

**SIAI** GPT powered chatbot

# Safety Education in the AI Era.

*Master thesis*

*Created by Anh Tuan Nguyen-s1816187*

*Industrial Design Engineering*

*Faculty of Engineering Technology*

*Supervisors: Mohammad Rajabali Nejad*

*Date: 17/4/2024*

textual and visual  
customization





# Contents

<b>1</b>	<b>Introduction</b>	<b>3</b>
<b>2</b>	<b>Methodology</b>	<b>7</b>
<b>3</b>	<b>Safety Standard Education</b>	<b>9</b>
3.1	Industry-based and Academic Routes to Learn Standards. . . . .	10
3.2	Safety Standards: Where Do They Fall Short? . . . . .	10
3.3	Prototype requirements: . . . . .	21
<b>4</b>	<b>The Gap in Current Methods</b>	<b>23</b>
4.1	Using Serious Games to Teach Safety Standards . . . . .	23
4.1.1	Benefits of Serious Games . . . . .	26
4.1.2	Limitations: Lack of Personalized Feedback and Guidance . . . . .	27
4.2	Generative AI in Safety Education . . . . .	28
4.2.1	Practical Uses and Advantages . . . . .	29
4.2.2	Limitations and criticism . . . . .	31
4.3	The Gap: Recap . . . . .	35
<b>5</b>	<b>The Exploration Journey</b>	<b>37</b>
5.1	Stage 1: Ideation and Early Brainstorming: . . . . .	39
5.2	Stage 2: Concept Development and Evaluation Against Requirements . . . . .	43
5.3	Stage 3: In-depth Research through Literature Review . . . . .	49
5.3.1	Safety Visualization: . . . . .	49
5.3.2	Customized Generative AI Technology . . . . .	56
5.3.3	Integration of GenAI and Visuals: . . . . .	57
5.4	Final thoughts: The promising direction for the thesis. . . . .	59
<b>6</b>	<b>The Design Prototype</b>	<b>61</b>
6.1	Solution architecture . . . . .	61
6.2	Prototype Development . . . . .	64
6.2.1	Technical Preparation . . . . .	64
6.2.2	Iteration Design of AI's behaviors . . . . .	69
6.2.3	Iteration Design of visual mockups . . . . .	74



6.3	Final prototype . . . . .	79
6.3.1	SAI (Safety Instructor AI): the AI chatbot systems . . . . .	79
6.3.2	Four Visual mockups: . . . . .	83
<b>7</b>	<b>What We Tested, What We Found</b>	<b>86</b>
7.1	Test 1: Robustness Evaluation . . . . .	87
7.2	Test 2: Comprehension test for visualization . . . . .	90
7.3	Test 3: Experimental Study . . . . .	94
<b>8</b>	<b>A Storyline Recap</b>	<b>101</b>
8.1	Answers for sub-research question 1: The Identified Problems . . . . .	101
8.2	Answers for sub-research question 2: The Gap . . . . .	102
8.3	Answers for sub-research question 3: Design Solution . . . . .	103
8.4	Answers for sub-research question 4: Tests and Results . . . . .	106
<b>9</b>	<b>Discussion</b>	<b>108</b>
<b>10</b>	<b>Conclusion</b>	<b>112</b>
<b>11</b>	<b>Acknowledgment</b>	<b>114</b>
<b>A</b>	<b>Detailed Results of Test 1</b>	<b>125</b>
<b>B</b>	<b>Detailed Results of Test 2</b>	<b>129</b>
<b>C</b>	<b>Results Analysis of Test 3</b>	<b>133</b>

# ***Abstract***

## ***Background***

Safety standards are essential in shaping a modern society that is safe, convenient, and trustworthy. However, their integration into the European education system remains limited.

## ***Problems***

One contributing factor is that safety standards are not well-suited as learning materials for a diverse student population, varying in culture, literacy, and expertise. As a result, students often struggle to find, understand, and apply these standards in their education.

## ***The Gap***

While current solutions—such as e-learning, professional training, university courses, and gamification—offer certain strengths, they often fall short in providing personalized learning experiences. This is where generative AI holds promise. However, generic AI models typically lack the specialized knowledge needed to deliver the precise information required by safety standards.

## ***Solution***

To bridge this gap, we developed SIAI (Safety Instructor AI), a system of customized GPT-based chatbots designed to teach safety standards. This system provides tailored textual and visual responses based on content from the Safety by Design course. The main goal is to support students in learning safety standards with more accessible and specialized guidance.

## ***Tests and Results***

The SIAI prototype was evaluated through three distinct tests. The first focused on robustness, demonstrating that the prototype could reliably deliver both general and domain-specific information aligned with NEN NTA 8287:2021. Its performance exceeded expectations, achieving 100% % in both accuracy and completeness for short-form answers. Additionally, SIAI successfully interpreted a variety of communication styles, showcasing its adaptability to different user inputs.

The second test assessed comprehension through visual mock-ups, highlighting the effectiveness of visual elements in enhancing user engagement and understanding. The results also indicated

that a hierarchical structure fits well within the chatbot format, presenting a strong alternative to traditional layouts based on headings and subheadings.

The third test was an experimental study comparing two groups: one using SIAI and the other using traditional safety standards and a conventional search browser like Google. The SIAI group outperformed the control group in all assigned tasks, including retrieving information from the NEN NTA 8287:2021, understanding technical terms, and accurately describing the sub-tasks in steps 1, 2, 4, and 6 of the method. Furthermore, students using SIAI demonstrated a stronger ability to apply NEN NTA 8287:2021 information to real-world design assignments.

## ***Conclusion***

SIAIs (Safety Instructor AI) present a promising approach to making safety information more accessible in educational settings, highlighting the potential of generative AI to transform how safety standards are taught and learned. As a result, further large-scale testing is recommended to validate these findings

# 1 Introduction

## *Step into the World of Standardization*

Driving a rental car in a foreign country would become very dangerous if the dashboard symbols were completely different from those drivers know at home. Faced with these unfamiliar icons, drivers may struggle to correctly interpret warnings and essential information, increasing the risk of accidents. To avoid such confusion and promote safety, the automotive industry has standardized dashboard symbols across countries, as seen in Figure 1, allowing drivers to operate vehicles more confidently (International Organization for Standardization, 2021). As a result, regardless of where the car was made, drivers can operate it safely.

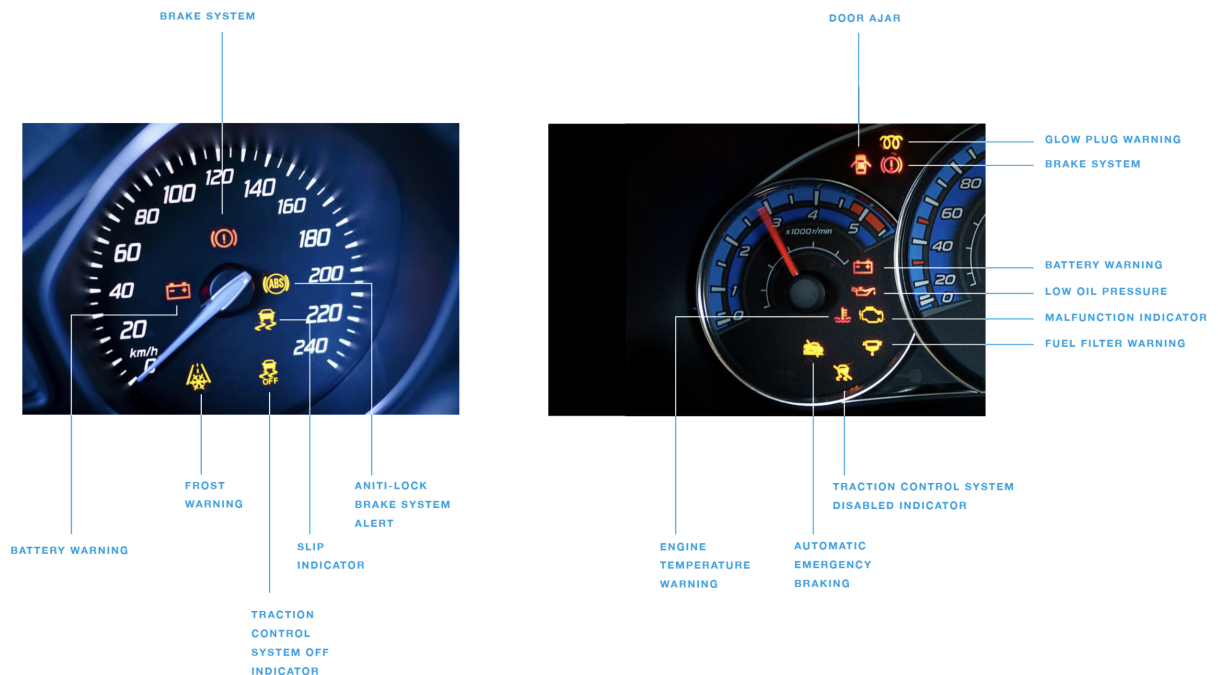


Figure 1: This figure displays the dashboard interfaces of BMW (left) and Subaru (right), highlighting a shared selection of warning symbols to ensure clarity and promote safety.

This example highlights how standardization plays a crucial role in shaping a safer and more convenient living environment. Standardization is the systematic process of developing and using technical standards to maximize efficiency, quality, and safety. These standards are established by authoritative bodies, such as the International Organization for Standardization (ISO), providing suggestions and guidelines that help products achieve optimal practice and performance (Nederlands Normalisatie Instituut, [n.d.](#)).

## Unpacking the Problems

Although standardization is essential to the development of our societies, education on this topic has long received limited attention, particularly within Europe. Data since 2007 showed that only 10 to 30 universities across Europe offered courses on standardization, while South Korea alone had 46 universities (de Vries & Egyedi, 2007). A few years later, Dr. Wilfried Hesser reported that around 1,000 European students took standardization courses in 2013, which was far fewer than in Japan, with 2,100 students in 2011, and South Korea, with 3,883 students in 2012 (Hesser, 2014). The issue of lacking standardization courses continued, as it was still highlighted in a 2022 report by the European Commission. Consequently, experts expressed serious concerns about a future shortage of qualified professionals in the field of standardization (European Commission, 2022).

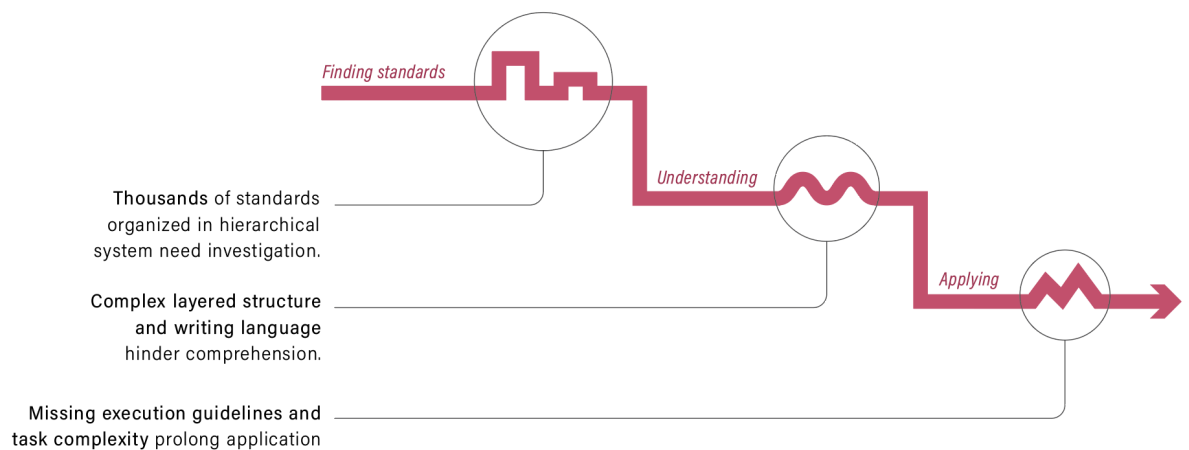


Figure 2: This figure illustrates the difficulties involved in finding relevant standards, understanding their content, and applying them in practical projects.

Despite many efforts to promote standardization, current teaching materials often fail to meet the needs of diverse learner groups regarding cultural background, literacy levels, and subject expertise (International Organization for Standardization, 2014). Figure 2 outlines key challenges related to safety standards in the educational context. The first difficulty lies in identifying relevant standards from among thousands of options, often buried within a complex hierarchical classification system. Then, as primary learning material, these standards offer precise and comprehensive information but are often highly abstract. Consequently, the documents frequently do not offer explicit implementation guidance for all applicable problems. As a result, these issues cause significant learning barriers for safety learners in finding, understanding, and applying standards in their education.

Several approaches, ranging from university courses at the undergraduate or graduate level, professional training, and life-long learning programs with e-learning or gamification techniques, have been used to enhance the effectiveness and engagement of safety standard education (Arezes & Swuste, 2012; Rajabalinejad, 2020; Rodeghiero Neto & Amaral, 2024). For example, *Safety by Design* at the University of Twente is a master's degree course that guides students to design safe products by applying theoretical safety standards in real-world situations (University of Twente, n.d.). However, these approaches usually do not offer instant, individualized, and context-relevant feedback on students' demands, making it challenging to study and put safety principles into practice.

Emerging technologies such as Generative Artificial Intelligence (GenAI) have the potential to address these challenges by offering personalized learning experiences (Onatayo et al., 2024). However, research has shown that such technologies may lack the specialized knowledge required for safety education (Izadi & Forouzanfar, 2024). Therefore, the effectiveness of GenAI models, like ChatGPT, is limited and more likely to present confusing and inaccurate results.

## ***Research Aim***

To bridge this gap, we developed a customized generative AI tool called Safety Instructor AI (SIAI), designed specifically to support safety standards education. With the specialized knowledge derived from the Safety by Design course materials, SIAI delivers tailored textual and visual outcomes, helping learners to find, understand, and apply safety standards. In other words, this thesis aims to answer the following research question:

***How can customized generative multi-modal artificial intelligence help learners find, understand, and apply safety standards in real-world situations?***

To cover all aspects of the main research question systematically, five sub-research questions have been formulated as below:

*1. What exact challenges do learners face in studying safety standards?*

To understand the specific difficulties learners have encountered, these challenges are framed as key targets that this thesis aims to address.

*2. How have the current methods dealt with these problems, and where is the gap?*

To pinpoint what related works have been done and, more importantly, what is still missing. Thereby, the need for innovation is justified.

*3. In what ways could generative multi-modal artificial intelligence solve these challenges?*

To explore the potential of the emerging technology in supporting standards learners and to develop a prototype based on those insights.

*4. To what extent does the prototype address the identified problems, compared to the traditional method?*

To validate whether the prototype provides substantial improvements compared to the conventional method.

## ***Thesis Structure***

This thesis outlines our journey in customizing generative artificial intelligence to support standard learners in engaging with and applying these documents. The next chapter explains the methodology used to address the main and sub-research questions (page 7). Then, the challenges in safety standards education are described in chapter 3 (page 9); followed by an analysis of the limitations in current solutions (chapter 4 - page 23).

The beginning of our search for a promising solution is described in chapter 5 (page 37), and continues with the presentation of the prototype (chapter 6 - page 61 ). Chapter 7 then focuses on testing and results (page 86 ), and is followed by a discussion in Chapter 9 (page 108). The thesis concludes with the Conclusion chapter on page 112.

## 2 Methodology

This thesis was carried out through a close collaboration between the instructor of the *Safety by Design* course and a current Master's student in the Industrial Engineering Design field. The study aims to explore how generative AI, an emerging technology, can be employed to facilitate safety education.

### ***Problem identification methods:***

At the beginning of the thesis, we employed two main methods to identify the core problems: a literature review and stakeholder interviews. The literature review emphasized broader issues of using safety standards as learning material. In contrast, the stakeholder interview offered more focused insights into the specific problems of learning safety standards in the *Safety by Design* course at the University of Twente. In particular, we collected insights from three main groups of stakeholders:

**1. Former students of the *Safety by Design* course:** Using our personal experience as former students of the course, we also consulted feedback from ten group projects to better understand the challenges they faced in finding, interpreting, and applying safety standards within their design projects.

**2. The course instructor:** We interviewed the instructor to gain insight into their teaching methods, objectives, and expectations regarding student performance through the weekly meetings.

**3. Students from a variety of disciplines:** Students from different fields, such as chemical engineering, computer science, business administration, and game design, were interviewed in order to explore how individuals outside the industrial design engineering field perceive content related to safety. The diversity of participants was intentionally chosen to reflect the wide audience of safety standards.

The insights from both the literature review and stakeholder interviews helped pinpoint specific problems in the use of safety standards in education. These insights served as the foundation for defining the prototype requirements, which in turn shaped the solution architecture developed later in the project.



## ***Prototype development:***

We embraced the principles of **Agile Usability Testing**, which involved the gradual process of finding, developing, and evaluating possible solutions through repeated cycles of user testing and feedback (Nielsen, [n.d.](#)). In particular, the master's students and the course instructor held weekly meetings to discuss insights gathered from literature reviews, stakeholder analysis, and user tests to explore potential directions for the study collaboratively.

In this thesis project, we also used course materials from the Safety by Design curriculum, including lecture notes, textbooks (Mohammad Rajabalinejad, [2020](#)), presentation slides, submitted student assignments, and, especially, the course game. These materials were employed as reference points to ensure the solutions aligned with the goals of the study.

## ***Assessment Approach:***

A set of evaluation tests was conducted to assess the effectiveness of the proposed solution in improving information searching, strengthening students' comprehension, and assisting the application of the Dutch national standard NEN-NTA 8287:2021—Safety Cube Method for the design of real-world systems and products.

### 3 Safety Standard Education

As mentioned in the Introduction, standardization is the key process to achieving safer and more efficient production, as the majority of standards are developed to support resilient infrastructure, sustainable industry, and safe innovation (International Standard Organization, [n.d.](#)). Figure 3 illustrates how standards align with the 17 Sustainable Development Goals outlined by the United Nations. Therefore, educators often use safety standards as reliable learning materials to present the most basic principles, proven frameworks, and best practices to guide learners in designing high-performance products (International Organization for Standardization, [2014](#)).



Figure 3: International Standards Across 17 United Nation Sustainable Development Goals. This figure shows the largest number of standards focused on industry, innovation, and infrastructure.

### 3.1 Industry-based and Academic Routes to Learn Standards.

In Europe, safety standards can be learned through different formats. A common approach is to attend a relatively short online or offline course provided by regulatory bodies, safety organizations, or consultancies (Arezes & Swuste, 2012). Online courses often last just a few hours and cover basic information about a single standard, while in-person courses usually span a few days and focus on more advanced level (BSI, n.d.).

Another approach is through university courses, where safety standards are integrated into lectures, explored through assignments, and used as mandatory reading materials (International Organization for Standardization, 2014). According to a study by Arezes and Swuste (2012), which analyzed 269 safety courses across 24 European countries, universities offer approximately 85% of these safety courses, primarily in Master's programs. Around 12% of the courses come from polytechnic institutions or specialized schools.

For example, the master's course *Safety by Design* at the University of Twente guides students in designing safe products through real-world, multidisciplinary projects. Throughout the course, students receive relevant knowledge based on the NEN NTA 8287: 2021, which helps them understand the different challenges and integrate safety aspects throughout an entire product life cycle (Mohammad Rajabalinejad, 2021).

### 3.2 Safety Standards: Where Do They Fall Short?

Many master's-level courses on safety and product design typically aim to give students a complete view of safety standards, covering from identifying relevant standards to understanding and implementing them throughout the design process (KTH, n.d.; Politecnico, n.d.; University of Twente, n.d.). The following subsections outline the challenges associated with finding, understanding, and applying safety standards in an educational context. These issues were identified through personal experience while participating in the Safety by Design course, combined with feedback from fellow students, insights from teachers, as well as results from the literature.

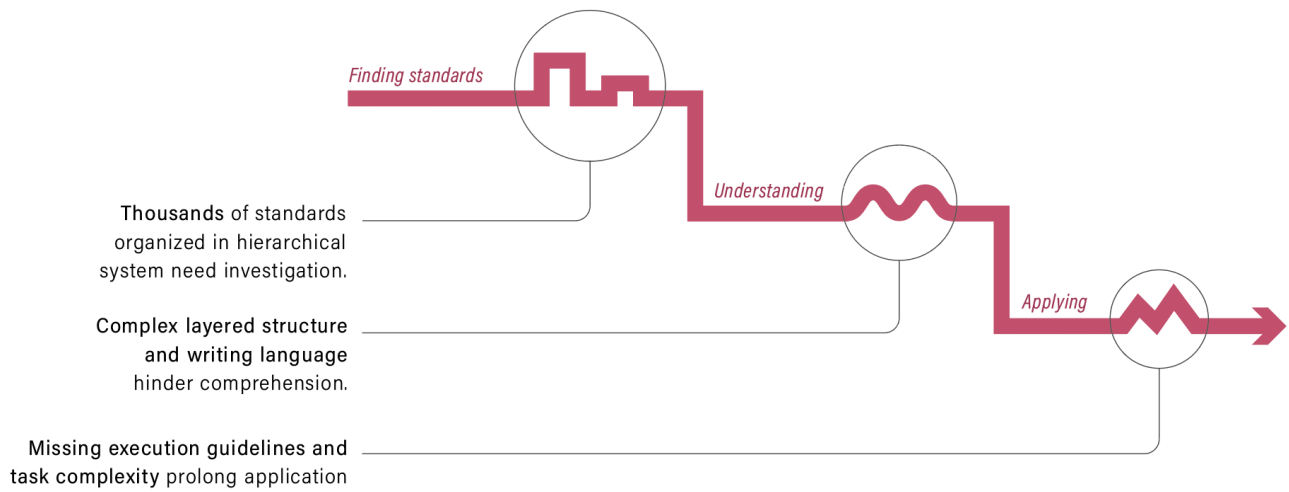


Figure 2 is reshowed here to summarize the identified problems

### **Challenges in Finding Safety Standards: Complex Navigation, Overwhelming Lists, and Lack of Selection Guidance:**

When engineers design a complex system comprising multiple components, such as a railway network, they must consult a wide range of standards across various technical areas. These standards ensure that each individual component is safe and can work seamlessly together. For example, as shown in Figure 4, building a reliable railway system in Europe requires compliance with multiple *Technical Specifications for Interoperability* (TSIs). These are technical and operational standards regulated by the European Union (EU), ensuring trains, tracks, signals, and other components are compatible across countries. Each TSI focuses on a specific area, such as operations, energy, or train wagons (European Union Agency for Railways, [n.d.](#)).



- A:** TSI-ENE – Energy
- B:** TSI-WAG – Wagons
- C:** TSI-PRM – Persons with Reduced Mobility
- D:** TSI-NOISE – Noise
- E:** TSI-INS – Infrastructure

Figure 4: The figure shows that creating a railway system in Europe requires consideration of multiple relevant standards.

The European Commission's website is considered a reliable source for providing a comprehensive list of the latest harmonized standards (European Commission, [n.d.](#); Procter Machine Safety, [2018](#)). Safety learners can use this platform to find different recommended standards relevant to their specific products.

However, these learners may face challenges in navigating the website due to its complex hierarchical structure. As can be seen in Figure 6, the website presents lists of standards under 12 distinct directives, which can be further divided into subcategories. Sometimes, the number of subcategories can go up to 11. As a result, the process of locating suitable standard lists is slowed down by the vast number of available pathways and frequently leads to learners' cognitive overload (Taky-eddine & Madaoui, 2024).

Another problem is that while the standard lists provided by the European Commission's website are comprehensive, they are also lengthy and lack clear selection guidance. For example, according to a 2017 EU evaluation report, under the Machinery Directive alone, there are over 800 standards, with 5 to 10 new/ revised standards published each month (Simmonds et al., 2017). The website presents these standards in a table format spanning over 200 pages. Without any built-in filtering/sorting mechanism and clear guidance on how to make selections, learners have no immediate way to distinguish between relevant and irrelevant standards (Procter Machine Safety, 2018). Figure 5 maps the two problems onto the user journey, from entering the European Commission's website to accessing the list of standards.

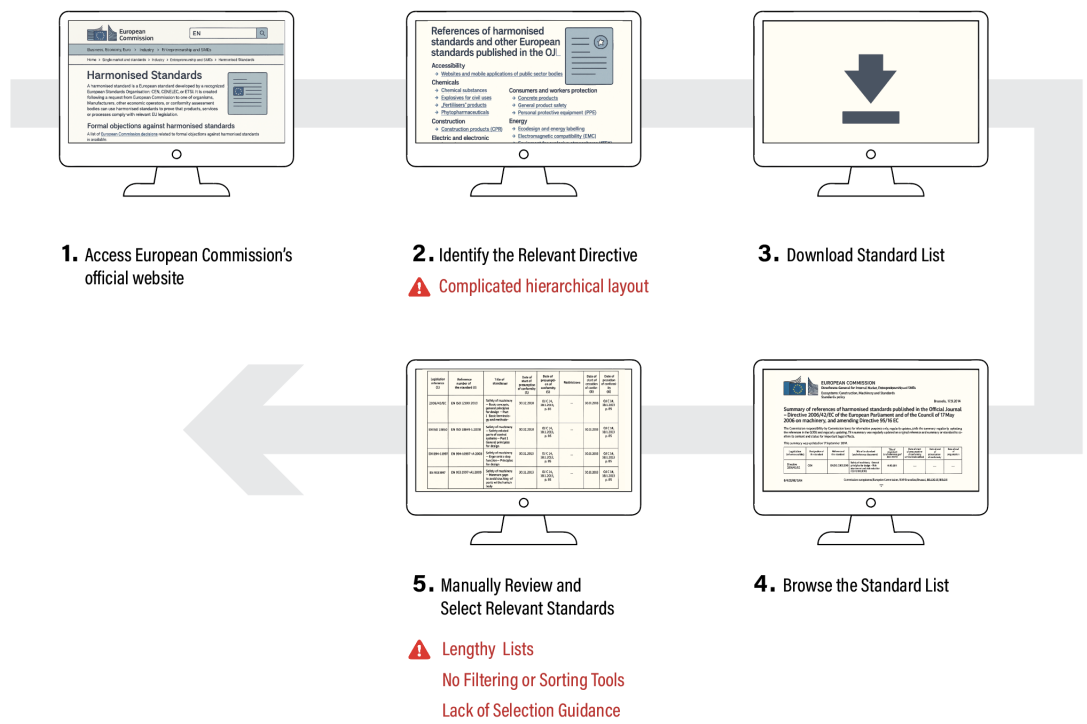


Figure 5: The figure shows the challenges users face when navigating from the European Commission website to the recommended safety standards list.

# References of harmonised standards and of other European standards published in the OJEU

- 1 **Accessibility**
  - [Websites and mobile applications of public sector bodies](#)
- 2 **Chemicals**
  - [Chemical substances \(REACH\)](#)
  - [Explosives for civil uses](#)
  - [Fertilising products](#)
  - [Pyrotechnic articles](#)
- 3 **Conformity assessment and management systems**
  - [New Legislative Framework \(NLF\) and Eco-Management and Audit Scheme \(EMAS\)](#)
- 4 **Construction**
  - [Construction products \(CPD/CPR\)](#)
- 5 **Consumers and workers protection**
  - [Cosmetic products](#)
  - [General product safety](#)
  - [Personal protective equipment \(PPE\)](#)
  - [Toy safety](#)
- 6 **Energy efficiency**
  - [Ecodesign and energy labelling](#)
- 7 **Electric and electronic engineering**
  - [Electromagnetic compatibility \(EMC\)](#)
  - [Equipment for explosive atmospheres \(ATEX\)](#)
  - [Low Voltage \(LVD\)](#)
  - [Radio Equipment \(RED\)](#)
  - [Restriction of the use of certain hazardous substances \(RoHS\)](#)
- 8 **Healthcare engineering**
  - [In vitro diagnostic medical devices \(Regulation 2017/746\)](#)
  - [Medical devices \(Regulation 2017/745\)](#)
  - [Active implantable medical devices \(Directive 90/385/EEC\)](#)
  - [In vitro diagnostic medical devices \(Directive 98/79/EC\)](#)
  - [Medical devices \(Directive 93/42/EEC\)](#)
- 9 **Measuring technology**
  - [Measuring instruments \(MID\)](#)
  - [Non-automatic weighing instruments \(NAWI\)](#)
- 10 **Mechanical engineering and means of transport**
  - [Cableway installations designed to carry persons](#)
  - [Equipment for explosive atmospheres \(ATEX\)](#)
  - [Gas appliances \(GAR\)](#)
  - [Inspection of pesticide application equipment](#)
  - [Lifts](#)
  - [Machinery \(MD\)](#)
  - [Pressure equipment \(PED\)](#)
  - [Rail system: interoperability](#)
  - [Recreational craft and personal watercraft](#)
  - [Simple Pressure Vessels \(SPVD\)](#)
  - [Unmanned aircraft systems \(UAS\)](#)
- 11 **Services**
  - [Community postal services](#)
- 12 **Sustainability**
  - [Packaging and packaging waste](#)
  - [Plastic caps and lids](#)

Figure 6: This figure shows the interface of the European Commission's website, where students must navigate a lengthy list of directives to locate the relevant standards.

### ***Barriers to Understanding: Hyper-specific Definitions, Complex Formatting, and Dense Writing Styles***

Even after identifying the relevant safety standards, learners often struggle to understand them quickly due to their formatting structure, vocabulary and their definitions, as well as writing style.

#### **Complex Information Hierarchy Structure:**

Regarding formatting structure, the documents typically adopt the traditional hierarchical structure with multiple nested sections, similar to the European Commission's website. In lengthy standards, deeply nested sections can increase learners' mental effort required to understand how the documents are organized and how different sections relate to each other (Taky-eddine & Madaoui, [2024](#)).

For example, as can be seen in Figure 7, ISO 12100, a fundamental standard under the Machinery Directive, features a highly complex structure with seven main sections and up to five layers of nested subsections. Furthermore, these subsections often cross-reference one another, requiring learners to flip through pages repeatedly, resulting in a disrupted reading experience (Pryor, [2019](#)). Our observation shows that once the detailed content begins, in Chapters 4 and 5 of ISO 12100 alone, there are nearly 20 instances of cross-referencing between sections.



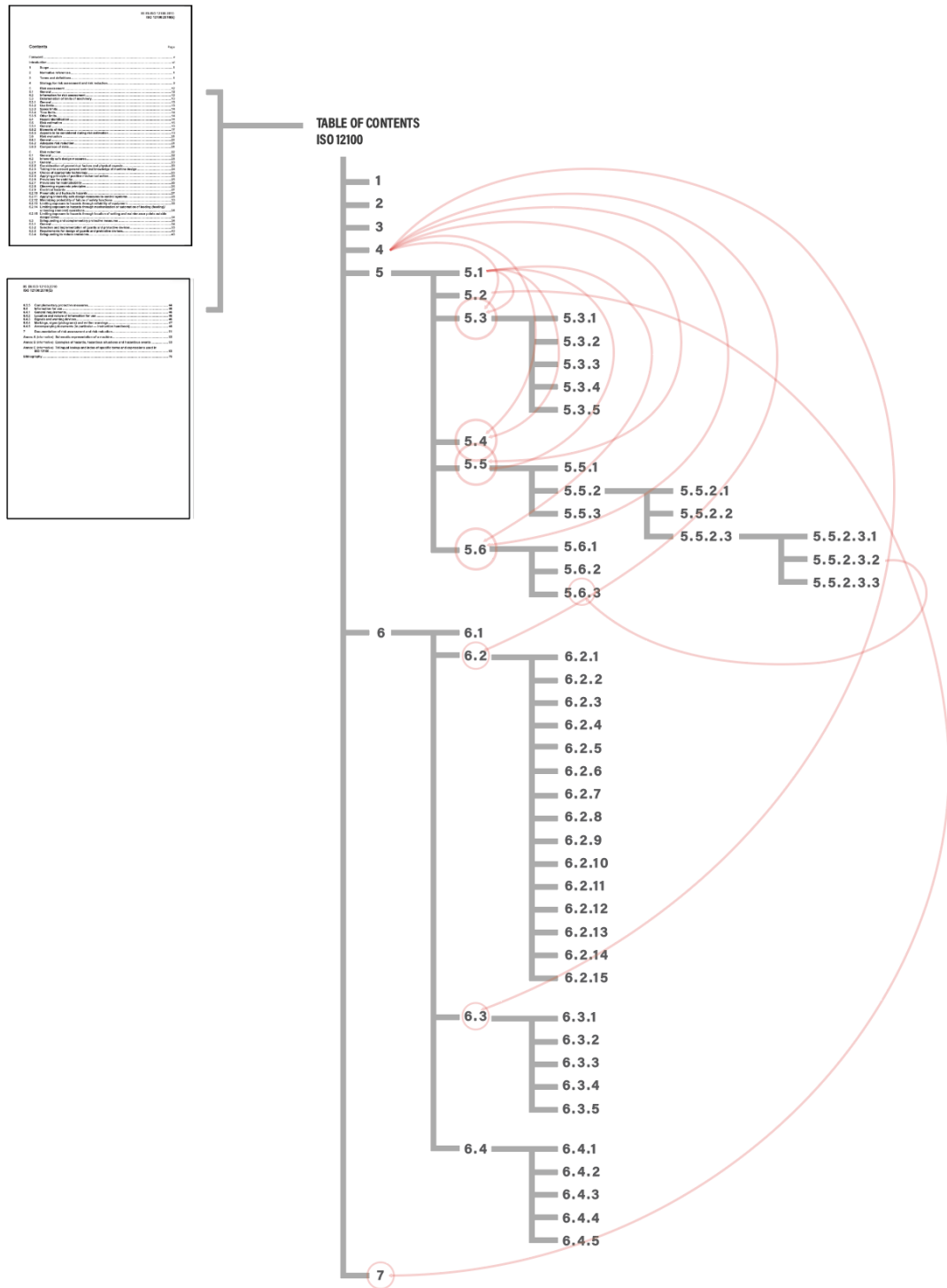


Figure 7: This figure illustrates the complex hierarchical structure of ISO 12100. The left images are snapshots of the original Table of Contents, while the right diagram visualizes this structure. The red lines indicate where each cross-reference starts and where it points to (the circles).

## Specialized Terms with Complex and Interlinked Definitions

To ensure that professionals from different industries can interpret safety principles and agreements consistently, standards are built around a set of specialized terminologies. For example, phrases like the system under consideration, critical safety function, and acceptable risk criteria can effectively create a shared understanding of complex safety concepts among experts. Yet, these phrases also create steep learning curves for new learners (Spychalski, 2022).

Recognizing this drawback, many guidelines advise standards writers to only use specialized terminology when readers are familiar with them, prefer them, or need to learn them to accomplish their goals (CEN-CENELEC, 2010; ISO Standards, n.d.; NEN-ISO, 2025). Additionally, when these terms are used, their definition must be presented in "*Terms and definitions*", a mandatory section in every standard ("Societal stakeholders and standards - 1.1 What are standards?" n.d.).

These definitions are carefully and precisely crafted to ensure they can be applied across various fields but, therefore, often become complex. Figure 8 presents an example of a "*human-system integration*" definition, structured as a complex noun phrase enriched with multiple modifiers (Mohammad Rajabalinejad, 2021).

### Definition of "human-system integration":

1. integrated and comprehensive 2. analysis, design, assessment of  
3. requirements, concepts and resources 4. for system manpower, personnel,  
training, safety and occupational health, habitability, personnel  
survivability, and human factors engineering

1.: adjectives describing the nouns (modifier 1)

2.: core nouns

3.: specifies what kind of assessment (modifier 2)

4.: tells the purpose or target of the whole phrase (modifier 3)

Figure 8: The figure shows the grammatical structure of the "*human system integration*" definition as a noun phrase.

Not only is each individual definition sometimes difficult to understand, but many also depend on each other for clarity (Spychalski, 2022). Figure 9 shows that in order to understand precisely and comprehensively the word "functional risk assessment", learners might also need to read the definitions of "functional" and then "function". This interconnecting approach creates a complex network of definitions, making it difficult for learners to fully understand even one single term (Spychalski, 2022).

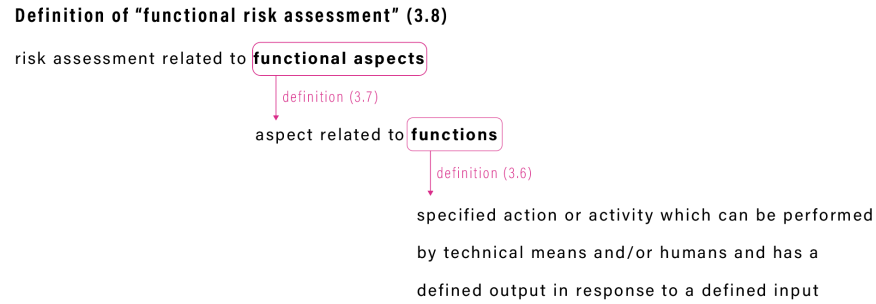


Figure 9: The figure illustrates how the definition of "functional risk assessment" relies on the definitions of "functional aspects" and "functions" for a complete explanation.

## Information-dense writing style

Many handbooks on standard writing recommend keeping sentences between 14 to 20 words for easy comprehension (ISO Standards, n.d.). Despite this advice, standard writers often prioritize completeness over readability by condensing all relevant information into a single statement. For example, ISO 45001 begins section 5.4, "Consultation and Participation of Workers", with sentences containing 43 words, presented in Figure 10.

The organization shall establish, implement and maintain a process for  
consultation and participation of workers at all applicable levels and functions,  
and, where they exist, workers' representatives, in the development, planning,  
implementation, performance evaluation and actions for improvement of the  
OH&S management system

Figure 10: This figure shows the extended opening sentence of Section 5.4 in ISO 45001

This writing style can reduce engagement and slow down understanding (Kadayat & Eika, 2020). Given that many individuals depend on this document for safety, a lack of thorough understanding could put them in substantial danger. Studies have found that individuals with low literacy levels are more likely to experience injuries in the construction field (Pejtersen & Holt, 2022). Therefore, this complex writing style can put them into a risky situation.

## ***Applying Without Support: Lack of Practical Guidance and Overwhelming Task Complexity***

Safety standards can vary widely in purposes and scope. Some might focus on general principles, while others provide more detailed instruction on specific contexts. For example, within the Machinery Directive, harmonized safety standards are typically categorized into three types: *Type A* contains fundamental standards explaining basic principles); *Type B* covers general safety aspects and protective devices; and *Type C* specifies safety requirements for specific machines or families of machines (Simmonds et al., 2017).

However, while the more general Type A and B standards clarify what must be achieved in terms of safety, they often lack a guide on how to achieve it. This includes a lack of step-by-step instruction, recommended tools, methods, or evaluation criteria. For example, a study by Björnsdóttir et al. (2022) shows that the majority of reviewed ISO standards (15 out of 18) offer no practical instructions or methodologies for carrying out risk analysis despite it being the critical component of the risk management process.

Similarly, Paques et al. (2007) also highlights the absence of standardized tools for performing risk assessments. As a result, many learners often face challenges in determining which tools best suit their specific needs. This problem also applies to companies, often forcing them to develop their own analytical methods and tools.

Another example is ISO 45001, an international standard for occupational health and safety. While this standard effectively provides a framework for organizations to manage risks and promote employee safety, it is also accused of lacking specific implementation guidelines for health sectors. Consequently, Turkish healthcare workers have been exposed to several intense risks in recent years (Yeşilgöz & Arga, 2025).

Additionally, studies also find that ISO 45001, require users to complete numerous complex safety management activities, making implementation overwhelming and difficult to handle. Many employees perceive the activities and requirements outlined in this standard as adding unnecessary bureaucracy (Podrecca et al., 2024). When employees feel this way, they are more likely to bypass or work around safety protocols (Clark et al., 2024).

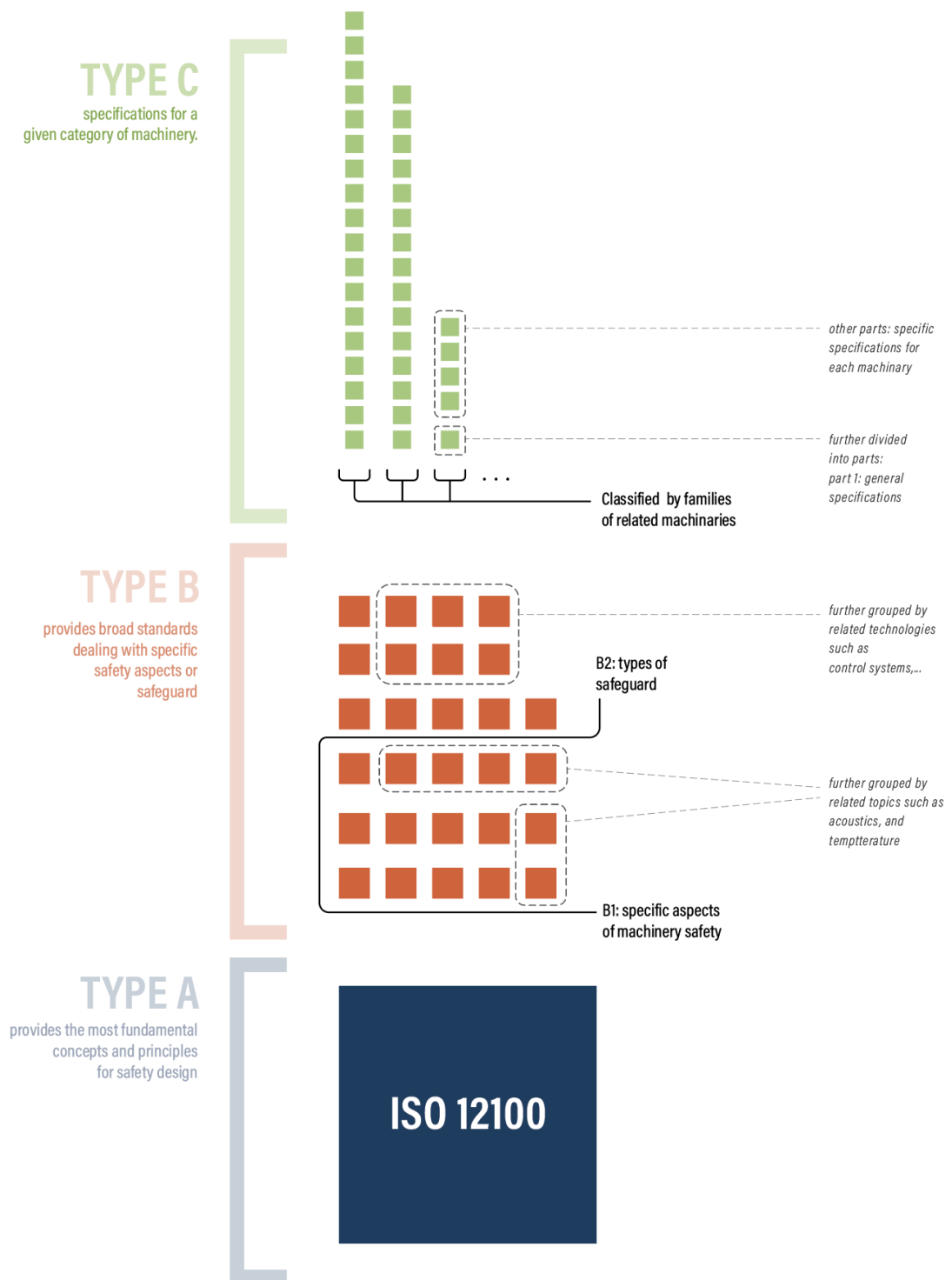


Figure 11: This figure illustrates how Type A, B, and C safety standards under the Machinery Directive are organized into groups and subgroups.

### 3.3 Prototype requirements:

The problems identified in the previous sections were gathered from multiple sources, including literature review, personal experience, and interviews with both students and the teacher of the *Safety by Design* course, as well as students from various academic backgrounds such as chemical engineering, computer science, interactive technology, and business administration. These insights were translated into a set of need statements that reflect the goals and challenges faced by both students and educators (presented in Figure 12).

Notably, we observed a strong similarity between the needs of students in the *Safety by Design* course and those from other disciplines. This could be explained by the multidisciplinary nature of the course itself, which attracts students from diverse fields such as industrial design, mechanical engineering, and psychology. As a result, the need statements are classified into two groups: students and teachers.

The need statements are formatted based on the framework suggested by Gibbons (2019), which contains three main components: 1) *a user*, 2) *a need*, and 3) *a goal*. Based on these need statements, we developed a set of product requirements, which follow a framework mentioned in the book of Mohammad Rajabalinejad (2020). The requirements contain three components: *Function or Capability + Condition + Constraints*, explaining what the product must do, in what condition, and what limitations. The requirement statements are also presented in Figure 12



Figure 12: This figure illustrates how the prototype requirements were derived from stakeholders' needs.

# 4 The Gap in Current Methods

## 4.1 Using Serious Games to Teach Safety Standards

University courses often employ various technologies and teaching methods to make learning safety standards more active and engaging. According to Rodeghiero Neto and Amaral (2024), problem-based learning, project-based learning, and gamification are the three main strategies commonly used in European safety courses to promote active learning.

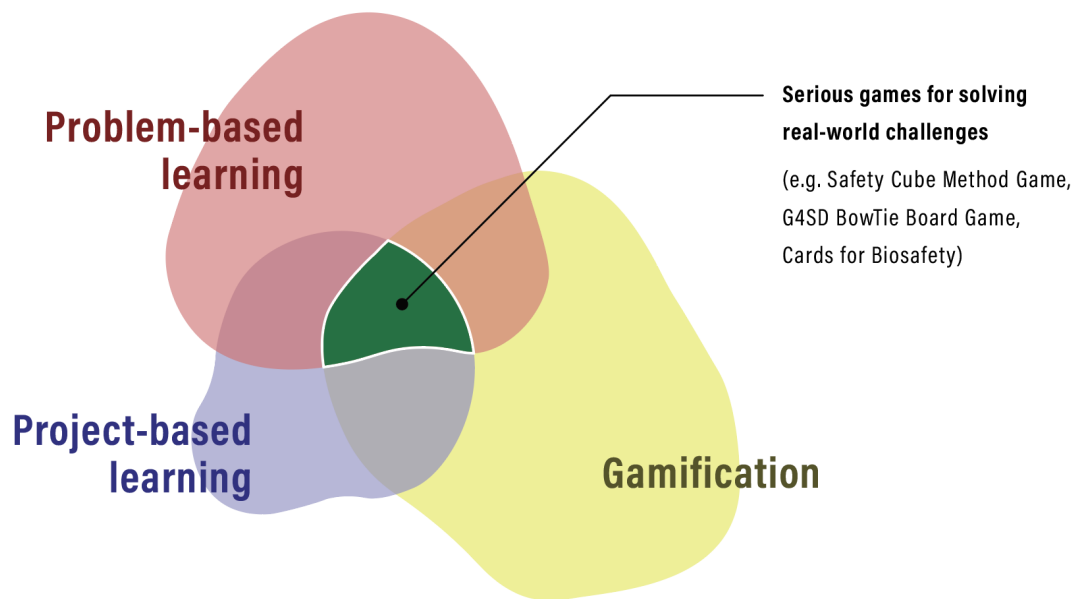


Figure 13: This figure presents the three most widely used active learning methods in European safety courses, based on data from the study by Rodeghiero Neto and Amaral (2024)

While problem-based learning emphasizes analyzing complex, open-ended problems to encourage critical and reflective thinking, project-based learning extends this by also guiding students to develop comprehensive solutions, from initial concept to completion. Finally, the most popular strategy, gamification, involves using game elements such as rewards, quests, and storylines, allowing students to learn through entertainment (Almeida, 2017; Rodeghiero Neto & Amaral, 2024).

Sometimes, these three strategies can be combined to create serious games that allow students to explore real-world challenges and create meaningful solutions in a playful and entertaining way (illustrated in Figure 13). For example, the following paragraphs introduce three serious games universities currently use. A commonality of these games is that they follow the same safety-by-design thinking process: start with real-life scenarios, help players recognize potential hazards and select suitable mitigation measures, and



sometimes end with a recap discussion to reinforce learning (Bouri et al., 2020; Rajabalinejad, 2020; Rodeghiero Neto & Amaral, 2024).

### Example 1: Safety Cube Method Game

The Safe Cube Method Game is a role-playing game where students work as a team of consultants to develop a design proposal. To do so, students need to complete 15 contracts, which are based on NEN NTA 8287:2021. Figure 14 shows the answer sheet prepared by the teacher for the 15 contracts students are expected to complete. Through this process, students briefly apply all theoretical principles of the standard, guiding students to explore key aspects of their product, such as stakeholders, components, and environmental factors, identify major risks, and propose suitable solutions. At the end of the game, each team presents their design proposal to the class, and the team that receives the most votes from the audience wins (Rajabalinejad, 2020).

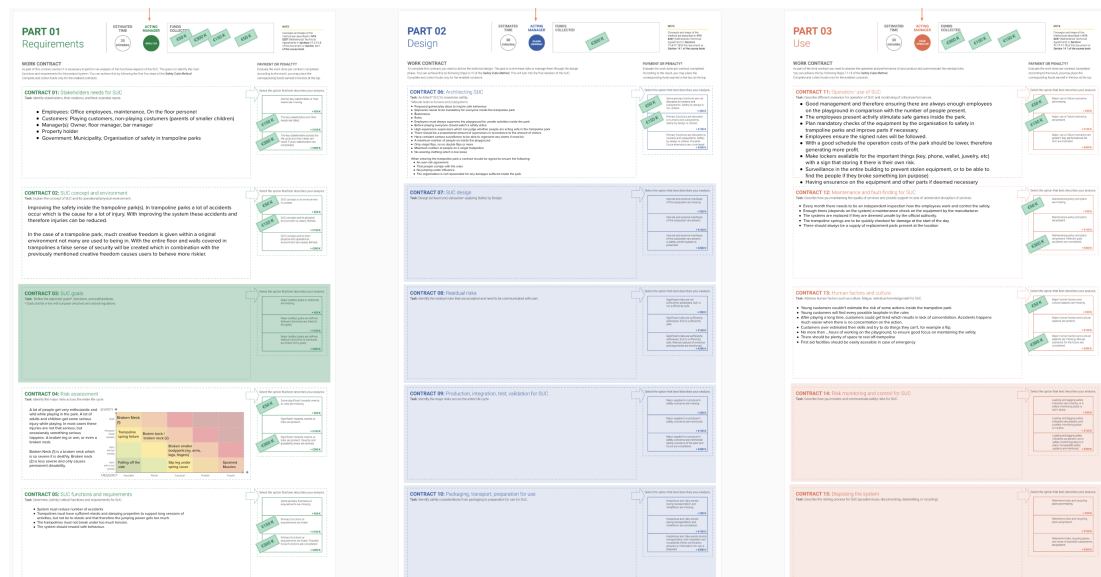


Figure 14: The figure shows the answer sheets used in the Safety Cube Method game.

### Example 2: ISO 45001 In-class Game

The second game is designed to help mechanical engineering students understand and apply the ISO 45001 standard. Students are introduced to a fictional construction company composed of several departments. Then, students are divided into teams of four, and each team is assigned to represent a specific department. Their task is to perform a detailed risk assessment and develop a questionnaire consultation system aligned with ISO 45001. After completing their work, each team presents their findings to the class through an oral presentation, facilitating discussions and interactive Q&A sessions to exchange

feedback. At the end of the session, students vote to determine the team that has delivered the most effective and insightful analysis as a winner. (Rodríguez-Martín et al., 2023).

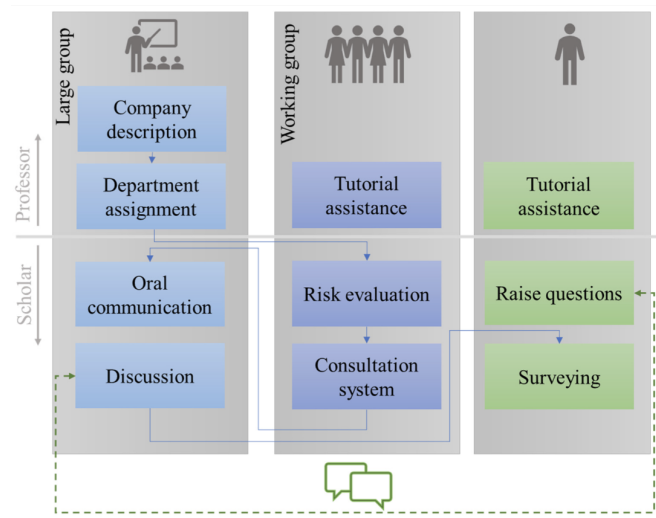


Figure 15: This figure shows the game's workflow, as presented in the paper by Rodríguez-Martín et al. (2023)

### Example 3: EnviRun- A Pedagogical Game for Learning the ISO 14001:2015

Envy Run is a board game in which players go clockwise around a board consisting of 24 boxes, each containing questions related to specific topics (See Figure 16). The dark blue boxes represent the topic of waste management; the grey boxes cover detailed knowledge of ISO 14001; the green boxes focus on general environmental knowledge and sustainable development; and the white boxes present real-life scenarios where players must collaborate to assess risks and propose appropriate technical solutions. Through this game, the designers aim to evaluate players' understanding of ISO 14001 and their general knowledge of environmental laws and regulations (Bouri et al., 2020).



Figure 16: This figure shows the game board for Envy Run.

### 4.1.1 Benefits of Serious Games

#### *Offering An Active Learning Approach*

Instead of passively reading extensive safety standards, students participating in these games are asked to actively apply their knowledge to new contexts. For example, in the Safety Cube Method Game, each team learns by manually applying tasks outlined in the standard to their chosen safety topics, such as redesigning a visual sign or developing an automated vehicle (Rajabalinejad, [2020](#)). By creating original content from their understanding and receiving peer feedback, students can engage with the material deeply. This active learning approach aligns closely with the ICAP (Interactive, Constructive, Active, and Passive) framework introduced in a study by Chi and Wylie ([2014](#)).

#### *Providing A Simple Problem-Solving Framework*

All three games embrace a problem-solving approach. With the given scenarios, the games guide participants from identifying risks and evaluating potential solutions to concluding with thoughtful reflection. By doing so, the games can connect theoretical safety knowledge to practical applications, effectively preparing students to tackle real-world challenges (Bouri et al., [2020](#); Rajabalinejad, [2020](#); Rodríguez-Martín et al., [2023](#)).

#### *Encouraging Learning through Collaboration*

All three games leverage collaboration to facilitate learning, which closely aligns with Vygotsky's principle of social constructivism. According to this theory, community is essential through the process of "making meaning", as learners can build their understanding from comparing and exchanging ideas with others (Negi, [2020](#)). These games embody this principle by encouraging students to work closely in a team and take turns to express their opinions. As a result, students can learn from each other and develop essential communication skills (Bouri et al., [2020](#); Rajabalinejad, [2020](#); Rodríguez-Martín et al., [2023](#)).

## 4.1.2 Limitations: Lack of Personalized Feedback and Guidance

### *The Absence of Individualized Feedback*

Despite their strengths, the games face several limitations, one of which is the lack of personalized feedback at the individual level. In both examples 1 and 2, most of the feedback students receive is directed toward group results, given during class presentations (Bouri et al., 2020). Therefore, the personal struggles each student encounters throughout group work are rarely addressed. This issue is further amplified in multidisciplinary settings, where individual difficulties vary widely due to their diverse background. As a result, general group feedback is often insufficient to address specific student needs.

### *Missing Personalization in Learning Support*

A second issue is that these games do not provide immediate support to bridge students' knowledge gap, which is required for effective gameplay. For instance, in example 2, students report significant difficulty in formulating effective questionnaires that both align accurately with the ISO 4500 standard and the specific scenario given in the game. For example, choosing the appropriate response scales or avoiding ambiguous wording (Bouri et al., 2020; Rajabalinejad, 2020). This issue can arise spontaneously and fall outside the teacher's expectations. As a result, although teachers might briefly answer students' questions, they were unable to provide comprehensive guidance to fully address this challenge.

### *Sacrificing Comprehensiveness for Engagement*

Sometimes, the content of safety standards is intentionally simplified or shortened to maintain engagement. After repeatedly receiving feedback that the game was too complex within two hours playtime, in the most recent year, the Safety Cube Method game reduced the number of contracts participants must complete from 15 to 8 (Rajabalinejad, 2020).

## 4.2 Generative AI in Safety Education

Generative Artificial Intelligence (GenAI) refers to a subclass of AI systems that can generate content typically created by humans, such as text, images, music, and videos (See Figure 17). Unlike traditional AI technology, which can only generate predictive outcomes, generative AI can create new content that has slightly similar patterns found in its vast training data (Harshvardhan et al., 2020).

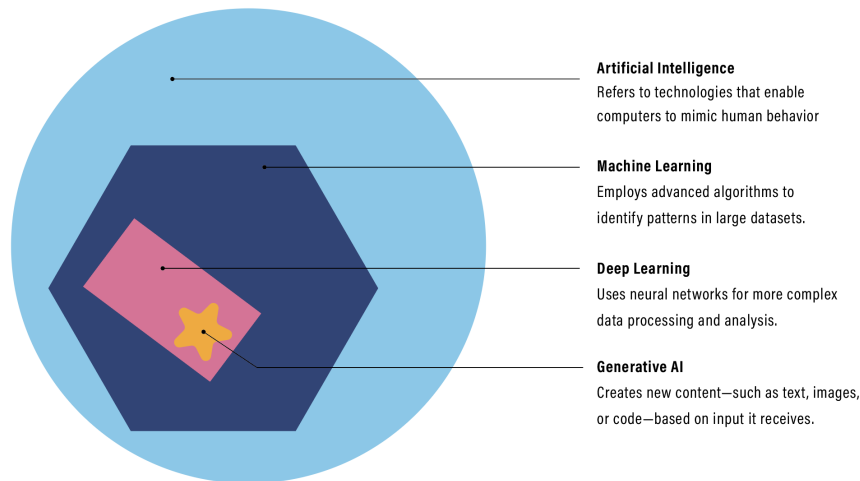


Figure 17: A (visually enhanced) comparative view of AI, Machine Learning, Deep Learning, and Generative AI proposed by Zhuhadar (2023)

### *The Rapid Rise of Generative AI in Education*

The last two years have witnessed a significant increase in the adoption of GenAI in education. A 2023 survey by Tyton Partners, which gathered data from 1,600 students across 600 U.S. institutions, found that student use of GenAI tools rose from 27% to 46% within six months (Coffey, 2023). Similar patterns in student adoption rates have been observed in multiple studies across different regions, suggesting a fast and global trend of using AI tools in education (Grove, 2024; Onatayo et al., 2024).

This increasing adoption is rooted in GenAI's ability to process vast datasets, generate meaningful insights, and, most importantly, offer context-aware assistance to enhance personalized learning experiences (Onatayo et al., 2024). AI chatbots and virtual tools like GPT-3, GPT-4, and Google Bark can help students enhance learning by answering inquiries, facilitating language acquisition, supporting research, and helping develop new skills (Bai et al., 2023). From educators' perspective, this technology can be used to create engaging simulations and immersive tours, design presentations, performance analyses, and study plans (Onatayo et al., 2024).

With these significant advancements, experts anticipate that GenAI can reshape the educational landscape in the near future (Onatayo et al., 2024). When widely integrated into classrooms, each student might have their own personal AI tutor, enabling individualized learning through human-like dialogues and interactive conversations. The role of teachers would transition into facilitators, focusing on guiding students and addressing more advanced questions (Extance, 2023).

#### 4.2.1 Practical Uses and Advantages

Although research is still limited, 2024 has seen some pioneering projects exploring how GenAI can be used to teach safety. These early works, as discussed below, emphasize the great possibility of employing this technology to enhance safety education. Figure 18 summarizes key advantages of generative AI, along with concerns that will be discussed in the following subsection.

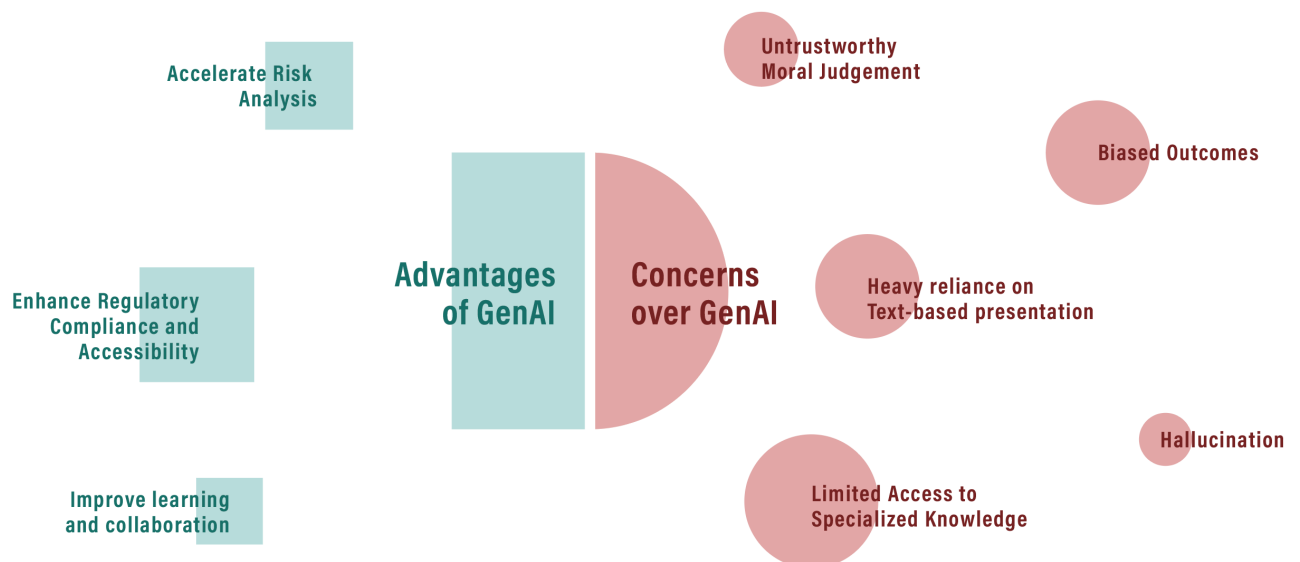


Figure 18: This figure summarizes the advantages and concerns of using generative AI in safety education.

#### ***Accelerate Risk Analysis***

Many studies have explored ChatGPT's applications in various risk analysis activities. For example, a study by Uddin et al. (2023) tested the performance of engineering students in recognizing construction hazards before and after using tools like ChatGPT. On average, the student's performance improved from below 35% to 60% after using ChatGPT. Furthermore, students have shown a high level of acceptance of this technology.

Another example is a study by Smetana et al. (2024) that applied GPT-3.5 to incident analysis in the highway construction industry. This study demonstrates that the large language model can enhance

data-driven safety analysis by identifying patterns and commonalities within the Severe Injury Reports database of Occupational Safety and Health Administration (OSHA). The findings suggest that by leveraging this technology, safety professionals can develop more effective incident prevention and intervention strategies by understanding the root causes of accidents.

Furthermore, GenAI has also been applied to classified risks, as demonstrated in a study by Mohamed Hassan et al. (2022). The researchers aimed to develop a method that could automatically classify injury narratives based on work activities, injury type, and injury severity, using the BERT model. The study reports achieving "state-of-the-art results", demonstrating AI's remarkable ability to process large volumes of injury reports and extract meaningful insights.

### ***Enhancing Regulatory Compliance and Accessibility***

The study by Liu et al. (2023) showcases the remarkable accuracy of GPT models in understanding and processing regulatory documents. In their research, they developed an automated compliance-checking system for building design specifications using GPT-based models. The results demonstrated that these models achieved 91% accuracy in interpreting and analyzing regulations. This highlights the potential of generative AI to facilitate compliance in the architecture, engineering, and construction industry.

GenAI not only comprehends safety materials but also can rewrite and reformat them effectively. Studies have shown that GenAI can rewrite materials using simple language and concise sentences, increasing the accessibility to a broader audience. For instance, healthcare organizations have started using GenAI to simplify complex safety guidelines, technical manuals, and regulatory documents for workers and patients (Howell, 2024).

### ***Improve learning and Collaboration***

Project-based learning is a widely used approach in the construction industry that challenges students to think critically by simulating real-world scenarios. Recent research has explored how ChatGPT can enhance learning and collaborate with students in this framework (Onatayo et al., 2024).

Studies indicate that ChatGPT can support students throughout the entire project life cycle, from conceptualization to implementation. For example, a study by (Saka & Chan, 2023) suggests that ChatGPT could be used to enhance incident documentation and evaluation, determine underlying causes, and propose safety measures. Additionally, it can provide guidance, generate design iterations, offer feedback, and enhance decision-making (Onatayo et al., 2024).

Furthermore, research in the construction and engineering industries suggests that using ChatGPT can enhance collaborative learning. For example, it can serve as a virtual discussion moderator, facilitating

group brainstorming and helping students explore design ideas. Studies show that this approach of collaborative learning offers significant benefits, including deeper understanding, increased self-confidence, and improved conflict resolution skills (Onatayo et al., 2024).

## 4.2.2 Limitations and criticism

Despite the growing popularity of generative AI among students, many experts have raised significant and valid concerns about its implications for education. The following paragraphs will examine the factors contributing to skepticism toward generative AI as an educational tool.

### *Limited Access to Specialized Knowledge*

One major drawback is their reduced effectiveness when providing highly accurate, domain-specific, and authoritative information—such as safety standards. (Chakraborty et al., 2023; Ling et al., 2024). For example, a study by Oviedo-Trespalacios et al. (2023) evaluated the accuracy and appropriateness of safety-related advice provided by ChatGPT. The findings raise serious concerns: experts warn that ChatGPT could produce incorrect and potentially harmful advice. Therefore, the study emphasizes the importance of expert verification and highlights the user's responsibility to critically assess AI-generated advice before acting on it.

This limitation stems from the fact that generic AI models lack direct access to specialized databases like ISO or CEN, which present all in-depth safety knowledge, best practices, and recommendations. When faced with topics outside its broad training data, these models are more likely to generate inaccurate and even confusing outcomes (Chakraborty et al., 2023; Ling et al., 2024).

### *Hallucination*

One of the most significant limitations of GenAI is hallucination, its ability to convincingly present incorrect information. However, the frequency of hallucinations varies widely across different AI models. A study by Chelli et al. (2024) evaluates the performance of GPT-4, GPT-3, and Gemini in retrieving real scientific references on a medical condition called shoulder rotator cuff pathology. While GPT-4 outperforms other models, it still shows low accuracy and generates fake references nearly 30% of the time. As a result, the authors show caution against relying on AI models as the primary tool for conducting literature reviews in healthcare and medical research.

In response, AI companies have been working to address this issue and have shown promising progress. For example, significant improvements are reported from GPT-3.5 to GPT-4. According to OpenAI's 2023 technical report, GPT-4 outperformed GPT-3.5 by approximately 20% across various topics, including history, learning, and math—based on internal adversarial factual evaluations—a test crafted to measure



the model's ability to generate accurate and truthful information. However, GPT-4's accuracy remains in the range of 70% to 80%. Therefore, the report clearly advises users to verify the information it provides (OpenAI, [2023a](#)).

### ***Heavy reliance on Text-based presentation***

ChatGPT and other AI models are constrained by their default reliance on text-based communication. Currently, generic AI models tend to structure their responses using hierarchical formats, such as nested headings, lists of bullet-point, and occasionally tables (See Figure [19](#)). While this format is suitable for general communication, it might not always be the most effective approach to explaining safety information (Abuiyada, [2018](#); Küchemann et al., [2025](#)). For example, safety communication is often delivered through standardized warning pictograms and signage, as they allow quick comprehension. Without the support of visuals, the models miss an opportunity to enhance understanding through multiple modalities — a proven strategy to improve learning and retention (Mayer, [2014](#)).

1. Simple Heading & Paragraph Style

Safety Analysis in Product Development

Safety analysis involves identifying hazards, assessing associated risks, and implementing mitigation measures. This process ensures product reliability, regulatory compliance, and user safety.

2. Bullet Point Style

Key Aspects of Safety Analysis

- Identifying potential hazards
- Evaluating risk severity and likelihood
- Implementing mitigation strategies
- Ensuring compliance with safety regulations

3. Numbered List Style

Steps in Safety Analysis

1. Define the system and its components
2. Identify potential hazards
3. Assess the risk levels
4. Implement safety measures
5. Monitor and update the system as needed

4. Table Style

Stage	Description
Hazard Identification	Recognizing potential risks in the system design
Risk Assessment	Evaluating the severity and probability of risks
Mitigation Planning	Implementing safety measures to reduce risks
Review & Monitoring	Regularly updating safety protocols

5. Flowchart Style

Safety Analysis Process

→ Identify Hazards → Assess Risks → Implement Controls → Monitor & Improve

6. Infographic Style (Icons + Text)

- 🔍 Step 1: Identify Hazards
- ⚠️ Step 2: Evaluate Risks
- ✅ Step 3: Apply Safety Measures
- 🔧 Step 4: Review & Improve

Figure 19: This figure illustrates GPT’s ability to present information in both hierarchical and non-hierarchical formats.

### ***Biased Outcomes***

Since GenAI operates solely based on patterns in its training dataset, any bias present in those data can significantly influence its outcomes. Research has highlighted that GPT 3.5 shows gender and ethnicity biases in its medical advice and symptom management recommendations for acute coronary syndrome. This bias stems from the lack of sufficient, balanced, and diverse data available in the medical field (Currie et al., [2024](#)). In the context of safety, relying on the biased output of GenAI could result in serious issues such as inaccurate diagnoses and recommendations, failure to address blind spots, and increased discrimination (Currie et al., [2024](#); Izadi & Forouzanfar, [2024](#)).

### ***Untrustworthy Moral Judgement***

Skepticism also arises when AI is tasked with making decisions and moral judgments. Ryan ([2020](#)) argues that humans should not fully trust GenAI models because they are fundamentally "a set of computational techniques", which do not have emotive states and a sense of responsibility for their actions. As a result, GenAI models cannot be held accountable for its actions as it cannot understand the impact and consequences of its recommendations. Hence, Ryan ([2020](#)) emphasized the importance of responsible practices when using GenAI, encouraging users to critically review AI outputs, especially when they involve moral judgments or ethical evaluations.

### 4.3 The Gap: Recap

Studies have shown that serious games using team-based, problem-solving techniques are among the most popular methods for teaching safety standards. Despite their benefits, these games often fail to provide personalized feedback and detailed guidance that would equip students with the additional knowledge needed to play effectively (Bouri et al., 2020; Rajabalinejad, 2020).

In the last few years, Generative AI has emerged as a promising technology capable of offering personalized learning experiences. With access to extensive knowledge bases, AI models could provide tailored feedback and answer specific questions posed by students (Onatayo et al., 2024). Therefore, this technology has increasingly been integrated into education, including the safety domain (Grove, 2024; Onatayo et al., 2024). However, Generative AI often struggles to provide coherently accurate, reliable, and domain-specific information, like safety standards, (Oviedo-Trespalcacios et al., 2023) , and thus does not fully meet the prototype requirements established in Section 3.3.

To address this gap (as illustrated in Figure 20), this thesis is conducted to explore potential solutions to present safety standard information in a **reliable, intuitive, and customized** manner. Accordingly, a diverse group of safety learners can effectively find, understand, and apply these documents in real-world settings.

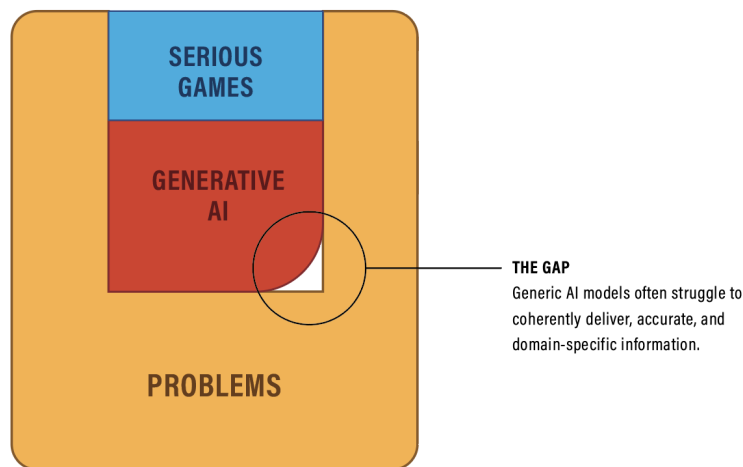


Figure 20: This figure illustrates the gap between serious games and generic AI models in relation to the identified problems



# 5 The Exploration Journey

The exploration journey to identify a promising solution for this thesis can be summarized in three main phases, as illustrated in Figure 21:

**Stage 1: Ideation and Early Brainstorming:** In the beginning, we started analyzing and grouping related requirements outlined in Section 3.3 together to identify key requirements that our prototype needed to manage. Then, we employed different brainstorming techniques, such as Mind Mapping and Forced Association, to generate a wide range of potential ideas. The primary goal during this stage was to rapidly explore a number of possible concepts without any judgment or restriction.

**Stage 2: Concept Development and Evaluation Against Requirements:** Following several weeks of brainstorming both independently and collaboratively with teachers during weekly meetings, we identified four concepts that both sparked our interest and had the most potential based on our judgments. We then evaluated these concepts against the original requirements in Section 3.3 to determine their potential fit. This evaluation indicated that a customized AI chatbot combined with visual aids offered the most potential for effectively addressing all the requirements.

**Stage 3: In-depth Research through Literature Review:** At this stage, we validated our idea for the potential solution by conducting an extensive literature review. During this process, we examined existing studies and relevant projects related to safety visualization, customized AI technology, and their combination to better understand prior work and theoretical foundations. These findings provided further evidence that multimodal generative AI with customized knowledge and visualization holds significant potential for addressing challenges in safety education.

The following sections describe this journey more deeply, including the results obtained along the way.

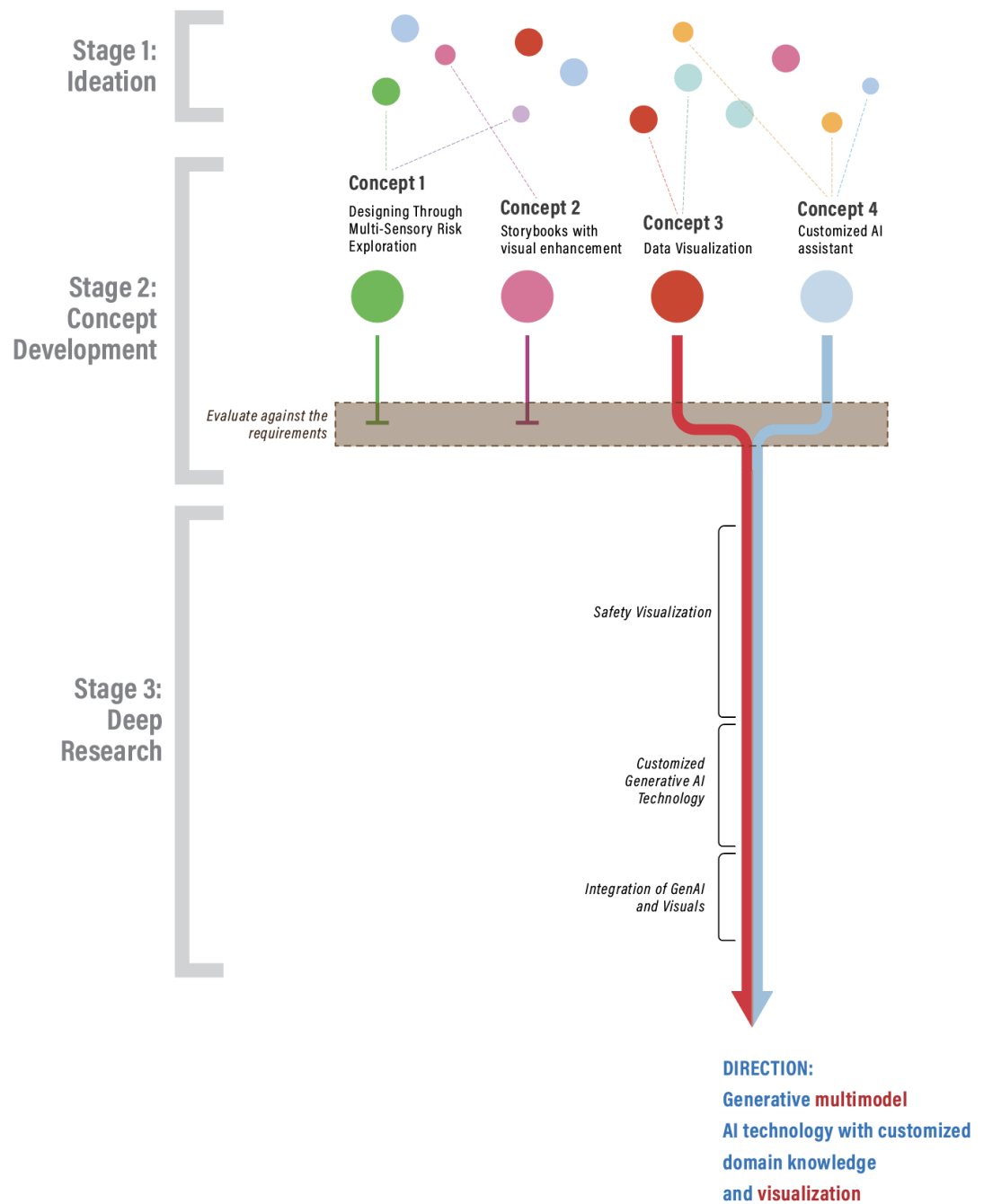


Figure 21: This figure illustrates the exploration process, showing how, after stage two, Concepts 3 and 4 were merged and further developed through in-depth research—ultimately leading to the confirmed direction of the thesis.

## 5.1 Stage 1: Ideation and Early Brainstorming:

The journey began by grouping the identified requirements outlined in Section 3.3 into similar themes, allowing us to pinpoint the essential requirements of the prototype. As a result, we ended up with three main themes: (1) sharing standard-related information, (2) organizing findings, and (3) providing evaluation, as illustrated in Figure 22.

Recognizing that the majority of requirements are about presenting and sharing standard information, we created a mind map to brainstorm teaching methods, technologies, or means that are designed to deliver information through different human senses (See Figure 23). For example, booklets communicate with humans through sight, while video tutorials combine visuals and audio.

Next, we applied a technique known as *"forced associations,"* where we created new ideas by combining two or three seemingly unrelated words together. We had two groups of words: first is a list of the technologies/ learning methods that we had identified through the mindmap; second is a list of verbs, for example, *"make"*, *"destroy"*, and *"reflect"*.

The aim was to pair words from groups 1 and 2 to create new ideas based on the connection between two words. For example, from 'AI' and 'cook,' we generated a metaphorical idea that considers AI as a private chef who "cooks" user inputs into a "dish." After that, users can "taste" (review) and refine it based on their preferences. This creative exercise sparked numerous ideas, some of which are depicted in Figure 24.



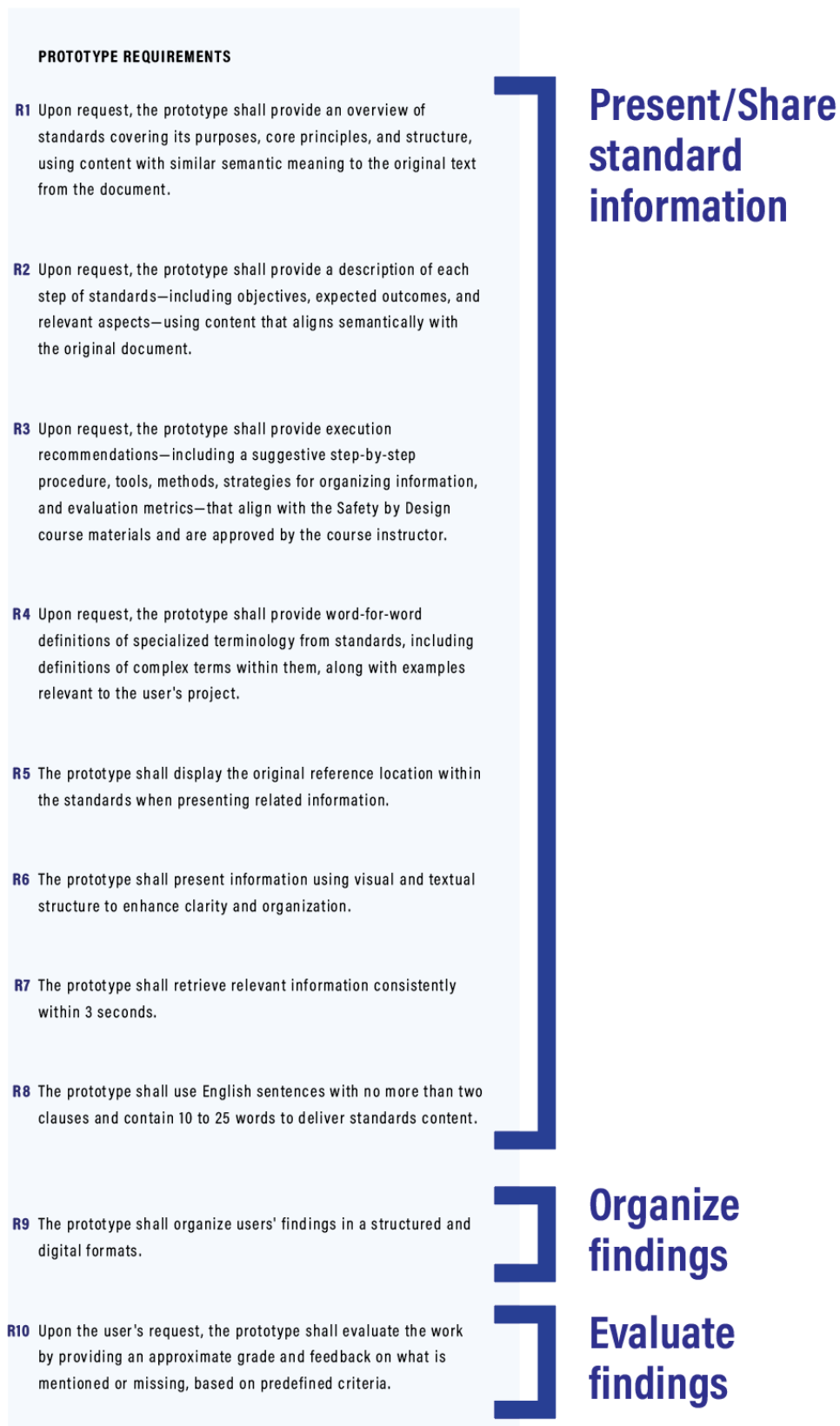


Figure 22: This figure presents the results of the requirement analysis, grouping similar requirements into common themes.

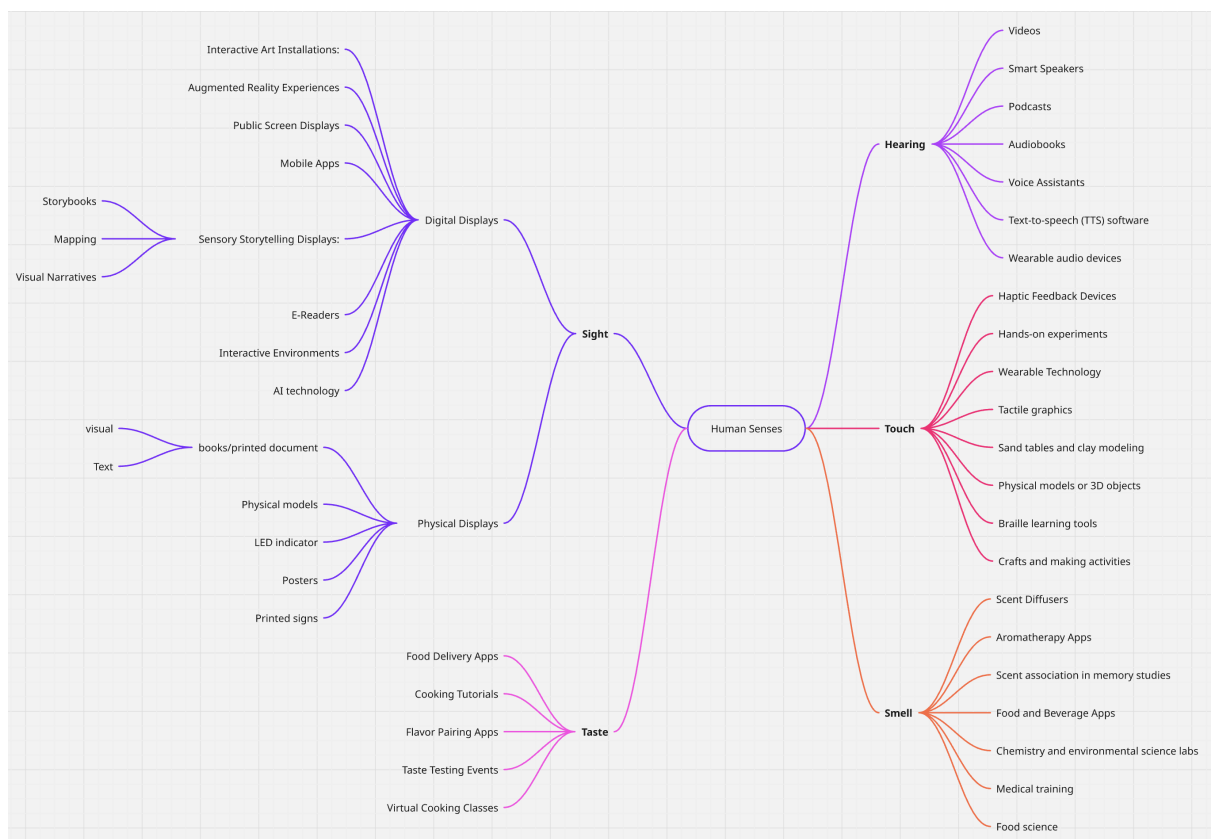
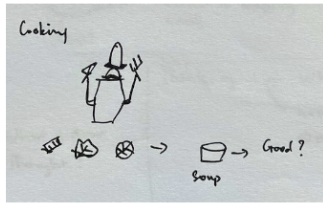
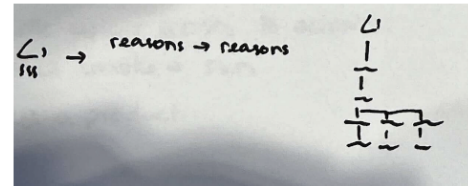


Figure 23: This mind map shows various technologies linked to human senses, exploring how information can be presented through different senses.



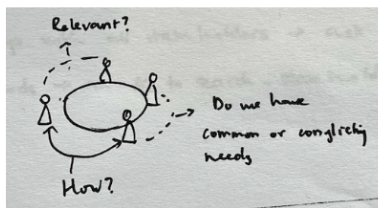
#### AI as a Chef

The AI acts like a chef, "cook" ideas/ inputs from users. Users taste/review the result and refine it through iteration.



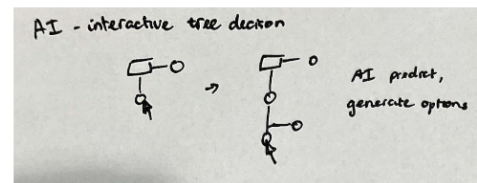
#### Smell-Based Reasoning

Users detect problems through multisensory cues, especially smell, and come up with a reason.



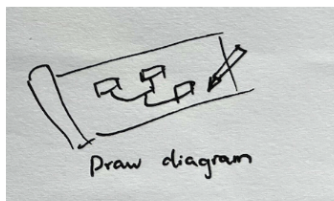
#### Group Dialogue with AI

A group of users discusses their shared or conflicting needs with an AI facilitator to reach a solution together.



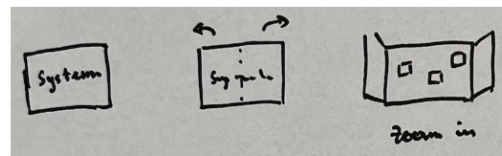
#### AI-Generated Tree Diagrams

The AI can expand the tree diagrams with example answers.



#### User-Drawn Risk Diagrams

Users create their own risk analysis diagrams in class



#### System-in-System Visualization

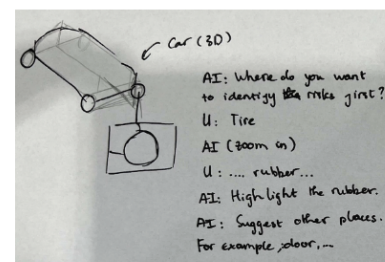
This idea visualizes how smaller systems components within larger ones, highlighting hierarchical relationships in a visual format.

Sensory experience helps people to:

- Detect the risk
- Remember the risk → brain good for training
- Control emotion: create VR experience
- Link sensor senses to action: smell smoke → run
- Test a product

#### Sensory and Risk Perception

The concept explores how sensory experiences, like touch or sound, can help users better understand and assess risks.



#### AI with 3D Model Interaction

Users interact with AI through voice while manipulating a 3D model, creating a more immersive and responsive learning experience.

Figure 24: This figure shows initial ideas generated using the force association method.

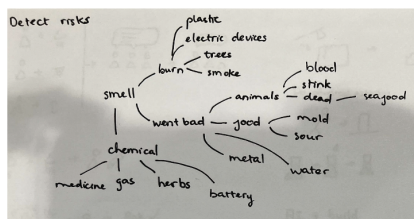
## 5.2 Stage 2: Concept Development and Evaluation Against Requirements

After several weeks of brainstorming both alone and together with the teacher, the four most engaging concepts emerged:

### Concept 1: Designing Through Multi-Sensory Risk Exploration

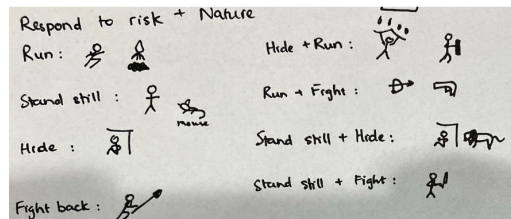
The first concept is to create a multi-sensory design activity in a classroom. Students will interact with broken physical items, such as a burnt phone, using their senses to identify potential issues and come up with solutions inspired by how humans respond to danger in nature, like hiding or running.

This concept was shaped by two inspirations. The first is the natural human ability to perceive risk through sensory cues (See Figure 25). For example, many people instinctively sniff their food to check if it has gone bad. The second inspiration is the ways humans respond to threats in nature, sketched in Figure 25. For instance, when faced with fear, human instinct often triggers a freezing response.



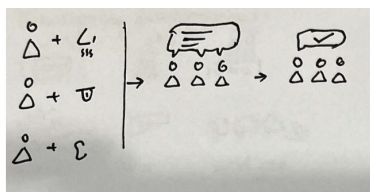
**Senses mindmap:**

The map shows in what situation human use smell to detect problems



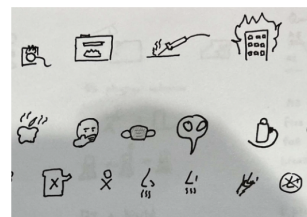
### Human Responses to Risks:

Humans and other animals might react differently in risky situations



**Multisensory activities:**

An activity where each student can identify a problem through different sense and then discuss for the solutions



**Smell-Icons:**

**Smell-Icons:**  
These sketches show how smell can be communicated visually

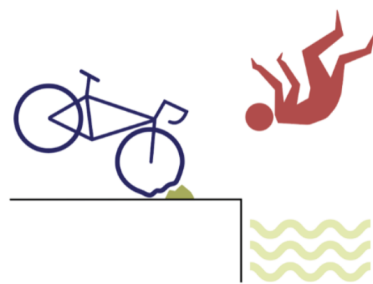
Figure 25: This figure shows sketches related to Concept 1, focusing on the connection between senses and safety.

The potential benefit of this concept is that it moves students away from complex risk assessment methods or technical terminologies. Instead, students just rely on their natural instincts and responses to create safety and solve problems. This way could create an immediate and relatable experience for all students, making risk perception more accessible.

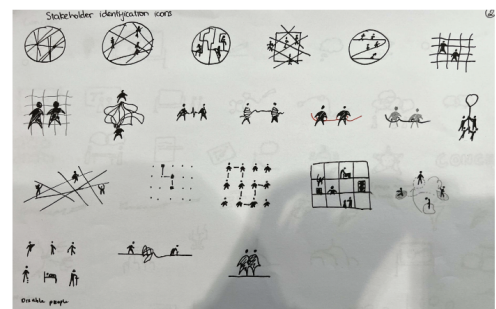
## Concept 2: Storybooks with visual enhancement

The second concept focuses on developing an engaging booklet that combines storytelling with illustration to present the core content of safety standards. Rather than relying on dense text, this booklet would use minimal text and clear visuals to convey information effectively.

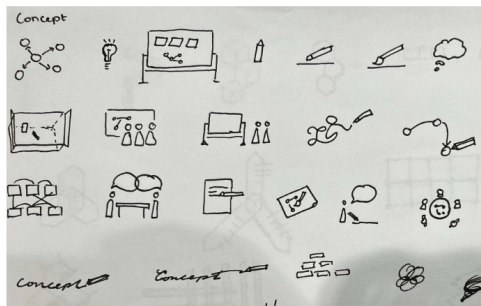
Each chapter would be structured around a story to help users grasp abstract ideas. For example, picture A in Figure 26 depicts a bike accident, providing a concrete scenario for users to analyze. The color code represents safety elements, such as red for humans, blue for the system/product, and green for the environment.



A



B



C

Figure 26: This figure shows materials for developing storybooks: A illustrates a bike accident, B presents stakeholder icons, and C visualizes the concept development process.

Each chapter would incorporate an icon-based system to visually represent essential concepts. For example, picture B in Figure 26 shows different icons to present stakeholders and their connection, while picture C shows how a concept can be made through different activities such as discussion, drawing, and constructing from building blocks.

At the end of each chapter, an answer sheet would guide users through their own projects. This section would include leading questions to help them apply the safety concepts.

### Concept 3: Data Visualization

The third concept focuses on developing a set of customized visualizations specifically designed to organize the results of risk analysis. To ensure consistency across different visualizations, a set of visual rules is established. For instance, human factors will always be represented by a red circle, a system by a blue rectangle, and the environment by a green diamond (See picture A in Figure 27). These elements will be consistently applied throughout various visualizations, as illustrated in pictures B, C, and D. These visualizations can serve multiple purposes, such as organizing information, showing user progress, and demonstrating the key concepts of safety standards.

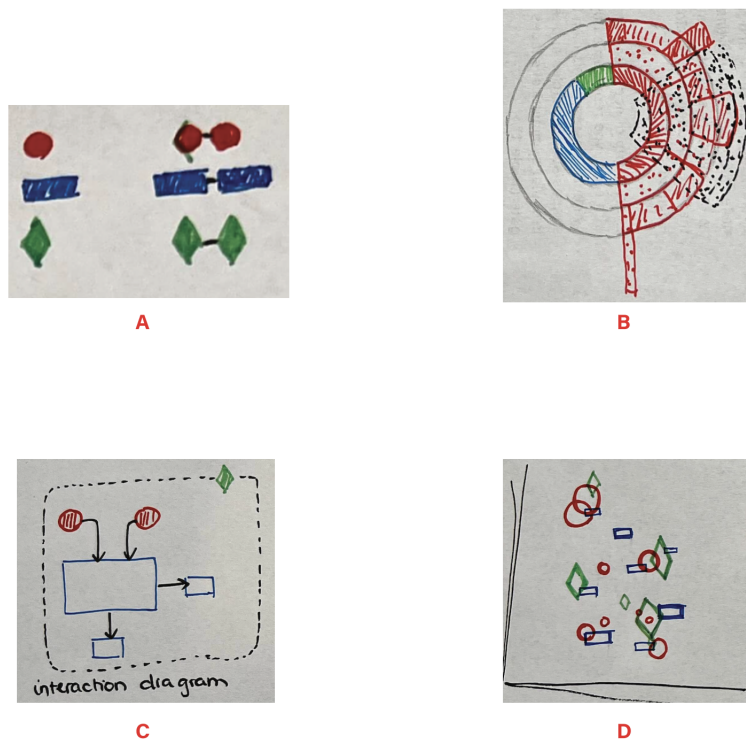


Figure 27: This sketch shows how elements of the human, safety system, and environment can be represented using red circles, blue rectangles, and green diamonds. These shapes can also be applied in various data visualizations to convey related information.

### Concept 4: Customized AI assistant

The fourth concept is creating an AI assistant in a classroom, specifically designed with specialized knowledge of safety standards. This assistant could engage with students in various. For instance, the AI could function as a tutor, providing targeted information, assigning relevant tasks, and offering feedback based on students' responses (See Figure 28).

Beyond this, the AI could even be a competitor, forcing students to defend their reasoning and test their understanding through argumentation. Another potential role is that of a thought guide, leading students through each step of a design task while encouraging reflective thinking and decision-making.



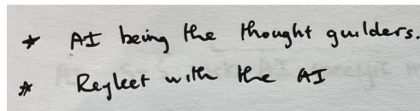
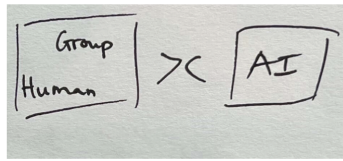


Figure 28: This figure presents initial sketches exploring ideas of how AI could interact with humans

### **Evaluation Against Requirements**

Based on our evaluation against the defined requirements, the concept of a customized AI assistant demonstrates strong potential to meet all criteria. This is primarily due to its ability to rely on domain-specific information provided by developers. As a result, the AI assistants are more likely to produce accurate and reliable outputs while minimizing the risk of hallucination.

However, as outlined in Section 4.2.2, generative AI still presents a limitation of presenting information in lengthy and text-heavy responses. This makes the information difficult to interpret, particularly for visually oriented users. Therefore, we came up with the idea of **combining customized generative AI and data visualizations (concept 3)** to present complex content in a more accessible format. This is because visualization can effectively present information, such as procedures and data structures, in a concise and easy-to-understand manner. Moreover, given the AI's capacity to convert text into visuals, it has the potential to generate customized visualizations tailored to user needs.

Table 1: This table presents the evaluation of different concepts based on the initial requirements.

<b>Requirements</b>	<b>Concept 1: Designing Through Multi-Sensory Risk Exploration</b>	<b>Concept 2: Storybooks with visual enhancement</b>	<b>Concept 3: Data Visualization</b>	<b>Concept 4: Customized AI assistant</b>
Upon request, the prototype shall provide an overview of standards covering its purposes, core principles, and structure, using content with similar semantic meaning to the original text from the document.	(– –), The concept struggles to convey complex ideas effectively. Students still dependent on text for deeper understanding.	(+) The storybook delivers information through multimedia formats, but it lacks innovation and adaptability.	(++) A simple diagram or visualization can effectively communicate the method overview.	(++) AI can be responsive and provide information through textual form
Upon request, the prototype shall provide in-depth content of standards—including objectives, expected outcomes, and relevant aspects—using content that aligns semantically with the original document.	(++) It can transform risk assessment and design exercises into a more engaging learning experience.	(+) This idea can provide in-depth content through text, but falls short when it comes to visual representation. Plus, the idea is too similar to safety standards	(– –) Using only visuals cannot deliver very complex and exact abstract ideas.	(++) It can provide in-depth content from its knowledge base
Upon request, the prototype shall provide execution recommendations—including a suggestive step-by-step procedure, tools, methods, strategies for organizing information, and evaluation metrics—that align with the Safety by Design course materials and are approved by the course instructor.	(++) Students can base their instinct to detect risks and brainstorm ideas to eliminate them.	(+) The execution recommendation can be explained well by visuals, such as diagrams.	(– –) Using only visuals cannot deliver very complex and exact abstract ideas. It might be able to visualize execution plan through diagrams	(++) Material of Safety by design course can be added to AI's knowledge base
Upon request, the prototype shall provide word-for-word definitions of specialized terminology from standards, including definitions of complex terms within them, along with examples relevant to the user's project.	(–) It cannot provide in-depth and complex theoretical ideas.	(– –) It can provide word-for-word definitions but no other benefits	(– –) It cannot provide in-depth and complex theoretical ideas.	(++) It can give personal examples together with the definitions
The prototype shall display the original reference location within the standards when presenting related information.	(++) Short information can be put in a tutorial	(++) The storybook can refer to the original standard	(++) Visualization can demonstrate location well, like map	(+) AI can provide this information from knowledge base



Requirements	Concept 1: Designing Through Multi-Sensory Risk Exploration	Concept 2: Storybooks with visual enhancement	Concept 3: Data Visualization	Concept 4: Customized AI assistant
The prototype shall present information using visual and textual structure to enhance clarity and organization.	(-- ) This concept does not address this requirement	(++) The story can organize information through story format, and visuals help to clarify	(++) Visualization can do the best in structure information in a clear manner	(+) The AI can only present information through textual organization
The prototype shall retrieve relevant information consistently within 3 seconds.	(-- ) This concept does not address this requirement	(-- ) this concept is not responsive	(-- ) This concept does not address this requirement	(++) AI can retrieve information fast
The prototype shall use English sentences with no more than two clauses and contain 10 to 25 words to deliver standard content.	(++) The instruction and discussion between students are in English	(++) The storybook can be written in simple English sentences	(+) the text in visualization can be written in English	(++) It can adapt communication style
The prototype shall organize users' findings in a structured and digital format.	(-- ) This concept does not address this requirement	(-) Students need to organise their own findings, the book can just give an answer sheet for their response. Plus, it is not in digital form.	(++) Interactive Visualization can organise finding in digital format	(++) AI can summarise and rewrite students findings
Upon the user's request, the prototype shall evaluate the work by providing an approximate grade and feedback on what is mentioned or missing, based on predefined criteria.	(-- ) This concept does not address this requirement	(-) no evaluation feature	(-- ) This concept does not address this requirement	(+) AI can provide suggestions
<b>Total</b>	<b>4</b>	<b>5</b>	<b>5</b>	<b>10</b>

## 5.3 Stage 3: In-depth Research through Literature Review

After identifying the promising idea of combining customized AI technology and visualization, we conducted a literature review to further explore this idea and confirm its potential. This section presents literature research on three topics: Safety visualizations, customized AI technology, and the integration of generative AI and visuals.

### 5.3.1 Safety Visualization:

Visuals have long been used as a powerful communication tool to define, explain, and represent safety information (Le Coze & Reiman, 2023). By leveraging variables such as color, shape, size, texture, position, and orientation, visuals can make complex theoretical concepts, safety requirements, and procedures more accessible, memorable, and enjoyable (Pieters & Wedel, 2004; Scharf, 2017; Shepard, 1967).

These visuals take various forms, including posters, warning signs, diagrams, drawings, maps, and more. The following paragraphs focus on some of these most widely used and impactful visual forms in safety communication.

#### *Safety Signage*

In response to the increasing dangers of early 20th-century industrial workplaces, the need for a clear system to communicate hazards emerged (Smith, 2013). This led to the introduction of the first national safety signage standard, ASA Z35.1, in 1941, which established standardized formats for various safety signs (See Figure 29). Since then, nearly a century later, safety signage has remained essential and continues to evolve. For instance, ISO 7010, an international standard, introduced new signs as recently as 2020, reflecting the ongoing development of visual safety communication (International Organization for Standardization, 2020).

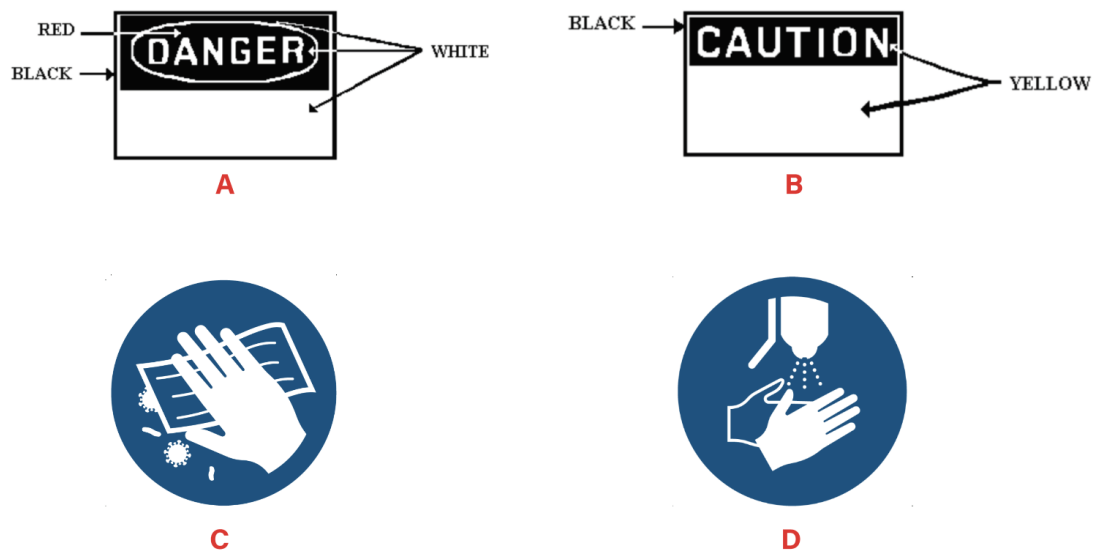


Figure 29: This figure shows four safety signs: A and B present the standard "danger" and "caution" signs introduced in ASA Z35.1; C and D show "Disinfect Surface" and "Disinfect your hand" signs introduced in ISO 7010

## Posters and Infographics

Posters and infographics have been long used to communicate more detailed and complex messages, such as safety practices. The use of posters in the Netherlands can be traced back to the early 20th century (See Figure 30). Initially, safety posters relied on fear-based messaging and emphasized family values. Over time, particularly from the 1970s onward, the approach shifted to non-moralistic. Jacques Castan's 1962 poster series exemplifies this transition, turning technical safety information into accessible and humorous designs to promote safer behavior (Le Coze & Reiman, 2023, pp. 21-32).



Figure 30: The left picture displays the first Dutch safety poster, *Why Did You Not Wear a Cap Like Me?*, reminding female workers of the dangers of loose hair near machinery. The right picture is in the poster series of Jacques Castan, *The Use of Dosimeter Films and Pens* (1962), using humorous methods to educate people to follow the safety protocol.

## Diagrams used to represent and analyze risks

In the domain of safety, visuals can also function as cognitive aids for analyzing risks by helping specialists process complex safety information more systematically. A well-known example is Fault Trees, a diagram that uses both text and visual elements such as nodes and links to identify, organize, and analyze system failures.

Fault Tree was first introduced in 1962 by H.A. Watson. The diagram has a tree-like structure, starting with one single failure placed centrally at the top. From there, several branches expand symmetrically downward;

each branch represents the causal propagation of the top failure. At the bottom, the circles represent the root causes (Patil et al., 2013) (See Figure 31).

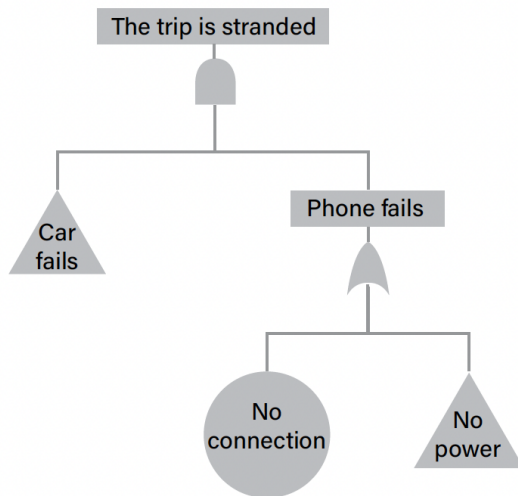


Figure 31: This is an example of a fault tree diagram, showing that the trip becomes stranded when both the car and the phone fail. The phone fails if there is either no connection or no power.

Another example is the Bow-tie diagram, which effectively uses visuals and text to visualize the relationship between a potential hazard and its causes and consequences (See Figure 32). The diagram uses lines, nodes, and spatial positioning to convey the sequential flow of triggers, barriers, a top event, and consequences (Brazier, 2017).

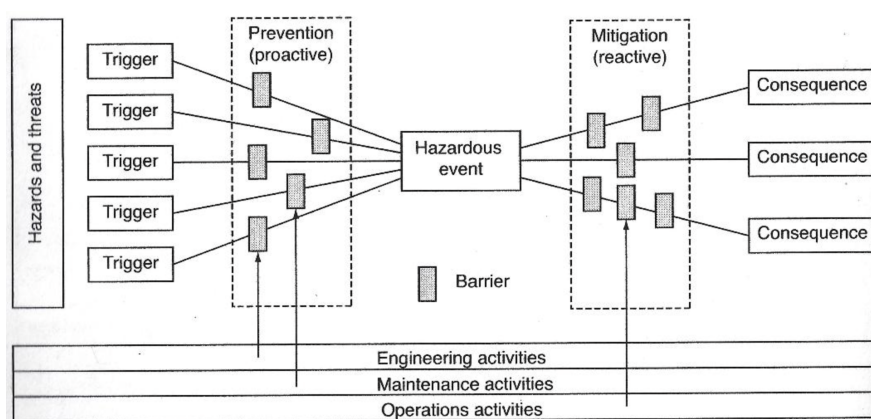


Figure 32: This diagram is an example of a Bowtie diagram, with the trigger event on the left, the central risk in the middle, and the potential consequences on the right.

# ***Benefits of visuals as information presenting method***

## ***Suitable for describing spatial structures and locations***

In 1980, Bartram conducted research to study how effective different information-presenting methods are for humans. Bartram showed 32 undergraduates the same information about bus routes but in 4 different ways: (a) as a traditional road map, (b) as a schematic map, (c) as a set of lists of bus stops in sequential order, and (d) as a set of lists of bus stops in alphabetical order. The participants needed to figure out which buses could be used to get from the 1st to the 2nd location (Bartram, [1980](#)).

Notably, the results found that participants could complete the given tasks much faster with conventional road maps and schematic maps than with the lists. Additionally, schematic maps are significantly more effective than the other maps. Bartram interpreted his findings to suggest that when information on spatial structure and location is conveyed visually, humans can process this information with little cognitive load, leading to shorter processing speed (Bartram, [1980](#)).

## ***Bridging Literacy, Age, and Educational Gaps with Visuals***

A study by Houts et al. ([2006](#)) reports that adding pictures to health instructions benefits all patients, particularly those with low literacy skills. This is because these patients can process information not only through text but also through visuals, allowing them to engage with both formats. This finding aligns with the core principles of multimedia learning (Mayer, [2014](#)). Similarly, Dretzke ([1993](#))'s research shows that mnemonic illustrations enhance learning across all age groups, including individuals between 17 and 84, supporting the Dual Coding Theory proposed in 1986 (Shepard, [1967](#)). Lastly, a study by A highlights the effectiveness of dynamic visuals for learners with limited prior knowledge. In particular, the study shows that these learners can understand the operation of pumps most effectively when text is supplemented with animated simulations, such as videos.

## ***Superior Processing Speed***

Many studies have shown that human brains process visuals significantly faster than textual content. In fact, visuals can be processed up to ten thousand times faster than text-based or verbal information (Orban, [2001](#); Scharf, [2017](#); Tong, [2003](#)). This ability can be explained by the human dependency on visual cues for survival throughout long evolution. Essential tasks, like recognizing faces, reading body language, spotting predators, and identifying food sources, require a highly efficient visual processing system (Ware, [2004](#)).

Compared to experience with texts, humans have only practiced writing and reading for a few thousand years.

Additionally, reading is not an innate ability but a skill that requires consistent practice over time to develop mastery (Johnson, 2014). As a result, some individuals become more proficient at reading than others, making text-based communication less universally accessible than visuals.

### ***Engagement and Attention***

The research of Pieters and Wedel (2004) studied the effect of graphics and texts on capturing attention, hoping to apply these findings to enhance advertising strategies. The study analyzed 1363 print advertisements on 3600 consumers and used infrared eye-tracking technology to measure their reactions. The study concluded that pictorial information is remarkably superficial at capturing attention compared to texts. However, textual information tends to hold attention longer, likely because users require more time to read and process the content.

In the book of Ware (2004), this phenomenon is explained by the fact that the human brain has evolved to prioritize visual information for survival. Quickly recognizing visual cues is essential for detecting threats, locating food, and interpreting social signals.

## ***Limitations of Visuals***

### ***Limiting in delivering precise abstract concepts.***

Even though visuals can effectively deliver spatial structures, locations, and physical appearance with a high level of detail, they often fall short in representing logical conditions, complex procedures, and abstract concepts (Ware, 2004). This statement is supported by a report of Ross (n.d.) , which presents the limitations of safety signs: While symbols can effectively portray the hazards and injury, they struggle at showing the severity, probability of the injury, and how to avoid the hazards.

Similarly, a study by Wogalter et al. (2006) also mentions that one simple symbol can rarely inform a complete procedure or "if-then" instruction. Therefore, using visuals alone could lead to risks of oversimplifying safety messages, omitting important context knowledge for users to fully understand, and reducing the depth of data analysis. The study suggests that conveying a complete instructional and procedural narrative requires the use of multiple symbols alongside supplementary text.

### ***Being challenging to produce and customize.***

Numerous studies have shown that factors such as cultural background, intentions and expectations, level of expertise, emotional states, and attitudes can significantly influence how viewers interpret visuals (Bagagiolo et al., [2019](#); Chan et al., [2009](#); Holsanova, [2014](#)). Due to these complexities, testing the comprehensibility of visuals requires a strict and extensive evaluation process. For example, according to standard ISO 9186, all safety symbols must pass two rounds of comprehensive testing before being published. The first one is the *Comprehensibility Judgment Test*, where symbols are presented with intended meaning, and participants need to assess how easily they can be understood (International Standard Organization, [2001](#)).

The second, the *Comprehension Test*, evaluates whether participants can correctly interpret the symbols without any explanation. Participants are asked to write down their answers on a paper. Each test must involve at least 100 participants from at least two countries to ensure reliability and cross-cultural validity. The threshold of the tests is around 70% to 85%, depending on the countries. If the symbols fail the tests or any changes in the symbols' design are made, the new symbols must retake the evaluation process from the beginning. (International Standard Organization, [2001](#)).



### 5.3.2 Customized Generative AI Technology

Compared to safety visualizations, customized generative AI is supported by a relatively limited literature, largely due to its recent emergence. However, there have been promising examples in the literature that demonstrate its application in delivering domain-specific information, suggesting potential for further development and exploration.

Customized AI technology offers a promising solution to improve AI reliability by extending their knowledge base through Transfer Learning. This technique uses domain-specific data in the training process, enabling the AI model to combine general and specialized knowledge. As a result, the chatbot becomes better equipped to address knowledge gaps and adapt to the needs of a specific field (Izadi & Forouzanfar, 2024).

#### *Example 1: Genie on Demand*

Recent studies have explored the potential of customized AI chatbots in education and shown some promising results. For example, in the study of Hakim et al. (2024), a customized knowledge AI chatbot called Genie on Demand was developed to assist electrical engineering students in applying safety knowledge in a lab setting. The results indicate that such AI tools can improve overall learning performance, boost self-efficacy, and increase technological acceptance among students.

#### *Example 2: Samantha*

Similarly, Samantha is a customized AI chatbot specifically designed to train users in preventing workplace hazards in the mining industry (Contreras Aguilar et al., 2024). The chatbot was fine-tuned with company protocols, procedures, and training manuals to answer questions about operational risk prevention. The study highlighted several advantages of using Samantha. For example, it could provide more precise and industry-specific safety information than general AI models like ChatGPT. Additionally, it enhanced user engagement, improved safety awareness, and received high satisfaction ratings from users.

#### *Example 3: Customised AI in surgical education*

Unlike systems such as Samantha and Genie on Demand, which primarily focus on providing educational information, a study by Sathe et al. (2024) demonstrates how a customized AI chatbot can be trained to perform specific tasks following domain-specific protocols within surgical education. Specifically, the study introduced a customized chatbot designed to support various tasks, including thematic analysis, survey creation, curriculum planning, and drafting research protocol tailored to the institutions' guidelines.

Additionally, a patient education chatbot is developed to offer post-support for individuals recovering from bariatric surgery.

### 5.3.3 Integration of GenAI and Visuals:

The combined usage of visual tools and generative AI remains a relatively new area of research. For example, a recent study by An et al. (2025), published in 2025, explored this intersection by evaluating the impact of generative AI in improving student performance in mindmap-based collaborative learning environments. Their findings reveal that the AI-supported group significantly outperformed those using only the traditional mind map approach in several tasks, including lesson planning and idea generation, teaching material development, and feedback provision. This study suggests a new and promising direction for enhancing education through AI and visual integration.

#### *Example 1: Miro AI Mindmap*

Not only in research but in the industry, real-world applications that combine AI and visual tools have been put into practice. The first example is Miro's automatically AI-generated mind map. Miro is a popular collaborative digital whiteboard application. In 2024, the Miro AI was introduced, which contains several AI-related features, including the ability to automatically expand a single chosen topic into an entire mind map with multiple AI-generated subtopics (See Figure 33) (Miro, n.d.). This functionality allows learners to quickly create a visual structure for their study content, facilitate ideation, and organize thoughts. Such automatically generated maps demonstrate how generative AI models can produce meaningful visual structures, offering new ways to support learning and collaboration.

#### *Example 2: AI chatbot assistant in Make.io*

Another example is the AI chatbot assistant in Make.io. Make.io is a visual platform that enables users to create, build, and automate workflows using node-link diagrams. As can be seen in Figure 34 each node represents a specific app or service, while the connecting links illustrate the flow of data between them. Users can construct their own workflows by simply dragging and dropping nodes into the canvas, creating customized automation paths. Recently, Make.io introduced a chatbot assistant designed to enhance user experience. This assistant can explain how different nodes function, assist with troubleshooting, and even generate entire workflows based on user input, making the experience more accessible and intuitive.

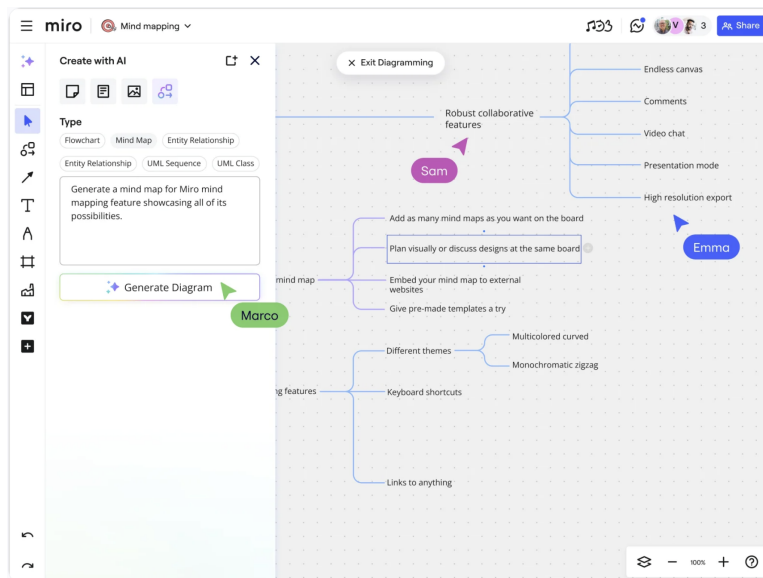


Figure 33: This figure shows the interface of Miro AI, which can generate diagrams based on user input.

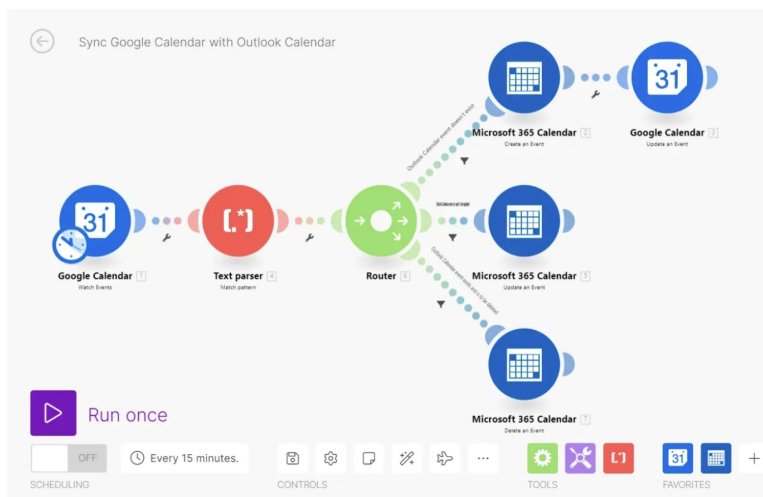


Figure 34: The figure illustrates a workflow created by Make.io, with the support of a chatbot that can assist in building and refining the workflow.

## 5.4 Final thoughts: The promising direction for the thesis.

All the academic literature gathered so far, from the use of visual tools in safety education to the advantages of customized generative AI and the integration of AI and visuals, shows that each technology offers unique strengths in addressing challenges within safety education. Customized AI technologies allow users to access domain-specific knowledge while still leveraging the broad capabilities of generative AI, such as answering questions, simplifying content, explaining concepts, and generating ideas. At the same time, visual tools offer an effective way to organize information, support quick information retrieval, and convey complex messages in a clear and concise manner.

These strengths are even complementary: while visualizations can address the limitations of generative AI's text-heavy outputs, in return, generative AI can generate quick and on-demand visual content to support users, especially those who do not have the skills to create effective visualization. Therefore, the combination of these two technologies not only enhances their individual benefits but also helps reduce their limitations.

These insights form the foundation of our approach to addressing the challenges identified in safety education: exploring how generative multimodal AI technology with customized domain knowledge and visualization can support students in learning safety standards. This approach opens up possibilities for more personalized, reliable, and intuitive educational experiences.

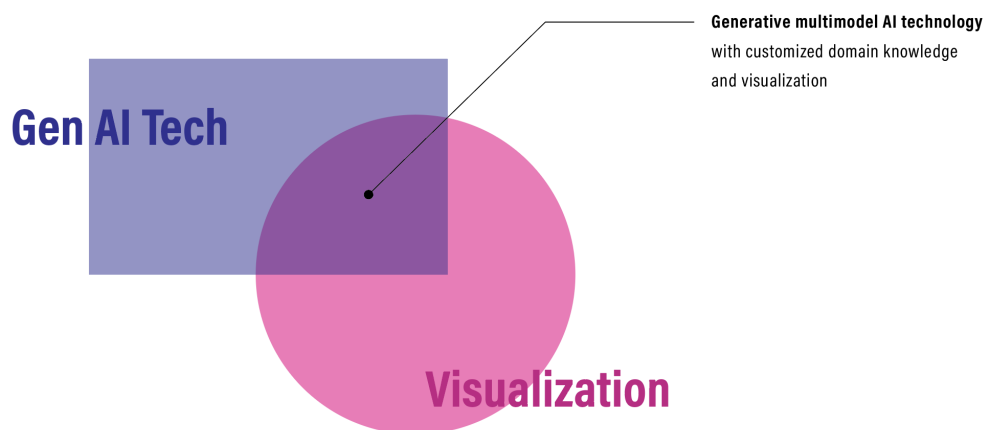


Figure 35: This figure illustrates that Generative Multi-Modal AI technology fits the niche of customizable AI with the added capability of producing visual outputs.



# 6 The Design Prototype

## 6.1 Solution architecture

At the end of the previous section, we concluded that multimodal generative AI technology with customized knowledge and visualizations has a strong potential to improve safety education.

As the next step, we began developing a solution architecture based on the prototype requirements established earlier and outlined in Section 3.3. This thesis introduces four main AI modules—DesignMate, AssessMate, MemoMate, and VisualMate— developed in response to the prototype requirements, as illustrated in Figure 36. These main modules are designed to interact with one another and connect with smaller supporting modules, as illustrated in Figure 37. Each of these modules is explained in more detail below here:



Figure 36: This figure illustrates how the AI solution architecture aligns with the prototype requirements, featuring a DesignMate AI module that handles most of those requirements.

## Main Modules

**DesignMate AI module:** This first module supports students by guiding them through the Safety by Design process. The step-by-step guidance is based on the NEN-NTA 8287:2021 method, integrating insights from lectures, teacher commentary, course reports, and online examples. All resources are carefully reorganized to reflect the structure of the NEN-NTA method, creating a coherent and thorough learning experience.

**AssessMate AI module:** Acting as a feedback assistant, this module evaluates student submissions, highlighting strengths and pinpointing areas that require further development.

**MemoMate AI module:** This module helps users save and organize findings drawing from the interactions with the other two AI modules.

**VisualMate AI module:** Designed to make insights more accessible, this module translates findings into visual formats—whether hierarchical, non-hierarchical, or sequential—making complex information easier to understand and communicate.

## Small Modules

**Knowledge base for DesignMate AI module:** contains information collected from NEN-NTA 8287:2021 method, lecture, and teacher comments to assist users in the design process.

**Knowledge base for AssessMate AI module:** contains the evaluation matrix used in the Safety by Design course, provided by the teacher, which allows users to self-assess their work before submission.

**Visual Library:** This is a digital file containing customized visual materials created specifically for this project. These materials help users organize their findings, such as stakeholders or technical components, after applying the method.

**External API:** The external API is used to transfer information from the MemoMate AI module to an external database

**External database:** A database for storing information from the MemoMate AI module.

**Interface:** A common ground that four main modules can interact/exchange data with each other

The four proposed AI models are intended to structurally offer personalized, precise, and field-specific responses to solve students' individual inquiries. Their overall goal is to improve the accessibility of safety standards by adapting to the varied needs of students from different disciplines in a user-friendly way.



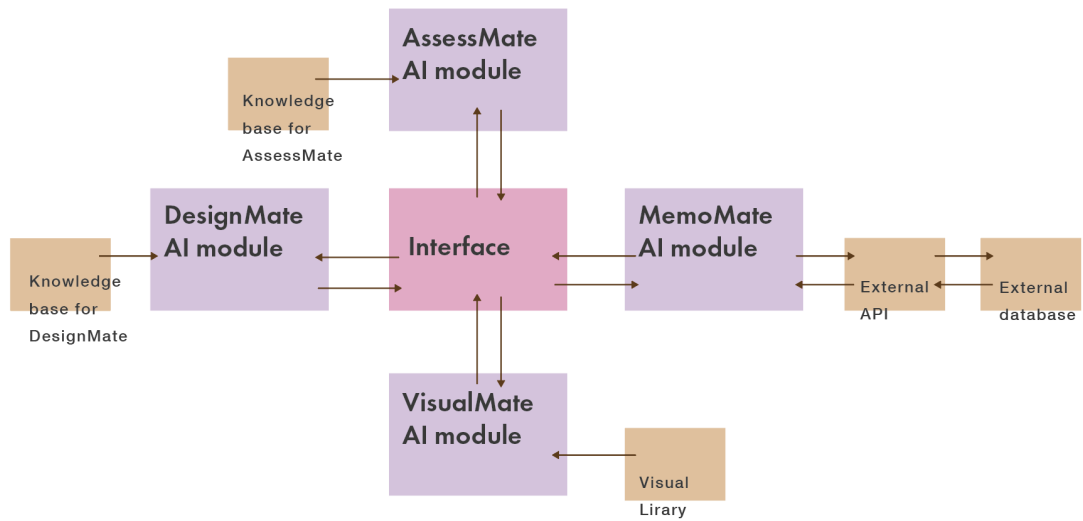


Figure 37: This figure shows the product architecture of the prototype, with the main modules in purple connected to smaller modules in yellow.

## 6.2 Prototype Development

The development of the prototype was carried out in three main stages. The first stage focused on *Technical Preparation*. This involved selecting the appropriate AI models, preparing the knowledge base, and determining the most effective format for translating content into a form the AI models could process, and a prompting technique.

The second stage concentrated on *Iteration Design of AI's behaviors*. Through a series of integration tests with users, we explored how the AI chatbots should interact in order to best support student learning. The final stage involved the *Iteration Design of Visual Mockups*. This stage describes how the initial hand-drawn sketches were refined into high-quality visuals using Illustrator, followed by user testing to evaluate how these visual elements could enhance understanding—eventually, choosing the best sets of visualizations. These three stages are explained in detail in the following sections.

### 6.2.1 Technical Preparation

#### *AI models selection*

This thesis began in early January 2024, at a time when customized GPTs was the most widely available and accessible option for developing tailored AI models (OpenAI, 2023b). Other major platforms, such as Gemini's customization tools and Microsoft Copilot Studio, were introduced later in mid-2024 and were not

available during the initial stages of this project (Google, 2024; Microsoft, 2024). As a result, customized GPT was the most practical and feasible choice at the time.

In addition, the decision to use customized GPTs was also influenced by its ease of use, extensive online tutorials, and lack of coding requirements (OpenAI, 2023b). These features make the model particularly suitable for users without a programming background.

### Knowledge base preparation

The learning materials, including the NEN NTA 8287:2021 standard, lecture notes, teacher insights, and additional online resources, were first semantically analyzed. The goal was to classify and organize these materials systematically according to the 15 steps outlined in the NEN NTA 8287:2021. These 15 steps correspond to the stages of the product life cycle, beginning with identifying stakeholders and their needs and concluding with product disposal and recycling (See Figure 38).

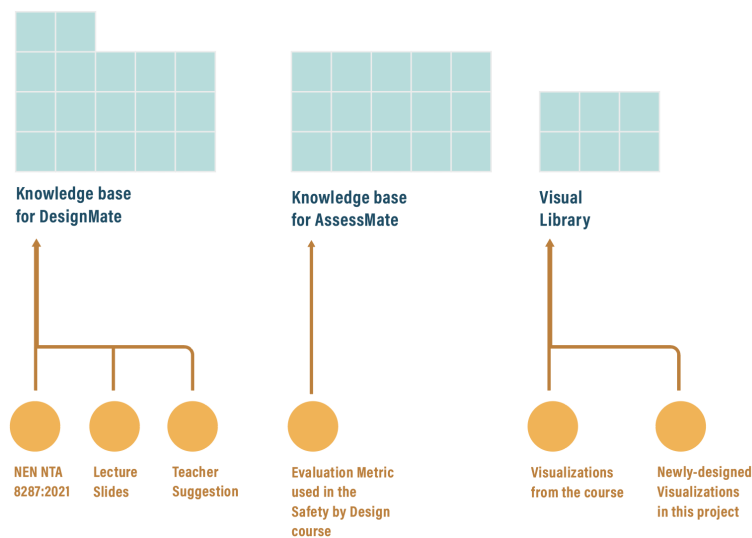


Figure 38: This diagram shows the sources of information used in the knowledge base for Design Mate, assessment, and the visual library.

This information formed the core knowledge used to guide students through the design process and was incorporated into the knowledge base of the DesignMate AI chatbot. The knowledge base consists of 17 files in total: 15 files correspond to the 15 steps of the NTA method, while the remaining two contain information on method overview and definitions of technical terms.

For the knowledge base of the AssessMate AI chatbot, the development process was less straightforward. Initially, we adopted an evaluation metric based on three criteria: completeness, quality, and presentation.

However, we encountered challenges in teaching the AI to recognize high-quality work, as this often involves subjective judgment and depends heavily on experience and interpretation of how the content is presented. Additionally, the AI struggled to accurately interpret visual elements in student submissions. As a result, we decided to explore alternative approaches.

After testing several evaluation approaches, we found that the metric used in the Safety by Design course produced the most effective results (See Figure 39). This evaluation metric was originally developed to help students self-assess their work. Then, we adopted it as the foundation for the AssessMate AI model.

PERFORMANCE RATING				
	Excellent (8-10)	Good (6.5-8)	Satisfactory (5.5-6.5)	Needs improvement (< 5.5)
<b>Stakeholders and their relations</b>	all the stakeholders across the entire life cycle are presented by relation diagram and prioritized, and their important needs are listed, the past and future stakeholders are analyzed	all the stakeholders across the entire life cycle are presented by relation diagram and prioritized, and their important needs are listed	major stakeholders are presented and prioritized, and their important needs are listed, the past and future stakeholders are analyzed	some stakeholders are presented, and their important needs are listed
<b>System concept and its environment</b>	system concept (subsystems and human interfaces) and its environment is clearly defined across the entire life cycle, comparable past systems and possible future system concepts are mentioned	system concept (subsystems and human interfaces) and its environment is clearly defined across the entire life cycle	system concept (subsystems and human interfaces) and its environment is clearly defined	system concept (subsystems and human interfaces) and its environment is unclear
<b>System (safety) objectives</b>	safety goals along with other system goals are defined and elaborated across the entire life cycle (e.g. by objective tree), relevant directives and standards (type A) are used and linked to goals, past and future goals are indicated	safety goals along with other system goals are defined and elaborated across the entire life cycle (e.g. by objective tree), relevant directives and standards (type A) are used and linked to goals	safety goals along with other system goals are defined and elaborated, relevant directives and standards (type A) are mentioned	safety goals along with other system goals are unclear, and important directives or standards (type A) are missing
<b>Risk assessment and acceptance criteria</b>	significant hazards (e.g. fire) are listed, major events in the entire life cycle (e.g. fire during the usage) are identified, past and possible future events are considered, risk matrix for the hazards/events is defined (using tailored severity and probability tables)	significant hazards (e.g. fire) are listed, major events in the entire life cycle (e.g. fire during the usage) are identified, risk matrix for the hazards/events is defined (using tailored severity and probability tables)	several hazards (e.g. fire) are listed, several events in the entire life cycle (e.g. fire during the usage) are identified, risk matrix for the hazards/events is defined (using tailored severity and probability tables)	significant hazards (e.g. fire) or events in the entire life cycle (e.g. fire during the usage) are missing, risk matrix for the hazards/events is unclear
<b>System functions and requirements</b>	primary (safety) functions are defined across the entire life cycle, past function failures are mentioned, and possible future functions are in consideration	primary (safety) functions are defined across the entire life cycle, past function failures are mentioned	primary (safety) functions are defined across the entire life cycle	some primary (safety) functions across the entire life cycle are missing
<b>System architecture is defined</b>	primary (safety) functions are allocated to humans and subsystems, and safety categories are assigned, past accidents and design failures (Section 17.5 of course book) and future functions are considered, safety by design is utilized	primary (safety) functions are allocated to humans and subsystems, and safety categories are assigned, safety by design is utilized	several primary (safety) functions are allocated to humans and subsystems, and safety categories are assigned, safety by design is utilized	several primary (safety) functions are allocated to humans and subsystems, but safety categories or safety by design is missing
<b>System design</b>	(at least) one subsystem is elaborated in further design focusing on internal interactions and safety control system (parts e.g. sensors, logic, output) to reduce the risks, relevant past design failures and possible future functions of the desired subsystem are considered	(at least) one subsystem is elaborated in further design focusing on internal interactions and safety control system (parts e.g. sensors, logic, output) to reduce the risks, relevant past design failures of the desired subsystem are considered	(at least) one subsystem is elaborated in further design focusing on internal interactions and safety control system (parts e.g. sensors, logic, output) to reduce the risks	elaborated design of at least one subsystem, internal interactions, and the safety control system (parts e.g. sensors, logic, output) is missing
<b>Residual safety risks</b>	sufficient safety has been achieved by reducing the technical, human, and environmental risks through the entire life cycle and demonstrated through the report, safety case, or GSN. Comparable safety reports are mentioned.	sufficient safety has been achieved by reducing the technical, human, and environmental risks through the entire life cycle and demonstrated through the report, safety case, or GSN.	sufficient safety has been achieved by reducing the technical, human, and environmental risks through the entire life cycle	sufficient safety have been achieved by reducing the technical, human, and environmental risks through the whole life cycle
<b>production</b>	suppliers', as well as producers' concerns, are considered, relevant cases from the past as well as future are considered, possible technical integration issues from the past and possible extension for future needs are considered	suppliers', as well as producers' concerns, are considered, relevant cases from the past as well as future are considered	suppliers', as well as producers' concerns, are considered	either suppliers' or producers' concerns are missing
<b>Packaging to preparation</b>	hazardous and risky events during transportation and installation are considered, past events are studied, the certification process is known, clear insight how to prepare it for use and what information is to be shared for use	hazardous and risky events during transportation and installation are considered, past events are studied, the certification process is known,	hazardous and risky events during transportation and installation are considered, past events are studied	hazardous and risky events during transportation and installation are missing
<b>Operation and performance</b>	start of operation and required performances are clear, use and failure scenarios are known for different hazards, past and possible future known-failures (Section 17.5) are considered and controlled or communicated	start of operation and required performances are clear, use and failure scenarios are known for different hazards, past and possible future known failures (Section 17.5) are considered	start of operation and required performances are clear, use and failure scenarios are known for different hazards	start of operation and required performances are not clear, use and failure scenarios are missing for different hazards
<b>Maintenance and support</b>	safety plans during maintenance and fault finding for the aforementioned subsystem (Step 7 of SCM) are in place, past and possible future incidents are considered	safety plans during maintenance and fault finding for the aforementioned subsystem (Step 7 of SCM) are in place, past incidents are considered	safety plans during maintenance and fault finding for the aforementioned subsystem (Step 7 of SCM) are in place	safety plans during maintenance and fault finding for the aforementioned subsystem (Step 7 of SCM) are missing
<b>Human factors and culture</b>	awareness about safety maturity and cultural aspects, misuse failure of the past and possible future (section 17.5) are considered, human factors at work and socio-technical aspects are considered	awareness about safety maturity and cultural aspects, misuse failures of the past (section 17.5) are considered, human factors at work and socio-technical aspects are considered	awareness about safety maturity and cultural aspects, misuse failures are considered, human factors at work are considered	awareness about safety maturity and cultural aspects is missing, or misuse failures are not considered, human factors at work are missing
<b>Safety-risk monitoring</b>	leading and lagging safety indicators are known, past and future safety monitoring systems are considered, similar safety monitoring systems are mentioned	leading and lagging safety indicators are known, similar safety monitoring systems are mentioned	leading and lagging safety indicators are known, a basic safety monitoring system is in place	leading and lagging safety indicators are missing, a basic safety monitoring system is not in place
<b>Retirement</b>	retirement, environmental, and socio-environmental risks and concerns of past and possible future are considered, recycling and reuse of (at least one) subsystem is considered	retirement, environmental, and socio-environmental risks and concerns are considered, recycling and reuse of (at least one) subsystem is considered	retirement and environmental risks and concerns are considered, recycling and reuse of (at least one) subsystem is discussed	retirement and environmental risks and concerns are missing, recycling and reuse of (at least one) subsystem is missing

Figure 39: This figure shows the evaluation metric from the Safety by Design course, also used for the prototype to evaluate student work.

## Translating Knowledge into AI-Usable Format

Throughout the development process, we experimented with various data formatting methods, including plain text, Markdown, and JSON. Initially, we began by directly copying information from the NEN NTA 8287:2021 into customized GPTs, but this approach resulted in vague, incomplete, and heavily paraphrased responses.

We then moved to Markdown, a lightweight formatting language that organized information hierarchically. However, this format cannot express deep hierarchical relationships, which are crucial for GPT to effectively locate and interpret connected information. To address this limitation, we ultimately adopted the JSON format, which allows for a more structured and layered representation of data, enabling more accurate and relevant responses from the AI.

Figure 40 presents three versions of the same content formatted in plain text, Markdown, and JSON. The key advantage of the JSON format is its ability to assign labels—such as “goal” or “requirement”—to different pieces of information. This structure allows GPT to understand the intended meaning and function of the content.



Figure 40: This figure shows three formats for presenting the same content: plain text, Markdown, and JSON. The JSON format provides more detailed classification, while also organizing the information hierarchically.

## ***GPT prompting techniques***

To ensure the GPTs behave according to our intended design, it is essential to provide them with clear instructions through prompting. This section outlines the main approach used to guide the AI chatbot's behavior. The technique we adopted is based on a prompting method introduced by OpenAI (OpenAI, [n.d.](#)). This technique contains four steps:

**1. Defining the Context of User Inquiries:** We started by informing GPT about the nature of the questions or inquiries it needs to handle. *"You, as GPT, will be asked about the definitions of specific safety technical terms."*

**2. Specifying the Behavior and Actions:** Instruct GPT on what actions it should perform to respond appropriately. This includes referring to a specific knowledge base where it can find relevant information and specifying carefully the desired outcomes.

*"Look into the prepared file 'definition.json' and extract the definition. Provide the definition word-for-word from the file. You should provide the following content for each term:*

- *Definition*
- *A list of difficult words in the definition with their definitions*
- *Examples that related to users' project*
- *Extra explanation (if available)"*

**2. Formatting the Response:** the format in which the response should be delivered. For example *"The term should be in bold, and the definition should be enclosed in triple quotes ("""). A list of difficult words, formatted as bullet points. Each difficult word should have its definition in a nested bullet point. Examples and extra explanations should be included in plain text.*

**3. Defining Restrictions:** We clearly outline what GPT should avoid doing. For example, *" GPT should not generate its own definitions if it cannot find the requested information in the file."*

### **6.2.2 Iteration Design of AI's behaviors**

The development of the AI's behavior followed an iterative integration approach described in the study by Nielsen ([n.d.](#)). We began with an initial chatbot based on our envisioned concept and conducted informal

tests with both students and teachers. Feedback from these sessions helped us identify key problems, which we addressed by refining the AI chatbot, adjusting its behavior, and adding new features as needed.

This process was repeated over two rounds of user testing, resulting in three distinct versions of the AI chatbots (as illustrated in Figure 41). Throughout the evaluation process, the most significant shift in the AI chatbot's behavior was its transformation from a tutor—providing step-by-step guidance and assigning tasks—to a collaborative tool that works alongside users to support them in finishing the given tasks. The following paragraphs introduce the three versions of the AI chatbot developed during the process, along with the results and insights gained from two rounds of user testing.

## ***Version 1***

### ***Description***

In this version, all 4 AI modules described in the Solution Architecture were integrated into one GPT chatbot. This chatbot functioned as a tutor, controlling the interaction by assigning users to design tasks and managing task flow. In particular, the customer journey was designed as follows:

1. The chatbot guides users through the 15 steps of NEN NTA 8287:2021, addressing each step individually.
2. Within each step, the chatbot presents task descriptions in a sequential manner, requiring users to complete one task before proceeding to the next. In the task descriptions, the chatbot also includes a digital link to customized visualizations
3. After each task, the chatbot evaluates users' responses based on predefined metrics. If the answers are incomplete or incorrect, the chatbot asks users to revise their responses before progressing. If the answers pass, the chatbot introduces the next tasks.
4. Once all the given tasks within one step are completed, the chatbot summarizes the findings into a small document and moves to the next step.

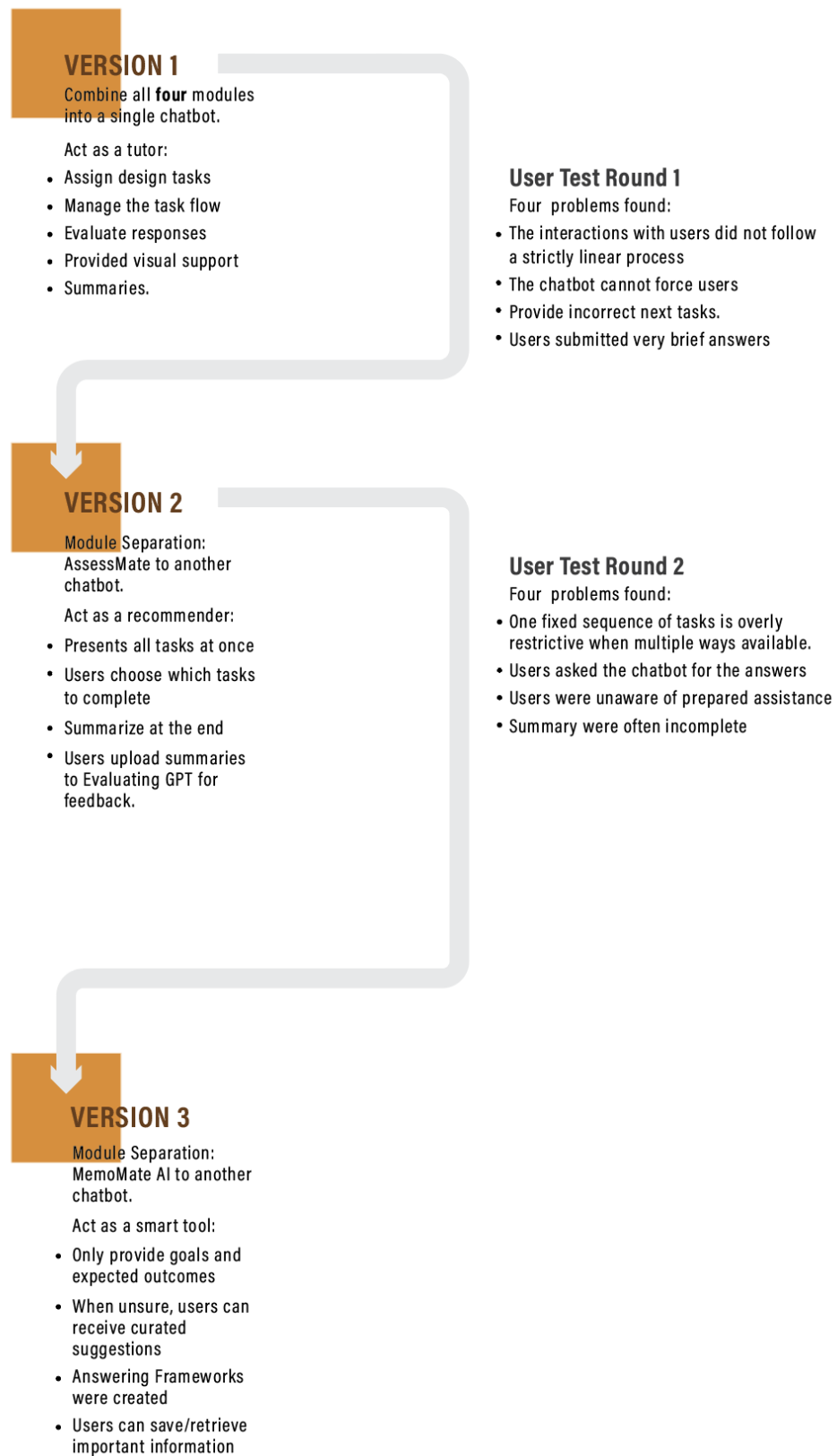


Figure 41: This figure illustrates the development of the AI chatbots with user feedback. The general trend moves from AI acting as a powerful tutor with more control, toward a tool-like role where users hold greater power and actively direct the AI's actions.



## *Results from User Testing Around 1*

After letting the chatbot guide two users through the first two steps of the NEN NTA 8287:2021 (Stakeholders and their needs; System concept, system definition, and system environment), we identified the following issues:

- **Deviation from Linear Progression:** The interactions with the chatbot did not follow a strictly linear process but were instead more dynamic and flexible. Users did not always respond immediately after a task was given, but they might ask for follow-up questions or even want to skip the given task.
- **Inability to Enforce Task Completion:** By default, OpenAI designed GPT to prioritize user autonomy (Ramel, 2024), making it challenging for the chatbot to force users to redo a task when users are unwilling to do so.
- **Limited Memory Capacity:** In longer interactions, the chatbot occasionally lost track of the conversation flow, leading to providing incorrect next tasks.
- **Minimal User Input:** Users often submitted very brief, minimal answers that do not accurately reflect the true knowledge of users.

## *Version 2*

### *Description*

In the second version, the AssessMate module is separated into another GPT called Evaluating, while the other three modules are still integrated into Design Guiding. The role of Design Guiding also changed:

1. Unlike the first version, the Safety by Design Tutor GPT presents all task descriptions for each step upfront, giving users a clear overview of what they need to accomplish from the start.
2. Instead of forcing users to do tasks, GPT asks users to choose the tasks users want to perform, promoting their autonomy and active participation.
3. After completing all tasks within a step, the Safety by Design Tutor GPT generates a concise summary of the answers and findings.

4. Users can then upload this summary to the Evaluating GPT to receive feedback on their work. In this way, users have enough time to refine their answers and produce more organized and comprehensive responses. Furthermore, the evaluation is more accurate as the Evaluating GPT needs to process shorter information.

## ***Results from User Testing Around 2***

Similar to round 1, we let two students and the teacher interact with the chatbots to complete steps 1 and 2 of the NEN NTA 8287:2021. The feedback of students and the teacher is summarized below:

- **Multiple Valid Approaches for Each Step:** The teacher noted that different fields may approach the same task in varied ways. Therefore, having the chatbot offer only a single, fixed sequence of tasks could be overly restrictive.
- **Letting GPT execute the tasks:** Users often asked the chatbot for the answers instead of performing the tasks themselves. They only checked and refined the given results
- **Hidden knowledge:** Although extensive support materials were prepared, such as examples and use cases, users were unaware of their availability and, therefore, rarely accessed them.
- **Summary Incompleteness:** Observations reveal that, at the end of steps, chatbot's summaries often miss key information. As a result, users need to edit and complete the summaries manually.

## ***Version 3***

### ***Description***

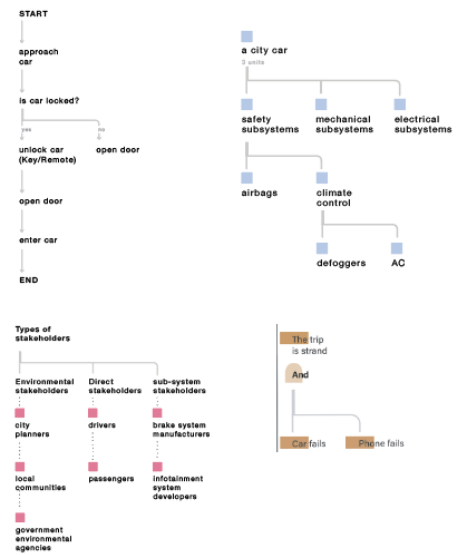
In the third version, the MemoMate AI module was separated into its own GPT, called Documenting GPT. As a result, the Design Guiding GPT included two modules—DesignMate and VisualMate—while AssessMate continued to be managed by the Evaluating GPT. The Design Guiding chatbot also changed its behavior:

1. In this version, the Design Guiding chatbot no longer enforces a fixed sequence of tasks for users. Instead, it provides an overview of each step, including the step goal, a brief description, recommended timing, required knowledge, expected outcomes, and suggested topics users can ask about. This gives users more freedom to choose different methods to complete tasks based on their preferences, maximizing their freedom and engagement. If users are unsure about which method to apply, the chatbot offers suggestions reviewed by the course instructor to guide them

2. Answering frameworks were developed to help the chatbot generate responses that align more closely with the structure and terminology of the NEN NTA 8287:2021. For example, templates were created to organize risks into functional, operational, and technical categories. These frameworks guide the AI in structuring its answers according to the logic and expectations of the NEN NTA 8287:2021 method.
3. Instead of generating summaries at the end of each step, users can ask the Documenting GPT during the conversation to save any details or results they find valuable in a separate database at any time. This information can be recalled as needed.
4. Once users finalize their answers or results, they can submit them to the Evaluating GPT, similar to the process in Version 2.

### **6.2.3 Iteration Design of visual mockups**

The journey of creating customized visualizations can also be divided into three distinct phases, summarized in Figure [42](#).



## Visual Language Refinement

One single visual language with adaptable variations to suit hierarchical, sequential, and non-hierarchical information structures



## Exploring Visual Connections

- Using the same square shapes with different color to represent safety elements
- Present information in various structures: non-hierarchical, hierarchical, sequential, and 3D
- The visualizations still look greatly different

## Early Brainstorming

A range of visualizations were created.

Two Primary Goals

- Clarify information from NEN NTA 8287:2021
- Organize user findings visually.

Figure 42: This figure summarizes the three-stage process of developing customized visualizations—starting broad and gradually refining into a unified visual language with adaptable variations.

## ***Early Brainstorming***

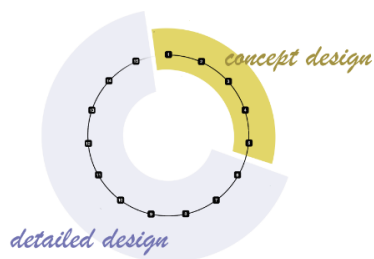
The first phase focused heavily on ideation. During this stage, we explored a wide variety of visual formats with two primary goals in mind. The first was to present specific information about the NEN NTA 8287:2021 document in a clear and accessible way. The second was to find an effective method for organizing users' findings visually. The initial ideas were first sketched on paper, then translated into digital visualizations using Adobe Illustrator. Some examples of the visualizations developed during this phase are shown in Figure 43.

However, when these visualizations were shared with potential users and teachers, it became clear that they lacked consistency. The visuals appeared disconnected from one another, with no unifying system in terms of color, shape, or typography to establish a coherent visual language.

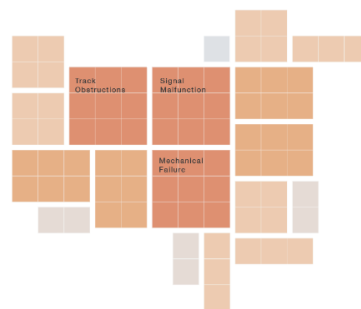
## ***Exploring Visual Connections***

After receiving feedback from users and the teacher, we introduced a more structured design approach. This involved reusing the same core visual elements across different contexts to create a more cohesive system. In this revised approach, each of the three key safety elements—human, system, and environment—is represented by a square shape, with distinct color coding: pink for human, blue for system, and green for environment. We also explored how these visual elements could be used to organize information across different structural formats, including non-hierarchical, hierarchical, sequential, and even three-dimensional visualizations. As a result, we developed a set of multiple visualizations, some of which are presented in Figure 44.

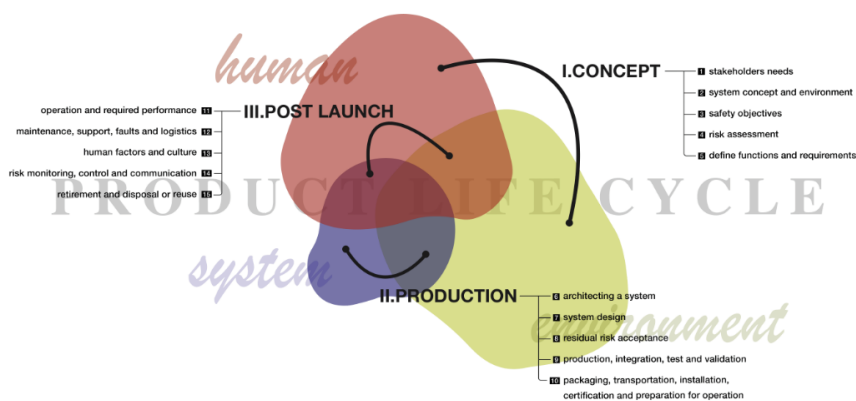
After introducing this set of visualizations to users, the feedback was mixed. While some visualizations were considered clearer than others, users consistently pointed to Picture A in Figure 44 as the most effective in conveying hierarchical information. Another key piece of feedback was that many users still felt there were too many options and approaches. The variety led some users to comment that the visualizations did not feel like they were part of a single, cohesive product.



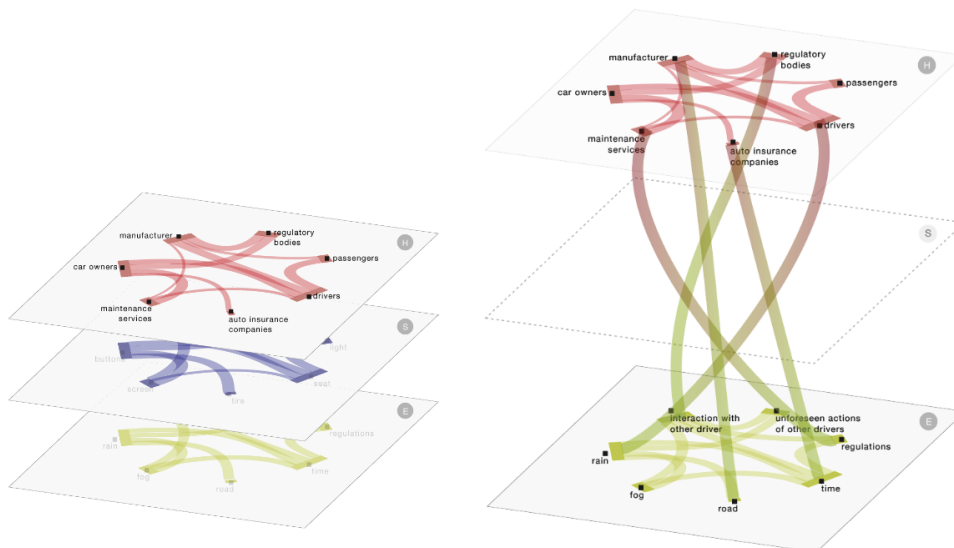
The visualization shows the first 5 steps of NTA 8287 focus on designing concepts and the other 10 steps focus on detailed design.



The visualization was inspired by word clouds, placing the most serious risks at the center and highlighting them with the most intense colors.



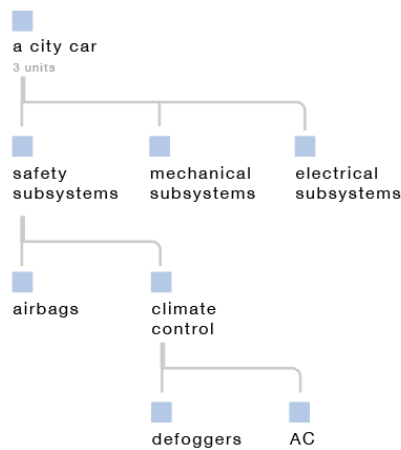
The visualization groups the 15 steps of NTA 8287 into three phases—concept, production, and post-launch—each focusing on different elements: human, system, environment, and their interactions.



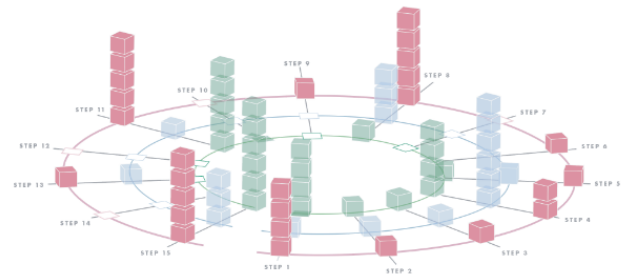
This visualization illustrates how human, system, and environmental factors can be organized using ribbon diagrams, with each factor represented on a separate layer.

The three diagram can connect to each other as well

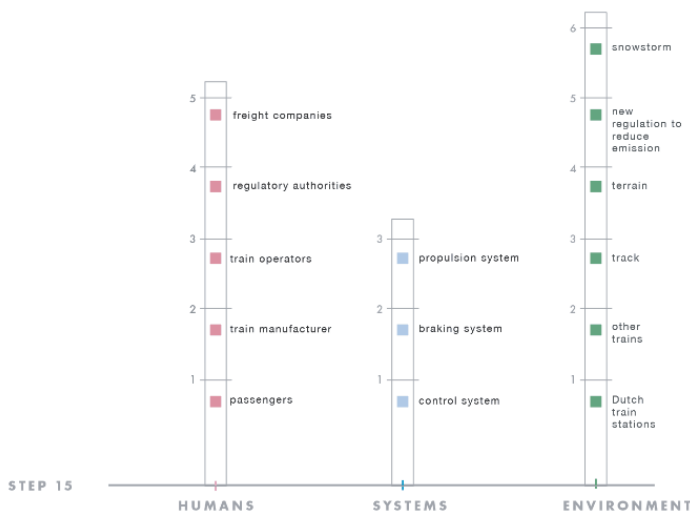
Figure 43: This figure shows early-stage visualizations created during the initial phase. Each was designed with a specific goal, but they differ significantly and lack a shared visual language.



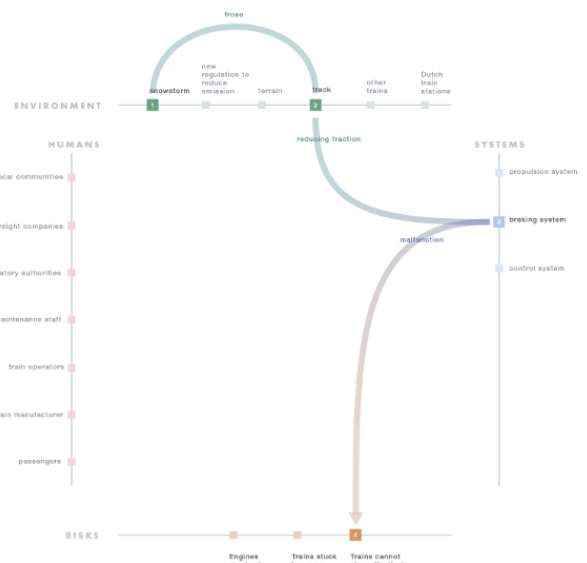
**A** The visualization demonstrates how various car components can be organized in a hierarchical structure.



Each human, system, and environment factor is represented by a colored cube, arranged in a circular layout from step 1 to step 15.



Each human, system, and environment factor is displayed in a bar chart, allowing users to easily compare their quantities.



The visualization depicts a sequence of interactions, starting from the environment, affecting the system, and ultimately leading to risks.

Figure 44: This figure shows visualizations from the second stage, where a shared visual language begins to emerge—using consistent colors and rectangles to present information across different structures.

## Visual Language Refinement

The insight from users shaped the third phase: refining the visualization into a single, adaptable visual language. Instead of offering a range of different visualizations, we focused on selecting one well-performing design and developing it further with variations. The goal was to establish a single visual language with multiple variations that could adapt to different information structures—hierarchical, sequential, and non-hierarchical. Therefore, visual identity can be maintained consistently and be recognizable. More information about this visual language is described in Section 6.3.2 on page 83.

## 6.3 Final prototype

### 6.3.1 SIAI (Safety Instructor AI): the AI chatbot systems

This thesis introduces SIAI (Safety Instructor AI), a system comprising three specialized GPT chatbots, each with its own distinct role. This system allows users to engage with all three chatbots in a single conversation or interact with each one individually. See Figure 6.3.1 for the SIAI interface and visual mock-up examples.

**Design Guiding Chatbot:** This chatbot integrates DesignMate and VisualMate AI modules, having the primary function of providing accurate safety information drawn from standard documents and presenting them in both textual explanations and visualizations. In addition to all the standard capabilities of a generic GPT model, this chatbot includes six specially designed features:

**1. Provide exact definitions with additional explanation:** the chatbot presents precise definitions of technical terms commonly found in standard documents. Additionally, the chatbot also clarifies complex words used within the main definitions by explaining their meanings and offering examples tailored to the user's project, helping them connect the terminology to their own work. An example of chatbot's response can be seen in Figure 46

**2. Present NEN NTA 8287:2021 overview and its application:** The chatbot offers a brief introduction to the NTA 8287 method, including its purpose, core principles, and overall structure. In addition, it suggests relevant steps based on the user's project focus. For instance, if the user is interested in technical aspects, the chatbot recommends exploring step 2 (*System concept, system definition and system environment*), step 6 (*Architecting a safe system*), and step 7 (*System (safety) design*).

**3. Deliver In-Depth Content of NEN NTA 8287:2021:** The chatbot provides detailed information on each of the 15 steps from the NTA 8287 document, structured to help users easily understand and apply them (See Figure 46). Each step is introduced through a consistent framework:



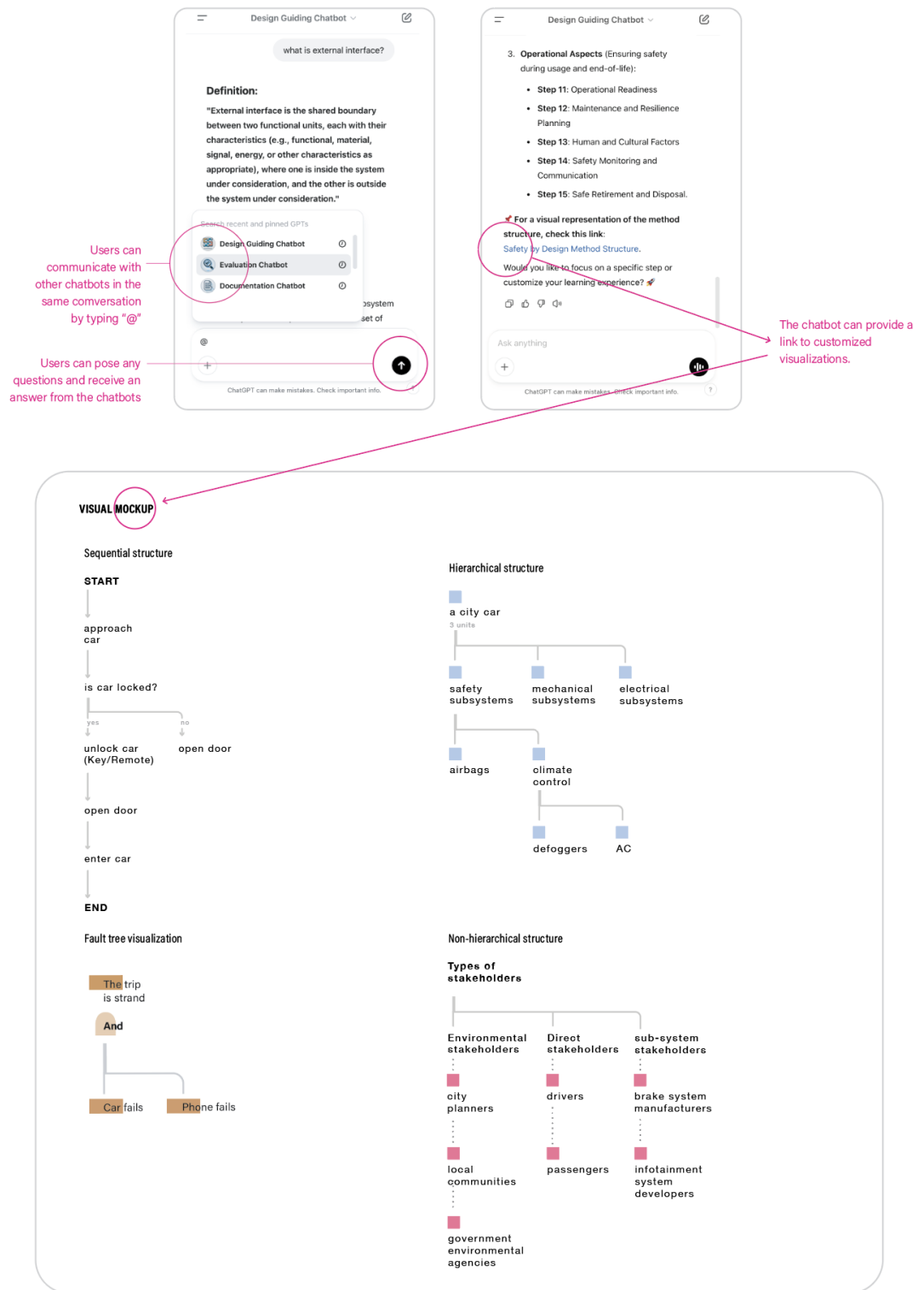


Figure 45: This figure illustrates the main features of SIAI and visual mockups. Blue rectangles denote the system and its subsystems, pink rectangles highlight the human factors considered in safety analysis, and orange rectangles mark intermediate risks identified in Fault Tree diagrams.

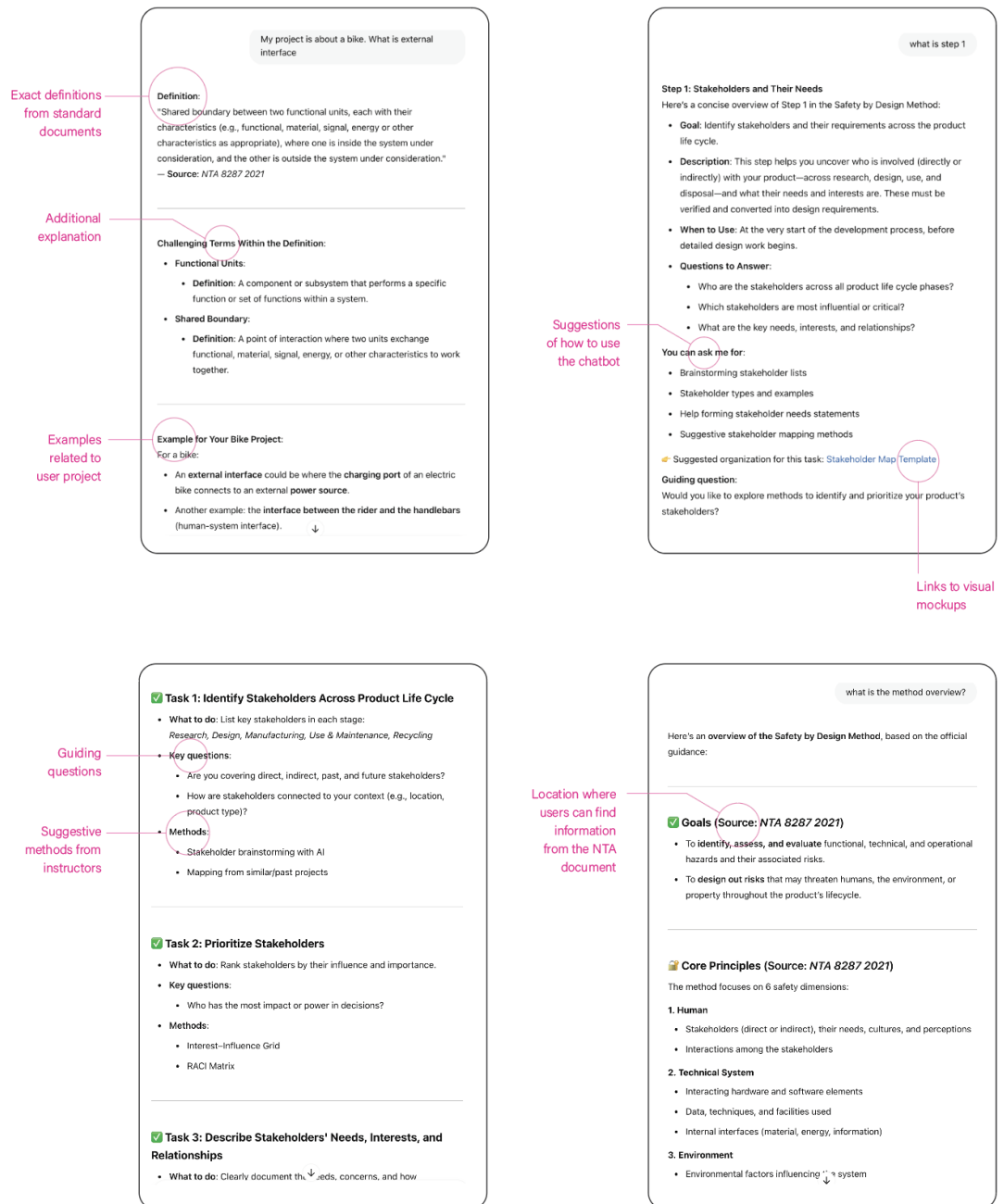


Figure 46: This figure presents key information generated by SIAI, when it provided definitions (top left), step descriptions (top right), execution guidelines (bottom left), and an overview of the method (bottom right).

- *Goals*: what users can achieve through the step
- *Short Description*: Provide contextual information on why the step is important
- *When to Apply*: highlights the most appropriate moment to use each step, showing its connection to other steps in the NTA 8287.
- *Question-to-Answer*: demonstrate expected outcomes in the form of questions. Instead of explicitly stating the required output, the use of guiding questions invites users to reflect, interpret, and engage more actively with the task
- *GPT Usage Tips*: hint at pre-prepared resources that can help users complete the task more efficiently.
- *Source reference*: mentions where users can find the information in the original standard content, facilitating information verification

**4. Propose execution guidelines:** This guideline breaks down the expected outcomes of each step into specific, actionable tasks. It also includes suggested methods for completing these tasks, drawing from instructors' experiences.

**5. Provide digital links to visual mockups:** Although the chatbot is not yet capable of converting live data into predefined visualizations, it can deliver access to visual mockups through digital links. These mockups offer a glimpse of how the chatbot could support participants in organizing and presenting their findings across various stages of the NEN-NTA 8287:2021, using hierarchical, non-hierarchical, and sequential structures. See Section 6.3.2 for more information about the customized visualizations.

**6 Support for Risk Analysis Tasks:** When users request the chatbot to perform a risk analysis task, the chatbot can deliver a comprehensive response using a predefined template that links with the NEN NTA 8287:2021. For example, if the task involves identifying stakeholders, the chatbot organizes them into five categories defined in the NTA document: environmental, direct, subsystem, past, and future stakeholders. In this way, users can see how the NTA 8287 standard applies to a specific context and task.

**Evaluating Chatbot:** This chatbot incorporates the functions of the AssessMate AI module, allowing users to submit their findings and receive feedback on both their strengths and areas for improvement. The evaluation of the chatbot is based on the metrics provided in the Safety-by-Design course for students to self-assess their work.

**Documenting Chatbot:** By integrating the MemoMate AI module, the chatbot enables users to save and retrieve key findings gathered while interacting with the Evaluating and Design Guiding chatbots. It offers a

fast and convenient way to capture key insights, making the documentation process easier and more efficient.

### 6.3.2 Four Visual mockups:

This section provides more information about the visual mockups available through the Design Guiding AI Chatbot. In this thesis, we developed a visual language that is constructed by three building blocks: nodes, explanatory text, and connecting lines. By mixing, matching, and adjusting these elements, users can present information in various structures: hierarchical, non-hierarchical, and sequential (See Figure 47).

The core of this visual language lies in its flexibility: the ability to adapt to different types of content and contexts. To demonstrate this flexibility, we took a step further by adapting it to a fault tree diagram, a widely used tool in risk analysis. This example not only showcases the visual language's versatility but also illustrates how it can be integrated with established visual methods to support a clear and consistent presentation.

The following paragraph outlines the design rationale about the nodes, connecting lines, explanatory text, and the design intentions behind the visually enhanced appearance of the Fault Tree diagram.

#### **Node**

In this visual language, nodes are presented as square shapes with distinct color coding: pink represents humans, blue indicates the system, product, or subsystem, and green represents environmental factors such as weather or regulations. These three elements—human, system, and environment—are the fundamental components that consistently appear across all safety situations. This is also the core theoretical foundation of the NEN NTA 8287: 2021.

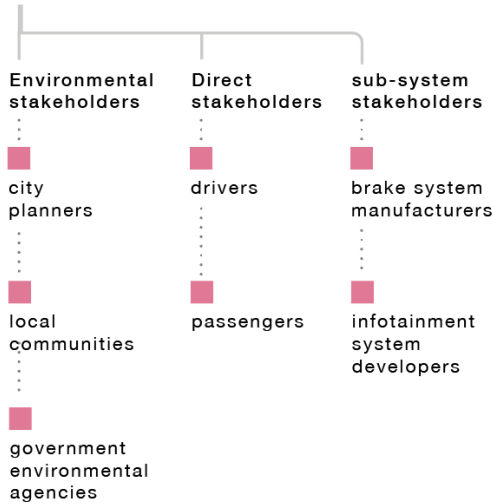
#### **Connecting Lines**

Lines indicate logical relationships between nodes. This thesis introduces various line styles to represent different types of logic, with the simplest form representing a hierarchical structure in Figure 47. The following bullet point outlines the visual design rationale used in lines:

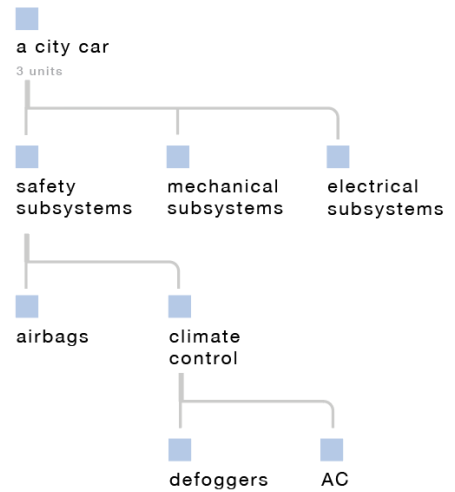
- *Thickened upper segment*: A short vertical line segment is thickened at the top to indicate the convergence of input nodes.

#### Non-hierarchical structure

##### Types of stakeholders

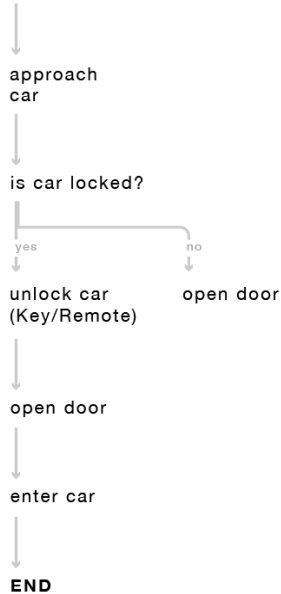


#### Hierarchical structure



#### Sequential structure

##### START



#### Fault Tree Visualization

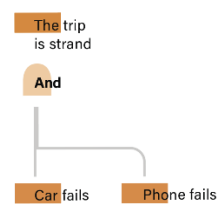


Figure 47: This figure displays four customized visualizations to structure information in non-hierarchical, hierarchical, and sequential formats. Together, they offer a simple and consistent approach to presenting information. The visual language can also be adapted to Fault Tree visualization, a widely used diagram in risk analysis

- *Indicating numbers and words:* As the number of input nodes increases, continuously thickening the line eventually transforms it into a rectangular shape, removing the impression of a slim and vertical appearance. Therefore, to maintain clarity when more than two units are involved, a number paired with the word "unit" is added to indicate the required quantity.
- *Rounded corner:* The corner of the rightmost line is rounded to signal that no additional inputs will connect to the gate.
- *Added Arrows:* When an arrow is added to a line, the logic shifts from a hierarchical relationship to a sequential one, indicating a progression or flow between nodes.
- *Dotted lines:* The dotted line is used exclusively in the non-hierarchical structure to visually group elements within the same category. It was chosen over a solid line to make it clear that there is no actual interaction between these elements. Instead, the dotted line serves as a gentle visual guide, "sewing" all elements that belong to the same category together. **Explanatory Text**

In the explanatory text, we use various typographic techniques to indicate hierarchy. For instance, bold text signals higher importance compared to lighter text, as it naturally draws more attention. Similarly, all-caps text conveys emphasis and a louder tone.

These visual cues help guide the user's attention. For example, in the sequential structure, such emphasized text is placed at the beginning and end to mark the start and the end of the information flow.

### **Fault Tree Diagram**

The Fault Tree Diagram can benefit from the same use of connecting lines and explanatory text as the other mockups. However, because the shape of each node carries specific meaning in fault tree notation, we must adhere to these rules and avoid using square shapes.

Instead, we retain the original shapes used in the Fault Tree Diagram—such as rectangles, circles, and other standard symbols—to preserve their established meanings. The only modification introduced is the use of the color orange to represent risks.

# 7 What We Tested, What We Found

Throughout the thesis, the prototype underwent three distinct rounds of testing: the first focused on the robustness of Safety Instructor AI (SIAI) in delivering NEN-NTA 8287:2021 information, the second evaluated how well users understood its visual mockups, and the final one assessed its effectiveness in fostering meaningful collaboration with users. The objectives of these tests were guided by a set of criteria known as Hallmarks proposed in Kozierok et al. (2021), (See Figure 48 for more information.)



Figure 48: A summary of the goals and results from the three tests, each addressing different evaluation criteria under the “Hallmark” framework.

## 7.1 Test 1: Robustness Evaluation

### **Objective**

The goal of this test is to evaluate whether the SIAI can reliably and accurately retrieve specific information about the NEN-NTA 8287:2021 standard from its knowledge base when prompted in four different styles: direct, paraphrased, indirect, and prompts containing typos.

*Direct prompts:* These prompts contain no more than 20 words and explicitly request the targeted information.

*Paraphrased prompts:* These prompts use synonyms and alternative phrasing while preserving the same semantic meaning as the direct prompts.

*Indirect prompts:* These prompts ask for the targeted information implicitly by expressing confusion or unfamiliarity. For example, *"I am new to the concept of external interface"*

*Prompts with typos:* These prompts contain incomplete information or misspellings.

### **Procedure**

The test, carried out by the thesis author, began by asking SIAI to define three technical terms, provide the method overview, and describe one specific method step along with its execution guidelines. All of this information is based on the NEN-NTA 8287:2021 standard (Mohammad Rajabalinejad, 2021). Each of these pieces of information was tested five times with the same direct prompt, five paraphrased versions, five different indirect prompts, and five prompts with typos. Three technical terms were classified under a *"brief answer"* group, while method overview, step description, and execution guidelines were placed in a *"long answer"* group, as generating these responses required more tokens—the unit representing the length and complexity of AI outputs.

Each interaction with SIAI was documented as a single log of human-machine exchange. Each interaction was evaluated for both accuracy and completeness, using criteria detailed in Table 2. Finally, the success rate was determined by calculating the proportion of correct and complete responses out of the total number of attempts.



Table 2: This table outlines the criteria used to evaluate AI responses in terms of accuracy and completeness.

	Definitions	Method Overview	Step Description	Execution Guideline
Accuracy	Whether the information was presented word-for-word, as defined in the knowledge base.			
Completeness	The answer is complete when it include: <ul style="list-style-type: none"> <li>• The exact definition</li> <li>• Explanation of difficult words</li> <li>• Examples related to user's project</li> </ul>	The answer is complete when it include: <ul style="list-style-type: none"> <li>• The method purposes</li> <li>• The core principles</li> <li>• Names of 15 steps</li> </ul>	The answer is complete when it include: <ul style="list-style-type: none"> <li>• Goals</li> <li>• Short Description</li> <li>• When to Apply</li> <li>• Question-to-Answer</li> <li>• GPT Usage Tips</li> <li>• Source reference</li> </ul>	The answer is complete when it include: <ul style="list-style-type: none"> <li>• All suggested tasks</li> <li>• Task goals</li> <li>• Their suggestive methods</li> </ul>

## Theoretical framework

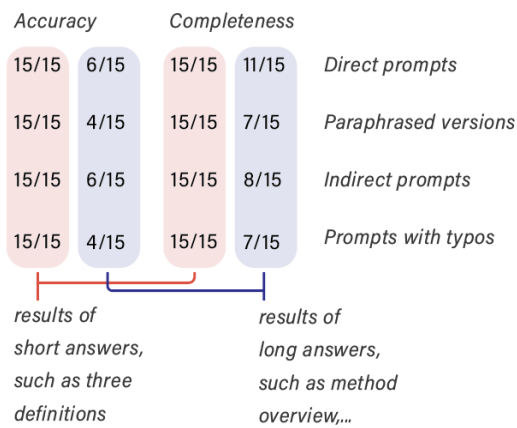
The first test followed two testing methods. The first one is a *Metamorphic Prompting Testing* method introduced in the study (X. Wang & Zhu, 2024). The method explained that instead of testing with one prompt several times, an AI model must be tested with multiple rephrased versions that should logically lead to the same answer. Through this testing method, the robustness and consistency of the AI model can be evaluated.

The second method is *Adversarial Testing* (J. Wang et al., n.d.). In this method, the AI model is intentionally fed challenging or deceptive inputs, such as vague and misspelled prompts, to evaluate its response. The method assesses whether the model can overcome this challenge and provides expected answers.

## Results

As can be seen in Figure 49, the results show that SIAI achieved the highest success rate (15/15 times - 100%) when responding to both direct and indirect prompts requesting brief answers, such as definitions. Variations in phrasing and the presence of typos did not impact its performance on brief answers, as reflected by the same success rates.

However, for more detailed responses, especially method overviews, SIAI started to produce incomplete and paraphrased content, resulting in lower success rates in both accuracy and completeness. Our observations show that, although the outcomes from the "long answer" group did not always match the knowledge base word-for-word, they conveyed content with the same semantic meaning, as illustrated in the left picture of Figure 50. However, it occurred five times that the step names for steps 8, 9, 10, and 11



### High success rate:

- short answers when requested by direct prompts; wording variations and typos did not influence the success rate. SIAI offers relevant definition when prompts indicate confusion.

### Lower success rate:

- long answers were paraphrased incorrectly and missing some information

Figure 49: The figure shows the results from the first test, accompanied by brief interpretations on the right-hand side.

were paraphrased incorrectly, as also shown in Figure 50. Furthermore, the most frequently omitted piece of information in the "long answer" group was the source citation, missing 15/60 times across different prompt types, especially in responses to the method overview.

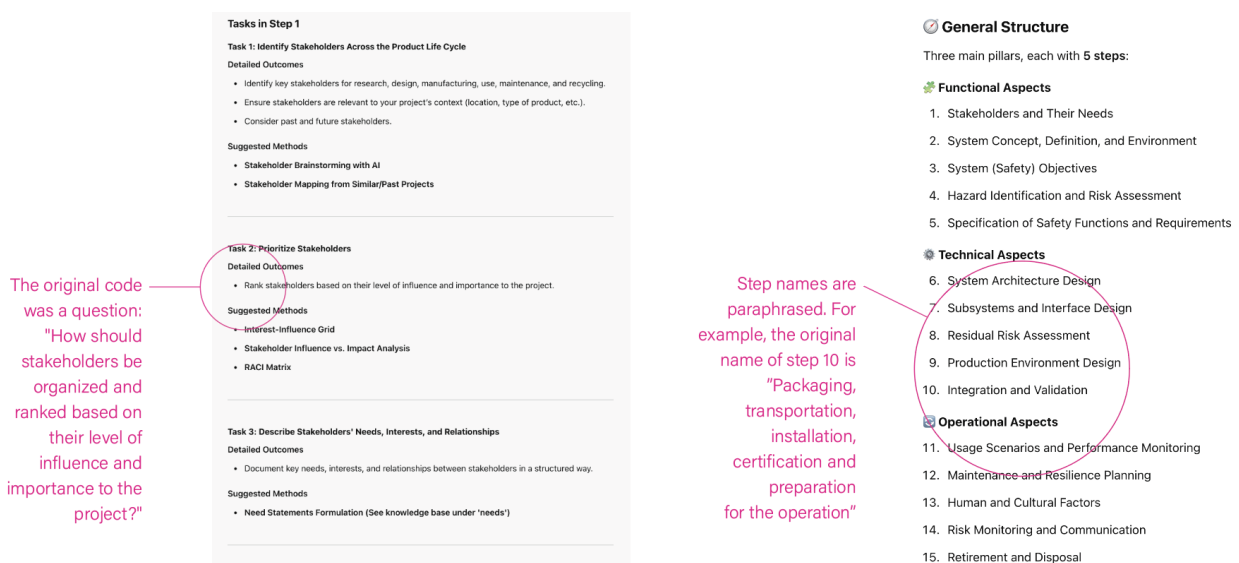


Figure 50: The left image shows the chatbot paraphrasing task outcomes from questions in the knowledge base into statements. The right image highlights incorrect paraphrasing of step names.

## 7.2 Test 2: Comprehension test for visualization

### *Objective*

This test evaluated participants' comprehension of visual mock-ups, assuming the Design Guiding chatbot could display live data within them.

### *Procedure*

Using the Thinking Aloud method, three participants reviewed four visualizations, each representing a different information structure (see Figure 47 ). As they interpreted the visuals, they verbalized their thoughts, and the interviewer asked questions to explore any comprehension difficulties. Their interpretations were then compared to the intended meaning defined by the thesis author, who is also the designer for the visualizations, to assess how accurately the visualizations conveyed the intended message.

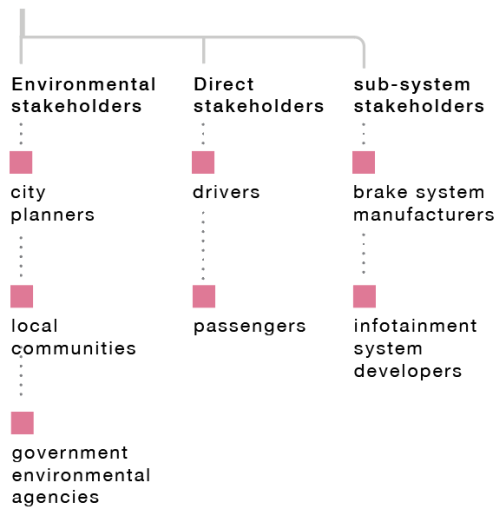
After that, participants conducted a small survey using a Likert scale to rate three aspects: the ease of interpreting the visualizations, the level of engagement they felt, and the suitability of using the visualization to support AI answers.

### *Theoretical framework*

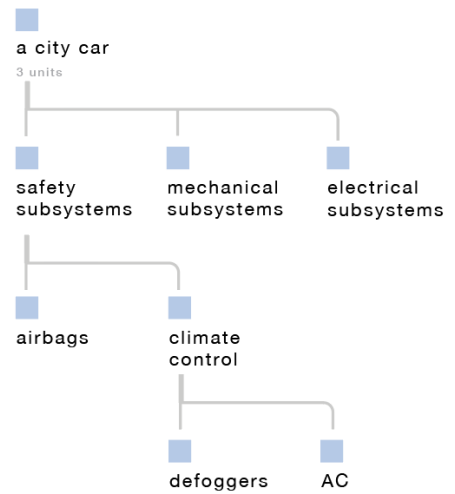
As mentioned earlier, this test followed the Thinking-Aloud method, a widely used approach for exploring user experience (Someren et al., [n.d.](#)). Additionally, the test was guided by principles from visualization comprehension assessments, as outlined in ISO 9186 (International Standard Organization, [2001](#)).

### Non-hierarchical structure

#### Types of stakeholders

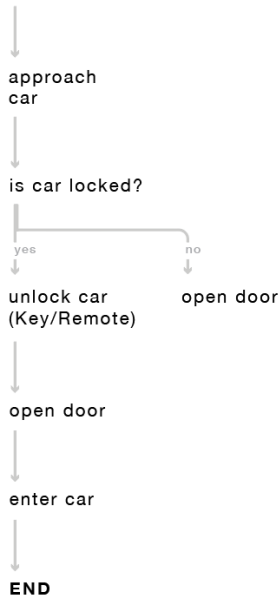


### Hierarchical structure



### Sequential structure

#### START



### Fault Tree Visualization

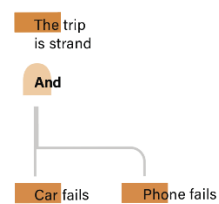


Figure 47 is re-shown here as the material used for the second test

## Data analysis

**Correct Interpretation Rate:** This rate was calculated as the percentage of correct answers out of the total questions answered by all three participants.

**The ease of interpretation score:** we calculated the average score of the Likert scale responses using the standard mean formula. This average was then converted into a percentage to make comparison easier, following this formula:

$$\text{Percentage} = \left( \frac{\sum (\text{Response value} \times \text{Frequency})}{\text{Total responses} \times \text{Maximum score}} \right) \times 100$$

**Engagement rate:** calculated similarly to the ease of interpretation score.

**The suitability as AI answer supporter level:** calculated similarly to the ease of interpretation score.

**Interview data:** The interview data was organized into two main themes: (1) explanations of the suitability of visualizations in supporting AI-generated responses, and (2) potential problems or limitations identified by participants.

## Results

The results of the four visualizations, non-hierarchical (NHV), hierarchical (HV), sequential (SV), and fault tree visualizations (FTV), are presented in Figure 51. With HV and SV, all participants were able to answer correctly all three questions regarding both the intended meanings of the visualizations and the relationships between the depicted elements, such as city planners and the local communities, resulting in 100% correct interpretations. In the case of the NHV, one participant gave an irrelevant answer to one of the three questions. As a result, the interpretation rate was 89%.

However, the FTV showed a significantly low correct interpretation rate. Only one participant answered one out of three questions correctly, resulting in a 16% correction rate. This low performance was due to participants' lack of understanding of the specific interpretation rules for fault tree diagrams. These findings suggest that when introducing specialized risk visualizations such as fault trees, additional explanation, and instruction are necessary to ensure proper interpretation.

In terms of ease of understanding, the HV received the highest rating, with a mean score of 4.6 (equivalent to 92% agree and strongly agree). This was followed by the non-hierarchical (M=4, 80%), sequential (M=3.6, 72%), and, lastly, the fault tree visualization (M=3, 60%), which received the lowest rating. These results aligned with the correct interpretation rate presented previously.

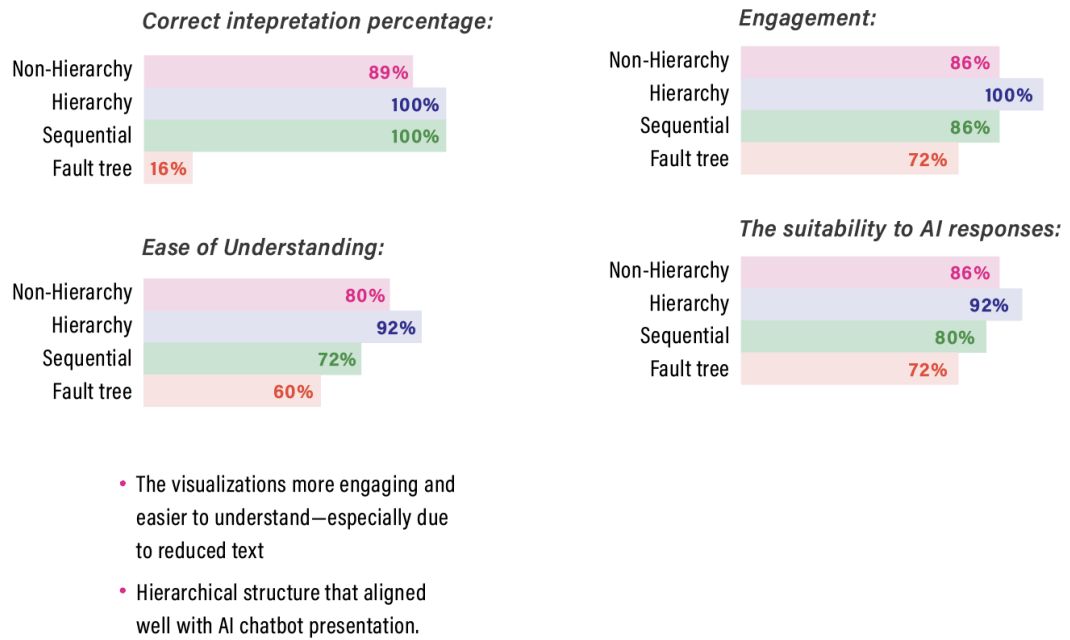


Figure 51: This figure summarizes the results of Test 2, comparing the four visualizations based on correct interpretation percentage, ease of understanding, engagement, and suitability for presenting AI-generated responses.

Regarding engagement level, the hierarchical visualization again received the highest rating, with a mean score of 5 (100%). This was followed by the non-hierarchical and sequential visualizations, both scoring a mean of 4.3 (86%). The fault tree visualization received the lowest engagement rating, with a mean score of 3.6 (72%).

Regarding the suitability of visualizations to support AI-generated answers, the hierarchical structure received the highest rating, with a mean score of 4.6 (92% agree and strongly agree). This was followed by the non-hierarchical visualization (M=4.3, 86%), the sequential visualization (M=4, 80%), and finally the fault tree diagram, which received the lowest rating at M=3.6, 72%.

Interview data revealed that participants agreed that all visualizations supported quicker understanding and were more engaging due to their reduced use of text. The use of color was also positively received. In particular, the hierarchical structure was highlighted as aligning well with the way AI chatbots organize information. Participants noted that this format could effectively support or even replace, the traditional text-based hierarchy of headings and subheadings.

## 7.3 Test 3: Experimental Study

### *Objective*

This final test aimed to assess SIAI's capability to deliver information, offer explanations and guide users in finding, understanding, and applying the NEN-NTA 8287:2021 standard within a practical context.

### *Procedure*

A controlled experimental setup was used, involving six participants divided into two groups: a control group (CG) and an experimental group (EG). Each group consisted of one bachelor's student, one master's student, and one professional with over five years of industry experience. None of the participants had prior knowledge of the Dutch safety standard NEN-NTA 8287:2021. Furthermore, their academic and professional backgrounds varied widely, including fields such as computer science, chemical engineering, game design, business administration, and industrial engineering management.

Participants in the CG worked with the original NEN-NTA 8287:2021 document and standard web search tools (e.g. Google), while the EG used the SIAI system. Within a 1.5-hour time limit, both groups were given the same set of tasks as stated below:

### *Finding critical information*

**Task description:** First, participants from both groups were asked to look up the NEN-NTA 8287:2021 overview, including method purpose, general principles, and method structure, and the descriptions of steps 1, 2, 4, and 6. These steps were chosen because they helped users analyze the safety context of the product, identify potential risks, and develop solution architectures. All of which prepared them for the redesign tasks introduced later in the test.

While the CG had to rely on the Table of Contents of the NEN-NTA 8287:2021 document to find this information, the EG could simply ask the prototype directly.

**Results:** As can be seen in Figure 52, on average, participants in the experimental group (EG) successfully located all 7 out of 7 required pieces of information related to NEN-NTA 8287:2021 (100%), whereas participants in the control group (CG) found approximately 5 out of 7 (71%) on average. The performance analysis of each participant is presented in Appendix C.

Observation showed that the CG spent over 10 minutes perfecting the first answer (method's purpose), leading to little time to explore its principles and structure. Furthermore, while the NEN-NTA 8287:2021 document covers Steps 1, 2, 4, and 6 in both Chapters 7 and 11, participants only found the content in

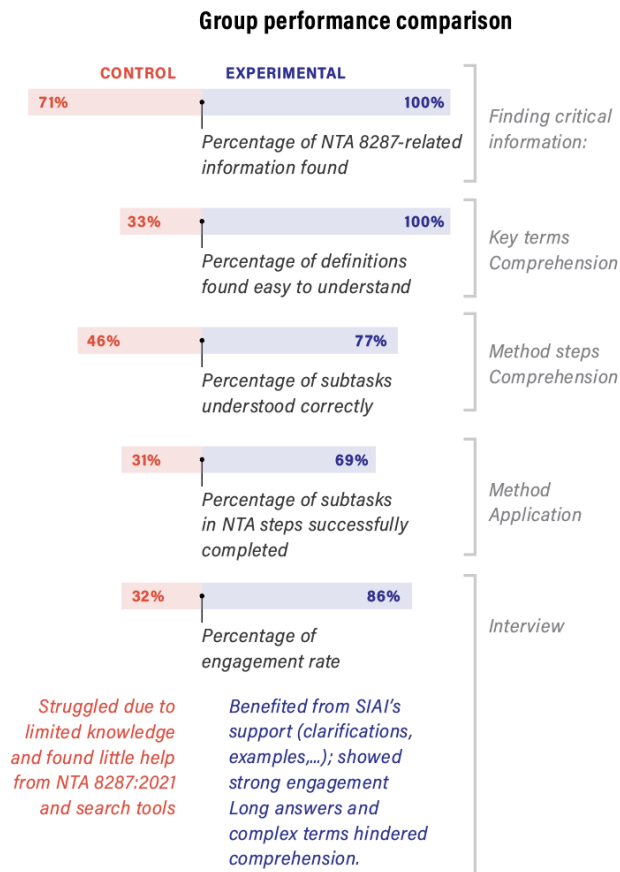


Figure 52: This figure summarizes the results of Test 3

Chapter 11. They completely overlooked Chapter 7, as its relevant information was not clearly indicated in the table of contents.

In contrast, participants in the experimental group could quickly access the NTA 8287:2021-related information by asking the prototype directly. The relevant content appeared within three seconds, helping them locate the information efficiently. However, in one case, a participant misunderstood the method's principles and challenged the prototype. After one attempt to correct the user, SIAI produced fabricated answers to satisfy them (as can be seen in Figure 53). This issue highlights the limited ability of customized AI technology to correct users' misunderstandings.

### Comprehension of key terms

**Task description:** Following the Retrospective Thinking Aloud approach, participants in the CG reviewed definitions of three technical terms (*external interface*, *system under consideration*, and *operational aspects*) by consulting the NEN-NTA 8287:2021 document. These terms were essential to be correctly interpreted for applying the standard.



Briefly name the key principles, with a single word starter

Here are the **key principles** of the **Safety by Design Method**, each starting with a single word:

1. **Human** – Stakeholders, needs, and interactions.
2. **Technical** – Hardware, software, and internal interfaces.
3. **Environment** – Regulations, infrastructure, and physical conditions.
4. **System** – Physical and non-physical interactions with the environment.
5. **Interaction** – Human relationships with systems and support tools.
6. **Regulation** – Compliance with policies, workplaces, and external systems.

Would you like further details on any of these principles?

Are those principles?

Not exactly—those are the **key aspects** considered in the **Safety by Design Method**. The actual **Safety by Design principles** focus on **how** to reduce risks through design choices.

Here are the **true principles**:

1. **Eliminate** – Remove hazards entirely through design.
2. **Reduce** – Minimize how often hazards occur.
3. **Prevent** – Lower the probability of hazards leading to accidents.
4. **Protect** – Implement safeguards and protective measures.
5. **Inform** – Communicate risks through warnings and user instructions.

Would you like examples of how these principles apply to a specific product?

Figure 53: The image shows an instance where the SIAI failed to correct the user's misunderstanding and instead reinforced it by providing hallucinated information

Figure 53: The image shows an instance where the SIAI failed to correct the user's misunderstanding and instead reinforced it by providing hallucinated information

Meanwhile, participants in the EG obtained the definitions by interacting with SIAI, which not only delivered accurate definitions but also explained them with definitions of complex words within the main definition and contextual examples. EG participants were encouraged to ask follow-up questions to clarify uncertainties.

Following this, the interviewer conducted a series of open-ended questions to uncover what characteristics of the NTA 8287 document and SIAI's features helped or hindered participants' grasp of the definition. An example question was, *"How would you describe the experience of interpreting these key terms?"*.

**Results:** On average, participants in the CG found only one out of three definitions (33%) easy to understand. The remaining two terms—*external interface* and *system under consideration*—were perceived as *"very"* challenging, primarily due to complex terminology and long, dense sentence structures presented in the definitions.

In contrast, participants in the EG reported that all three definitions (100%) were clear and understandable. At first, these participants also struggled to understand the definitions, similar to those in the CG. However, the SIAI's ability to simplify language, respond to follow-up questions, and especially provide relevant examples helped them gain clarity. One participant even mentioned: *"I did not fully understand the definition*

until I read the examples.”. Two participants also noted that providing definitions of complex terms did not make the original definition clearer. Instead, they sometimes made the explanation even more confusing.

### ***Understanding and coverage of the steps described in the method***

**Task description:** In the next task, participants were asked to briefly explain what they were expected to perform in each subtask within steps 1, 2, 4, and 6, reflecting their understanding of the NEN-NTA 8287:2021. For example, they were asked to briefly describe how they would approach risk evaluation in a shared Google Doc.

Participants in the CG had to interpret and understand the method independently with the help of NEN-NTA 8287:2021 and the search browser. In contrast, those in the EG received guidance and real-time feedback from SIAI, which was designed to support their comprehension and walk them through each step. The number of accurately described subtasks was recorded for analysis.

**Result:** On average, participants in the CS group correctly described 7 out of 16 sub-tasks (46%), whereas those in the EG group explained 12 out of 16 (77%), indicating a better grasp of steps 1, 2, 4, and 6. However, observations revealed that in some cases, SIAI was unable to respond accurately to questions about NEN-NTA 8287:2021 due to gaps in its knowledge base, resulting in incorrect or unclear answers.

### ***Applying the NEN-NTA 8287:2021 in a redesign assignment***

**Task description:** Finally, participants of both groups were tasked with redesigning a Dutch traffic system, which involved analyzing the safety scenario, identifying potential risks, and creating a product architecture for their solution idea. These tasks were outlined in steps 1, 2, 4, and 6 of the NEN-NTA 8287:2021 standard.

While participants in CG needed to do this task independently, those in EG were guided by SIAI and collaborated with the prototype to finish the assignment. At the end, participants in EG submitted their results to the SIAI for feedback. This subtask was designed to evaluate how thoughtfully participants engaged with and applied the AI-generated feedback.

**Results:** Observations revealed that participants in the EG collaborated with SIAI by refining its outputs to have their final results. In contrast, CG participants relied more on their own ideas, typically after struggling to find useful information through traditional search engines.

On average, EG participants completed 11 out of 16 subtasks (68%), compared to just 5 (31%) in the CG, and the responses from EG tended to be more detailed (See the performance comparison in Appendix C). Furthermore, while EG participants generally found the feedback from SIAI helpful, they cautiously approached it and selectively used the suggestions that only made sense to them.

## Interview

**Task description:** At the end of the test, each participant took part in a one-on-one interview to reflect on their experience, discussing both the challenges they faced when using the NEN-NTA 8287:2021 document and SIAI, as well as the positive aspects of their interaction. Furthermore, they were also asked to rate their level of engagement on a Likert scale ranging from 1 (very low engagement) to 5 (very high engagement).

## Results:

### Experimental Group

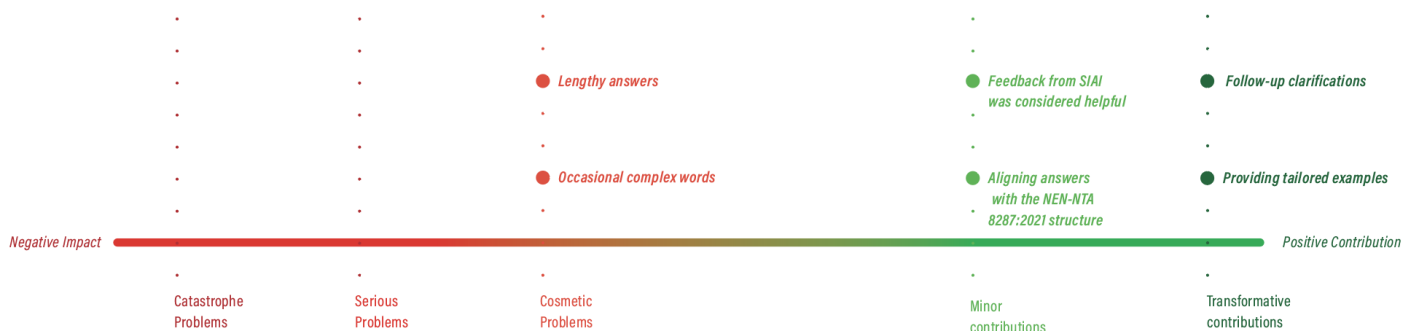


Figure 54: This figure compares the positive contributions and negative impacts of the SIAI versus traditional standard-based methods, based solely on interview responses. Positive contributions are categorized as minor (briefly mentioned or considered “nice to have”) and transformative (frequently mentioned and highly valued). Negative impacts are classified into cosmetic problems (minor irritations with no effect on performance), serious problems (notable issues that users managed to overcome), and catastrophic problems (issues that prevented task completion).

A comparison of both groups' answers revealed a clear pattern that helps explain the performance gap observed in earlier tasks. A summary can be seen in Figure 54. Participants in EG highlighted several valuable contributions from SIAI that made task completion easier. These included:

- **Providing tailored examples:** A highly appreciated feature was the prototype's ability to provide project-relevant examples. This helped participants connect the abstract content of NEN-NTA 8287:2021 to their own work, setting the prototype apart from traditional search tools: *"I understood the method better after reading the examples. That was why, in the next steps, I kept asking for more."*
- **Follow-up clarifications:** Another valuable feature of the prototype is its ability to respond to follow-up clarification questions. As one participant said, *"It was not always clear how the prototype explained the method at first, especially the definitions, but after asking a few times, I eventually understood what it (the prototype) meant"*
- **Presenting answers aligned with the structure and content of the NEN-NTA 8287:2021 method:** Participants also appreciated how SIAI arranged information in a way that directly related to the NEN-NTA 8287:2021. For example, SIAI classified risks into functional, technical, and operational groups, helping participants clearly understand how the method can relate to their projects
- **Feedback from SIAI was considered helpful:** Participants found the evaluation feature guided them toward areas needing improvement. As one participant said: *"After working with SIAI for a while, I started to feel a bit lost and overwhelmed by all the information. However, the evaluation feature helped me get back on track and refocus on the areas that still needed improvement."* However, all participants expressed some skepticism about the grades provided by the chatbot. One participant expressed, "I did not believe my work on this task deserved a high grade like 7, as the prototype judged. But it gave me detailed feedback on what I had included and what I had not. Just the grade felt odd."

Besides its valuable features, SIAI also has some limitations that participants pointed out:

- **Lengthy answers:** Participants reported that the prototype sometimes gives overly long answers with multiple points at once. However, this issue was considered cosmetic, as it caused slight irritation and did not affect the participant's performances. Furthermore, this problem was easily resolved, as one participant asked the prototype to provide a more concise response.
- **Occasional complex words:** Since none of the participants had prior knowledge of the NEN-NTA 8287:2021, they were unfamiliar with some of the terms used by the prototype, such as functional, operational, and technical concepts. Although the prototype provided explanations alongside these complex words, some participants expressed a preference for avoiding the use of such terms altogether. They believed that simpler language would be even more helpful as it eliminated all barriers to understanding.

Interviews with the control group did not provide any positive feedback, but revealed two main reasons behind the participants' struggles:

- **Inadequate knowledge of the assignment and NEN-NTA 8287:2021:** All participants agreed that the main reason the redesign assignment was so challenging was their limited knowledge of the Dutch traffic system and the NEN-NTA 8287:2021. Participants noted that several new safety concepts were explained in the standard using technical terms and long, complex sentences. These factors made it difficult to fully understand the method and apply it within the one-hour time frame. As one participant put it, "It seems like this document is not designed for a total beginner like me."

- **Little support from NEN-NTA 8287:2021 and search browser:** Participants found little to no guidance in the NEN-NTA 8287:2021 for applying it to the redesign assignment. As a result, they relied on Google, which often failed to provide relevant information, especially on operational, functional, and technical concepts, as well as stakeholder examples and risks. Ultimately, they had to rely on their own understanding.

**Participant engagement:** The Likert scale results reflected a clear difference in engagement between the two groups. The experimental group scores were 5, 4, and 4, averaging 4.3 out of 5 (86%), while the control group averaged just 1.6 (32%), with individual scores of 1, 2, and 2.

These scores aligned with participants' reported experiences through the interview. In the experimental group, one participant was impressed by the AI-generated answers by stating, "*Without the AI, I would not have completed this task within one hour at this level of detail.*" Another expressed enthusiasm when saying they would be willing to continue working with the prototype for another hour to refine their responses.

In contrast, the control group's low engagement was not only evident in the scores but also in their behavior. For example, one participant expressed the desire to end the test early, while another repeatedly asked how much time was left.

# 8 A Storyline Recap

## 8.1 Answers for sub-research question 1: The Identified Problems

Safety standards are often used to introduce students to the concept of standardization. They offer structured guidance, recommendations, and principles for designing high-performance products. However, learners frequently face several difficulties when engaging with these materials, as detailed below:

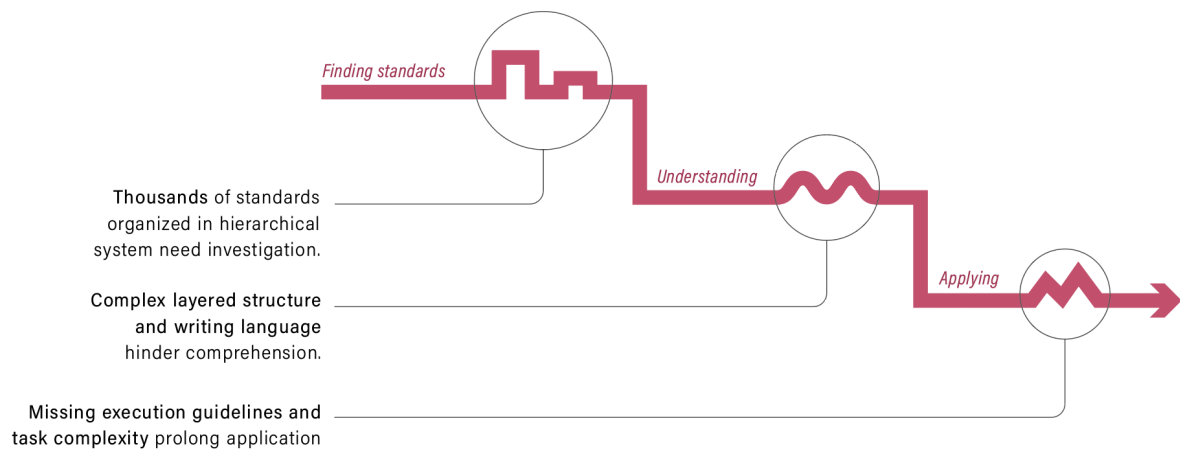


Figure 2 is reshown here to summarize the identified problems

**1. Challenges in Finding Safety Standards: Complex Navigation, Overwhelming Lists, and Lack of Selection Guidance:** Learners seeking safety standards often need to visit the official websites of regulatory authorities, for example, the European Commission. However, navigating these websites can be challenging due to their complex hierarchical structure. Information is typically organized under a number of directives and subgroups, resulting in an overwhelming number of choices that slow down the search process.

Even after identifying the correct directive and locating the recommended lists, learners often face a lengthy table containing hundreds of standards. For instance, the Machinery Directive includes up to 800 standards spread across more than 200 pages. These lists lack built-in filtering tools or clear selection guidance, making it difficult for learners to distinguish relevant from irrelevant information and efficiently find what they need.

**2. Barriers to Understanding: Hyper-specific Definitions, Complex Formatting, and Dense Writing Styles:** Even after identifying the relevant standards, understanding them quickly remains a challenge. These documents are typically structured in a layered, hierarchical format with numerous sections and

subsections that frequently refer to one another. As a result, the reading experience is often disjointed, requiring users to constantly flip between sections to find key information.

Additionally, the use of technical terminology is necessary for precision communication, but they are often explained by lengthy and complex definitions. Many of these definitions are interdependent, making them even harder to grasp without understanding related concepts first.

Finally, safety standards tend to rely on dense, information-heavy sentences to convey exact procedures, logic, or abstract principles. While this ensures accuracy, it often reduces learner engagement, leading to confusion and slower comprehension.

### ***3. Applying Without Support: Lack of Practical Guidance and Overwhelming Task Complexity:***

Some safety standards focus primarily on abstract principles, often lacking practical implementation guidance such as step-by-step instructions, recommended tools, or evaluation criteria. As a result, students, particularly those new to the subject, have limited references to help them apply these standards effectively. Furthermore, the complexity and sheer number of required safety analysis tasks can make the application process feel overwhelming.

## **8.2 Answers for sub-research question 2: The Gap**

***Gamification as One of The Most Popular Teaching Methods:*** To help learners better study safety standards, university courses often use a range of instructional strategies and technologies, including gamification, team-based learning, and project-based learning (Bouri et al., 2020; Rajabalinejad, 2020). For instance, the Safety by Design course at the University of Twente combines these methods by creating a game called the Safety Cube Method game. While these approaches tend to make learning more interactive and enjoyable through collaborative discussions and problem-solving activities, they often lack the flexibility to offer real-time, personalized support tailored to learners with varying levels of expertise, literacy, and background (Bouri et al., 2020; Rajabalinejad, 2020).

***Rapid Integration of Generative AI in Safety Education:*** In recent years, the use of generative AI in education has grown significantly worldwide (Onatayo et al., 2024). Despite ongoing concerns and critiques, studies have shown its promise in delivering personalized learning experiences and supporting safety analysis-related tasks. For example, generative AI tools such as ChatGPT have been used for activities like hazard identification, risk management, compliance support, and classroom-based safety training (Liu et al., 2023; Onatayo et al., 2024; Uddin et al., 2023)

**The Remaining Gap:** However, general-purpose AI systems often fall short when it comes to providing accurate, coherent, and domain-specific information, as required by safety standards (Oviedo-Trespalcacios et al., 2023). Therefore, this study addresses that limitation by investigating a solution designed to help a diverse group of safety learners engage with safety standards in a more **reliable, intuitive**, and **customized** way.

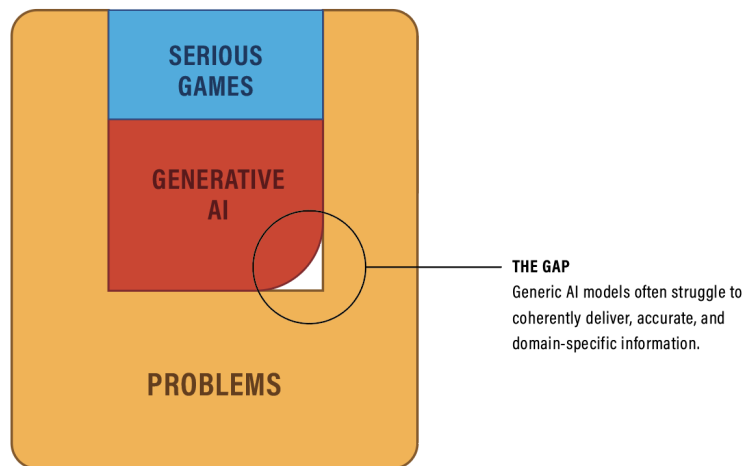


Figure 20 is re-uploaded here for illustration purposes

## 8.3 Answers for sub-research question 3: Design Solution

**The Solution Direction:** Following the extensive exploration detailed in Chapter 5, we concluded that generative multi-modal AI, enhanced with customized knowledge and visualizations, holds strong potential for improving safety education and training. By equipping AI with domain-specific learning content and clear instructions, the system can generate tailored responses that deliver precise, relevant information in a clear and visually engaging manner.

**The Solution Architecture:** Then, Section 6.1 introduces four AI modules designed to act as virtual assistants for students. The selection of these modules aligns closely with stakeholder needs and the requirements established earlier in Section 3.3. Figure 37 illustrates how the four modules, Design-Mate, Assessment-Mate, Memo-Mate, and Visual-Mate, interact with one another and with smaller supporting models. Further explanation is present as below:



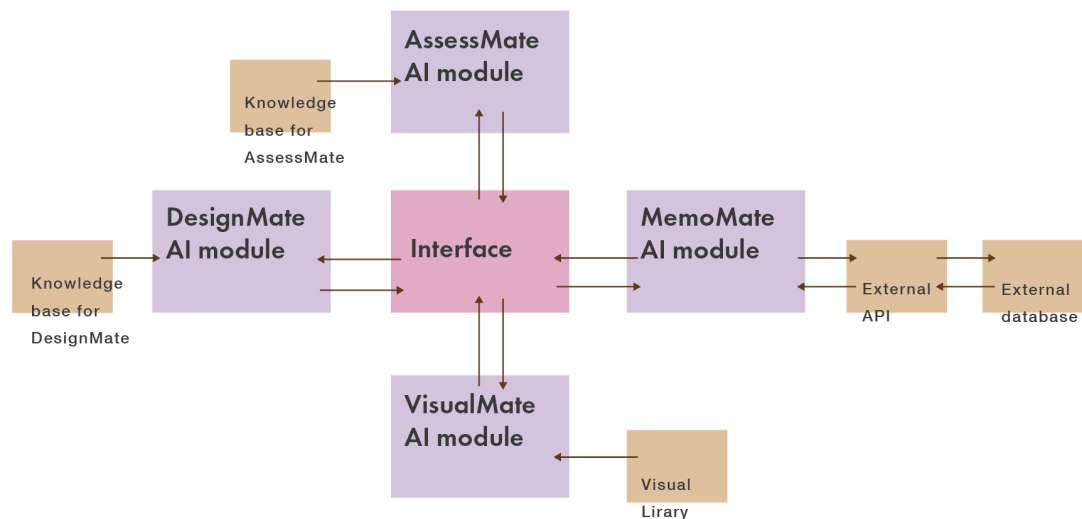


Figure 37 is reshown here for reference

- **DesignMate AI module:** This module supports students in navigating the safety-by-design process by offering structured guidance derived from NEN-NTA 8287:2021. It also integrates lecture content, instructor insights, and examples from course reports or online sources.
- **AssessMate AI module:** Acting as an evaluation assistant, this module reviews student submissions and delivers constructive feedback, highlighting areas that require further development.
- **MemoMate AI module:** This module helps participants efficiently record their findings, capturing insights and progress from their interactions with the other AI tools.
- **VisualMate AI module:** This module transforms information from the other modules into visual representations, using hierarchical, non-hierarchical, and sequential formats to enhance clarity and understanding.

**Final prototype:** After a long iteration design process with both AI chatbot and visualizations, this thesis introduces SIAL (Safety Instructor AI), a GPT-based chatbot system. The core purpose is to provide accurate safety information drawn from standards. It is designed to present exact definitions of technical terms, an overview of the method, and deliver detailed content with step-by-step execution guidance informed by the instructor's experience. When users request a risk analysis, SIAL responds using a structured template. It also includes links to visual mockups to help users organize their findings and allows them to save their work directly to a Google document. Additionally, it offers personalized feedback on student submissions (see Figure 6.3.1 for the interface and mock-up examples).

The four visual mock-ups proposed in this project were intentionally designed to be adaptable across different scenarios. This adaptability is grounded in the use of nodes that represent three commonly appearing safety elements: human, system, and environment. In addition, the information was structured

using three common formats—non-hierarchical, sequential, and hierarchical—to accommodate and organize a wide range of content types effectively. Additionally, the information is organized using three familiar structures—non-hierarchical, hierarchical, and sequential—to accommodate and present a wide range of content types effectively. The main purpose of these mockups is to provide both users and the chatbot with a simple, consistent yet flexible framework for presenting information

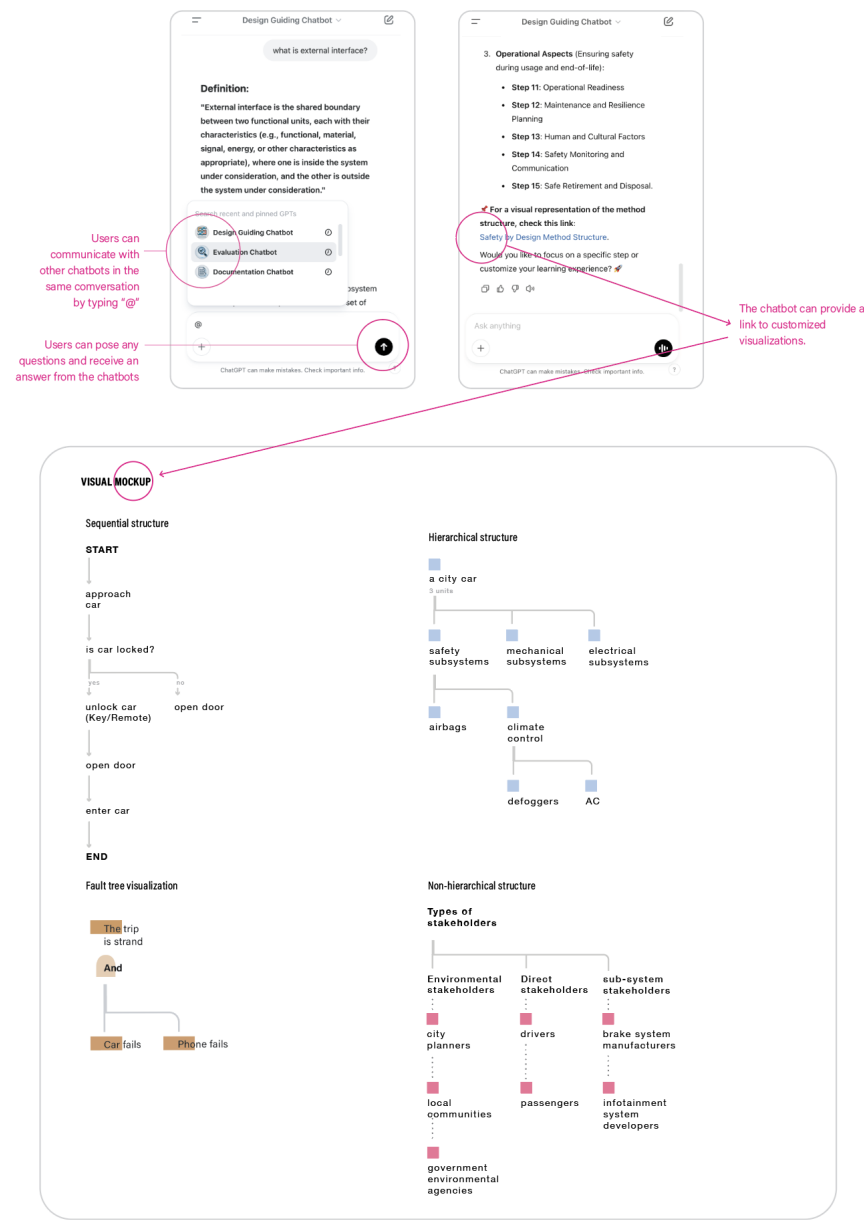


Figure 6.3.1 is shown again here for reference.

# 8.4 Answers for sub-research question 4: Tests and Results

To evaluate the prototype, we applied the Hallmarks framework proposed by Kozierok et al. (2021). The testing goal and outcomes are summarized in Figure 48.



Figure 48 is re-uploaded here for references

**Test 1: Robustness Evaluation:** Test 1 focused on whether SIAI (Safety Instructor AI) could reliably and accurately provide NEN-NTA 8287:2021-related information, regardless of prompt variation.

We asked the prototype to define three technical terms, summarize the method overview, and explain one specific method step with its execution guideline. Each task was tested using four prompt types: direct, paraphrased, indirect, and those containing typos—five times each.

Every interaction counted as one test instance. Performance was measured by evaluating accuracy and completeness, with the success rate calculated as the ratio of correct, complete responses to total attempts.

**Test 2: Comprehension test for visualization:** This test assessed how well participants understood visual mock-ups, assuming the Design Guiding chatbot could display live data within them.

Three participants used the Thinking Aloud method to review four visualizations, each showing a different information structure (see Figure 6.3.1). As participants reviewed the visuals, they described their thinking process and responded to follow-up questions to uncover any misunderstandings. Their interpretations were then compared to the intended message. Finally, they completed a brief survey rating the visuals on clarity, engagement, and usefulness in supporting AI responses.

**Test 3: Experimental Study:** The purpose of this final test was to evaluate how effectively SIAI could deliver information, provide clear explanations, and guide users in navigating and applying the NEN-NTA 8287:2021 safety standard in a real-world scenario.

A controlled experiment with six participants, split into a control and an experimental group. Each included a bachelor's student, a master's student, and an experienced professional. All participants were unfamiliar with NEN-NTA 8287:2021 and came from diverse fields such as engineering, business, and design.

The control group used the original NEN-NTA 8287:2021 document along with standard web tools like Google, while the experimental group interacted with the SIAI system. Both groups were given 1.5 hours to complete the same set of tasks:

**1. Finding Critical Information:** participants were asked to locate key information from the NEN-NTA 8287:2021 standard, including the method's purpose, general principles, structure, and detailed content of steps 1, 2, 4, and 6.

**2. Comprehension of key terms:** Using the Retrospective Thinking Aloud method, participants looked up three key terms—*external interface*, *system under consideration*, and *operational aspects* using NTA:8287 or SIAI. Afterwards, the interviewer asked open-ended questions to explore which aspects of the document or SIAI supported or hindered their understanding.

**3. Understanding and coverage of the steps described in the method:** Next, participants needed to outline what each subtask in steps 1, 2, 4, and 6 requires them to perform, showing their understanding of NEN-NTA 8287:2021.

**4. Applying the NEN-NTA 8287:2021 in a redesign assignment:** Finally, participants from both groups redesigned a Dutch traffic system by analyzing safety, identifying risks, and outlining a solution based on steps 1, 2, 4, and 6 of the NEN-NTA 8287:2021 standard.

# 9 Discussion

This thesis explored how a customized GPT-based AI prototype, SIAI (Safety Instructor AI), could support a diverse group of safety learners in finding, understanding, and applying the safety standard NEN-NTA 8287:2021. In designing the solution architecture, special attention was given to the varied needs of stakeholders, including differences in culture, literacy levels, and domain expertise. The prototype delivers customized textual and visual guidance to assist learners in designing products with an emphasis on safety, using specialized materials from the *Safety by Design* course at the University of Twente. To evaluate its impact, a series of three tests were conducted, including a robustness evaluation of AI responses, a comprehension test for visual mockups, and experimental study and interviews. All offer a comprehensive view of the AI's effectiveness. The results of these tests are summarised in Figure 48

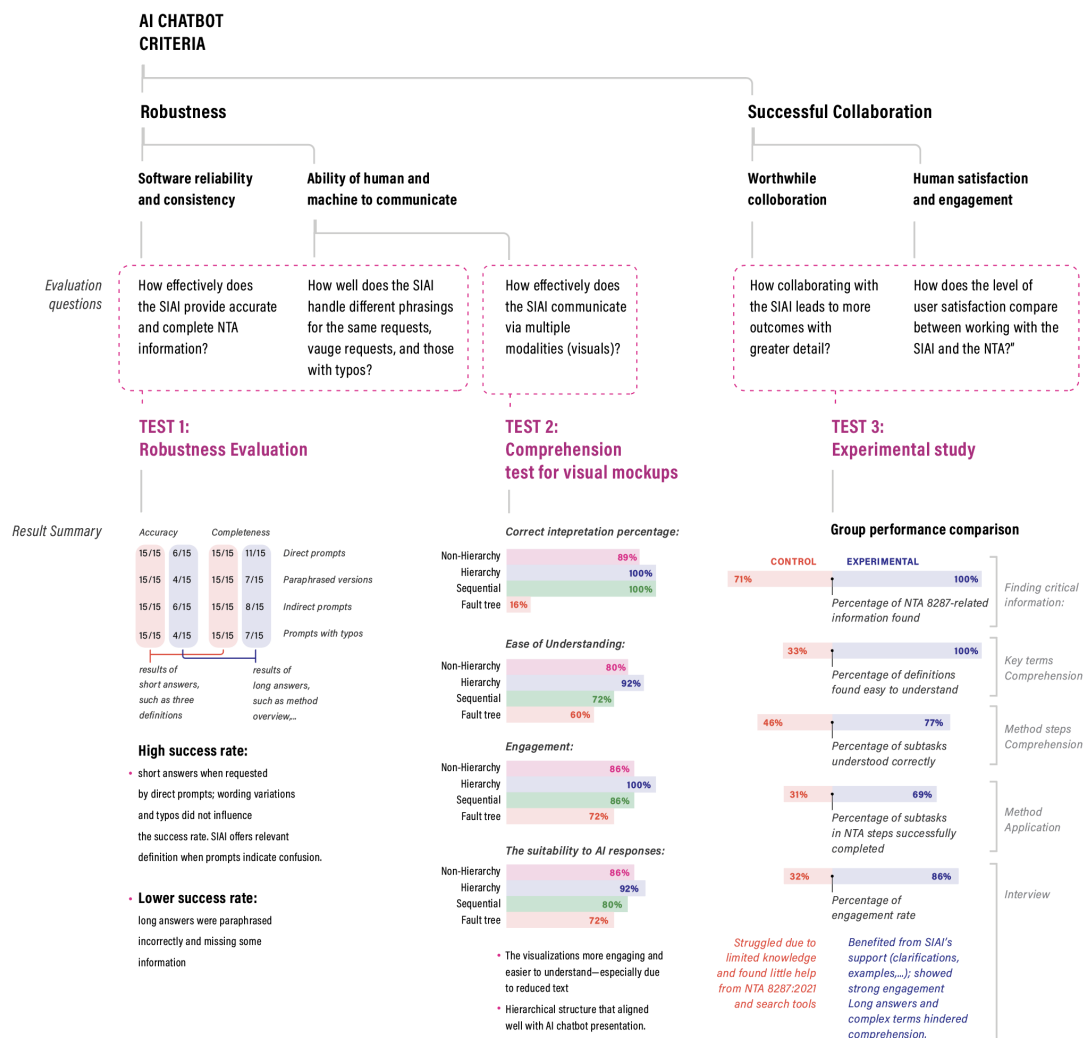


Figure 48 is re-uploaded here for references

### ***Test 1: Robustness Evaluation***

The results from the robustness evaluation revealed that SIAI was capable of delivering both general and NEN-NTA 8287:2021-related information in all trials across four prompt styles (direct, paraphrased, indirect, and prompt with typos). In particular, the prototype could provide definitions of technical terms, summarize the method, and deliver in-depth content, such as step descriptions and execution guidelines, all directly based on the NTA document and the Safety by Design course materials. At the same time, SIAI could offer examples tailored to users' projects, using general wisdom. This capability of SIAI reflects the performance observed in the customized AI chatbot described in studies by Hakim et al. (2024) and Contreras Aguilar et al. (2024), reinforcing the idea that customized AI technologies can effectively deliver both general and domain-specific knowledge.

Furthermore, when asked for brief answers regardless of different prompt styles, SIAI could deliver with the highest degree of accuracy and completeness. The test found that SIAI provided precise, word-for-word definitions and their relevant explanations in all 60 test cases, resulting in a success rate of 100% in both accuracy and completeness, surpassing our expectations. This high level of performance was not influenced by word paraphrasing or minor typos in input prompts. Even when the author did not ask directly but simply expressed confusion or unfamiliarity, SIAI could still provide the targeted information. These findings suggest that SIAI is capable of understanding a range of communication styles and recognizing users' underlying needs. These results reinforce earlier observations from the study by Contreras Aguilar et al. (2024), highlighting the intuitive usability of AI chatbots.

When generating longer responses, SIAI began to reorganize and paraphrase the source material, and more importantly, it omitted information (e.g., citations). This paraphrasing behavior is a commonly known trait of generative AI models, rooted in how they operate: these models find information from their database, reconstruct, and express it in their own ways (Hassanipour et al., 2024). Therefore, the way the information is presented may vary slightly each time. Furthermore, the omissions of information could possibly be explained by the limited context window of AI models—the maximum amount of data the models can process at once. When responses exceed this limit, the models must prioritize what they think is the most relevant, resulting in overlooking certain details (Dong et al., 2024). These findings reflect the current limitation of the generative AI model in sustaining absolute accuracy and completeness when producing extended responses, aligning with a study by Hatia et al. (2024).

### ***Test 2: Comprehension test for visualization***

The comprehension test of the visual mock-ups demonstrated their strong ability to improve both user engagement and understanding during interactions with the chatbot. Among the four visualizations evaluated—non-hierarchical, hierarchical, sequential, and fault tree—the hierarchical and sequential formats stood out with 100% correct interpretation rates. Specifically, participants were able to clearly understand

the intended meanings and accurately identify the relationships between elements depicted in the visualizations. The non-hierarchical format also showed strong performance, with an 89% correct interpretation, affected only by one irrelevant response.

A clear trend emerged, showing that the hierarchical visualization consistently ranked as the most understandable (92%), most engaging (100%), and most compatible with AI-generated responses (92%). Interview feedback further highlighted the benefits of rapid interpretation across all visual formats, especially hierarchical visualization. One participant noted that the hierarchical structure closely mirrors the logical organization commonly found in chatbot content.

However, the fault tree visualization received a notably low interpretation rate, with only 16% of responses answered correctly. This was likely due to participants' lack of familiarity with the specific rules of interpretation used in fault tree analysis. As a result, participants also reported lower levels of engagement and ease of understanding. During the interview, once the interpretation rules were explained, participants quickly understood the visualization, noting that *"now it makes sense"*. This finding highlights the importance of providing users with additional explanation or instructional support when introducing specialized risk visualizations, particularly those not commonly understood by individuals outside the safety domain.

### ***Test 3: Experimental Study***

The third test revealed that the experimental group (EG) using SIAI outperformed the control group (CG), which relied on the NEN NTA 8287:2021 standard and traditional search engines. Particularly, the test showed that SIAI enabled participants to access a greater amount of NTA-related information. While all participants in the EG successfully identified all seven required pieces of information (100%), those in the CG managed to find only 5 out of 7 on average (71%). This improved performance was attributed to SIAI's ability to quickly deliver relevant information, consistent with the findings reported in the studies by Uddin et al. (2023) and Onatayo et al. (2024).

In addition, SIAI proved helpful in clarifying technical terms, leading to a higher number of words being perceived as understood by participants. While those in the control group reported understanding only one out of three terms (33%), participants in the experimental group confirmed understanding all three (100%). Interview data showed that this improvement was credited to the SIAI's ability to simplify complex terms using plain language and illustrate them with examples directly related to the participants' own projects. This ability of SIAI aligns with the advantages of Generative AI models in explaining difficult and field-specific information (Howell, 2024; Liu et al., 2023).

Also, under the theme of comprehension, on average, participants in the EG could accurately explain what was expected in 12 out of 16 subtasks (77%) while the number of CG was 7/16 (46%). This suggests a clear understanding of steps 1, 2, 4, and 6 in the NEN-NTA 8287:2021. However, observations revealed that

in one case, SIAI offered inaccurate information about NEN-NTA 8287:2021 and failed to correct the participant's misunderstanding. Hence, it is recommended that future students maintain access to the original NEN-NTA 8287:2021 document while working with SIAI. In other words, SIAI should be viewed as a supplement, **not** a replacement, for the NTA document.

The task of applying NEN-NTA 8287:2021 to solve a real-world challenge demonstrated a successful collaboration between participants and the SIAI, leading to higher task completion rates and more detailed answers. While, on average, participants in CG could briefly address 5 out of 16 subtasks (31%), those in EG addressed 11/16 (68%). The responses of the EG group were also more well-written and detailed, such as including a larger number of stakeholders and risks. The prototype was especially effective in supporting participants through tasks such as identifying stakeholders, environmental factors, risks, and brainstorming solutions. These outcomes align closely with findings from recent research in the field (Hakim et al., 2024; Onatayo et al., 2024; Uddin et al., 2023). The improved performance in the application task highlights AI's potential to translate theoretical safety principles explained in NEN-NTA 8287:2021 into real-world practice.

Additionally, observation showed that participants did not blindly follow the feedback provided by SIAI. While they appreciated the level of detail in SIAI's feedback, they engaged critically with the suggestions and were cautious about its reliability. These behaviors reflect a high level of digital literacy among participants and an awareness of the limitations of generative AI. It also suggests that participants are learning to use AI tools more responsibly and transparently. Furthermore, interview data also indicate higher engagement and satisfaction among participants in EG. They rated their experience on average 4.3 out of 5 (86%), reflecting a very high level of engagement. In contrast, the control group scored an average of 1.6 (32%).

Importantly, observation also revealed the growing dependence of participants on this technology. All participants in EG co-created results by refining SIAI's responses rather than manually executing the tasks as the CG group did after failing to find relevant information from Google. This suggests that dependency on AI technology among students should be expected when such technology becomes more integrated into future education. As a result, there is an urgent need to develop a new evaluation framework that can more effectively assess both the collaborative process, learning outcomes, and students' contribution in AI-supported environments.



# 10 Conclusion

SIAI (Safety Instructor AI) demonstrates a highly potential solution for enhancing the accessibility of safety-related content in educational settings. Its development focused on delivering accurate information derived from standard documents, including precise definitions of technical terms, overviews of safety methods, in-depth content, detailed step-by-step guidance informed by the instructor's expertise, and visual mockups for information structure. By customizing and presenting both general and discipline-specific materials in textual and visual formats, SIAI supports safety learners in finding, understanding, and applying the Dutch safety standard NEN-NTA 8287:2021.

Through a series of three tests, we evaluated different aspects of the SIAI prototype. The first test, focused on robustness, demonstrated that SIAI can reliably provide both general and specific information from the NEN-NTA 8287:2021, particularly in short-form responses. It also showed the AI's ability to adapt to varied communication styles used by participants.

The second test assessed participants' comprehension of visual mock-ups, revealing the strong potential of using visual elements to enhance both engagement and understanding in chatbot interactions. In particular, the hierarchical structure aligned well with existing chatbot formats, offering a promising alternative to traditional layouts based on headings and subheadings.

The third experimental study examined the collaboration between SIAI and users, showing that customized AI technology effectively supported learning by providing examples, using simplified language, answering follow-up questions, offering feedback, and delivering information quickly. This enabled participants to quickly locate relevant safety-related content, better understand the NEN-NTA 8287:2021, and apply it within a design assignment. This prototype illustrates how generative AI could influence future approaches to teaching safety standards.

## **Limitations**

Despite the promising outcomes, this thesis also acknowledges several important limitations. First, it does not offer conclusive evidence regarding the reliability of SIAI in generating accurate information across a wide range of fields and safety-related topics. The evaluation focused solely on content from the NEN-NTA 8287:2021 standard, and therefore, the findings may not apply to other safety standards.

Second, the issue of potential plagiarism was not covered in this study because of the limited testing time. Participants were not asked to check their answers against academic sources or verify the information provided by the AI prototype. However, it is important to highlight that in real situations, students are

responsible for using AI-generated outcomes for their submitted work to teachers. Additionally, the students need to explain the extent of their usage of AI across the project.

Third, this thesis focuses only on the short-term benefits of using the AI chatbot and does not provide evidence of its long-term impact on learning. Important aspects such as creativity, independent thinking, and social collaboration skills were not explored and require further research to address these potential issues.

Finally, the current technological limitations of customized GPT models did not allow this thesis to assess SIAI's ability to offer a fully integrated multimedia learning experience. As a result, the impact of when visuals and text could work together in real-time to support learning was not evaluated.

## ***Recommendations***

In terms of implementation, we recommend that SIAI be used only as a supplementary tool to support student learning. It is not a replacement for official safety standards or other instructional materials.

Additionally, SIAI's outputs, particularly its recommendations, should be critically reviewed, as they are not intended to replace the teacher's professional judgment. This is particularly important given that educators typically assess the quality of student work based on their expertise, experience, and intuitive understanding of clarity and persuasiveness—elements that are difficult to replicate through AI.

We also suggest that teachers consider asking students to submit their chatbot conversations alongside their assignments. Based on our observations, students' understanding can often be reflected in how they formulate prompts and how deeply they engage with AI to seek accurate and meaningful responses.

For future development of SIAI, we recommend combining a rule-based chatbot with an AI-generated chatbot. This hybrid approach could address issues related to inconsistency and paraphrasing in long responses. The rule-based chatbot excels at following predefined rules and decision trees, making it ideal for delivering precise, word-for-word information. This ensures that key content from the NEN NTA 8287 is presented consistently and remains under the full control of the developers. By integrating this with a generative AI component, the chatbot can also adapt and apply that information to user-specific project contexts, offering personalized support while maintaining accuracy and reliability.

Additionally, since current no-code customized GPTs are not capable of generating real-time visualizations based on specific references, it is worth exploring the development of a local AI chatbot using GPT APIs combined with other technologies to make this functionality possible.

### *Last words*

The outcomes of this thesis, both positive and negative, offer valuable insights into the ongoing discussion on integrating AI solutions in safety education. The results suggest that generative AI can help educators provide customized support for students in finding, understanding, and applying safety standards. The effectiveness is even stronger for those with limited knowledge or lower literacy levels. Nonetheless, additional research is necessary to address the study's limitations and strengthen these initial findings.

## 11 Acknowledgment

During the preparation of this work, I used **ChatGPT and Consensus** to generate ideas, provide suggestions and feedback, find literature, and refine writing. This included grammar checking, finding synonyms, and improving sentence structure to enhance the overall flow and readability.

After using this tool/service, we thoroughly reviewed and edited the content as needed, taking full responsibility for the final outcome

# References

- Abuiyada, R. (2018). How Students' Multiple Intelligences Differ in Terms of College and Gender. International Journal of Business and Social Science.  
[https://www.academia.edu/59102746/How\\_Students\\_Multiple\\_Intelligences\\_Differ\\_in\\_Terms\\_of\\_College\\_and\\_Gender](https://www.academia.edu/59102746/How_Students_Multiple_Intelligences_Differ_in_Terms_of_College_and_Gender)
- Almeida, F. L. F. (2017). Learning Entrepreneurship with Serious Games - A Classroom Approach.  
<https://arxiv.org/abs/1710.04118v1>
- An, S., Zhang, S., Guo, T., Lu, S., Zhang, W., & Cai, Z. (2025). Impacts of generative AI on student teachers' task performance and collaborative knowledge construction process in mind mapping-based collaborative environment. Computers & Education, 227, 105227.  
<https://doi.org/10.1016/J.COMPEDU.2024.105227>
- Arezes, P. M., & Swuste, P. (2012). Occupational Health and Safety post-graduation courses in Europe: A general overview. Safety Science, 50(3), 433–442.  
<https://doi.org/10.1016/J.SSCI.2011.10.003>
- Bagagiolo, G., Vigoroso, L., Caffaro, F., Cremasco, M. M., & Cavallo, E. (2019). Conveying Safety Messages on Agricultural Machinery: The Comprehension of Safety Pictorials in a Group of Migrant Farmworkers in Italy.  
International Journal of Environmental Research and Public Health 2019, Vol. 16, Page 4180, 16(21), 4180. <https://doi.org/10.3390/IJERPH16214180>
- Bai, L., Liu, X., & Su, J. (2023). ChatGPT: The cognitive effects on learning and memory. Brain-X, 1(3). <https://doi.org/10.1002/BRX2.30>
- Bartram, D. J. (1980). Comprehending spatial information: The relative efficiency of different methods of presenting information about bus routes. Journal of Applied Psychology, 65(1), 103–110. <https://doi.org/10.1037/0021-9010.65.1.103>
- Björnsdóttir, S. H., Jensson, P., de Boer, R. J., & Thorsteinsson, S. E. (2022). The Importance of Risk Management: What is Missing in ISO Standards? Risk Analysis, 42(4), 659–691.  
<https://doi.org/10.1111/RISA.13803>
- Bouri, M., Chraïbi, L., & Sefiani, N. (2020). A Serious Game for Evaluating the Competencies of Environmental Consultants. <https://doi.org/10.11159/icmie20.124>
- Brazier, A. (2017). Bowtie diagrams and human factors. [www.abrisk.co.uk](http://www.abrisk.co.uk)
- BSI. (n.d.). BSI Training - Fundamentals of ISO 45001. <https://bsi.learncentral.com/shop/Course.aspx?id=26578&name=Fundamentals+of+ISO+45001>

- CEN-CENELEC. (2010). CEN-CENELEC GUIDE 17 Guidance for writing standards taking into account micro, small and medium-sized enterprises (SMEs) needs. [www.cen.eu](http://www.cen.eu)
- Chakraborty, C., Pal, S., Bhattacharya, M., Dash, S., & Lee, S. S. (2023). Overview of Chatbots with special emphasis on artificial intelligence-enabled ChatGPT in medical science. *Frontiers in Artificial Intelligence*, *6*, 1237704. <https://doi.org/10.3389/FRAI.2023.1237704>
- Chan, A. H., Han, S. H., Ng, A. W., & Park, W. (2009). Hong Kong Chinese and Korean comprehension of American security safety symbols. *International Journal of Industrial Ergonomics*, *39*(5), 835–850. <https://doi.org/10.1016/J.ERGON.2009.02.009>
- Chelli, M., Descamps, J., Lavoué, V., Trojani, C., Azar, M., Deckert, M., Raynier, J. L., Clowez, G., Boileau, P., & Ruetsch-Chelli, C. (2024). Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis. *Journal of Medical Internet Research*, *26*, e53164. <https://doi.org/10.2196/53164>
- Chi, M. T., & Wylie, R. (2014). The ICAP Framework: Linking Cognitive Engagement to Active Learning Outcomes. *Educational Psychologist*, *49*(4), 219–243. <https://doi.org/10.1080/00461520.2014.965823>
- Clark, D., Lawton, R., Baxter, R., Sheard, L., & O'Hara, J. K. (2024). Do healthcare professionals work around safety standards, and should we be worried? A scoping review. *BMJ Quality and Safety*. <https://doi.org/10.1136/BMJQS-2024-017546>
- Coffey, L. (2023). Most students outrunning faculty in AI use. <https://www.insidehighered.com/news/tech-innovation/artificial-intelligence/2023/10/31/most-students-outrunning-faculty-ai-use>
- Contreras Aguilar, D., Medina, F., Oyanedel, M., Salamó, M., & Sánchez-Marrè, M. (2024). SAMANTHA: A chatbot to assist users in training tasks to prevent workplace hazards. *INTERACCIÓN 2024: proceedings of the XXIV International Conference on Human-Computer Interaction*. <https://doi.org/10.1145/3657242.3658587>
- Currie, G. M., Hawk, K. E., & Rohren, E. M. (2024). Generative Artificial Intelligence Biases, Limitations and Risks in Nuclear Medicine: An Argument for Appropriate Use Framework and Recommendations. *Seminars in Nuclear Medicine*. <https://doi.org/10.1053/J.SEMNUCLMED.2024.05.005>
- de Vries, H. J., & Egyedi, T. M. (2007). Education about Standardization: Recent Findings. *International Journal of IT Standards and Standardization Research (IJITSR)*, *5*(2), 1–16. <https://doi.org/10.4018/JITSR.2007070101>

- Dong, Z., Li, J., Men, X., Zhao, W. X., Wang, B., Tian, Z., Chen, W., & Wen, J.-R. (2024). Exploring Context Window of Large Language Models via Decomposed Positional Vectors. <https://arxiv.org/abs/2405.18009v2>
- Dretzke, B. J. (1993). Effects of pictorial mnemonic strategy usage on prose recall of young, middle-aged, and older adults. *Educational Gerontology*, *19*(6), 489–502. <https://doi.org/10.1080/0360127930190602>
- European Commission. (n.d.). Harmonised Standards. [https://single-market-economy.ec.europa.eu/single-market/european-standards/harmonised-standards\\_en](https://single-market-economy.ec.europa.eu/single-market/european-standards/harmonised-standards_en)
- European Commission. (2022). *An EU Strategy on Standardisation - Setting global standards in support of a resilient, green and digital (tech. rep.)*. <https://ec.europa.eu/docsroom/documents/48598>
- European Union Agency for Railways. (n.d.). Technical Specifications for Interoperability (TSIs). [https://www.era.europa.eu/domains/technical-specifications-interoperability\\_en](https://www.era.europa.eu/domains/technical-specifications-interoperability_en)
- Extance, A. (2023). ChatGPT has entered the classroom: how LLMs could transform education. *Nature*, *623*(7987), 474–477. <https://doi.org/10.1038/D41586-023-03507-3>
- Gibbons, S. (2019). User Need Statements - NN/g. <https://www.nngroup.com/articles/user-need-statements/>
- Google. (2024). Google Gemini updates: Custom Gems and improved image generation with Imagen 3. <https://blog.google/products/gemini/google-gemini-update-august-2024/>
- Grove, J. (2024). British Academics Despair as ChatGPT-Written Essays Swamp Grading Season. <https://www.insidehighered.com/news/global/2024/06/21/academics-dismayed-flood-chatgpt-written-student-essays>
- Hakim, V. G. A., Paiman, N. A., & Rahman, M. H. S. (2024). Genie-on-demand: A custom AI chatbot for enhancing learning performance, self-efficacy, and technology acceptance in occupational health and safety for engineering education. *Computer Applications in Engineering Education*, *32*(6), e22800. <https://doi.org/10.1002/CAE.22800>
- Harshvardhan, G., Gourisaria, M. K., Pandey, M., & Rautaray, S. S. (2020). A comprehensive survey and analysis of generative models in machine learning. *Comput. Sci. Rev.*, *38*. <https://doi.org/10.1016/J.COSREV.2020.100285>
- Hassanipour, S., Nayak, S., Bozorgi, A., Keivanlou, M.-H., Dave, T., Alotaibi, A., Joukar, F., Mellatdoust, P., Bakhshi, A., Kuriyakose, D., Polisetty, L. D., Chimpiri, M., & Amini-Salehi, E. (2024). The Ability of ChatGPT in Paraphrasing Texts and Reducing

- Plagiarism: A Descriptive Analysis. *JMIR medical education*, *10*(1), e53308.  
<https://doi.org/10.2196/53308>
- Hatia, A., Doldo, T., Parrini, S., Chisci, E., Cipriani, L., Montagna, L., Lagana, G., Guenza, G., Agosta, E., Vinjolli, F., Hoxha, M., D'Amelio, C., Favaretto, N., & Chisci, G. (2024). Accuracy and Completeness of ChatGPT-Generated Information on Interceptive Orthodontics: A Multicenter Collaborative Study. *Journal of Clinical Medicine*, *13*(3), 735.  
<https://doi.org/10.3390/JCM13030735>
- Hesser, W. (2014). *Memorandum on Standardization in higher Education in Europe* (tech. rep.). Helmut Schmidt Universitat. Hamburg, Germany. <http://www.wisestandard.org/>
- Holsanova, J. (2014). In the eye of the beholder: Visual communication from a recipient perspective. *Visual Communication*, 331–355.  
<https://doi.org/10.1515/9783110255492.331>
- Houts, P. S., Doak, C. C., Doak, L. G., & Loscalzo, M. J. (2006). The role of pictures in improving health communication: a review of research on attention, comprehension, recall, and adherence. *Patient education and counseling*, *61*(2), 173–190.  
<https://doi.org/10.1016/J.PEC.2005.05.004>
- Howell, M. D. (2024). Generative artificial intelligence, patient safety and healthcare quality: a review. *BMJ Quality & Safety*, *33*(11), 748–754.  
<https://doi.org/10.1136/BMJQS-2023-016690>
- International Organization for Standardization. (2014). *Teaching Standards Good practices for collaboration between National Standards Bodies and universities*.  
<https://www.iso.org/files/live/sites/isoorg/files/store/en/PUB100354.pdf>
- International Organization for Standardization. (2020). *Graphical symbols — Safety colours and safety signs — Registered safety signs — Amendment 7 (Standard No. 3641:2019)* (tech. rep.). <https://www.iso.org/standard/86774.html>
- International Organization for Standardization. (2021). *Road vehicles-Symbols for controls, indicators and tell-tales (ISO Standard No. 2575:2021)* (tech. rep.). [www.iso.org](http://www.iso.org)
- International Standard Organization. (n.d.). Sustainable Development Goals.  
<https://www.iso.org/sdg>
- International Standard Organization. (2001). *Graphical symbols — Test methods for judged comprehensibility and for comprehension (Standard No. 10441:2001)* (tech. rep.).
- ISO Standards. (n.d.). How to Write Standards.

Izadi, S., & Forouzanfar, M. (2024). Error Correction and Adaptation in Conversational AI: A Review of Techniques and Applications in Chatbots. *AI 2024*, Vol. 5, Pages 803-841, 5(2), 803–841. <https://doi.org/10.3390/AI5020041>

Johnson, J. (2014). Designing with the Mind in Mind: Simple Guide to Understanding User Interface Design Guidelines, Second Edition. *Designing with the Mind in Mind: Simple Guide to Understanding User Interface Design Guidelines*, Second Edition, 1–234. <https://doi.org/10.1016/C2012-0-07128-1>

Kadayat, B. B., & Eika, E. (2020). Impact of sentence length on the readability of web for screen reader users. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12188 LNCS, 261–271. [https://doi.org/10.1007/978-3-030-49282-3\\\_18/FIGURES/3](https://doi.org/10.1007/978-3-030-49282-3\_18/FIGURES/3)

Kozierok, R., Aberdeen, J., Clark, C., Garay, C., Goodman, B., Korves, T., Hirschman, L., McDermott, P. L., & Peterson, M. W. (2021). Assessing Open-Ended Human-Computer Collaboration Systems: Applying a Hallmarks Approach. *Frontiers in Artificial Intelligence*, 4, 670009. <https://doi.org/10.3389/FRAI.2021.670009/BIBTEX>

KTH. (n.d.). MF2144 Safety by Design or Mechatronics. <https://www.kth.se/student/kurser/kurs/MF2144/?l=en>

Küchemann, S., Avila, K. E., Dinc, Y., Hortmann, C., Revenga, N., Ruf, V., Stausberg, N., Steinert, S., Fischer, F., Fischer, M., Kasneci, E., Kasneci, G., Kuhr, T., Kutyniok, G., Malone, S., Sailer, M., Schmidt, A., Stadler, M., Weller, J., & Kuhn, J. (2025). On opportunities and challenges of large multimodal foundation models in education. *NPJ Science of Learning*, 10(1), 11. <https://doi.org/10.1038/S41539-025-00301-W>

Le Coze, J.-C., & Reiman, T. (2023). *Visualising Safety, an Exploration* (J.-C. Le Coze & T. Reiman, Eds.). Springer International Publishing. <https://doi.org/10.1007/978-3-031-33786-4>

Ling, C., Zhao, X., Deng, C., Zheng, C., Wang, J., Chowdhury, T., Pei, J., Yang, C., Zhao, L., Zhang, X., Zhao, T., Panalkar, A., Mehta, D., Pasquali, S., Cheng, W., Wang, H., Liu, Y., Chen, Z., Lu, J., . . . Gu, Q. (2024). Domain Specialization as the Key to Make Large Language Models Disruptive: A Comprehensive Survey. <https://doi.org/10.1145/nnnnnnnn.nnnnnnnn>

Liu, X., Li, H., & Zhu, X. (2023). A GPT-based method of Automated Compliance Checking through prompt engineering.

Mayer, R. E. (2014). The Cambridge Handbook of Multimedia Learning. *The Cambridge Handbook of Multimedia Learning*, Second Edition, 1–930. <https://doi.org/10.1017/CBO9781139547369>



- Microsoft. (2024). New and planned features for Microsoft Copilot Studio, 2024 release wave 1 | Microsoft Learn. <https://learn.microsoft.com/en-us/power-platform/release-plan/2024wave1/microsoft-copilot-studio/planned-features>
- Miro. (n.d.). AI Mind Map Generator | Ideate Entities Faster. <https://miro.com/ai/mind-map-ai/>
- Mohamed Hassan, H. A., Marengo, E., & Nutt, W. (2022). A BERT-Based Model for Question Answering on Construction Incident Reports. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 13286 LNCS, 215–223. [https://doi.org/10.1007/978-3-031-08473-7\\\_20](https://doi.org/10.1007/978-3-031-08473-7\_20)
- Mohammad Rajabalinejad. (2020, December). Safety by Design: Engineering Products and Systems. Brave New Books.
- Mohammad Rajabalinejad. (2021). NTA 8287 (en) Safety Cube Method for design, engineering and integration of systems and products (tech. rep.).
- Nederlands Normalisatie Instituut. (n.d.). About NEN. <https://www.nen.nl/en/about-nen>
- Negi, D. S. (2020). Constructivist approach of Vygotsky for innovative trends in learning and teaching. International Journal of Advanced Academic Studies, 2(1), 349–353. <https://doi.org/10.33545/27068919.2020.V2.I1F.662>
- NEN-ISO. (2025). NEN-ISO 24495-1:2025 nl. <https://www.nen.nl/nen-iso-24495-1-2025-nl-333677>
- Nielsen, J. (n.d.). Iterative User-Interface Design.
- Onatayo, D., Onososen, A., Oyediran, A. O., Oyediran, H., Arowoia, V., & Onatayo, E. (2024). Generative AI Applications in Architecture, Engineering, and Construction: Trends, Implications for Practice, Education & Imperatives for Upskilling—A Review. Architecture 2024, Vol. 4, Pages 877-902, 4(4), 877–902. <https://doi.org/10.3390/ARCHITECTURE4040046>
- OpenAI. (n.d.). Prompt engineering - OpenAI API. <https://platform.openai.com/docs/guides/prompt-engineering>
- OpenAI. (2023a). GPT-4 Technical Report. <https://cdn.openai.com/papers/gpt-4.pdf#:~:text=GPT,designed%20factuality%20evaluations>
- OpenAI. (2023b). Introducing GPTs | OpenAI. <https://openai.com/index/introducing-gpts/>
- Orban, G. A. (2001). Imaging image processing in the human brain. Current opinion in neurology, 14(1), 47–54. <https://doi.org/10.1097/00019052-200102000-00008>
- Oviedo-Trespalacios, O., Peden, A. E., Cole-Hunter, T., Costantini, A., Haghani, M., Rod., J., Kelly, S., Torkamaan, H., Tariq, A., Newton, J. D. A., Gallagher, T., Steinert, S.,

- Filtiness, A., & Reniers, G. (2023). The Risks of Using ChatGPT to Obtain Common Safety-Related Information and Advice. SSRN Electronic Journal.  
<https://doi.org/10.2139/SSRN.4346827>
- Paques, J. J., Gauthier, F., & Perez, A. (2007). Analysis and classification of the tools for assessing the risks associated with industrial machines.  
International journal of occupational safety and ergonomics : JOSE, 13(2), 173–187.  
<https://doi.org/10.1080/10803548.2007.11076720>
- Patil, R. B., Waghmode, L., Chikali, P. B., & Mulla, T. S. (2013, February). An Overview of Fault Tree Analysis (FTA) Method for Reliability Analysis & Life Cycle Cost (LCC) Management.  
[https://www.researchgate.net/publication/277971116\\_An\\_Overview\\_of\\_Fault\\_Tree\\_Analysis\\_FTA\\_Method\\_for\\_Reliability\\_Analysis\\_Life\\_Cycle\\_Cost\\_LCC\\_Management](https://www.researchgate.net/publication/277971116_An_Overview_of_Fault_Tree_Analysis_FTA_Method_for_Reliability_Analysis_Life_Cycle_Cost_LCC_Management)
- Pejtersen, J. H., & Holt, H. (2022). Literacy and risk of occupational injury.  
International Archives of Occupational and Environmental Health, 95(10), 1971–1978.  
<https://doi.org/10.1007/S00420-022-01898-W/METRICS>
- Pieters, R., & Wedel, M. (2004). Attention Capture and Transfer in Advertising: Brand, Pictorial, and Text-Size Effects. Journal of Marketing, 68(2), 36–50.  
<https://doi.org/10.1509/JMKG.68.2.36.27794>
- Podrecca, M., Molinaro, M., Sartor, M., & Orzes, G. (2024). The impact of ISO 45001 on firms' performance: An empirical analysis.  
Corporate Social Responsibility and Environmental Management, 31(5), 4581–4595.  
<https://doi.org/10.1002/CSR.2782>
- Politecnico. (n.d.). Industrial Safety and Risk Engineering.  
<https://www.polimi.it/en/education/laurea-magistrale-programmes/programme-detail/industrial-safety-and-risk-engineering>
- Procter Machine Safety. (2018). Review of PD CEN ISO/TR 22100-1:2017.  
<https://www.machinesafety.co.uk/news/pd-cen-iso-tr-22100-12017/>
- Pryor, P. (2019). Developing the core body of knowledge for the generalist OHS professional.  
Safety Science, 115, 19–27. <https://doi.org/10.1016/J.SSCI.2019.01.013>
- Rajabalinejad, M. (2020, October). SAFETY CUBE METHOD: A Visual Guide. Brave New Books.  
<https://research.utwente.nl/en/publications/safety-cube-method-a-visual-guide>
- Ramel, D. (2024). OpenAI Reveals Internal Rules for ChatGPT: 'Don't Try to Change Anyone's Mind' – Virtualization Review.  
<https://virtualizationreview.com/Articles/2024/05/10/chatgpt-rules-spec.aspx>

- Rodeghiero Neto, I., & Amaral, F. G. (2024). Teaching occupational health and safety in engineering using active learning: A systematic review. *Safety Science*, 171, 106391. <https://doi.org/10.1016/J.SSCI.2023.106391>
- Rodríguez-Martín, M., Rodríguez-González, P., & Domingo, R. (2023). Asking about Risk in a Company: A New Approach to Learning ISO 45001 in Engineering Programs. *Standards* 2023, Vol. 3, Pages 21-30, 3(1), 21–30. <https://doi.org/10.3390/STANDARDS3010003>
- Ross, K. (n.d.). Revised Safety Label Standards Create Opportunity and Risk.
- Ryan, M. (2020). In AI We Trust: Ethics, Artificial Intelligence, and Reliability. *Science and Engineering Ethics*, 26(5), 2749–2767. <https://doi.org/10.1007/S11948-020-00228-Y/TABLES/1>
- Saka, A. B., & Chan, D. W. (2023). BIM divide: an international comparative analysis of perceived barriers to implementation of BIM in the construction industry. *Journal of Engineering, Design and Technology*, 21(5), 1604–1632. <https://doi.org/10.1108/JEDT-07-2021-0348/FULL/XML>
- Sathe, T. S., Roshal, J., Naaseh, A., L'Huillier, J. C., Navarro, S. M., & Silvestri, C. (2024). How I GPT It: Development of Custom Artificial Intelligence (AI) Chatbots for Surgical Education. *Journal of surgical education*, 81(6), 772–775. <https://doi.org/10.1016/J.JSURG.2024.03.004>
- Scharf, R. E. (2017). Pictures tell their own story. *Hämostaseologie*, 37(3), 181–183. <https://doi.org/10.1055/S-0037-1619833>
- Shepard, R. N. (1967). Recognition memory for words, sentences, and pictures. *Journal of Verbal Learning and Verbal Behavior*, 6(1), 156–163. [https://doi.org/10.1016/S0022-5371\(67\)80067-7](https://doi.org/10.1016/S0022-5371(67)80067-7)
- Simmonds, P., Brown, N., & Rentel, M. (2017). Evaluation of Directive 2006/42/EC on Machinery - Final Report - September 2017. [www.technopolis-group.com](http://www.technopolis-group.com)
- Smetana, M., Salles de Salles, L., Sukharev, I., & Khazanovich, L. (2024). Highway Construction Safety Analysis Using Large Language Models. *Applied Sciences* 2024, Vol. 14, Page 1352, 14(4), 1352. <https://doi.org/10.3390/APP14041352>
- Smith, T. (2013). The Evolution of Safety Signs – Occupational Health & Safety. <https://ohsonline.com/Articles/2013/02/01/The-Evolution-of-Safety-Signs.aspx>
- Societal stakeholders and standards - 1.1 What are standards? (n.d.). <https://www.standards4all.eu/index2.php#/lessons/AmzJJwDL23bgz-AHDetau8pMEc5JIYu4>

- Someren, V., Barnard, M. W. ; & Sandberg, Y. F. ; (n.d.). The think aloud method: a practical approach to modelling cognitive processes. <https://dare.uva.nl>
- Spychalski, B. (2022). Threats for ISO 9001. *EUROPEAN RESEARCH STUDIES JOURNAL*, 25(Special Issue 3), 107–118. <https://doi.org/10.35808/ERSJ/3023>
- Taky-eddine, O., & Madaoui, R. (2024). International Journal of English Language Studies Cognitive Overload in the Hypertext Reading Environment. <https://doi.org/10.32996/ijels>
- Tong, F. (2003). Primary visual cortex and visual awareness. *Nature reviews. Neuroscience*, 4(3), 219–229. <https://doi.org/10.1038/NRN1055>
- Uddin, S. M., Albert, A., Ovid, A., & Alsharif, A. (2023). Leveraging ChatGPT to Aid Construction Hazard Recognition and Support Safety Education and Training. *Sustainability*, 15(9). <https://doi.org/10.3390/SU15097121>
- University of Twente. (n.d.). Safety By Design. <https://www.utwente.nl/en/et/safety-by-design/>
- Wang, J., Hu, X., Hou, W., Chen, H., Zheng, R., Wang, Y., Yang, L., Ye, W., Huang, H., Geng, X., Jiao, B., Zhang, Y., & Xie, X. (n.d.). On the Robustness of ChatGPT: An Adversarial and Out-of-distribution Perspective. <https://github.com/microsoft/robustlearn>
- Wang, X., & Zhu, D. (2024). Validating LLM-Generated Programs with Metamorphic Prompt Testing. <https://arxiv.org/abs/2406.06864v1>
- Ware, C. (2004). Information Visualization: Perception for Design: Second Edition. [https://www.researchgate.net/publication/224285723\\_Information\\_Visualization\\_Perception\\_for\\_Design\\_Second\\_Edition](https://www.researchgate.net/publication/224285723_Information_Visualization_Perception_for_Design_Second_Edition)
- Wogalter, M. S., Silver, N. C., Leonard, S. D., & Zaikina, H. (2006). Handbook of Warnings. *Handbook of Warnings*. <https://doi.org/10.1201/9781482289688>
- Yeşilgöz, P., & Arga, K. Y. (2025). A Health-Sector-Specific Occupational Health and Safety Management System Model. *Healthcare* 2025, Vol. 13, Page 271, 13(3), 271. <https://doi.org/10.3390/HEALTHCARE13030271>
- Zhuhadar, L. (2023). A comparative view of AI, machine learning, deep learning, and... | Download Scientific Diagram. [https://www.researchgate.net/figure/A-comparative-view-of-AI-machine-learning-deep-learning-and-generative-AI-source\\_fig1\\_373797588](https://www.researchgate.net/figure/A-comparative-view-of-AI-machine-learning-deep-learning-and-generative-AI-source_fig1_373797588)

\*Appendices

# A Detailed Results of Test 1

Table 3: The evaluation result for AI chatbot performance in providing NTA 8287 related information with direct prompts

	External interface	Functional risk	Operational aspects	Method overview	Step overview	Execution guidance
Accuracy						
Does the information use the extract words as it appears in the knowledge base?	[5/5]	[5/5]	[5/5]	[0/5]. Wrong step names	[3/5] Questions changed into statements. But keep the same meanings	[3/5] Questions changed into statements. But keep the same meanings
Completeness						
Is the definition explained with an example align with the context of users project, and citation?	[5/5]	[5/5]	[5/5]			
Does the overview include all topics, including purposes, core principles, a general structure, citation, and visual links?				[1/5] Lack citation		
Does each step include all topics, including expected outcomes, goal, what to ask, question to answer,...					[5/5]	
Does the guidance include all topics, including expected outcomes, suggestive methods						[5/5]

Table 4: The evaluation result for AI chatbot performance in providing NTA 8287 related information with paraphrased prompts

	External interface	Functional risk	Operational aspects	Method overview	Step overview	Execution guidance
Accuracy						
Does the information use the extract words as it appears in the knowledge base?	[5/5]	[5/5]	[5/5]	[0/5]. Wrong step names	[2/5] Questions changed into statements. But keep the same meanings	[2/5] Questions changed into statements. But keep the same meanings
Completeness						
Is the definition explained with an example align with the context of users project, and citation?	[5/5]	[5/5]	[5/5]			
Does the overview include all topics, including purposes, core principles, a general structure, citation, and visual links?				[2/5] Lack citation		
Does each step include all topics, including expected outcomes, goal, what to ask, question to answer,...					[3/5] Lack citation	
Does the guidance include all topics, including expected outcomes, suggestive methods						[2/5] Lack citation and visual support

Table 5: The evaluation result for AI chatbot performance in providing NTA 8287 related information with indirect prompts

	External interface	Functional risk	Operational aspects	Method overview	Step overview	Execution guidance
Accuracy						
Does the information use the extract words as it appears in the knowledge base?	[5/5]	[5/5]	[5/5]	[0/5]. Wrong step names	[3/5] Questions changed into statements. But keep the same meanings	[3/5] Questions changed into statements. But keep the same meanings
Completeness						
Is the definition explained with an example align with the context of users project, and citation?	[5/5]	[5/5]	[5/5]			
Does the overview include all topics, including purposes, core principles, a general structure, citation, and visual links?				[2/5] Lack citation		
Does each step include all topics, including expected outcomes, goal, what to ask, question to answer,...					[3/5] Lack citation	
Does the guidance include all topics, including expected outcomes, suggestive methods						[3/5] Lack citation



Table 6: The evaluation result for AI chatbot performance in providing NTA 8287 related information with prompts with typos

	External interface	Functional risk	Operational aspects	Method overview	Step overview	Execution guidance
Accuracy						
Does the information use the extract words as it appears in the knowledge base?	[5/5]	[5/5]	[5/5]	[0/5]. Wrong step names	[2/5] Questions changed into statements. But keep the same meanings	[2/5] Questions changed into statements. But keep the same meanings
Completeness						
Is the definition explained with an example align with the context of users project, and citation?	[5/5]	[5/5]	[5/5]			
Does the overview include all topics, including purposes, core principles, a general structure, citation, and visual links?				[2/5] Lack citation		
Does each step include all topics, including expected outcomes, goal, what to ask, question to answer,...					[3/5] Lack citation	
Does the guidance include all topics, including expected outcomes, suggestive methods						[2/5] Lack citation

# B Detailed Results of Test 2

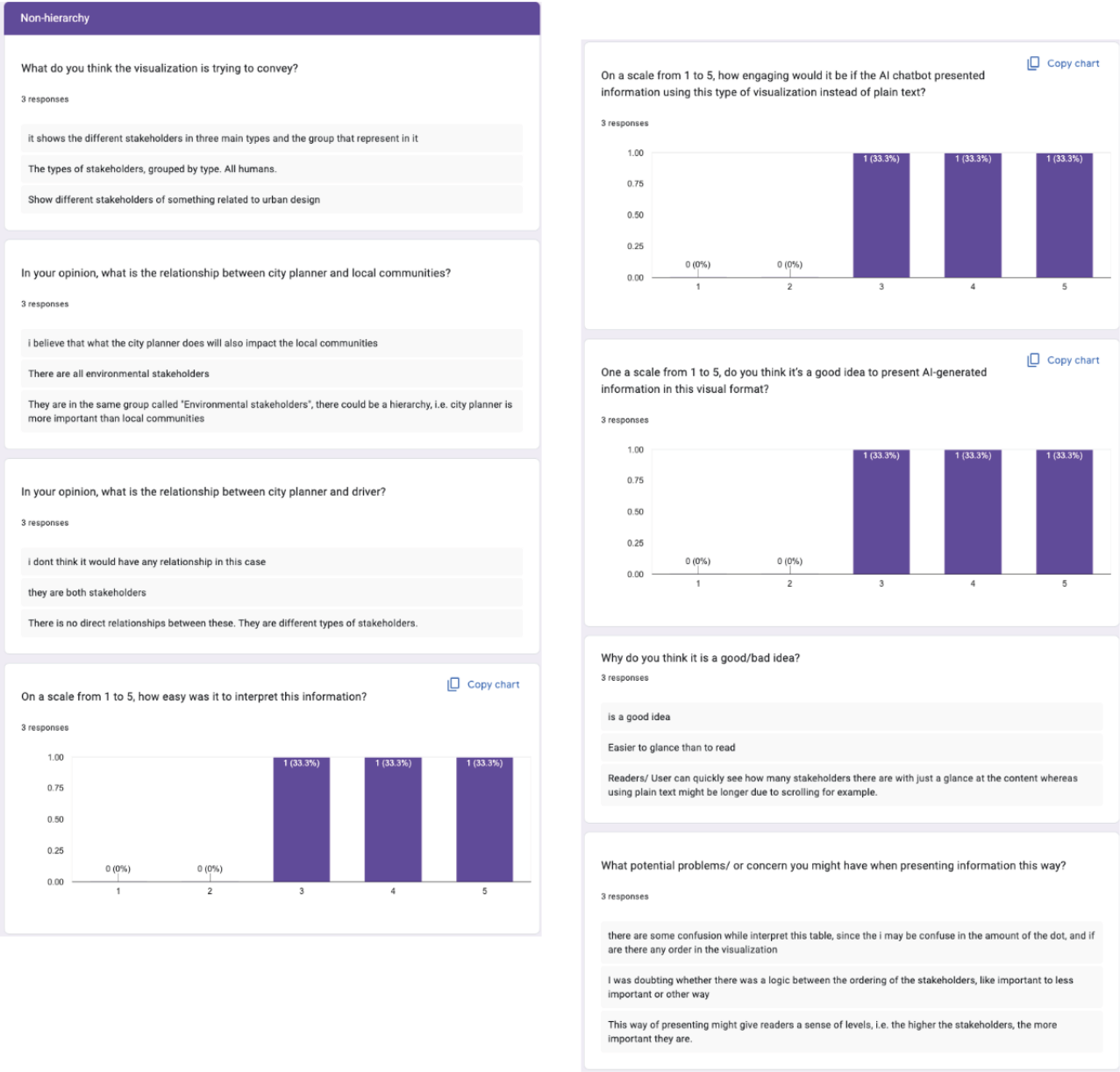


Figure 55: The comprehension results of non-hierarchy visualization

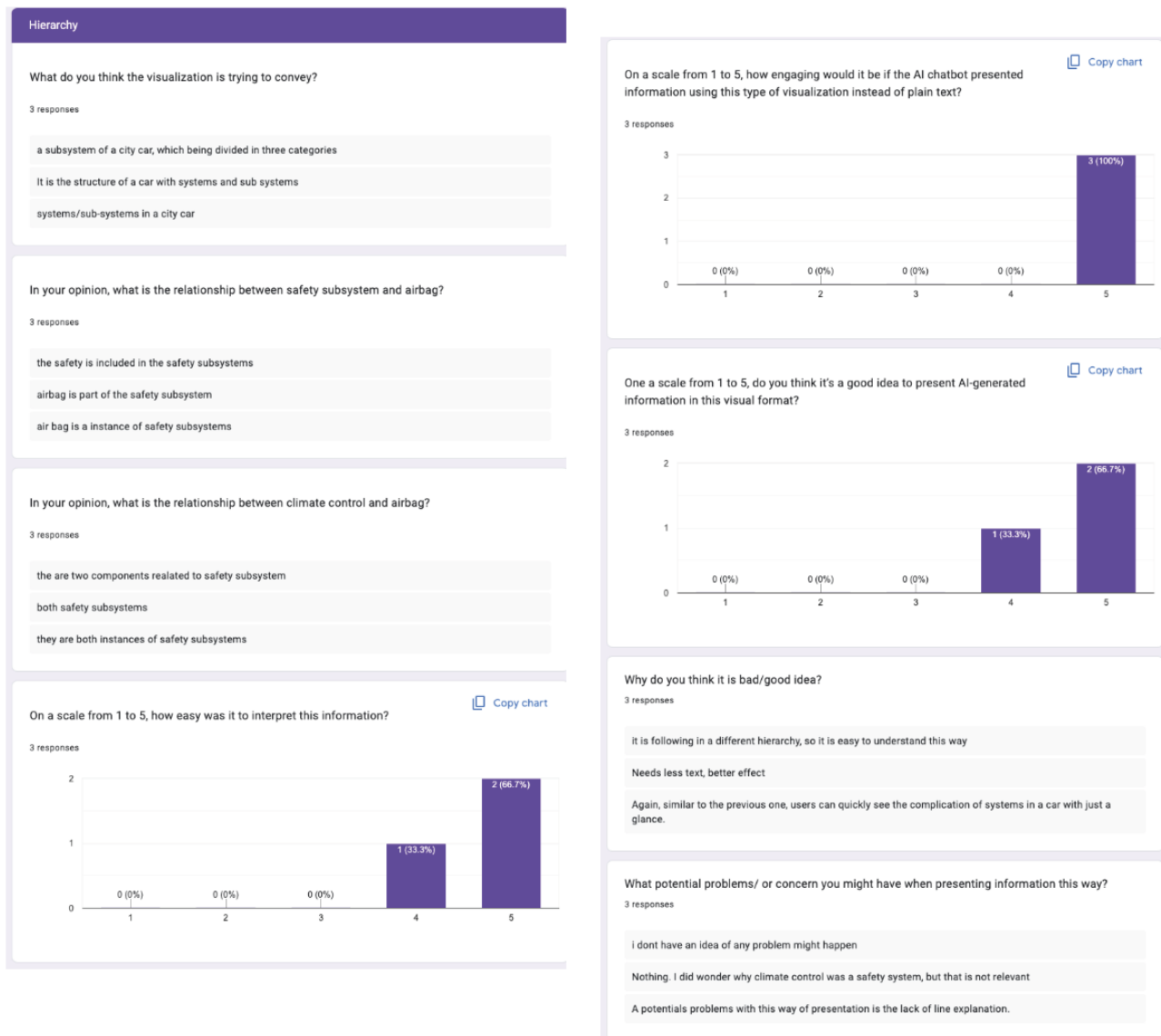


Figure 56: The comprehension results of hierarchy visualization

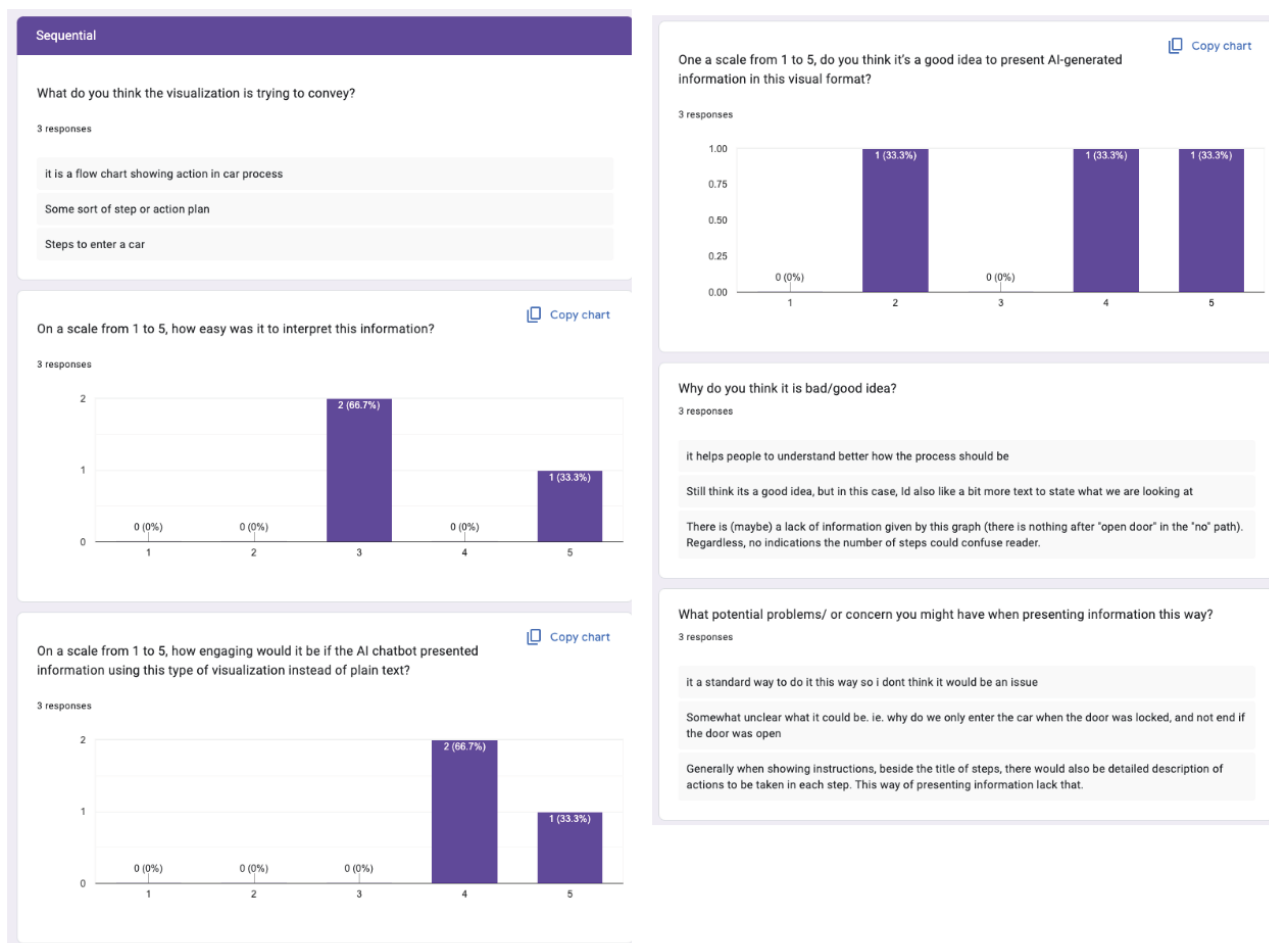


Figure 57: The comprehension results of sequential visualization

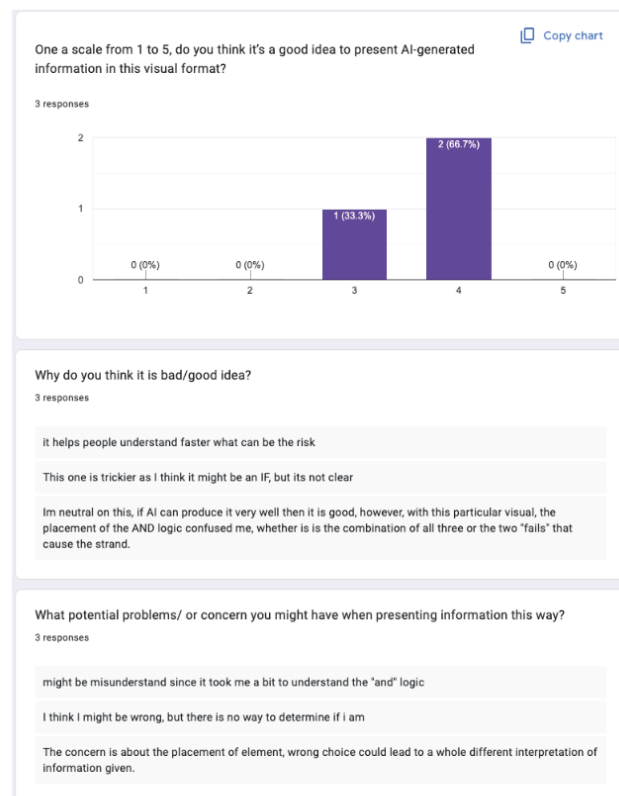
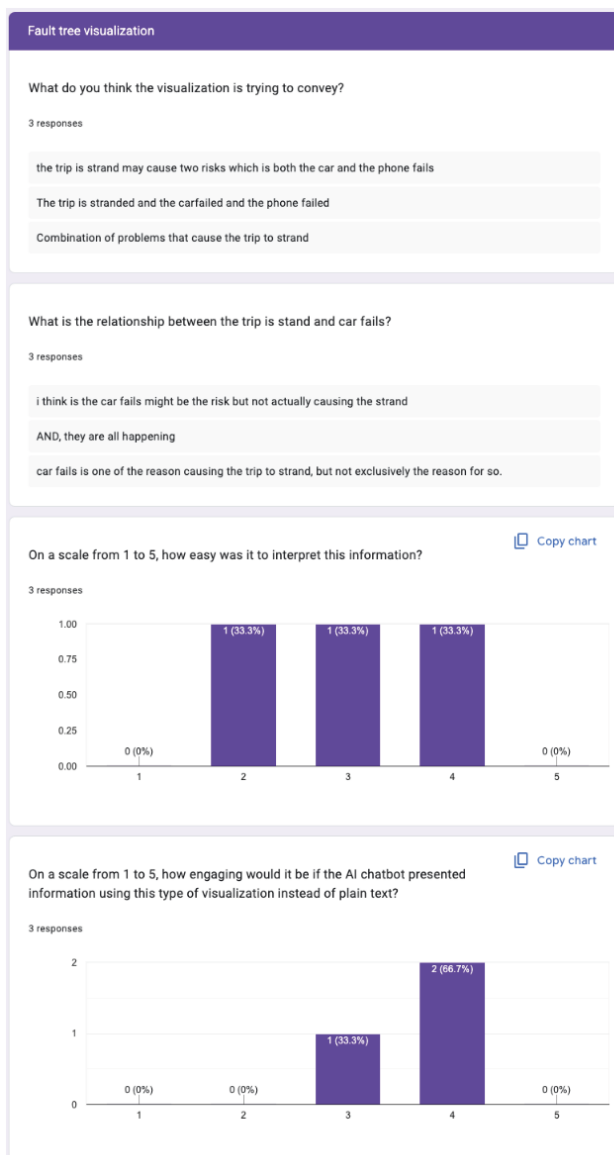


Figure 58: The comprehension results of fault tree visualization

# C Results Analysis of Test 3

The results of the experimental group are presented in tables [7](#) and [8](#) . The results of the control group are presented in tables [9](#) and [10](#). Comparative theme analysis of interview data from control and experimental groups is presented in table [11](#)

Table 7: The results of the experimental group in finding critical information and key terms comprehension

	Participant 1	Participant 2	Participant 3
Finding critical information			
What is the NEN NTA 8287:2021?	✓ Found a short description and what the method is about and two purposes	✓ Found a short description and what the method is about and two purposes	✓ Found a short description and what the method is about and two purposes
What are the key principles of the NEN NTA 8287:2021?	✓ Found 6 main aspects the method mention	✓ Found 6 main aspects the method mention. But misunderstood and then the Chatbot gave wrong answer	✓ Found 6 main aspects the method mention
What are the main steps involved in the NEN NTA 8287:2021?	✓ Found all 15 steps	✓ Found all 15 steps	✓ Found all 15 steps
What is the description of step 1?	✓ Found both description and execution recommendation	✓ Found both description and execution recommendation	✓ Found both description and execution recommendation
What is the description of step 2?	✓ Found both description and execution recommendation	✓ Found both description and execution recommendation	✓ Found both description and execution recommendation
What is the description of step 4?	✓ Found both description and execution recommendation	✓ Found both description and execution recommendation	✓ Found both description and execution recommendation
What is the description of step 6?	✓ Found both description and execution recommendation	✓ Found both description and execution recommendation	✓ Found both description and execution recommendation
Total	7	7	7
Comprehension of key terms			
External interface	✓ confirm understood after reading examples and simplified definitions	✓ confirm understood after reading examples. But needed to re-read two times.	✓ confirm understood after reading examples and asking clarified questions
functional risk assessment	✓ confirm understood after reading examples	✓ confirm understood after reading examples	✓ confirm understood after reading examples and extra explanation generated by AI
operational aspect	✓ confirm understood after reading examples and extra explanation generated by AI	✓ confirm understood after reading examples	✓ confirm understood after reading examples and extra explanation generated by AI
Total	3	3	3

Table 8: The results of the experimental group in understanding the NTA method and applying it in redesign assignment

	Participant 1	Participant 2	Participant 3
Understanding and coverage of steps described in the method:			
Step 1	Describe correctly (3/3): ✓ Stakeholder identification ✓ Formulate need statements ✓ Stakeholder prioritization	Describe correctly (2/3): ✓ Stakeholder identification ✓ Formulate need statements ✗ Stakeholder prioritization	Describe correctly (3/3): ✓ Stakeholder identification ✓ Formulate need statements ✗ Stakeholder prioritization
Step 2	Describe correctly (3/4): ✓ Operational concept ✓ Technical concept ✗ Functional concept ✓ Environmental factors	Describe correctly (3/4): ✓ Operational concept ✓ Technical concept ✗ Functional concept ✓ Environmental factors	Describe correctly (2/4): ✗ Operational concept ✓ Technical concept ✗ Functional concept ✓ Environmental factors
Step 4	Describe correctly (3/4): ✓ Identify Risks ✓ Analyze Past Accidents ✓ Risk Assessment ✗ Risk Acceptance Criteria	Describe correctly (3/4): ✓ Identify Risks ✓ Analyze Past Accidents ✗ Risk Assessment ✓ Risk Acceptance Criteria	Describe correctly (3/4): ✓ Identify Risks ✓ Analyze Past Accidents ✗ Risk Assessment ✓ Risk Acceptance Criteria
Step 6	Describe correctly (4/5): ✓ Identify Safety-Critical Functions ✓ Allocate Functions to Subsystems & Humans ✗ Categorize Safety Risks ✓ Integrate with the Environment ✓ Document & Visualize Architecture	Describe correctly (4/5): Identify Safety-Critical Functions ✓ Allocate Functions to Subsystems & Humans ✗ Categorize Safety Risks ✓ Integrate with the Environment ✓ Document & Visualize Architecture	Describe correctly (4/5): Identify Safety-Critical Functions ✓ Allocate Functions to Subsystems & Humans ✗ Categorize Safety Risks ✓ Integrate with the Environment ✓ Document & Visualize Architecture
Total	13	12	11
Applying the method			
Step 1	Finished (3/3): ✓ Stakeholder identification ✓ Formulate need statements ✓ Stakeholder prioritization	Finished (3/3): ✓ Stakeholder identification ✓ Formulate need statements ✓ Stakeholder prioritization	Finished (3/3): ✓ Stakeholder identification ✓ Formulate need statements ✓ Stakeholder prioritization
Step 2	Finished (3/4): ✓ Operational concept ✓ Technical concept ✗ Functional concept ✓ Environmental factors	Finished (2/4): ✗ Operational concept ✓ Technical concept ✗ Functional concept ✓ Environmental factors	Finished (2/4): ✗ Operational concept ✓ Technical concept ✗ Functional concept ✓ Environmental factors
Step 4	Finished (3/4): ✓ Identify Risks ✗ Analyze Past Accidents ✓ Risk Assessment ✓ Risk Acceptance Criteria	Finished (3/4): ✓ Identify Risks ✗ Analyze Past Accidents ✓ Risk Assessment ✓ Risk Acceptance Criteria	Finished (3/4): ✓ Identify Risks ✗ Analyze Past Accidents ✓ Risk Assessment ✓ Risk Acceptance Criteria
Step 6	Finished (3/5): ✓ Identify Safety-Critical Functions ✓ Allocate Functions to Subsystems & Humans ✗ Categorize Safety Risks ✓ Integrate with the Environment ✗ Document & Visualize Architecture	Finished (3/5): ✓ Identify Safety-Critical Functions ✓ Allocate Functions to Subsystems & Humans ✗ Categorize Safety Risks ✓ Integrate with the Environment ✗ Document & Visualize Architecture	Finished (3/5): ✓ Identify Safety-Critical Functions ✓ Allocate Functions to Subsystems & Humans ✗ Categorize Safety Risks ✓ Integrate with the Environment ✗ Document & Visualize Architecture
total	12	11	11



Table 9: The results of the control group in finding critical information and key terms comprehension

	Participant 1	Participant 2	Participant 3
Finding critical information			
What is the NEN NTA 8287:2021?	✓ Found a short description and what the method is about and two purposes	✓ Found a short description and what the method is about and two purposes	✓ Found a short description and what the method is about and two purposes
What are the key principles of the NEN NTA 8287:2021?	✗ Do not have enough time to perform this task	✗ Do not have enough time to perform this task	✗ Do not have enough time to perform this task
What are the main steps involved in the NEN NTA 8287:2021?	✗ Do not have enough time to perform this task	✗ Do not have enough time to perform this task	✗ Do not have enough time to perform this task
What is the description of step 1?	✓ only found information in Chapter 11 and not chapter 7	✓ only found information in Chapter 11 and not chapter 7	✓ only found information in Chapter 11 and not chapter 7
What is the description of step 2?	✓ only found information in Chapter 11 and not chapter 7	✓ only found information in Chapter 11 and not chapter 7	✓ only found information in Chapter 11 and not chapter 7
What is the description of step 4?	✓ only found information in Chapter 11 and not chapter 7	✓ only found information in Chapter 11 and not chapter 7	✓ only found information in Chapter 11 and not chapter 7
What is the description of step 6?	✓ only found information in Chapter 11 and not chapter 7	✓ only found information in Chapter 11 and not chapter 7	✓ only found information in Chapter 11 and not chapter 7
Total	5	5	5
Comprehension of key terms			
External interface	✗ do not understand because the definition is too complex	✗ do not understand because the definition is written complicatedly	✗ do not understand because the definition is too complex
functional risk assessment	✗ they still do not understand what is risk assessment	✗ they still do not understand what is risk assessment	✗ they still do not understand what is risk assessment
operational aspect	✓ confirm understood because they understand the meaning of the word operational	✓ confirm understood because they understand the meaning of the word operational	✓ confirm understood because they understand the meaning of the word operational
Total	1	1	1

Table 10: The results of the control group in understanding the NTA method and applying it in redesign assignment

	Participant 1	Participant 2	Participant 3
Understanding and coverage of steps described in the method:			
Step 1	Describe correctly (2/3): ✓ Stakeholder identification ✓ Formulate need statements ✗ Stakeholder prioritization	Describe correctly (2/3): ✓ Stakeholder identification ✓ Formulate need statements ✗ Stakeholder prioritization	Describe correctly (2/3): ✓ Stakeholder identification ✓ Formulate need statements ✗ Stakeholder prioritization
Step 2	Describe correctly (2/4): ✗ Operational concept ✓ Technical concept ✗ Functional concept ✓ Environmental factors	Describe correctly (2/4): ✗ Operational concept ✓ Technical concept ✗ Functional concept ✓ Environmental factors	Describe correctly (2/4): ✗ Operational concept ✓ Technical concept ✗ Functional concept ✓ Environmental factors
Step 4	Describe correctly (2/4): ✓ Identify Risks ✓ Analyze Past Accidents ✗ Risk Assessment ✗ Risk Acceptance Criteria	Describe correctly (2/4): ✓ Identify Risks ✓ Analyze Past Accidents ✗ Risk Assessment ✗ Risk Acceptance Criteria	Describe correctly (2/4): ✓ Identify Risks ✓ Analyze Past Accidents ✗ Risk Assessment ✗ Risk Acceptance Criteria
Step 6	Describe correctly (1/5): ✗ Identify Safety-Critical Functions ✓ Allocate Functions to Subsystems & Humans ✗ Categorize Safety Risks ✗ Integrate with the Environment ✗ Document & Visualize Architecture	Describe correctly (0/5): ✗ Identify Safety-Critical Functions ✓ Allocate Functions to Subsystems & Humans ✗ Categorize Safety Risks ✓ Integrate with the Environment ✓ Document & Visualize Architecture	Describe correctly (0/5): ✗ Identify Safety-Critical Functions ✗ Allocate Functions to Subsystems & Humans ✗ Categorize Safety Risks ✗ Integrate with the Environment ✗ Document & Visualize Architecture
Total	7	7	6
Applying the method			
Step 1	Finished (3/3): ✓ Stakeholder identification ✓ Formulate need statements ✓ Stakeholder prioritization	Finished (2/3): ✓ Stakeholder identification ✓ Formulate need statements ✗ Stakeholder prioritization	Finished (1/3): ✓ Stakeholder identification ✗ Formulate need statements ✗ Stakeholder prioritization
Step 2	Finished (1/4): ✓ Operational concept ✓ Technical concept ✗ Functional concept ✗ Environmental factors	Finished (1/4): ✗ Operational concept ✓ Technical concept ✗ Functional concept ✗ Environmental factors	Finished (3/4): ✗ Operational concept ✓ Technical concept ✓ Functional concept ✓ Environmental factors
Step 4	Finished (1/4): ✓ Identify Risks ✗ Analyze Past Accidents ✗ Risk Assessment ✗ Risk Acceptance Criteria	Finished (1/4): ✓ Identify Risks ✗ Analyze Past Accidents ✗ Risk Assessment ✗ Risk Acceptance Criteria	Finished (1/4): ✓ Identify Risks ✗ Analyze Past Accidents ✓ Risk Assessment ✓ Risk Acceptance Criteria
Step 6	Finished (0/5): ✗ Identify Safety-Critical Functions ✗ Allocate Functions to Subsystems & Humans ✗ Categorize Safety Risks ✗ Integrate with the Environment ✗ Document & Visualize Architecture	Finished (0/5): ✗ Identify Safety-Critical Functions ✗ Allocate Functions to Subsystems & Humans ✗ Categorize Safety Risks ✗ Integrate with the Environment ✗ Document & Visualize Architecture	Finished (0/5): ✗ Identify Safety-Critical Functions ✗ Allocate Functions to Subsystems & Humans ✗ Categorize Safety Risks ✗ Integrate with the Environment ✗ Document & Visualize Architecture
total	5	4	5

Table 11: Theme analysis of interview data comparing responses from the control and experimental groups.

	Control	Experiment
Support in risk analysis tasks	- <b>Internet Search Difficulty:</b> Users noted challenges in searching for information online, which added to their overall difficulties.	- <b>Great impression on Stakeholder Identification:</b> Users were impressed with stakeholder identification, highlighting aspects such as user burden or tourists, which they hadn't initially considered.
	<b>Challenges with Online Searching:</b> Google searches frequently left users feeling stuck and frustrated due to difficulties in finding relevant information. Even when they found correct information, they immediately copied and pasted it, showing excitement but avoiding deeper reading.	- <b>Desire for AI Automation:</b> Some users prefer that the AI chatbot complete entire tasks for them, leaving their role limited to verifying or reviewing results.- <b>Desire for AI Automation:</b> Some users prefer that the AI chatbot complete entire tasks for them, leaving their role limited to verifying or reviewing results.
		- <b>Structured and Organized Responses:</b> Users liked that the AI's answers consistently followed the method's structure, making information easier to organize, particularly regarding risks (functional, technical, operational) and stakeholders (five distinct types).
Comprehension supports	- <b>Internet Search Difficulty:</b> Users noted challenges in searching for information online, which added to their overall difficulties.	- <b>Frequent Follow-Up Questions:</b> Users appreciated the frequent follow-up questions from the AI, as these helped them gain a deeper understanding of the problem.
	<b>Challenges with Online Searching:</b> Google searches frequently left users feeling stuck and frustrated due to difficulties in finding relevant information. Even when they found correct information, they immediately copied and pasted it, showing excitement but avoiding deeper reading.	
	<b>Struggles Connecting Specific Systems (Traffic Sign Example):</b> Difficulty understanding the "traffic sign system" significantly complicated completing the stakeholder analysis.	- <b>Useful Examples and Ideas:</b> Users found AI's examples highly valuable, especially when explaining unfamiliar concepts like traffic signs and connecting theory to real-life cases. They were also pleasantly surprised by the AI's ability to introduce ideas they hadn't previously thought of.
	<b>Misunderstanding Specific Steps (Operational Concept):</b> Users specifically misunderstood Step 2, particularly the operational concept, largely because Google did not provide relevant information.	- <b>Tables as Concise Summaries:</b> Users found tables particularly helpful for summarizing information succinctly.
		- <b>(-) Hallucination: Incorrect Presentation of Principles:</b> The AI chatbot incorrectly included regulations when presenting the principles, creating confusion about the scope and content.
		- <b>Need for Visual Support:</b> Users emphasize a strong need for visual aids to balance the extensive amount of text, improving clarity and comprehension.

Feedback support		- Inaccurate and Shallow Evaluation: The evaluation process did not accurately reflect the quality or completeness of user work. Users who missed or made errors in steps (particularly Step 6) still received overly positive evaluations, and the AI chatbot failed to identify critical gaps in user
Engagement	<b>Difficulty Leading to Desire to Give Up:</b> Several times, users expressed a desire to give up because the material was too challenging. They found information very abstract, especially when unfamiliar with safety concepts, and struggled to grasp new terms like the "system by design philosophy" due to limited time.	- <b>Perceived Time Efficiency:</b> Users reported that time seemed to pass more quickly when using
		- <b>Positive Engagement Through Emojis:</b> Emojis were well-received as they enhanced engagement and made complex instructions feel friendlier and more approachable.
		One user remarked, "With AI, I can do 1 hour 30 minutes in no time," emphasizing its engaging nature and consistent utility.
Accessibility and Usability		- <b>(-) Overuse of Complex and Technical Language:</b> The AI chatbot frequently employed complicated phrasing and excessive technical terms, making the information difficult and uninviting for users.
		- <b>(-) Information Overload and Lengthy Texts:</b> Users felt overwhelmed due to excessively long responses from the AI chatbot, causing them to skim rather than read carefully. This resulted in superficial engagement rather than deep understanding.
		- <b>(-) Missing Connection Between Steps :</b> The AI chatbot failed to clearly explain why certain steps were necessary or how they fit into the overall process. Users expressed a strong need for an overview or clear narrative connecting the steps.
		- <b>Messy Interactions with Multiple Steps:</b> Users encountered difficulties when interacting with the AI chatbot about multiple steps simultaneously, leading to messy or confusing exchanges.
		- <b>Messy, lack focus- Multiple Solutions Instead of Focused Answers:</b> When users requested solutions, the AI chatbot tended to offer numerous options rather than providing a clear, singular

		<p>- <b>Messy, lack focus- Multiple Solutions Instead of Focused Answers:</b> When users requested solutions, the AI chatbot tended to offer numerous options rather than providing a clear, singular</p>
		<p>- <b>Dislike of Extensive Bullet Points:</b> Users found long lists of bullet points unengaging and difficult to read, indicating a preference for more concise and visually appealing presentations.</p>
Ineffective/unuse features		<p>- <b>Low usage of Documentation Chatbot:</b> Users initially found the Documentation Chatbot impressive, especially its ability to automatically save information into Google Docs. However, observations showed that participants rarely used it in practice, instead manually copying and pasting information into their documents.</p>
		the definition of difficult words
chatbot limitations		<p>- <b>Reduced Effectiveness with Multiple Chatbots:</b> When users interacted with both the Evaluation Chatbot and the Design Guiding Chatbot within the same conversation, the quality and accuracy of chatbot responses decreased. The Evaluation Chatbot particularly struggled to assess users' work properly, possibly due to confusion or difficulty in managing multiple functions simultaneously.</p>
		<p>- <b>Preference for Short Answers vs. Information Loss:</b> Users frequently request shorter answers, but shortening responses sometimes leads to the omission or oversimplification of critical information.</p>
		<p>- <b>Poor Awareness of when giving Abstract and when giving Detailed Information:</b> The AI chatbot struggled to tailor responses to the user's request for brevity, often providing exhaustive detail rather than concise summaries or overviews.</p>
Users behaviour when approaching the tasks	<p><b>Preference for Visual Information:</b> Instead of reading text-heavy sources, users preferred looking at visuals for quicker comprehension.</p>	<p>- <b>Need for Step-by-Step Instructions and Clear Guidance:</b> Users often attempted to tackle multiple tasks simultaneously rather than following a structured, step-by-step approach (e.g., identifying stakeholders before defining their needs). Instructions should clearly outline sequential steps.</p>
	<p><b>Immediate Reliance on Online Answers:</b> When working on stakeholder analysis, users often skipped brainstorming, immediately turning to Google to find pre-existing answers, reflecting a habit of quick information retrieval rather than deep engagement.</p>	<p>- <b>Direct Copy-Pasting of Instructions to AI Chatbot:</b> Users frequently copied instructions directly into the AI chatbot to obtain immediate solutions, highlighting the importance of clear, carefully crafted instructions.</p>