | p              | $\eta_p^2$ | Sig. | Direction          |
|----------------------------------------------------------------------------------|--------|---------|----------------|------------|------|--------------------|
| <i>Diagnostic Correctness</i>                                                    |        |         |                |            |      |                    |
| Correctness                                                                      | LLM    | 11.08   | <b>0.003</b>   | 0.345      | **   | LLM > No-LLM       |
|                                                                                  | Labels | 0.02    | 0.887          | 0.001      | n.s. | —                  |
| <i>Task Completion Time</i>                                                      |        |         |                |            |      |                    |
| Time                                                                             | LLM    | 8.68    | <b>0.008</b>   | 0.292      | **   | LLM > No-LLM       |
|                                                                                  | Labels | 2.21    | 0.152          | 0.10       | n.s. | —                  |
| <i>Trust &amp; Decision Confidence</i>                                           |        |         |                |            |      |                    |
| Trust                                                                            | LLM    | 0.21    | 0.648          | 0.010      | n.s. | —                  |
|                                                                                  | Labels | 2.67    | 0.118          | 0.11       | n.s. | —                  |
| Confidence                                                                       | LLM    | 3.12    | 0.092          | 0.129      | n.s. | —                  |
|                                                                                  | Labels | 7.26    | <b>0.014</b>   | 0.26       | *    | Labels > No-Labels |
| <i>Cognitive Load (NASA-TLX)</i>                                                 |        |         |                |            |      |                    |
| Mental demand                                                                    | LLM    | 8.10    | <b>0.010</b>   | 0.28       | *    | No-LLM > LLM       |
| Frustration                                                                      | Labels | 4.58    | <b>0.044</b>   | 0.18       | *    | No-Labels > Labels |
| Physical demand, temporal demand, performance, effort: all n.s. for both factors |        |         |                |            |      |                    |
| <i>User Experience (UEQ-S)</i>                                                   |        |         |                |            |      |                    |
| Hedonic quality                                                                  | LLM    | 16.05   | < <b>0.001</b> | 0.43       | ***  | LLM > No-LLM       |
| Pragmatic quality                                                                | LLM    | 18.53   | < <b>0.001</b> | 0.47       | ***  | LLM > No-LLM       |
| Overall Usability                                                                | LLM    | 29.85   | < <b>0.001</b> | 0.59       | ***  | LLM > No-LLM       |
| Labels factor: all UEQ-S subscales n.s.                                          |        |         |                |            |      |                    |

\*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$

## 6. Discussion

### 6.1. Interpretation of Results

#### 6.1.1. H1: LLM Support (Supported for Correctness, Cognitive Load, and User Experience)

The large and statistically reliable benefit of LLM assistance on diagnostic correctness ( $F(1, 21) = 11.08$ ,  $p = 0.003$ ,  $\eta_p^2 = 0.345$ ) supports Hypothesis H1 for the primary outcome. The effect size exceeds the conventional threshold for a large effect (Cohen, 1988), indicating that the availability of LLM assistance markedly raised the proportion of correct diagnoses (see Figure 5.1). The LLM system (a RAG pipeline built on Groq’s hosted llama-3.1-8b-instant model with domain-relevant retrieval documents) was able to provide contextually appropriate guidance that participants could use to refine or correct their initial fault hypotheses, and the spoken interaction modality removed any requirement for keyboard input in VR, lowering the barrier to querying the assistant.

A design decision central to interpreting this effect is that the correct diagnosis for each scenario was supplied to the LLM as an internal background note in the system prompt. This choice is easily misread as the model simply “being handed the answer.” The injection was introduced deliberately as a *hallucination guardrail*: its purpose was to keep the assistant grounded so that it could not invent non-existent faults or actively send participants down a wrong track. The system was not built to recite the answer on demand. It was prompted to give the correct diagnosis *when explicitly asked*, but otherwise to explain its reasoning, to respond to follow-up and off-topic questions, and to engage in dialogue rather than to issue a verdict. Participants were also told during the briefing that the LLM could be wrong, even if it would usually be right, so they were not instructed to defer to it unconditionally. The correctness gain therefore reflects *grounded, reliable assistance that participants still had to elicit, interpret, and choose to accept*, not passive transcription of a known answer. The ecological cost of this design (a real deployment would require an automated fault-labeling pipeline to achieve equivalent grounding) is discussed as a limitation in Section 6.4.4.

How participants made use of the assistant varied considerably from person to person, and these interaction strategies are themselves an informative result. Some participants queried

the LLM immediately (“what do you make of the problems here?”) and then worked to reconstruct and verify its reasoning. Others reasoned first on their own and consulted the LLM only afterwards, frequently by naming the two fault categories they were torn between and asking the assistant to adjudicate. Several participants actively *debated* with the LLM, and were observed to end up deciding either for or against its suggestion. The intensity of use also differed markedly: some leaned on the assistant heavily, others sparingly, apparently as a matter of individual disposition. Part of this variance may stem from a residual inhibition threshold around speaking aloud to a virtual assistant inside an unfamiliar VR study setting; the option to interact in either German or English was provided specifically to lower this barrier and made the assistant more accessible to participants who were less confident in English.

The LLM’s benefit came at a cost in task completion time: LLM conditions took significantly longer than the corresponding No-LLM conditions ( $F(1, 21) = 8.68$ ,  $p = 0.008$ ,  $\eta_p^2 = 0.292$ ), the opposite of the direction predicted in H1. This is a sensible speed–accuracy trade-off. The additional time reflects both the interactive latency of the voice pipeline (formulating a spoken query, waiting for the Groq response and Cartesia TTS synthesis, and listening to the answer) and, just as importantly, the verification behavior described above: many participants did not accept the assistant’s output at face value but cross-checked it against the sensor readings or solicited a second opinion before committing. The magnitude of the time cost suggests that latency reductions, for instance through on-device inference or streaming responses, would make the system more competitive for time-critical maintenance work.

LLM assistance also significantly reduced perceived mental demand ( $F(1, 21) = 8.10$ ,  $p = 0.010$ ,  $\eta_p^2 = 0.28$ ). This is consistent with the interpretation that offloading part of the diagnostic reasoning to the assistant freed cognitive resources that would otherwise be spent integrating and interpreting sensor readings. No significant LLM effect emerged for the remaining five NASA-TLX subscales (all  $p > 0.10$ ), indicating that the cognitive benefit was specific to the mental-load dimension rather than a general workload reduction.

The absence of a significant effect on trust ( $p = 0.648$ ) is noteworthy, and admittedly ran counter to expectation: intuitively, a system that demonstrably improves correctness might be expected to earn higher trust. Two complementary explanations reconcile the null result. First, the assistant occasionally caused confusion — one participant (VP10) reported that it suggested diagnoses that did not exist and remained stubborn when challenged — and such episodes can dampen trust even when average performance is good. Second, and more constructively, the behavioral pattern is consistent with *calibrated reliance* rather than naive trust: participants used the assistant’s output as one input among several, verified it, and sometimes overruled it, which is precisely the appropriate-reliance profile that the automation-trust literature distinguishes from automation bias (Lee & See, 2004). Trust attitudes are also known to develop gradually through repeated interaction and are unlikely

to shift substantially within a single session (Lee & See, 2004); the explicit briefing that the LLM could err further encouraged measured rather than uniform reliance. The moderate, stable trust ratings should therefore not be read as a failure of the assistant but as evidence of a healthy, skeptical engagement with it.

The very large UEQ-S effect ( $\eta_p^2 = 0.59$ ,  $p < 0.001$ ) shows that the LLM voice assistant substantially improved participants' subjective user experience, across both pragmatic quality ( $\eta_p^2 = 0.47$ ) and hedonic quality ( $\eta_p^2 = 0.43$ ). This has direct relevance to technology adoption: a system that users find more engaging and better-designed is more likely to be accepted in practice, independent of its objective correctness benefit. The avatar-based interaction, with real-time lip synchronization and a realistic appearance, plausibly contributed to the hedonic advantage by making the interface feel more natural and socially engaging, in line with evidence that embodied agents with real-time lip synchronization (Peixoto et al., 2021) heighten social presence and perceived interaction quality in VR (Garau et al., 2003; Lim et al., 2025).

### 6.1.2. H2: Multimodal Visualization (Not Supported)

The absence of a significant effect of multimodal VR visualization on diagnostic correctness ( $p = 0.887$ ,  $\eta_p^2 = 0.001$ ) is the most striking finding of this study. Hypothesis H2 predicted that the temperature heatmap, the ghost-robot vibration overlay, and haptic controller feedback would reduce extraneous cognitive load and enable faster, more correct, and more confident fault diagnosis than text labels alone. This prediction was informed by Cognitive Load Theory (Sweller, 1988) and the principle of natural mapping in multimodal data representation (Norman, 2013; Ware, 2012). The data do not support this mechanism for the present implementation and participant population: the multimodal interface was neither faster nor more correct, and on the two subjective measures where the representation factor did reach significance, it performed *worse* than the label interface.

The most plausible explanation for the null result is *unfamiliarity* with the novel form of representation rather than a lack of domain knowledge. All participants received a detailed, structured introduction in which both the label interface and the full multimodal layer were explained and demonstrated before the counterbalanced conditions began, so a basic understanding of what each cue encoded was established for everyone. What could not be established in a short briefing was *fluency* in reading a previously unseen visual language. The heatmap color gradients and the amplitude-amplified vibration overlay are unconventional encodings, and interpreting them quickly enough to support a diagnostic decision is a skill that develops with exposure. A clear, if observational, indication of this came from the study sessions themselves: the longer participants worked with the multimodal representation, the more confidently and accurately they appeared to interpret it. This learning trajectory is not

directly quantified in the present data (the study was not designed to isolate a within-session learning curve), but it was a consistent first-hand observation across sessions and aligns with the participants’ own comments (see Section 5.6.2). Several participants asked clarifying questions about the visualization and, in particular, about what the color *transitions* signified, especially during their first encounter with the multimodal display. The implication is that the multimodal interface carries a *novelty overhead*: cognitive resources that participants had to invest in decoding the unfamiliar display partly counteracted any extraneous-load reduction the encoding might otherwise have provided.

This interpretation is reinforced, and made internally consistent, by the two significant representation effects, which both point against the multimodal interface. First, decision confidence was significantly *lower* in the multimodal conditions ( $F(1, 21) = 7.26$ ,  $p = 0.014$ ,  $\eta_p^2 = 0.26$ ): numerical labels give participants a concrete, familiar anchor (“the value is X”), which feels unambiguous, whereas the perceptual cues left participants less certain about how to translate what they saw into a fault category. Second, and consistently, frustration was significantly *higher* in the multimodal conditions ( $F(1, 21) = 4.58$ ,  $p = 0.044$ ,  $\eta_p^2 = 0.18$ ): the effort of interpreting an unfamiliar encoding under task pressure was experienced as more frustrating than reading off explicit numbers. Far from being a paradox, the two effects tell one coherent story: for this non-expert, first-exposure sample, the multimodal layer was simultaneously harder to interpret (more frustrating) and less reassuring (lower confidence), without buying any objective correctness or time advantage in return.

A secondary contributing factor lies in the design of the fault scenarios themselves. The scenarios were deliberately realistic but, in several cases, perceptually similar to one another, and the no-fault (“no-error”) scenario in particular proved difficult for participants to recognize with confidence. When the underlying signal is subtle or ambiguous, a continuous heatmap gradient offers little discriminatory advantage over a direct numerical reading, and may even introduce additional interpretive uncertainty. This scenario similarity likely suppressed any latent benefit of the multimodal encoding and is revisited as a limitation in Section 6.4.3.

The preference and helpfulness data corroborate this account at the participant level. No participant chose the multimodal No-LLM condition as their preferred option, 17 of 22 participants (77 %) preferred the label interface combined with the LLM, and 16 of 22 rated that same Labels + LLM condition as the most helpful (Section 5.6.1). Participants, in other words, did not perceive the richer visualization as more useful than the explicit numerical interface.

Taken together, the null H2 result does not imply that multimodal VR visualization is generally ineffective for maintenance tasks. Rather, it suggests that its benefit is conditional on *familiarity*: the perceptual cues must become a fluent visual language before they can translate into faster or more correct diagnosis. Whether trained maintenance technicians

would benefit is, on the present evidence, genuinely open. Two opposite outcomes are conceivable depending on the technicians' openness to new technology: a skeptical practitioner may reject both the LLM and the unfamiliar visualization, whereas a technology-open expert, already fluent in reading sensor signatures, might extract real value from the multimodal encoding. This remains speculation that only a dedicated expert study could resolve (cf. Section 6.5).

### 6.1.3. Interaction Effects and Additivity

The consistent absence of significant Visualization  $\times$  LLM interaction effects across all dependent variables indicates that the two design dimensions operated independently. The LLM correctness benefit was equally present in the label-only interface and in the multimodal interface, and the null visualization effect on correctness was replicated under both LLM conditions.

This additive pattern has a direct implication for system design: the two components can be evaluated, validated, and deployed independently. There is no evidence of synergy (the combination of both does not unlock benefits beyond the sum of its parts) nor of interference, where one component would undermine the other. On the present evidence, a practitioner seeking to improve diagnostic correctness in a VR digital twin context should prioritize LLM integration, which is the component carrying the measurable benefit. This recommendation is not an argument to abandon multimodal visualization: as argued in Section 6.1.2, its lack of measurable benefit here is plausibly tied to unfamiliarity rather than to an intrinsic limitation, and a more mature visualization paired with better onboarding and example scenarios in a larger study may well perform differently. The honest reading of the current data, however, is that the LLM is the primary driver and the multimodal layer is, for this population and this first exposure, not beneficial.

It is also possible that a true synergistic interaction exists but was not detectable at  $N = 22$ , for instance if the multimodal display and the spoken LLM interaction compete for attentional resources simultaneously. With all interaction effects non-significant in the present data, however, this remains a conjecture that a larger sample would be needed to confirm or rule out.

## 6.2. Comparison with Related Work

The diagnostic correctness observed in the No-LLM conditions of this study was relatively low in absolute terms (see Figure 5.1). It should be read against the chance level, however: with nine diagnostic categories, random guessing would yield only about 11 % correct, so the

No-LLM correctness already lies well above chance, and LLM assistance raised it further. A direct comparison with the VR/AR maintenance literature is difficult, however, because that work evaluates procedural *training* rather than open-ended diagnosis, and its performance effects are strongly modality- and task-dependent. Gavish et al. (2015), for instance, found that augmented-reality training reduced errors on an industrial assembly task relative to a conventional baseline, whereas the virtual-reality condition did not show a comparable advantage, a reminder that immersive technology does not uniformly improve task performance. Substantial differences in task complexity, participant expertise, and what counts as a “correct” response therefore make cross-study accuracy comparisons unreliable. The fault categorization task used here required selecting one of nine error types with no pre-filtering, making it inherently more demanding than the binary go/no-go judgments common in the maintenance literature. With LLM support, correctness rose substantially (Section 5.1), consistent with the general finding that well-designed advisory decision-support tools can improve operator performance (Sheridan & Parasuraman, 2005).

The LLM correctness benefit observed here ( $\eta_p^2 = 0.345$ ) is large in absolute terms and exceeds the more modest effects often reported for decision-aid and automation-support interventions in the human-factors literature. As discussed above, part of this magnitude is attributable to the grounded RAG design, in which the assistant was kept reliable through an injected diagnostic hint; a purely retrieval-based system without such grounding would likely produce a smaller, though plausibly still meaningful, effect.

The null visualization result stands in partial contrast to studies reporting accuracy improvements with augmented-reality overlays in industrial maintenance (Palmarini et al., 2018). A key difference is that AR maintenance studies typically target *procedural* task guidance (step-by-step instruction overlays), where spatial alignment of the overlay with the physical component reduces search time and errors. The task in this study is *diagnostic* rather than procedural: participants must infer a hidden fault category from sensor readings, not execute a visually cued procedure. The natural-mapping advantage of multimodal visualization may simply be stronger for procedural guidance than for diagnostic inference, which depends more on interpretation and reasoning (and, as this study suggests, on familiarity with the encoding) than on perceptual search efficiency.

## 6.3. Practical Implications

### 6.3.1. Design Recommendations for VR Maintenance Systems

Several evidence-based recommendations emerge for practitioners developing VR digital twins for industrial maintenance applications.

**Prioritize LLM voice integration as the primary correctness lever.** The LLM assistant produced a large correctness gain and a very large user-experience improvement across both interface conditions. Given that the multimodal layer showed no significant correctness benefit in the present study, the development effort most likely to improve diagnostic outcomes is the integration of a reliable, low-latency conversational assistant. Future systems should prioritize reducing voice-interaction latency (e.g., through on-device inference or streaming generation) and improving speech-recognition robustness in noisy industrial environments.

**Treat multimodal visualization as a maturing component, not a failed one.** The null correctness result does not justify abandoning the multimodal interface. Its weak showing here is plausibly an artifact of unfamiliarity and first exposure rather than an intrinsic shortcoming. The interface should be refined and given the chance to be *learned*: a dedicated onboarding with worked example scenarios, clearer encoding of state transitions (for instance an explicit gradient or direction indicator, as one participant proposed), and a larger study that allows familiarity to build would all give the visualization a fairer test. Designers should expect its value to depend strongly on how accustomed users are to the representation.

**Include a familiarization phase.** No formal warm-up scenario was included in the present study, and observation suggested that participants took longer to understand the interface in their first condition. Future implementations should incorporate at least one unscored practice scenario per interface modality, with experimenter feedback, so that learning the interface does not compete with the diagnostic task itself.

**Calibrate trust communication.** Despite large correctness and UEQ-S gains, trust did not increase with LLM availability. Rather than treating this as a problem to be “fixed” by inflating trust, systems should aim for *appropriate* reliance: communicating LLM uncertainty explicitly, for example by framing spoken responses with confidence qualifiers (“based on the sensor data this is most consistent with X, though I’d verify joint 2 temperature”), helps users build well-calibrated trust rather than either uniform deference or blanket skepticism.

### 6.3.2. Guidelines for LLM Integration in Industrial HCI

The RAG-based architecture employed here illustrates a practical pattern for integrating language models into safety-relevant industrial contexts. Several guidelines follow from the implementation and results, extending the general human-AI design principles of Amershi et al. (2019) to this domain.

**Ground the model to suppress hallucination.** Combining domain-specific retrieval documents (robot maintenance manuals and custom fault-description sheets) with in-context scenario information kept responses anchored to the fault types actually present and markedly reduced the risk of invented diagnoses. The diagnostic-hint injection used in this study is an

aggressive form of such grounding; in a real deployment, the equivalent role would be played by a fault-labeling pipeline fed from physical sensor data. Practitioners should maintain curated, regularly updated retrieval corpora and treat grounding as a first-class reliability mechanism rather than an afterthought.

**Constrain output length and temperature.** The LLM was configured with a maximum output length of 120 tokens and a temperature of 0.3, producing concise, comparatively deterministic responses suitable for spoken delivery. Verbose answers increase listening time and cognitive load; shorter responses preserve time efficiency and hold the user’s attention.

**Plan for graceful degradation.** The cloud-based API architecture is vulnerable to connectivity loss; isolated backend-connection drops were observed during data collection, though none were severe. Industrial deployments should provide local fallback models or cached response libraries for common queries so that the system remains partially functional when connectivity degrades.

**Support multiple languages.** Allowing users to interact in their preferred language (English or German here) lowered interaction friction and made the assistant accessible to participants less confident in English, a consideration of particular relevance for industrial workforces with diverse language backgrounds.

## 6.4. Limitations

### 6.4.1. Sample Size and Generalizability

The most significant limitations of this study concern its sample. With  $N = 22$  participants recruited from HAW Hamburg’s student and researcher community, the study relied on a convenience sample that is demographically unrepresentative of professional robot maintenance technicians. The mean age of 30.9 years and the exclusively academic background limit the generalizability of the findings to industrial practitioners, who bring hands-on fault-diagnosis experience that would qualitatively change the dynamics of both the visualization and the LLM effects. In particular, the null H2 finding might not replicate with trained technicians, for whom familiar sensor-cue patterns could carry stronger discriminatory information.

The sample size also bears on statistical power. It was sufficient to detect the large LLM correctness effect reliably, but the medium-range visualization effects on time and cognitive load ( $\eta_p^2 \approx 0.09$ ) fell below significance and may represent genuine but underpowered effects. A post-hoc estimate suggests that detecting a medium effect ( $f = 0.25$ ) at  $\alpha = 0.05$  and  $1 - \beta = 0.80$  would require roughly  $N = 28\text{--}34$  participants, depending on the correlation

structure of the repeated measures. A larger and more diverse sample is the single most important improvement for any replication.

### **6.4.2. No Familiarization Phase**

The second major limitation is the absence of a formal warm-up or practice scenario. While the experimenter-led briefing explained both interfaces and all fault types in detail, participants still had to learn the mechanics of each interface while simultaneously performing the cognitive task of fault diagnosis. This is especially consequential given the central role that unfamiliarity plays in interpreting the null H2 result (Section 6.1.2): without a practice phase, the multimodal interface was effectively evaluated at the steepest point of its learning curve, which likely suppressed its measured performance and inflated the relative standing of the already-familiar label interface. The first-condition disadvantage may also have interacted with learning effects across the session. A future replication should include at least one unscored warm-up scenario per interface modality before the counterbalanced conditions begin.

### **6.4.3. Ecological Validity and Scenario Design**

The controlled laboratory setting introduced several ecological-validity limitations. Real maintenance decisions are made under time pressure, in noisy conditions, and with tangible consequences for downtime and worker safety, factors absent here. Participants knew their decisions carried no real-world consequences, which may have influenced their risk tolerance and decision strategies. The study also used pre-defined sensor trajectories generated by the digital twin rather than live feeds from a physical robot; this is a comparatively mild limitation, however, since the simulated scenarios were designed to be realistic, even if idealized and free of the noise and cross-coupling artifacts of real hardware.

A more consequential issue concerns the scenarios themselves. Several scenarios were perceptually similar to one another, and the no-fault (“no-error”) scenario in particular proved difficult for participants to recognize with confidence. This similarity made some diagnoses genuinely ambiguous and likely added noise to the correctness measure, while also blunting any discriminatory advantage that the multimodal encoding might otherwise have offered. Future studies should construct scenarios with more clearly separable signatures, or explicitly balance scenario difficulty across conditions.

Relatedly, because each participant completed twelve trials drawn from only nine distinct scenarios (Section 3.2.4), individual scenarios recurred across conditions. A participant who learned a scenario’s signature in an early condition could in principle recognize it again

later, introducing a potential memory effect that is partially, but not fully, controlled by the Latin-square counterbalancing of condition order. This confound is most pronounced for the “Low Urgency” severity category, which contains only two scenarios (*No Error* and *Lubrication Needed*) and therefore recurs most often. Because such repetition is distributed across both LLM and No-LLM conditions by the counterbalancing, it is unlikely to fully explain the LLM main effect, but it cannot be ruled out as a contributing source of variance; future designs should use a larger, non-overlapping scenario pool.

Finally, the single-session design (approximately 60 to 90 minutes), while manageable, does not capture the fatigue that would accumulate over a full maintenance shift. Participant fatigue toward the end of the session is the most relevant such effect; extended VR use is known to increase sensory conflict and cognitive fatigue, which could interact with the effectiveness of both the visualization and the LLM assistant over longer horizons.

#### **6.4.4. LLM Variability and Design Choices**

Two further limitations are specific to the LLM. First, the stochastic nature of generation, even at a temperature of 0.3, means that different participants received different responses to semantically similar queries, introducing uncontrolled variance into the LLM conditions that is absent from the deterministic visualization conditions. The earlier-mentioned VP10 report, that the assistant “sometimes suggested diagnoses that did not exist” and “remained stubborn” when challenged, illustrates two concrete failure modes (hallucination and insufficient responsiveness to user evidence) that could erode correctness and trust in a real deployment. The hosted Groq model may additionally have been subject to silent updates over the data-collection period, potentially introducing response-quality drift; future studies should lock model versions and log all interactions for post-hoc analysis.

Second, the diagnostic-hint injection that grounded the assistant, while a deliberate and defensible reliability mechanism (Section 6.1.1), does limit ecological validity: the assistant had privileged access to ground truth that would not be available in a real deployment without a continuous fault-labeling pipeline driven by physical sensors. The reported correctness gain should therefore be read as an upper-bound estimate of what well-grounded LLM assistance can deliver, rather than as the figure a purely retrieval-based system would achieve in the field. Even with this grounding, participants still had to elicit, interpret, and decide whether to accept the guidance, so the effect is not reducible to the model simply stating the answer.

### 6.4.5. Trust Measurement and a Shifting Referent

A further measurement limitation concerns the trust scale. The TAS was administered after every condition, but the “system” it refers to is not the same object across conditions: in the No-LLM conditions it denotes the digital twin and its visualization alone, whereas in the LLM conditions it additionally encompasses the voice assistant. The trust ratings therefore compare attitudes toward two partly different referents, which complicates a clean attribution of the null trust effect to the LLM as such. The calibrated-reliance interpretation offered in Section 6.1.1 should thus be read alongside this caveat: part of the stability of the trust ratings may simply reflect that participants were rating a broader, shifting target rather than a single, well-defined system. Future studies should phrase the trust items to name the specific component under evaluation, or collect separate trust ratings for the visualization and the assistant.

## 6.5. Future Work

Several directions follow naturally from the findings and limitations of this study.

**Field study with domain experts.** The most immediate priority is a replication with professional maintenance technicians operating the system on a real or production-grade physical robot. Such a study would test directly whether the null visualization finding is an artifact of unfamiliarity, by giving expert users, already fluent in reading sensor signatures, a fair opportunity to benefit from the multimodal encoding, and would provide more ecologically valid correctness baselines. It would also reveal how technology-open versus technology-skeptical technicians differ in their uptake of both the LLM and the visualization.

**Familiarization and longitudinal designs.** Given the central role of familiarity in this study, a multi-session or longitudinal design with explicit training scenarios would clarify how quickly the multimodal interface becomes a fluent visual language and whether its performance catches up to or overtakes the label interface with practice. The same design would address the static trust finding by tracking how reliance patterns and decision correctness evolve over repeated LLM interactions, including whether users learn to query the assistant selectively for ambiguous cases and whether over-reliance develops over time.

**On-device and streaming LLM inference.** Reducing interaction latency is the most actionable engineering improvement identified here. Future work should evaluate on-device small language models deployed on the headset or a co-located edge device, comparing latency, correctness, and user experience against the cloud-based architecture used in this study.

**Adaptive fault-salience highlighting.** Rather than relying on participants to interpret continuous visualizations unaided, a future system could integrate ML-based anomaly detection to automatically highlight joints with anomalous readings, combining the perceptual richness of the multimodal display with explicit attentional guidance. Such guidance might resolve the null visualization result by directing attention to the most diagnostic cues, and would also reduce the onboarding burden identified above.

**Mixed-reality overlay on physical robots.** An AR variant that overlays the heatmap and vibration animation directly onto a physical robot arm would test whether the spatial alignment afforded by AR (as opposed to the side-by-side VR digital twin) produces correctness gains absent in the fully virtual condition.

**Domain-specific fine-tuning and hallucination mitigation.** Finally, replacing the general-purpose model with one fine-tuned on curated fault records from the specific robot platform, combined with structured output constraints (e.g., requiring responses to cite retrieved passages), could improve reliability and reduce the risk of confident but incorrect advice, moving the grounding mechanism from an injected hint toward a genuinely learned, deployable competence.

# 7. Conclusion

## 7.1. Summary of Contributions

This thesis set out to investigate whether combining multimodal data representation with LLM-supported analysis improves fault diagnosis in digital twins of industrial robots. Its central contribution is conceptual as much as empirical: it brings together three technologies that are rarely integrated in a single working system (immersive virtual reality, classical predictive maintenance on an industrial robot, and a large language model as an advanced decision-support layer) and evaluates the resulting combination in a controlled study. Three interrelated contributions can be distinguished.

**First, a novel integrated VR digital twin concept and prototype.** A Unity-based VR application was designed and implemented for the Meta Quest standalone headset, providing an interactive digital twin of a five-axis industrial robot arm sourced from the robotics laboratory of HAW Hamburg. The system unites a numerical text-label interface with a multimodal visualization layer (a temperature heatmap shader, a ghost-robot vibration overlay with amplitude amplification, and haptic controller feedback) and pairs both with a spoken LLM assistant. The LLM subsystem combines a Groq-hosted llama-3.1-8b-instant model with a RAG pipeline (VoyageAI embeddings, a ChromaDB vector store, and a Python FastAPI backend) and delivers responses through Cartesia text-to-speech via a realistic, lip-synchronized avatar. The demonstration that this combination runs on consumer-grade standalone hardware, without specialized infrastructure, is itself a contribution: it shows that the integration of immersive visualization, predictive maintenance, and conversational AI is technically feasible today.

**Second, empirical evidence from a controlled 2×2 within-subject study.** Twenty-two participants completed all four conditions in a single session, yielding data on diagnostic correctness, decision time, decision confidence, perceived cognitive load (NASA-TLX), system trust (TAS), and user experience (UEQ-S). To the author’s knowledge, this is the first controlled evaluation of the *joint* effects of immersive multimodal visualization and LLM voice assistance on fault diagnosis performance in a VR digital twin, and the factorial design makes it possible to attribute the observed benefits to the responsible component.

**Third, evidence-based design guidelines.** The results motivate concrete recommendations for future VR maintenance systems: prioritize LLM voice integration as the primary correctness lever, treat multimodal visualization as a maturing component that must be learned rather than as a failed one, include a familiarization phase, communicate LLM uncertainty to support calibrated rather than blind reliance, and architect for latency reduction and graceful degradation under connectivity constraints.

## 7.2. Answers to Research Questions

The central research question asked to what extent multimodal data representation and LLM-supported analysis *jointly* improve efficiency, diagnostic quality, and trust in fault diagnosis compared to isolated solutions. The results give a clear and differentiated answer, and they identify a concrete best-performing configuration. Across every objective and subjective measure, the strongest combination was *not* the fully immersive multimodal variant that the design had been built around, but the **text-label interface paired with the LLM assistant**: this configuration produced the highest diagnostic correctness, the lowest mental demand, the best user-experience ratings, and the highest decision confidence, and it was the variant participants overwhelmingly preferred (17 of 22) and rated most helpful (16 of 22). The central finding of this thesis is therefore not that the combination of technologies fails, but that one specific combination clearly wins, and that the decisive ingredient is the conversational LLM layer rather than the immersive visualization. The per-factor picture behind this result is as follows.

**H1 (Effect of LLM Assistance): Supported for correctness, mental demand, and user experience; not supported for trust or time.** LLM voice assistance produced a large, reliable improvement in diagnostic correctness ( $F(1, 21) = 11.08$ ,  $p = 0.003$ ,  $\eta_p^2 = 0.345$ ) and significantly reduced perceived mental demand ( $p = 0.010$ ). User-experience ratings improved by a very large margin ( $\eta_p^2 = 0.59$  for overall UX). These gains came at the cost of longer decision times ( $p = 0.008$ ), reflecting both voice-pipeline latency and the verification behavior many participants engaged in. Trust ratings were uniformly moderate and unaffected by any factor ( $p = 0.648$ ), a pattern best read not as a shortcoming but as evidence of calibrated reliance, since participants verified and sometimes overruled the assistant rather than deferring to it. The correctness figure should be understood as an upper bound enabled by a deliberately grounded (hallucination-guarded) assistant, which participants nonetheless still had to query, interpret, and choose to accept.

**H2 (Effect of Visualization Modality): Not supported.** The multimodal VR representation did not significantly improve diagnostic correctness, reduce task completion time, or lower perceived cognitive load compared to the text-label interface ( $p > 0.15$  for all three). The

two significant modality effects both ran *against* the multimodal interface: it yielded lower decision confidence ( $p = 0.014$ ,  $\eta_p^2 = 0.26$ ) and higher frustration ( $p = 0.044$ ,  $\eta_p^2 = 0.18$ ) than the label interface. The most plausible explanation is unfamiliarity rather than a lack of domain knowledge: all participants received a detailed introduction to both interfaces, but reading an unconventional visual encoding fluently is a skill that develops with exposure, and the study captured the multimodal interface at the steepest part of its learning curve. The finding therefore does not negate the value of immersive visualization in principle; it indicates that its benefit is conditional on familiarity and is unlikely to surface on first exposure in a non-expert sample.

**Interaction: Additive, not synergistic.** No significant Visualization  $\times$  LLM interaction emerged for any dependent variable. The two design dimensions proved *additive* rather than synergistic: the LLM benefit replicated equally under both interfaces, and the visualization effect held under both LLM conditions. This additivity is itself a constructive result, because it means the two components can be developed, validated, and deployed independently, and that the winning Labels + LLM configuration owes its advantage to a single, well-identified driver. The measurable improvement in diagnostic quality and user experience was carried almost entirely by the LLM; the multimodal layer added no performance benefit (and a small subjective cost) for this first-exposure, non-expert population, while introducing no negative interference. The combination therefore remains a viable and interference-free target for further development, but on the present evidence it is the LLM that makes the difference, and a more mature, *learned* visualization layer is the clearest opportunity to widen the margin further.

### 7.3. Final Remarks

If a single message is to be taken from this thesis, it is that LLM-based decision support holds substantial potential for improving fault diagnosis in industrial maintenance, particularly for non-expert operators, who in this study were able to reach markedly better decisions with conversational access to grounded domain knowledge than without it. Newer immersive visualization techniques, by contrast, do not provide an immediate benefit: they are valuable in principle but depend on familiarity, and their value must be established through onboarding, practice, and further study rather than assumed. More sensory richness does not automatically translate into better decisions, and effective human-machine interfaces must be designed for the actual perceptual and cognitive state of their intended users, not for an idealized expert who reads novel encodings effortlessly.

These findings sit within a broader and increasingly relevant context. Predictive maintenance is gaining importance both economically and as a contribution to sustainability: in the

narrow sense of extending equipment lifetime and reducing unplanned downtime, and in the wider sense of using resources more responsibly. Tools that make complex robotic systems easier to diagnose, and that lower the expertise barrier to maintaining them, therefore have a role to play beyond mere convenience. Without overstating the case, the integration demonstrated here, retrieval-augmented reasoning delivered through an embodied avatar inside an immersive digital twin, is a plausible building block for that future, and one whose central component, the LLM assistant, already delivers measurable value today. As language models become smaller, faster, and more domain-adaptable, and as the latency constraints identified in this work recede, the case for such AI-augmented digital twins in industrial maintenance is likely only to strengthen. This thesis offers both an empirical foundation and a concrete implementation reference from which that development can continue.

# Bibliography

- Afzal, U., Prouzeau, A., Lawrence, L., Dwyer, T., Bichinepally, S., Liebman, A., & Goodwin, S. (2022). Investigating cognitive load in energy network control rooms: Recommendations for future designs. *Frontiers in Psychology*, 13, 812677. <https://doi.org/10.3389/fpsyg.2022.812677>
- Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13. <https://doi.org/10.1145/3290605.3300233>
- Bickmore, T. W., & Picard, R. W. (2005). Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction*, 12(2), 293–327. <https://doi.org/10.1145/1067860.1067867>
- Bommasani, R., et al. (2021). On the opportunities and risks of foundation models. <https://arxiv.org/abs/2108.07258>
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., et al. (2020). Language models are few-shot learners. <https://arxiv.org/abs/2005.14165>
- Coban, M., Bolat, Y. I., & Goksu, I. (2022). The potential of immersive virtual reality to enhance learning: A meta-analysis. *Educational Research Review*, 36, 100452. <https://doi.org/10.1016/j.edurev.2022.100452>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.
- Fuller, A., Fan, Z., Day, C., & Barlow, C. (2020). Digital twin: Enabling technologies, challenges and open research. *IEEE Access*, 8, 108952–108971. <https://doi.org/10.1109/ACCESS.2020.2998358>
- Garau, M., Slater, M., Vinayagamoorthy, V., Brogni, A., Steed, A., & Sasse, M. A. (2003). The impact of avatar realism and eye gaze control on perceived quality of communication in a shared immersive virtual environment. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 529–536. <https://doi.org/10.1145/642611.642703>
- Gavish, N., Gutiérrez, T., Webel, S., Rodríguez, J., Peveri, M., Bockholt, U., & Tecchia, F. (2015). Evaluating virtual reality and augmented reality training for industrial maintenance and assembly tasks. *Interactive Learning Environments*, 23(6), 778–798. <https://doi.org/10.1080/10494820.2013.815221>

- Grieves, M. (2014). *Digital twin: Manufacturing excellence through virtual factory replication* (White Paper). Florida Institute of Technology.
- Grieves, M., & Vickers, J. (2017). Digital twin: Mitigating unpredictable, undesirable emergent behavior in complex systems. In F.-J. Kahlen, S. Flumerfelt, & A. Alves (Eds.), *Transdisciplinary perspectives on complex systems* (pp. 85–113). Springer. [https://doi.org/10.1007/978-3-319-38756-7\\_4](https://doi.org/10.1007/978-3-319-38756-7_4)
- Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., De Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*, 53(5), 517–527. <https://doi.org/10.1177/0018720811417254>
- Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Advances in psychology* (pp. 139–183, Vol. 52). North-Holland. [https://doi.org/10.1016/S0166-4115\(08\)62386-9](https://doi.org/10.1016/S0166-4115(08)62386-9)
- Ismail, L., Abdelmoti, A., Basu, A., Berini, A. D. E., & Naouss, M. (2025). *A systematic review of digital twin-driven predictive maintenance in industrial engineering: Taxonomy, architectural elements, and future research directions*. arXiv: 2509.24443 [cs.SE]. <https://arxiv.org/abs/2509.24443>
- Jardine, A. K. S., Lin, D., & Banjevic, D. (2006). A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mechanical Systems and Signal Processing*, 20(7), 1483–1510. <https://doi.org/10.1016/j.ymssp.2005.09.012>
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), 1–38. <https://doi.org/10.1145/3571730>
- Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, 4(1), 53–71. [https://doi.org/10.1207/S15327566IJCE0401\\_04](https://doi.org/10.1207/S15327566IJCE0401_04)
- Kritzinger, W., Karner, M., Traar, G., Henjes, J., & Sihn, W. (2018). Digital twin in manufacturing: A categorical literature review and classification. *IFAC-PapersOnLine*, 51(11), 1016–1022. <https://doi.org/10.1016/j.ifacol.2018.08.474>
- Lasi, H., Fettke, P., Kemper, H.-G., Feld, T., & Hoffmann, M. (2014). Industry 4.0. *Business & Information Systems Engineering*, 6(4), 239–242. <https://doi.org/10.1007/s12599-014-0334-4>
- Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1), 50–80. [https://doi.org/10.1518/hfes.46.1.50\\_30392](https://doi.org/10.1518/hfes.46.1.50_30392)
- Lei, Y., Liu, H., Li, N., Cao, J., Qiao, Y., & Wang, H. (2025). Condition monitoring and fault diagnosis of industrial robots: A review. *Science China Technological Sciences*, 68(1), 1110301. <https://doi.org/10.1007/s11431-024-2810-2>
- Lei, Y., Yang, B., Jiang, X., Jia, F., Li, N., & Nandi, A. K. (2020). Applications of machine learning to machine fault diagnosis: A review and roadmap. *Mechanical Systems and Signal Processing*, 138, 106587. <https://doi.org/10.1016/j.ymssp.2019.106587>

- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474. <https://arxiv.org/abs/2005.11401>
- Lim, S., Schmälzle, R., & Bente, G. (2025). Artificial social influence via human-embodied AI agent interaction in immersive virtual reality (VR): Effects of similarity-matching during health conversations. *Computers in Human Behavior: Artificial Humans*, 5, 100172. <https://doi.org/10.1016/j.chbah.2025.100172>
- Lu, Y. (2017). Cyber physical system (CPS)-based industry 4.0: A survey. *Journal of Industrial Integration and Management*, 2(3), 1750014. <https://doi.org/10.1142/S2424862217500142>
- Makransky, G., & Lilleholt, L. (2018). A structural equation modeling investigation of the emotional value of immersive virtual reality in education. *Educational Technology Research and Development*, 66(5), 1141–1164. <https://doi.org/10.1007/s11423-018-9581-2>
- Mavridis, N. (2015). A review of verbal and non-verbal human–robot interactive communication. *Robotics and Autonomous Systems*, 63, 22–35. <https://doi.org/10.1016/j.robot.2014.09.031>
- Milgram, P., & Kishino, F. (1994). A taxonomy of mixed reality visual displays. *IEICE Transactions on Information and Systems*, E77-D(12), 1321–1329.
- Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56(1), 81–103. <https://doi.org/10.1111/0022-4537.00153>
- Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 72–78. <https://doi.org/10.1145/191666.191703>
- Nee, A. Y. C., Ong, S. K., Chryssolouris, G., & Mourtzis, D. (2012). Augmented reality applications in design and manufacturing. *CIRP Annals*, 61(2), 657–679. <https://doi.org/10.1016/j.cirp.2012.05.010>
- Norman, D. A. (2013). *The design of everyday things* (Revised and Expanded). Basic Books.
- Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654), 187–192. <https://doi.org/10.1126/science.adh2586>
- OpenAI. (2023). GPT-4 technical report. *arXiv preprint arXiv:2303.08774*. <https://arxiv.org/abs/2303.08774>
- Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist*, 38(1), 1–4. [https://doi.org/10.1207/S15326985EP3801\\_1](https://doi.org/10.1207/S15326985EP3801_1)
- Palmarini, R., Erkoyuncu, J. A., Roy, R., & Torabmostaedi, H. (2018). A systematic review of augmented reality applications in maintenance. *Robotics and Computer-Integrated Manufacturing*, 49, 215–228. <https://doi.org/10.1016/j.rcim.2017.06.002>
- Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2), 230–253. <https://doi.org/10.1518/001872097778543886>

- Peixoto, B., Melo, M., Cabral, L., & Bessa, M. (2021). Evaluation of animation and lip-sync of avatars, and user interaction in immersive virtual reality learning environments. *2021 International Conference on Graphics and Interaction (ICGI)*. <https://doi.org/10.1109/ICGI54032.2021.9655283>
- Qin, Q., Liu, Z., Zhong, R., Wang, X. V., Wang, L., Wiktorsson, M., & Wang, W. (2026). Robot digital twin systems in manufacturing: Technologies, applications, trends and challenges. *Robotics and Computer-Integrated Manufacturing*, 97, 103103. <https://doi.org/10.1016/j.rcim.2025.103103>
- Radianti, J., Majchrzak, T. A., Fromm, J., & Wohlgenannt, I. (2020). A systematic review of immersive virtual reality applications for higher education: Design elements, lessons learned, and research agenda. *Computers & Education*, 147, 103778. <https://doi.org/10.1016/j.compedu.2019.103778>
- Schrepp, M., Hinderks, A., & Thomaschewski, J. (2017). Design and evaluation of a short version of the user experience questionnaire (UEQ-S). *International Journal of Interactive Multimedia and Artificial Intelligence*, 4(6), 103–108. <https://doi.org/10.9781/ijimai.2017.09.001>
- Schwab, K. (2016). *The fourth industrial revolution*. World Economic Forum.
- Shafto, M., Conroy, M., Doyle, R., Glaessgen, E., Kemp, C., LeMoigne, J., & Wang, L. (2010). *Modeling, simulation, information technology & processing roadmap: Technology area 11* (Technology Roadmap). National Aeronautics and Space Administration (NASA).
- Sheridan, T. B., & Parasuraman, R. (2005). Human-automation interaction. *Reviews of Human Factors and Ergonomics*, 1(1), 89–129. <https://doi.org/10.1518/155723405783703082>
- Slamani, M., Nubiola, A., & Bonev, I. A. (2012). Modeling and assessment of the backlash error of an industrial robot. *Robotica*, 30(7), 1167–1175. <https://doi.org/10.1017/S0263574711001287>
- Slater, M. (2003). A note on presence terminology. *Presence Connect*, 3(3), 1–5.
- Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2), 257–285. [https://doi.org/10.1207/s15516709cog1202\\_4](https://doi.org/10.1207/s15516709cog1202_4)
- Tao, F., Zhang, M., Liu, Y., & Nee, A. Y. C. (2018). Digital twin driven prognostics and health management for complex equipment. *CIRP Annals*, 67(1), 169–172. <https://doi.org/10.1016/j.cirp.2018.04.055>
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). LLaMA: Open and efficient foundation language models. <https://arxiv.org/abs/2302.13971>
- Turk, M. (2014). Multimodal interaction: A review. *Pattern Recognition Letters*, 36, 189–195. <https://doi.org/10.1016/j.patrec.2013.07.003>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
- Ware, C. (2012). *Information visualization: Perception for design* (3rd ed.). Morgan Kaufmann.

- White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with ChatGPT. <https://arxiv.org/abs/2302.11382>
- Witmer, B. G., & Singer, M. J. (1998). Measuring presence in virtual environments: A presence questionnaire. *Presence: Teleoperators and Virtual Environments*, 7(3), 225–240. <https://doi.org/10.1162/105474698565686>
- Zhang, Y., Wu, J., Gao, B., Xia, L., Lu, C., Wang, H., & Cao, G. (2025). Fault types and diagnostic methods of manipulator robots: A review. *Sensors*, 25(6), 1716. <https://doi.org/10.3390/s25061716>

# A. Questionnaire Materials

## NASA-TLX

The NASA Task Load Index (Hart & Staveland, 1988) was administered after each experimental condition in its raw (unweighted) form. Each of the six subscales was rated on a 21-point scale (1 = very low, 21 = very high); the pairwise weighting procedure of the original instrument was omitted, in line with common practice in recent human-factors research. The six items, with their endpoint anchors, were presented as follows:

**Mental Demand** How mentally demanding was the task? (Very Low – Very High)

**Physical Demand** How physically demanding was the task? (Very Low – Very High)

**Temporal Demand** How hurried or rushed was the pace of the task? (Very Low – Very High)

**Performance** How successful were you in accomplishing what you were asked to do? (Very Good – Very Poor)

**Effort** How hard did you have to work to accomplish your level of performance? (Very Low – Very High)

**Frustration** How insecure, discouraged, irritated, stressed, and annoyed were you? (Very Low – Very High)

## TAS – Trust in Automated Systems Scale

Trust was measured using the Trust in Automated Systems scale (Jian et al., 2000). Each of the twelve items was rated on a 7-point scale (1 = not at all, 7 = extremely). Items 1–5 capture distrust and are reverse-scored; items 6–12 capture trust. In the survey, the word “system” referred to the digital twin together with, where present, the LLM assistant.

1. The system is deceptive.
2. The system behaves in an underhanded manner.

3. I am suspicious of the system's intent, action, or outputs.
4. I am wary of the system.
5. The system's action will have a harmful or injurious outcome.
6. I am confident in the system.
7. The system provides security.
8. The system has integrity.
9. The system is dependable.
10. The system is reliable.
11. I can trust the system.
12. I am familiar with the system.

## User Experience Questionnaire Short (UEQ-S)

User experience was assessed with the short version of the User Experience Questionnaire (Schrepp et al., 2017). The instrument consists of eight items, each a 7-point semantic differential between two opposing terms. The first four items form the *pragmatic quality* subscale, the latter four the *hedonic quality* subscale; the mean of all eight yields an overall user-experience score.

| Item | Left anchor     | Right anchor | Subscale  |
|------|-----------------|--------------|-----------|
| 1    | obstructive     | supportive   | Pragmatic |
| 2    | complicated     | easy         | Pragmatic |
| 3    | inefficient     | efficient    | Pragmatic |
| 4    | confusing       | clear        | Pragmatic |
| 5    | boring          | exciting     | Hedonic   |
| 6    | not interesting | interesting  | Hedonic   |
| 7    | conventional    | inventive    | Hedonic   |
| 8    | usual           | leading edge | Hedonic   |

## Decision Confidence Item

Decision confidence was captured with a single item presented after each scenario, rated on a 7-point Likert scale:

*How confident are you in your diagnosis?*

(1 = not at all confident, 7 = extremely confident)

The three per-scenario ratings collected within each condition were averaged to yield one confidence score per condition (see Section 4.5.1).

## B. Informed Consent Form

The following information sheet and consent form was presented to and signed by every participant before any data were collected. It is reproduced here verbatim in the original German wording in which it was administered.

## Study: Immersive Data Visualization and LLM-Supported Decision-Making in Digital Twins for Predictive Maintenance of Industrial Robots

HAW Hamburg, FTZ Digital Reality

## Informationen und Einwilligungserklärung für Studienteilnehmende

**Studienleitung:**

Fakultät Informatik und Digitale Gesellschaft

Moritz Treu

## I. Inhalte und Daten

Die Studie untersucht, wie Virtual Reality (VR), digitale Zwillinge und KI-gestützte Assistenzsysteme die Entscheidungsfindung bei der vorausschauenden Wartung (Predictive Maintenance) von Industrierobotern unterstützen können. Dazu werden Sie ein VR Head-Mounted Display (HMD) aufsetzen und in einer simulierten Industrieumgebung Wartungsentscheidungen für einen virtuellen KUKA-Roboter treffen. Sie durchlaufen insgesamt vier verschiedene Systembedingungen, in denen sich die Darstellungsform (Label-basiert oder multimodal) und das Vorhandensein eines KI-Assistenten unterscheiden. Nach jedem Durchlauf füllen Sie einen digitalen Fragebogen aus.

## II. Ablauf der Teilnahme

Direkt nach dem Unterschreiben und Abgeben der Einwilligungserklärung werden Sie von der Studienleitung in den Ablauf der Studie eingewiesen. Sie werden ein VR HMD aufsetzen und nacheinander vier Bedingungen durchlaufen. In jeder Bedingung werden Ihnen Sensordaten (Temperatur- und Vibrationswerte) eines virtuellen Industrieroboters präsentiert, auf deren

Basis Sie eine Wartungsentscheidung treffen. Je nach Bedingung steht Ihnen dabei ein KI-Assistent zur Verfügung, den Sie per Sprachbefehl befragen können. Sie werden nicht gefilmt oder fotografiert. Nach jedem Durchlauf setzen Sie das VR HMD kurz ab und füllen einen digitalen Fragebogen auf einem bereitgestellten Gerät aus. Der Ablauf der gesamten Studie wird circa 60–90 Minuten dauern. Anschließend sollten Sie sich 15 Minuten hinsetzen und ausruhen, bevor Sie andere Tätigkeiten ausführen, um das nachträgliche Eintreten möglicher Nebenwirkungen (siehe nächster Abschnitt) besser einschätzen zu können.

### **III. Unannehmlichkeiten und Risiken (Warnhinweise)**

Die Studie wird mit Virtual Reality Technologie durchgeführt. Daher sind an dieser Stelle allgemeine Warnhinweise bezüglich VR Head-Mounted Displays angebracht: Die Benutzung von VR kann Übelkeit und Schwindel auslösen. Die Nähe zu virtuellen Objekten kann sich in einem unangenehmen Körpergefühl äußern oder sogar Platzangst auslösen. Körperliche Abwehrreflexe werden unter Umständen verstärkt auftreten. Lichtblitze und optische Effekte können erschrecken und bedrohlich wirken. Bei Menschen mit einer Neigung zu Krampfanfällen können diese Effekte unter Umständen akute Krampfanfälle auslösen. Sollten Sie während der Untersuchung mit diesen oder ähnlichen Symptomen konfrontiert sein oder andere unangenehme Effekte spüren, brechen Sie den Studiendurchgang bitte unverzüglich ab. Es kann auch unmittelbar nach der Nutzung von VR zu Übelkeit, Kopfschmerzen oder Schwindel kommen. In seltenen Fällen kann es im Nachhinein auch zu Migräne oder Orientierungsschwierigkeiten kommen. Bitte bedienen Sie deshalb unmittelbar nach der Studiendurchführung keine Maschinen oder gefährliche Gegenstände. Nehmen Sie erst am Straßenverkehr teil, wenn sich diese Symptome eindeutig gelegt haben. Falls es während oder nach der VR-Exposition zu einigen der genannten Nebenwirkungen kommt, sollten Sie sich ausruhen (hinsetzen oder hinlegen) und Wasser trinken (wird bereitgestellt). Die Studienleitung wird ein Fenster öffnen, um Sie mit Frischluft zu versorgen. Bei länger anhaltenden Nebenwirkungen sollten Sie einen Arzt aufsuchen. In Notfällen wird die Studienleitung einen Notarzt rufen.

### **IV. Ausschlusskriterien**

Der Aufruf zur Teilnahme an der Studie richtet sich ausschließlich an volljährige Menschen. Menschen, die eine vorliegende Herzschwäche oder Angststörungen aufweisen oder unter Krampfanfällen leiden, werden aus Sicherheitsgründen nicht in Betracht gezogen. Von der Teilnahme an der Studie während einer Schwangerschaft wird abgeraten. Bitte nehmen Sie auch nicht an der Studie teil, wenn Sie akut unter Übelkeit, Erbrechen, Kopfschmerzen, Schwindel oder einer Migräne leiden. Sollten diese Symptome oder Schmerzen in den Augen oder Ohren während der Durchführung der Studie auftreten, muss die Untersuchung unverzüglich abgebrochen werden. Bitte nehmen Sie nicht an der Studie teil, wenn Sie unter dem Einfluss von Alkohol oder anderen Drogen oder wahrnehmungseinschränkenden bzw. -verzerrenden Medikamenten stehen.

## **V. Freiwilligkeit**

Die Teilnahme an dieser Studie geschieht freiwillig. Sie können Ihre Teilnahme an der Studie jederzeit und ohne Angabe von Gründen abbrechen, ohne dass persönliche Nachteile für Sie entstehen. Auch die Studienleitung kann den Studiendurchgang jederzeit abbrechen.

## **VI. Datenschutz**

Im Rahmen der Studie werden keine personenbezogenen Daten (wie z. B. Name, Anschrift, Geburtsdatum) der Studienteilnehmenden erhoben. Sie erhalten zu Beginn der Studie eine anonyme Teilnehmer-ID (z. B. "P01"), unter der alle Ihre Antworten und Entscheidungen gespeichert werden. Die erhobenen Daten können nur für den Zweck dieses Forschungsvorhabens verwendet werden. Die Daten werden ausschließlich lokal auf Datenträgern der Hochschule für Angewandte Wissenschaften Hamburg gespeichert und nicht an externe Dienste übermittelt. Die Fragebogendaten werden als CSV-Dateien, die Entscheidungsdaten als Log-Dateien gespeichert. Die Daten werden für die Dauer von 10 Jahren gespeichert. Bei der Anonymisierung werden Name und andere Identifikationsmerkmale nicht gespeichert, sodass die Daten den Studienteilnehmenden nicht mehr zugeordnet werden können. Die Auswertung, Nutzung, Weitergabe und Veröffentlichung der Daten durch die Studienleitung und ihre Mitarbeitenden erfolgt somit ausschließlich in anonymisierter Form. Daher können die Daten einzelner Studienteilnehmenden auch im Nachgang nicht mehr gelöscht werden. Ein Widerruf der Studienteilnahme ist aufgrund der Anonymisierung nach Abschluss der Datenerhebung nicht möglich. Verantwortung für die sachgemäße Datenverarbeitung trägt die Studienleitung.

## **VII. Einwilligungserklärung**

Hiermit erkläre ich, dass ich über Ziele und Ablauf der Studie informiert wurde und diese verstanden habe. Ich hatte ausreichend Gelegenheit, mich bei der Studienleitung über die Studie zu informieren, sowie auftretende Fragen zu stellen. Diese wurden mir von der Studienleitung verständlich beantwortet. Ich habe die Ausschlusskriterien gelesen und verstanden und bestätige, dass keiner der aufgeführten Punkte auf mich zutrifft. Ich habe verstanden, dass die Teilnahme an der Studie freiwillig ist und ich jederzeit ohne Angabe von Gründen aus der Studie ausscheiden kann, ohne dass persönliche Nachteile für mich entstehen. Auch die Studienleitung kann die Studie jederzeit beenden. Mir ist bekannt, dass diese Studie in erster Linie der Wissenserweiterung dient und gegebenenfalls auch keinen persönlichen Vorteil für mich bringen kann. Da keine personenbezogenen Daten erhoben werden, habe ich verstanden, dass nach Abschluss der Datenerhebung prinzipiell keine Zuordnung mehr zwischen den Daten im Datensatz und meinen personenbezogenen Daten [Name, Adresse und genaues Geburtsdatum] möglich ist. Mir ist bewusst, dass daher nach Abschluss dieser Datenerhebung keine gezielte Löschung meines persönlichen Datensatzes mehr möglich ist, da dieser nicht zugeordnet werden kann. Eine Löschung meiner persönlichen Daten ist ausschließlich direkt nach der Durchführung der Untersuchung möglich. Eine Löschung aller

erhobenen Originaldaten ist nach Ablauf des Aufbewahrungszeitraums von 10 Jahren vorgesehen. Ich habe verstanden, dass die Ergebnisse und Daten dieser Studie als wissenschaftliche Publikationen veröffentlicht werden können und dass dies in anonymisierter Form geschieht. Ich bestätige hiermit, dass ich freiwillig an der Studie teilnehmen möchte.

---

Ort, Datum

---

Unterschrift der an der Studie teilnehmenden  
Person

### **VIII. Erklärung Studienleitung**

Ich, Moritz Treu (Studienleitung) erkläre hiermit, dass ich verantwortlich für die Durchführung des Projekts bin. Alle Beteiligten sind dazu verpflichtet, alle hier erwähnten Ziele einzuhalten und die Studienteilnehmenden darüber zu informieren, wenn im Studienverlauf Änderungen erfolgen sollten, die sich auf Inhalte der Einwilligungserklärung auswirken sollten.

### **Erklärung zur Schweigepflicht und zum Datenschutz**

Ich bin darüber aufgeklärt worden, dass ich über alle Daten und Informationen, die mir im Rahmen meiner Mitarbeit beim Projekt über einzelne Studienteilnehmende zugänglich werden, zur Verschwiegenheit verpflichtet bin. Diese Daten und Informationen dürfen nur direkten Projekt-Mitarbeitenden der HAW zugänglich gemacht werden. Insbesondere unterliegen alle personenbezogenen Daten dem Datenschutz und dürfen nicht an dritte Personen oder Institutionen weitergegeben werden. Ich darf keine eigenen Aufzeichnungen personenbezogener Daten über die Dauer meiner Projektmitarbeit hinaus aufbewahren. Sofern ich während meiner Projektmitarbeit vorübergehend personenbezogene Daten verarbeite und/oder aufbewahre, verpflichte ich mich, Missbrauch oder die unbefugte Weitergabe dieser Daten zu verhindern.

---

Ort, Datum

---

Unterschrift der Studienleitung

# Eigenständigkeitserklärung

Hiermit versichere ich, dass ich die vorliegende Masterarbeit mit dem Titel

## **Immersive Data Visualization and LLM-Supported Decision-Making in Digital Twins for Predictive Maintenance of Industrial Robots**

selbstständig und nur mit den angegebenen Hilfsmitteln verfasst habe. Alle Passagen, die ich wörtlich aus der Literatur oder aus anderen Quellen wie z. B. Internetseiten übernommen habe, habe ich deutlich als Zitat mit Angabe der Quelle kenntlich gemacht.

Hamburg, 30. Juni 2026