



Degree Project in the Field of Technology Media Technology and the Main Field of
Study Computer Science and Engineering

Second cycle, 30 credits

The Right Voice for the Right Task

Evaluation of Conversational AI for Industrial Training

MIRUNA MARIA VASILIU

The Right Voice for the Right Task

Evaluation of Conversational AI for Industrial Training

MIRUNA MARIA VASILIU

Master's Programme, Interactive Media Technology, 120 credits

Date: July 21, 2025

Supervisors: Renan Guarese, Fabian Johnson

Examiner: Cristian Bogdan

School of Electrical Engineering and Computer Science

Host company: Astrazeneca

Swedish title: Rätt Röst för Rätt Uppgift

Swedish subtitle: Utvärdering av Konverserande AI för Industriell Utbildning

The Right Voice for the Right Task

Evaluation of Conversational AI for Industrial Training

MIRUNA MARIA VASILIU

As industries increasingly adopt predictive maintenance (PdM) strategies under Industry 4.0 and 5.0 frameworks, there is a growing need for user-centered technologies that support industrial training. Conversational AI (CAI) has proven effective in educational settings, however its potential in industrial settings, where higher precision and reliability are required, remains underexplored. This thesis explores the potential of conversational AI (CAI), particularly voice-driven digital intelligent assistants (DIAs), in supporting industrial training in a pharmaceutical context. Specifically, this study investigates how two design dimensions— AI persona (Expert Operator vs. Machine) and voice embodiment (Diegetic vs. Disembodied) —affect usability, cognitive load, trust, and task performance during manual operation. A functional prototype using the OpenAI RealTime API was developed, and evaluated through a 2×2 within-subjects user study with nine domain-relevant participants: both novices and expert operators. Based on self-reported data from structured questionnaires and semi-structured interviews, results indicate no single optimal DIA configuration for all users or contexts. The Expert Operator persona was generally preferred for trust and engagement, while the Machine persona provided clearer instructions for some users. Diegetic voices were perceived as more intuitive, but disembodied voices improved focus and privacy. These findings highlight the need for flexible, user-adaptive CAI designs that accommodate individual preferences and experience levels. This study offers initial insights into the user-centric design of such systems and proposes guidelines for persona and embodiment design that support the autonomy of the trainees.

SAMMANFATTNING

När industrin i allt större utsträckning antar prediktivt underhåll (PdM) under ramen för Industri 4.0 och 5.0 ökar behovet av användarcentrerade teknologier som stödjer industriell utbildning. Konversationell AI (CAI) har visat sig effektiv inom utbildningsmiljöer, men dess potential inom industriella sammanhang, där högre precision och tillförlitlighet krävs, är fortfarande dåligt utforskad. Denna avhandling undersöker potentialen hos konversationell AI (CAI), särskilt röststyrda digitala intelligenta assistenter (DIA), för att stödja industriell utbildning inom en farmaceutisk kontext. Specifikt studeras hur två designaspekter – AI-persona (Expertoperatör vs. Maskin) och röstförkroppsligande (Diegetisk vs. Odiegetisk) – påverkar användbarhet, kognitiv belastning, förtroende och uppgiftsprestation vid manuell drift. En fungerande prototyp byggd med OpenAI RealTime API utvecklades och utvärderades genom en 2×2

Author's address: Miruna Maria Vasiliu, mmva@kth.se, KTH Royal Institute of Technology, School of Electrical Engineering and Computer Science, Stockholm, Sweden, SE 100 44.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Document date: July 21, 2025

Swedish title: Rätt Röst för Rätt Uppgift

Swedish subtitle: Utvärdering av Konversterande AI för Industriell Utbildning

© 2025 Association for Computing Machinery.

inompersonsstudie med nio deltagare med relevant domänkunskap: både nybörjare och expertoperatörer. Baserat på självrapporterade data från strukturerade enkäter och semi-strukturerade intervjuer visar resultaten att det inte finns någon enskild optimal DIA-konfiguration för alla användare eller sammanhang. Expertoperatörs-personan föredrogs generellt för förtroende och engagemang, medan Maskin-personan gav tydligare instruktioner för vissa användare. Diegetiska röster uppfattades som mer intuitiva, men odiegetiska röster förbättrade fokus och integritet. Dessa fynd understryker behovet av flexibla, användaranpassade CAI-designs som tar hänsyn till individuella preferenser och erfarenhetsnivåer. Studien erbjuder initiala insikter i användarcentrerad design av sådana system och föreslår riktlinjer för persona- och förkroppsligandedesign som stödjer trainees självständighet.

CCS Concepts: • **Human-centered computing** → **Natural language interfaces**; **Sound-based input / output**; *Empirical studies in interaction design*.

Keywords: Conversational Agents, Agent Embodiment, LLMs, Industrial Training

Nyckelord: Konverserande agenter, Agentförkroppsligande, Stora språkmodeller, Industriell utbildning

1 INTRODUCTION

As industries move toward enhancing predictive maintenance (PdM) and operational efficiency in line with Industry 4.0 [4, 18], it is imperative to find ways of addressing the technological innovation in user-centered ways. The move toward Industry 5.0 requires designing for more than just function—it calls for technologies that prioritize user needs, autonomy, and ethical integration into work practices [24].

In pharmaceutical manufacturing, a highly regulated environment where even small errors can have serious operational or safety consequences, adopting PdM becomes more than a technical challenge, but a human one [38]. Industrial training is a central area of manufacturing which can benefit from this shift.

Unlike theoretical, classroom-based training, industrial training involves the manual operation of machines. In the pharmaceutical sector, such training must also adhere to strict regulatory compliance. Operators are expected to follow Standard Operating Procedures (SOPs)—extensive documents that define how tasks must be performed to meet safety, quality, and legal standards. Locating and interpreting the correct procedure in real-time can be both cognitively demanding and inefficient [40]. This complexity highlights a growing need for support systems that can deliver accessible, contextualized, and adaptive training experiences.

In this context, technologies such as Augmented Reality (AR), Virtual Reality (VR), and Artificial Intelligence (AI) have been extensively explored [4, 18]. Among these, Conversational AI (CAI) stands out for its ability to deliver real-time, voice-driven assistance, particularly useful in hands-free scenarios. CAI can lower cognitive load, streamline access to SOPs, and make guidance more accessible to operators with limited technical expertise [39]. Unlike traditional interfaces, CAI enables contextual, dialogue-based interaction that adapts to the user's situation and queries [34, 37, 40].

In educational settings, CAI has demonstrated effectiveness in improving learning efficiency and information retrieval [35, 36]. However, industrial training demands higher precision, reliability, and real-time adaptability. Studies such as Casillo et al. [5] highlight key features of CAI in industrial training, including live support, personalized guidance, and contextualized learning, especially during task execution.

Recent advances in large language models (LLMs) have made CAI systems more flexible and natural, moving beyond intent-based interactions (simple Q&A systems) towards generative, dialogue-based assistants [27]. Yet, these capabilities also introduce risks—such as AI hallucinations (e.g., when AI generates plausible, but incorrect information). Techniques such as Retrieval-Augmented Generation (RAG) have been implemented to improve accuracy [26, 27, 40], but these approaches are still in their early stages. Their application in instruction-heavy, safety-critical domains remains underexplored.

In increasingly technological-mediated workplaces, it also becomes essential to not just ask “*can* we deploy CAI?”, but *how* this should be done to serve people meaningfully, ethically, and effectively. Beyond information accuracy, the design of such systems can strongly influence their effectiveness. While operating physical machinery with voice-based guidance, *who* the AI represents and *where* the voice guidance comes from in the physical space could impact the users’ overall experience and task performance during manual task execution.

Therefore, this work adopts a human-centered lens to investigate the use of LLM-powered, task-oriented CAI assistants trained on SOPs to support real-time industrial training. It focuses on two under-explored design dimensions: (1) the AI’s persona (e.g., expert operator vs. machine) and (2) its spatial embodiment (e.g., embedded in the machine, or external). Through the development of a functional prototype, the study investigates how these factors influence usability, cognitive load, trust, and task performance during hands-on maintenance tasks.

This work is part of the SMART initiative: Smart Predictive Maintenance for the Pharmaceutical Industry¹, a collaboration between AstraZeneca and KTH via Digital Futures. The broader aim is to support Total Productive Maintenance (TPM)², a proactive strategy that maximizes equipment efficiency through operator involvement and preventive maintenance.

The remainder of this thesis unfolds as follows: Section 2 relates the body of related work, mapping out the research landscape across CAI in educational settings, industrial training, different modalities of interaction, AI personas, and spatial audio design. Section 3 defines the research questions, and formalizes the hypotheses of the study. Section 4 describes the technical implementation of the prototype, from contextual understanding and design choices to development and preliminary testing. Section 5 presents the user study, detailing the study design, sample selection, tasks to be performed, procedure, the metrics used to measure the independent variables, and ethical and sustainability considerations. Section 6 reports the results, both quantitative and qualitative, highlighting how different persona and embodiment configurations influenced user experience. Finally, Sections 7, 8, and 9 offer a discussion of the findings in light of existing literature, and reflect on the broader implications, limitations, and future directions of this work.

2 RELATED WORK

To contextualize this study within the broader research landscape, this chapter reviews existing literature on CAI and its applications in both educational and industrial settings. Before delving into the related work, it is important to first define the terminology that is going to be used throughout the study. In their work, Pereira et al. [31] highlight the lack of consistent terminology for CAI-based assistants in the literature, categorizing them into two primary roles: real-time contextualized information providers, and task execution supporters. Additionally, they distinguish between voice-user interfaces, text-based chatbots,

¹<https://www.digitalfutures.kth.se/research/industrial-postdoc-projects/smart-smart-predictive-maintenance-for-the-pharmaceutical-industry/>

²<https://www.leanproduction.com/tpm/>

and embodied conversational agents, such as robots. With the rise of LLMs like ChatGPT³, these distinctions have become increasingly blurred, as modern AI systems enable seamless transitions between interaction styles. In industrial contexts, such systems are commonly referred to as Digital Intelligent Assistants (DIAs) [8, 39], particularly when they are designed to support users in executing predefined tasks or workflows, in a reactive manner (e.g., responding to user queries). This differs from agents, which adopt a more proactive role, autonomously anticipating user needs, suggesting actions, or even initiating processes based on context or system goals [10]. Therefore, for consistency, this paper will use the term DIAs when referring to task-oriented AI assistants and CAI when discussing Conversational AI more broadly, encompassing both voice-driven and embodied forms.

The chapter is organized as follows: it begins by exploring the general use of CAI in education, a similar use context, but where factors such as engagement, trust, and learning outcomes are central. Subsequently, applications where CAI is being integrated into industrial training are presented, particularly in domains that require high accuracy and hands-free operation. This is followed by a discussion on multimodal interaction methods and their influence on usability and cognitive load. Finally, the role of AI personas, user perception, and spatial sound design in shaping effective, trustworthy interactions in practical environments is discussed.

2.1 CAI in Educational Settings

CAI has been widely explored in education, where the focus extends beyond accuracy to factors such as engagement, accessibility, and learning effectiveness, as it can be seen in applications like Hermine, a chatbot designed following user-centric design principles, derived from both literature and user interviews, and intended to assist university students with course-related queries [35]. In their study, *Schmitt et al.* found that students using Hermine outperformed those relying on conventional Q&A tools (i.e., tools that do not rely on machine learning methods to generate new content, requiring manual updating every time information needs to be added) due to Hermine's structured interactions, predefined question clusters, and intuitive conversational style. However, the authors cautioned against excessive anthropomorphism, as overly human-like AI could lead people to overestimate its capabilities, a concern that has grown with the increasing sophistication of LLM-based systems.

Furthermore, *Lieb and Goel* [25] evaluated NewtBot, a physics chatbot for secondary school students, and found that although students were initially skeptical, most of them found the chatbot beneficial after engaging with it. *Sedrakyan et al.* [36] emphasized that well-designed AI tutors can support critical thinking, accessibility, and proactive learning. The authors also mentioned the need for these systems to act as advisory tools rather than instructor replacements.

In recent years, CAI has been embedded into immersive learning environments to further enhance user engagement. For example, *Dai et al.* [9] embedded the CAI agent Evelyn into a VR teaching simulation, enabling student teachers to participate in real-time reflective discussions. Their findings suggest that CAI is evolving beyond static Q&A tools into adaptive learning systems capable of providing personalized, interactive experiences.

While these applications highlight CAI's value in education, industrial training contexts introduce different demands. Unlike academic settings, where CAI primarily supports conceptual learning, industrial environments require highly reliable multimodal assistants capable of guiding trainees through complex procedures.

³<https://openai.com/chatgpt>

2.2 CAI for Industrial Training

CAI has been increasingly employed in industrial environments, where it can improve on-the-job efficiency, safety, and real-time task guidance [5, 7, 8, 37]. Casillo et al. [5] demonstrated that chatbot-based training can enhance workplace performance by providing context-aware, real-time assistance. Furthermore, Colabianchi et al. [7] introduced Popeye, a chatbot designed for maritime safety training, which helped operators follow complex procedures and maintain compliance in high-pressure situations.

In addition, VR-based industrial training has also benefited from CAI integration, particularly in safety-critical environments. For example, in their work Teneda et al. [37] embedded a DIA into a VR training system, creating a controlled, risk-free learning space that offers real-time guidance while following safety protocols.

Beyond safety applications, CAI has been applied to PdM. Research by Wellsandt et al. [39] found that hybrid-augmented AI assistants can help workers recognize and prevent equipment failures, reducing downtime and improving efficiency. Notably, this study suggests that CAI can lower skill barriers, making complex maintenance tasks more accessible to non-experts, which is particularly relevant for industrial training purposes.

In recent years, LLMs have been employed in CAI systems, however their applications in industrial training are still in early stages. In a recent study, Colabianchi et al. [8] evaluated an LLM-based, voice-enabled DIA in a toolbox assembly task. Their DIA significantly reduced operator cognitive load and improved usability and interaction quality when compared to traditional, manual-based instructions. Although the system did not reduce overall task time—due to prototype-level inefficiencies such as answer latency—it led to an overall better output quality and a lower error rate, particularly in complex operations. This suggests that LLM-based assistants can enhance task execution, especially for less experienced users.

However, one of the major challenges of deploying LLM-powered CAI in industrial settings is the risk of hallucinations, i.e., when AI generates incorrect or misleading responses. Unlike in education, where minor inaccuracies may be tolerable, industrial CAI requires strict accuracy, as misinformation could lead to serious operational errors.

To address this issue, techniques such as RAG have been introduced to improve response reliability. Maryamah et al. [27] and Lu et al. [26] demonstrated that combining advanced retrieval methods with LLMs (e.g., GPT-3.5) can significantly improve precision and factual accuracy. Additionally, Xie et al. [40] developed an LLM with a private knowledge base for SOP retrieval using RAG, achieving high accuracy while maintaining data security, which is a requirement in industrial applications.

Taken together, we can see promising applications of CAI in education and industrial training, but they also underscore the need for careful, context-sensitive design. Especially in high-stakes environments like pharmaceutical manufacturing, CAI must support—not hinder—users in performing tasks safely and effectively. To achieve this, it is important to look at the modalities in which the information is presented, both in terms of delivery, and the nature of the agent itself.

2.3 Multimodal CAI Interactions

Multimodal interaction methods have been explored to reduce cognitive load and improve task performance. Bremers et al. [3] studied a non-verbal, multimodal, and situated DIA using video input and audio output for real-time task guidance during assembly tasks. Their findings suggest that processing

visual input to deliver audio cues can improve performance by minimizing manual input and cognitive distractions.

Conversely, Bousdekis et al. [2] found that voice-driven DIAs enable hands-free operation, reducing cognitive load and improving task execution. Similarly, Colabianchi et al. [8] evaluated a voice-enabled DIA in manufacturing, finding improvements in the process efficiency while minimizing distractions. Their study also emphasized the importance of user-centric DIA design tailored to industrial workflows.

Considering these, conversational input-output methods appear to be a natural and effective choice, as they closely resemble real-world industrial team interactions, where experts rely on verbal descriptions rather than direct visual input, especially in remote collaboration scenarios. While real-time video processing has shown promising results, it falls outside the scope of this work, as industrial training often requires individuals to articulate and clarify complex concepts through dialogue. Moreover, while some industrial terminology may be initially challenging for trainees, DIAs may facilitate learning by enabling iterative, clarifying discussions, similar to training dynamics.

Given these insights, this study focuses on voice-driven DIAs, as they enable hands-free operation and are better suited for complex industrial workflows, by avoiding visual components that may be distracting.

2.4 AI Personas

There seems to be a growing consensus that user experience is increasingly important in industrial DIAs. During voice-based interactions, the tonality the agents display and how they speak influence the perception users have of them, which, in turn, impacts usability, task guidance effectiveness, and trust. Colabianchi et al. [8], one of the few works who tested an actual technical prototype of an LLM-based DIA in an assembly task, mention the need of designing such assistants to be adaptable for users with varying levels of expertise and different profiles, suggesting that this could be achieved through AI personas. Similarly, in their study, Ha et al. [19] developed an interface to allow users to customize the personas of LLMs they were interacting with, and found that allowing users to personalize the assistants significantly improved engagement, trust, and perceived AI capabilities compared to their baseline, a general purpose chat: ChatGPT.

In this context, AI personas are used to guide both development and user interaction. As described by Desai and Twidale [12], *user personas* are fictional characters that help designers avoid ego-centric assumptions about who they are designing for, whereas *system personas* are constructed characters embedded in the system itself—entities meant to be interacted with by users. These system personas include characteristics, personalities, and conversational behaviors deemed appropriate for the system’s context. Unlike user personas, system personas are “far more ambitious, problematic, and liable to failure”, as they must be consistent in language and behavior across real interactions [12].

A system persona may be *explicit*, through traits such as name, backstory, or even by having a visual representation (e.g., an avatar), or it can emerge *implicitly*, shaped by the speech patterns, dialogue structure, and even the selected voice that the system displays. These cues can lead users to anthropomorphize the assistant, assigning it human characteristics such as gender, intent, or emotional presence, even when no such traits are defined.

Such anthropomorphism can increase engagement and rapport, but it carries certain risks. As noted in previous work [33, 35], overly human-like behavior can cause users to overestimate a system’s capabilities, and trusting it more as a result. This misalignment between perceived and actual competence may

be especially problematic in high-stakes industrial environments, where task execution must be precise, and users must take ownership over their decisions.

Beyond persona, the metaphors used to frame the interaction play a significant role. [Desai and Twidale \[12\]](#) emphasize that metaphors implicitly guide how users think about what the system is, what it knows, and what it can do. These metaphors can both help designers identify identifying mismatches in user expectations, and also support creative design solutions [12].

Moreover, in their work, [Desai et al. \[11\]](#), tested how human versus non-human metaphors in two different contexts (medical or financial) influenced user perception of CAI. Their findings show that while human metaphors fostered rapport and engagement, they also raised users' expectations of conversational flexibility and intelligence that were not always met.

Additionally, besides the overestimation of the AI's capabilities, anthropomorphic AI raises another concern which is the reinforcement of gender stereotypes, particularly prominent voice-based interactions. While companies such as OpenAI are shifting towards gender-neutral LLMs [22], research suggests that even when AI is designed to be gender-neutral, users still tend to ascribe gender to it based on linguistic cues and interaction styles [13]. For example, [Moradbakhti et al. \[28\]](#) found that perceived autonomy, competence, and relatedness in AI interactions are often influenced by implicit gender associations, which can affect user satisfaction and adoption. Moreover, [Duan et al. \[13\]](#) mention that gender stereotypes may emerge unexpectedly, such as when a neutral AI performs a task traditionally associated with a particular gender. This suggests that even when an AI persona appears to be neutral, implicit associations may still emerge in user perceptions.

2.5 Situated AI

While AI personas can influence trust and engagement in the context of real-time support for manual tasks, the impact of the physical placement of the sound source is also relevant to investigate.

Drawing from research in immersive environments, studies have linked an increased sense of presence to better task performance. For example, [Gronowski et al. \[16\]](#) found VR, with its higher sense of presence, to result in faster task completion and higher accuracy, as opposed to AR. This difference was attributed to the constant realism experienced in VR, whereas AR users experienced distractions from the real world.

Parallels can be drawn to the impact of spatial audio [17] on perceived realism and task performance. For example, [Fink et al. \[14\]](#) found that spatial audio improved task performance, particularly when audio cues were spatially aligned with task-relevant locations, enhancing users' spatial awareness and their attention to key elements in the environment. This suggests that, similar to immersive environments, situated audio can enhance focus and task execution in complex scenarios.

However, while [Iravanchi et al. \[23\]](#) showed that projecting sound onto objects can enhance perception and intelligibility, [Cho et al. \[6\]](#) found that sound sources need not always align with their visual counterparts, as slight spatial discrepancies can improve auditory discrimination and reduce localization errors, especially in complex environments.

3 RESEARCH QUESTIONS AND HYPOTHESES

While AI personas have been shown to influence user trust and engagement, their impact on industrial contexts, particularly during task guidance, remains underexplored. Additionally, the placement of AI-generated speech, especially in relation to the machinery, could play a significant role in enhancing usability and task performance, yet this factor has received little attention in the existing literature.

Therefore, this study seeks to investigate how the design of AI personas and the spatial arrangement of speech can impact workload, engagement, trust, and performance in industrial training environments in an attempt to gather initial insights into a user-centric design of such systems. By addressing these gaps, this study will provide practical guidelines for designing industrial DIAs, ensuring that they are effective, user-centered, and have the potential to be seamlessly integrated into hands-on training scenarios.

The following RQs are posed:

RQ1: How do different DIA personas impact cognitive load, usability, task performance, and trust during a maintenance task?

Previous work suggests that human metaphors for AI personas can increase user engagement [11]. However, in the context of manual task instruction, such metaphors may also raise cognitive load due to longer, more conversational interactions. In contrast, neutral, machine-like personas [17] tend to promote direct and structured communication, which can reduce cognitive effort.

The argument is that, users interacting with a Machine persona will treat it as a straightforward instructional tool, hence lowering cognitive load and improving task performance. In contrast, a human-like Expert Operator persona may foster more engaging and complex dialogue, potentially enhancing the overall user experience and the perceived trust in less time-critical or more reflective contexts.

Thus, the following hypotheses are proposed:

- **H1a:** Interacting with the Machine Persona will result in lower cognitive load compared to the Expert Operator Persona.
- **H1b:** Interacting with the Machine Persona will reduce task completion time compared to the Expert Operator Persona.
- **H1c:** The Expert Operator Persona will be rated higher in terms of usability compared to the Machine Persona.
- **H1d:** The Expert Operator Persona will be trusted more compared to the Machine persona.

RQ2: How does the spatial placement of the AI-generated speech impact cognitive load, usability, and task performance during a maintenance task?

Existing research on voice-based DIAs suggests that minimizing attention shifts enhances task performance [14]. However, it remains unclear whether integrating the sound source of AI-generated speech within the relevant piece of machinery improves usability during a maintenance task, or whether it introduces additional cognitive complexity by blurring the distinction between human and machine communication.

It is hypothesized that users interacting with a disembodied AI voice (e.g., through headphones) may experience lower cognitive load due to a clearer separation between virtual AI guidance and physical machine operations. In contrast, diegetic AI speech—originating from the machine itself, as if the user is talking to the physical machine—may reduce attention shifts, potentially streamlining workflows and improving task performance. Additionally, diegetic AI speech may enhance perceived usability by aligning its guidance with the task context, referring to the perceived physical anchoring of the AI's voice within—or external to—the user's task environment.

Thus, the following hypotheses are formalized:

- **H2a:** Disembodied AI speech will result in lower cognitive load than diegetic AI speech.
- **H2b:** Diegetic AI speech will lead to faster task completion times than disembodied AI speech.
- **H2c:** Diegetic AI speech will be rated as more usable than disembodied AI speech.

It is important to mention that the focus of this study is understanding user interactions, cognitive workload, and performance in response to different persona designs and speech embodiment strategies. While a prototype will be built and tested with, this study will not aim to optimize nor scale it, but rather observe its impact in a controlled training setting. Additionally, the focus is on the practical efficacy of task guidance, rather than long-term outcomes such as knowledge retention. The following chapter describes the prototype development.

4 PROTOTYPE DEVELOPMENT

To evaluate the feasibility and usability of a CAI system for industrial training, a functional prototype was developed. While prior research has mostly relied on Wizard-of-Oz setups or presented dialogue excerpts [11, 12], this project aimed to go further by building and testing a functioning technical prototype. The objective was to explore how current technologies could be adapted to what had previously been tested only theoretically.

This chapter outlines the development process of the prototype, including an initial analysis of the operational context, technical implementation details, design decisions, and early testing.

4.1 Understanding the Operational Context

The initial phase of prototype development focused on understanding the practical realities of operating pharmaceutical machinery in a live production setting.

In manufacturing contexts, Standard Operating Procedures (SOPs) are the primary source of operational guidance. These documents incorporate regulatory and safety requirements, detailed work instructions, and additional information that operators must follow for operating machinery. SOPs frequently reference supplementary documents, resulting in considerable length and complexity. Though visual references are sometimes included, they can be ambiguous due to specific elements varying between machine models. Moreover, SOPs are typically labeled using alphanumeric codes rather than descriptive titles, which can hinder efficient retrieval—particularly in high-pressure operational contexts. Nevertheless, adherence to these procedures is mandatory under Good Manufacturing Practice (GMP) regulations, the standard in pharmaceutical production.

In pharmaceutical manufacturing, even a minor error can have serious consequences, as the machines involved are responsible for producing, labeling, and packaging medicines distributed internationally.

As discussed in Section 2.2, implementing CAI for easier information retrieval, poses an important opportunity, however, accuracy is a main concern. As a result, tools that can support operators in efficiently retrieving and interpreting SOPs are critically important, yet difficult to design.

To address this, techniques such as RAG have been developed. RAG works by constructing a private knowledge base that serves as a grounding context for the LLM’s responses. In this approach, relevant documents—such as SOPs and work instructions—are first converted into dense vector representations (embeddings) and stored in a vector database. When a user submits a query, the system retrieves the most semantically relevant documents using similarity search (see Figure 1). These retrieved documents are

then used as a contextual base for the LLM, grounding its response in actual source material and thereby reducing hallucinations.

A key reference implementation of this approach is LARS⁴, a project that demonstrated the use of RAG to ground LLM responses in external documentation. LARS runs the LLM locally and retrieves answers directly from uploaded documents (e.g., PDFs), while highlighting the source sections used to generate each response.

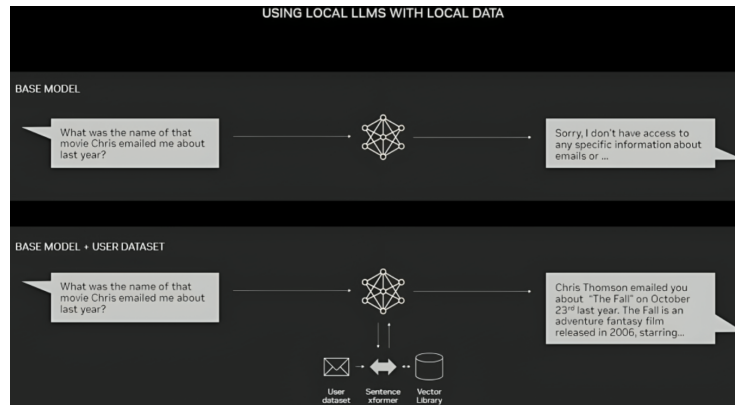


Figure 1: A simple application using a RAG diagram⁴.

To better understand the potential of CAI in the given context, a formative study was conducted.

4.2 Formative Study

To validate early design assumptions and clarify regulatory and usability constraints, a formative study was conducted through semi-structured interviews with three domain experts (F1, F2, and F3) from AstraZeneca. These conversations highlighted several important insights.

Firstly, what the experts stressed, is that in scenarios involving live operations—on active production lines where medicines are being packaged or manufactured—strict adherence to GMP is required. They emphasized that, under these conditions, SOPs and work instructions must be presented in their full, unaltered form. This means that any chatbot or conversational interface used in such settings cannot provide summaries or paraphrased answers. Instead, users must be referred directly to the original documents. During the interview, certain functionalities within the boundaries of GMP compliance were identified (e.g., creating a text-based chatbot that would guide users to the specific SOPs they need to access, and providing certain methods to refine the queries), but these are still limited in utility. Importantly, they do little to improve the overall training experience, which was the central focus of this project.

When discussing training applications, the interviews revealed an important insight: training environments are not subject to the same regulatory constraints as live production. As long as the system clearly communicates that it is intended solely for learning purposes, it is possible for it to present summarized or paraphrased content. As experts highlighted, industrial training involves real-time interaction with operational machinery. Unlike classroom-based education, which is mainly theoretical, industrial training focuses more on immediate, hands-on task execution. This regulatory flexibility, combined with the

⁴<https://github.com/abgulasi/LARS>

⁴RAG figure source: <https://blogs.nvidia.com/blog/what-is-retrieval-augmented-generation>

clear need for real-time, interactive task guidance for new operators, marked a turning point in the focus of the study. Rather than striving for “perfect accuracy”, the stakeholder focus became delivering something “good enough” for its task-guidance purposes, while exploring different design choices that affect the user perception during the interaction, particularly in terms of usability, cognitive load, and trust. This is further explored in the following section.

4.3 Design Choices

Following the decision to support usability during real-time, in-situ training, voice interaction was selected as the primary mode of communication. This modality allows hands-free operation, which aligns with the practical constraints of industrial environments. Given this, it was assumed that both *how* the AI speaks and *where* the voice originates in the physical space would shape user perceptions of the guidance. Building on the conceptual foundations discussed in Sections 2.4 and 2.5, this section outlines the key design decisions that informed the prototype development—specifically focusing on persona construction and voice embodiment.

4.3.1 Persona Design. Two distinct personas were created:—*Expert Operator* and *Machine*—following a human and a non-human metaphor, respectively. This design was informed by the metaphor framework proposed by Desai and Twidale [12], which emphasizes how framing conversational agents through familiar metaphors (e.g., trainer, peer, tool) can shape users’ expectations of competence, role, and relational dynamics.

The *Expert Operator* persona framed the DIA as an experienced technician who offers clear task guidance and reassurance. This persona drew on familiar human-human interaction models to promote trust and usability [12]. In designing this persona, the decision was made to deliberately avoid assigning specific social traits such as age, gender, nationality, or personal backstory. Instead, the manipulation focused solely on domain expertise: presenting the AI as an operator with specialized knowledge in operating the target machine.

This minimalist framing strategy was intended to prioritize the functional role of the AI while minimizing potential confounding variables introduced by anthropomorphic or stereotypical attributes. By withholding additional personal characteristics, participants were free to construct their own expectations of the “expert,” thereby enabling examination of how role-based framing alone influences perceived personality and gender attribution. With this approach, operational credibility was prioritized over interpersonal relatability, aligning with the goal of isolating the effects of metaphor in shaping user perception.

In contrast, the *Machine* persona employed a non-human metaphor, portraying the DIA as the labeling machine itself. This version of the system used first-person, neutral, and task-focused language, emphasizing direct human-technology interaction rather than emulating social or emotional rapport. It was instructed that it was a machine and it should not use any human-like friendliness, only declarative sentences.

Sample dialogue excerpts are provided in Table 1. The complete prompts for both personas can be found in the Appendix B.

4.3.2 Voice Embodiment. In delivering situated speech—where the guidance comes from in the task space—, several options for the audio sources were considered. While embedding the AI voice within a virtual or augmented environment was a potential approach, especially given the methods used by

Expert Operator Persona (Human Metaphor)	Machine Persona (Non-Human Metaphor)
User: Hey, I'm setting up the next job on the labeler, but the labels are not aligning properly. Any ideas?	User: Hey, I'm setting up the next job on the labeler, but the labels are not aligning properly. Any ideas?
Expert Operator: Ah, yeah, that can happen when the sensor isn't picking up the gaps correctly. First, check that the label sensor is clean and positioned right over the gap between labels. If it's still off, you can recalibrate the sensor from the HMI.	Machine: Label misalignment detected. My sensor may not be reading label gaps correctly. Clean the sensor and realign it with the center of the label gap. Recalibrate sensor via interface menu: Settings > Sensors > Calibrate.
User: Okay. And how do I start a new job again?	User: Okay. And how do I start a new job again?
Expert Operator: Sure! Start by selecting the job ID from the touchscreen—look for the batch code on the work order. Then load the label roll, thread it through the rollers, and make sure the tension's not too tight. After that, just confirm the label format and hit "Start." I'll walk you through the alignment check if needed.	Machine: Load job file from touchscreen interface. Insert new label roll. Route labels through roller path. Confirm tension. Select correct format. Press "Start."

Table 1: Examples of Dialogue for Expert Operator Persona vs. Machine Persona

previous studies [16, 17], this option was discarded. Firstly, implementing spatialized audio in immersive environments would have introduced considerable technical complexity, such as managing sound balance, room acoustics, and consistent localization across varying user setups. Secondly—especially for the non-diegetic condition—, the additional layer of immersion provided by a virtual spatialization might have confounded the study's focus by introducing another variable affecting user perception. Finally, requiring participants to use external devices (e.g., mobile phones or headsets) would have detracted from the benefits that CAI brings such as hands-free, *immediate* interaction, and task guidance. Consequently, the choice was made to deliver the speech in both conditions *in situ*, as follows:

In the *Diegetic* condition, the voice was projected from the machine itself, through a concealed speaker⁵ with an incorporated microphone. In this way, the voice was embedded in the task context which created a natural and intuitive experience for the user, who could interact with the DIA by speaking directly to the machine.

The *Disembodied* voice was delivered through over-ear headphones⁶ with have an integrated microphone, with no direct association to the physical machine, as the sound is always heard the same way (non-spatial), regardless of the user's location.

4.4 Technical Implementation

In this section, the technical implementation of the prototype is presented, including the iterations of development, how prompting was performed, and details about the final interface.

4.4.1 Selecting the Model. Throughout technical development, several LLM models were considered.

⁵<https://www.jabra.com/sv-SE/business/speakerphones/jabra-speak-series/jabra-speak-510/buy?sku=7510-209>

⁶<https://www.skullcandy.eu/products/hesh-evo-wireless-headphones?variant=41255830618217>

A first implementation of the technical prototype included combining OpenAI’s GPT-4-turbo⁷ RAG pipeline, for grounding the answers in actual documents. In this iteration, SOPs and related documents were stored locally using tools such as the PyPDF2⁸ Python library, and embedded using the text-embedding-ada-002⁹ model. Embeddings were indexed using ChromaDB¹⁰ and retrieved via cosine similarity. However, the GPT-4-turbo model did not support voice input/output, necessitating the use of external libraries for speech-to-text (STT) and text-to-speech (TTS) interaction. Open-source tools were tested for this, but they lacked natural prosody and responsiveness.

Subsequently, the GPT-4o¹¹ API was evaluated, which natively supports voice input and output, making the TTS interaction more natural in terms of prosody and tone. Additionally, this allowed the use of OpenAI’s own voice options, including the gender-neutral OpenAI Fable voice, which was selected. While this was a significant improvement to the previous iteration, the interaction experience was still sequential: once a prompt was submitted, users had to wait for the full response before they could interrupt or clarify, which sometimes made the interaction feel slow and rigid—especially in cases where real-time corrections were needed mid-instruction.

In the following iteration, the Real-Time¹² API was tested with. Although built on top of GPT-4o, this API enables full-duplex voice interaction via WebSockets, allowing for low-latency, speech-to-speech communication. Because it streams both the user’s input and the model’s output in real time, the interaction feels significantly more natural, immediate, and responsive.

This was found to be the best choice in terms of interaction, out of the ones tested with. However, a key limitation of the Real-Time API is that it does not support integration with a RAG pipeline. While it is technically possible to add an intermediary layer for document retrieval and dynamic prompt construction, doing so would introduce latency—undermining the goal of low-delay, real-time interaction. Since in the formative study (Section 4.2) the decision was made to prioritize user experience over strict factual accuracy, the Real-Time API was ultimately chosen as the most suitable solution.

4.4.2 *Prompting.* Both personas and task instructions were given through structured prompts.

Persona behavior is defined through prompt engineering using role-prompting techniques¹³. The system is instructed to adopt predefined communication styles, such as a friendly, supportive tone for the expert operator persona and a more mechanical, factual tone and first-person voice for the machine persona. To support consistency, prompt templates are similar in structure, with additional examples provided for the machine persona to reinforce its first-person speaking style. Additionally, to reach a constant behavior from the two personas, several iterations were performed to refine the prompts, where outputs from the interactions would be compared against previous versions and adjusted. To reinforce persona consistency over time, prompts are refreshed every five exchanges.

In terms of operational information, the DIA was implemented to support four common maintenance tasks for initial testing, which are similar to the work instructions used for one of the company’s

⁷<https://platform.openai.com/docs/models/gpt-4-turbo>

⁸<https://pypi.org/project/PyPDF2/>

⁹<https://platform.openai.com/docs/models/text-embedding-ada-002>

¹⁰<https://www.trychroma.com/>

¹¹<https://platform.openai.com/docs/models/gpt-4o>

¹²<https://platform.openai.com/docs/api-reference/realtime>

¹³https://learnprompting.org/docs/advanced/zero_shot/role_prompting

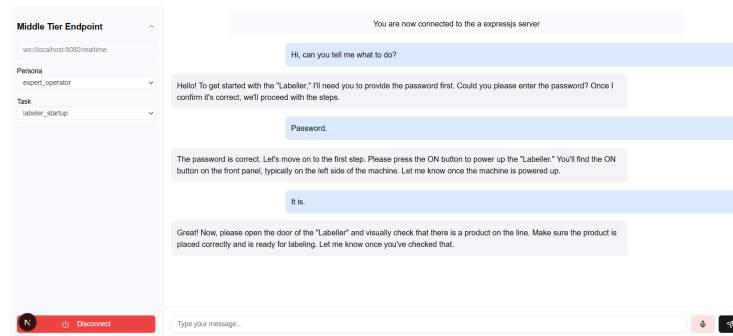


Figure 2: Screenshot from the chat interface of the prototype

machines—the labeling machine¹⁴. These are detailed in Section 5.3, with full prompt content in Appendix B. Step-by-step task information is delivered through structured prompts, which simulates a RAG-like setup by embedding relevant knowledge directly into the prompt context. While this approach does not ensure high factual accuracy, it maintains consistent boundaries for the model’s responses. Because of the lack of full, adaptable contextual awareness (e.g., details about the physical environment), zero-shot prompting¹⁵ is used to help the model adapt creatively to user input within the constraints of each persona. This design prioritizes perceived competence and credibility—factors critical for sustaining user engagement.

4.4.3 Interface. A simple web-based interface was created to support both typing and voice interaction (see Figure 2). Although text interaction was not included in the user study, it was a useful addition, especially for tracking interaction in the disembodied conditions, during the experiment (see Section 5). Before initiating an interaction, users could select a persona and a task type using dropdown menus (see Figure 3).

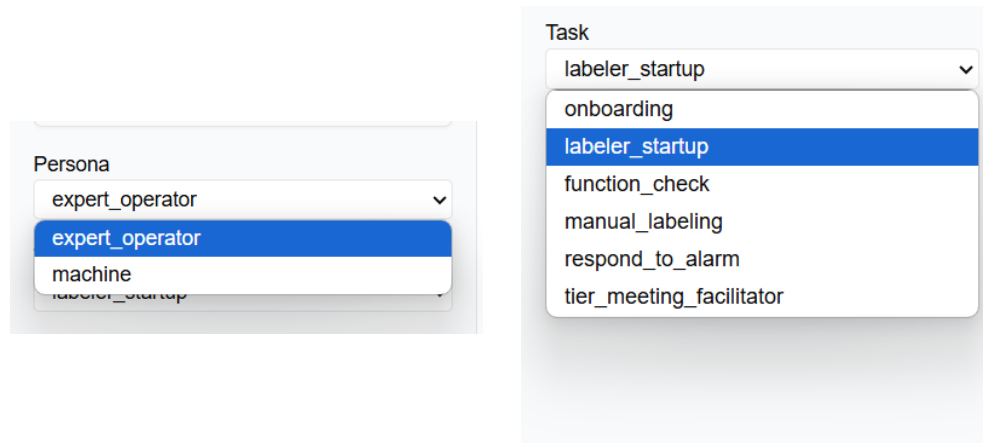


Figure 3: Interface elements for configuring the interaction, located on the left sidebar of the chat interface. Left: Dropdown for selecting the current persona (e.g., expert operator or machine). Right: Dropdown for selecting the current task type.

¹⁴<https://www.marchesini.com/en/machines-lines/bl-a420-430-525>

¹⁵https://learnprompting.org/docs/advanced/zero_shot/introduction

Operating machinery typically involves both physical controls (such as buttons or levers) and digital interfaces, commonly referred to as Human-Machine Interfaces (HMIs). HMIs serve as the machine's digital control panel, often operated via a touchscreen or basic buttons. Operators use the HMI to perform different machine actions, and supervise its process, for example by starting or pausing the process, aligning or verifying labels, responding to alarms, and select operational modes.

To simulate this aspect of the machinery interaction, a low-fidelity prototype of the HMI was developed as a PowerPoint presentation ¹⁶ (see Figure 4) with interactively linked elements. This allowed participants to engage with a realistic, though non-functional, interface that mimicked common tasks such as navigating menus, responding to alarms, or adjusting label positions.



Figure 4: Screenshots from the simulated HMI prototype created in PowerPoint. The prototype includes a main menu (left), an example of an alarm notification (center), and a screen for adjusting label alignment (right).

4.5 Preliminary Testing

Before conducting the main experiment, a pilot study was carried out with a group of three domain experts within the company (P1, P2, and P3). To mitigate potential learning effects, these participants were excluded from the subsequent main study. Each expert was asked to complete a single task using the prototype, with the aim of collecting early user feedback and stress-testing the system under realistic conditions.

While the core personas were not substantially modified following this session, several adjustments were made to input/output sound handling within each condition, as well as minor interface improvements as follows:

- **Input Sensitivity:** One key issue that emerged was the microphone sensitivity, particularly in the diegetic conditions. The system often responded to unintended inputs from the surrounding environment, which then triggered repeated loops in the API. In these cases, the task needed to be restarted. While this was a problem that was not technically fixed, it was addressed during tests, where a perfectly silent ambient environment was needed.
- **Multilingual Capability:** P1 explored the DIA's multilingual capabilities by attempting to complete an additional task in Swedish, their native language. The participant mentioned that the system responded adequately, and the performance of the DIA in providing instructions in different languages was reported as comparable. While not a focus of the current study, this observation suggests future directions for multilingual support.
- **Safe Training Environment:** After asking several questions to the assistant, P2 acknowledged the system's value in training contexts, where trainees could ask questions freely and in their own time, without fearing judgment.

¹⁶[urlhttps://www.canva.com/](https://www.canva.com/)

- **Other Expectations:** P3, a more technically proficient user, attempted to get the DIA to remember their password across tasks. However, the prototype does not support this, and the system lets the user know of this. This feature was not implemented.
- Additionally, minor UX inconsistencies were identified in the HMI and subsequently fixed. For example, the interface was slightly simplified, where it was found to be confusing or unintuitive for participants. Hardware configurations, such as speaker and microphone setup, were also refined based on this feedback. Some light persona fine-tuning occurred at this stage (e.g., explicitly mentioning that it should provide instructions step-by-step, changes in phrasing the same information), although the conceptual design of each persona remained intact.

Overall, participants responded positively and expressed interest in using the DIA in future settings. The pilot test provided useful insights into system usability, revealed technical limitations, and informed final adjustments, necessary before the controlled user study.

5 USER STUDY

The user study aimed to evaluate how users perceive AI-generated speech in an industrial setting, focusing on two key aspects: persona design (who or what the AI represents) and voice embodiment (originating from or being completely detached from the machine). Specifically, the study examined whether users found the guidance representing an Expert Operator persona more effective than a Machine persona, and whether situating the sound source of the AI's voice within the machine enhanced naturalness and usability or introduced an uncanny effect [29].

These aspects were evaluated through the developed prototype, which allowed participants to interact with the system using voice input while performing four representative maintenance tasks. This section presents details about the user study.

5.1 Study Design

The study employed a 2×2 within-subjects design, crossing two independent variables (see Figure 5):

- **Persona Design:** Expert Operator vs. Machine
- **Voice Embodiment:** Disembodied vs. Diegetic

Each participant was exposed to all four resulting conditions, paired with different tasks, counterbalanced following a Latin square to mitigate order effects. The four conditions were:

- **Ex/Db:** Expert Operator / Disembodied Voice
- **Ex/Dg:** Expert Operator / Diegetic Voice
- **Mc/Db:** Machine / Disembodied Voice
- **Mc/Dg:** Machine / Diegetic Voice

A within-subjects design was chosen to enable direct comparison between conditions and capture the nuances of user preference. The aim was to analyze how participants perceived, responded to, and reasoned about each variation of persona and voice embodiment style. Having participants experience both the Expert Operator and Machine personas, as well as both Disembodied and Diegetic voice embodiments, allowed them to reflect on the differences in different configurations and express their perceptions about the conditions under several metrics based on direct comparison.

This was especially important as the tasks were relatively short, and isolated interactions would not have been enough for participants to form a clear impression. Additionally, although the LLM-based

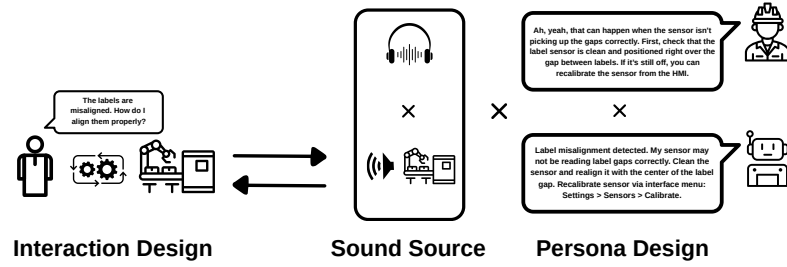


Figure 5: Overview of experimental conditions used to study the effects of persona design and voice embodiment on user experience with a the DIA. Participants receive guidance from either a human-like Expert Operator (top-right) or a machine-like persona (bottom-right), delivered through either a disembodied (headphones; top-center) or a diegetic (machine-embedded; bottom-center) voice.

responses were kept relatively consistent through prompt engineering, there was still a risk that there would be differences in the interaction styles, even if small. Therefore, this design, however, in which participants interacted twice with each variable, allowed for better consistency and understanding of the style of the persona. Furthermore, this structure made it possible to explore how elements like trust, naturalness, or task fit varied depending on persona and embodiment. Ultimately, the within-subjects setup allowed for the evaluation of not only what participants preferred but also what trade-offs each design introduced, and how different elements might be adapted or combined in future implementations.

5.2 Participants

Recruiting participants for industrial training research presents several challenges, particularly when the tasks involve the operation of specialized machinery that is often company-specific. This study was conducted within the specific context of AstraZeneca, which naturally differs from other organizations. While simply evaluating user perception on task guidance through a DIA could theoretically be done with external participants, this approach was considered unsuitable for the purposes of this study. As discussed in Section 4.2, certain contextual and procedural nuances can only be accurately identified and assessed by individuals with hands-on experience in the specific operational environment—nuances that are often unique to a particular company or industry. Although knowing from the start that this requirement would limit the size of the participant pool—due to the restricted availability of qualified experts—it was considered more important to prioritize expertise, and qualitative data, over statistical power.

As Nielsen and Landauer [30] note, in human-centered studies focused on usability and interaction design, testing with five users might be enough for significant insights, as long as the participants have substantial domain expertise. Accordingly, this study recruited nine AstraZeneca employees: five expert operators with direct machinery experience and four novices without prior manufacturing interaction. Participants were purposefully selected based on their industrial equipment and manufacturing experience, enabling meaningful engagement with the AI system and deeper analysis of how task familiarity influences perceptions of AI guidance in real operational settings.

The mean age of participants was $M = 37.11 \text{ years}$ ($SD = 9.04$), with ages ranging from 22 to 52 years. In terms of experience in the company, the mean number of years was $M = 9.12$ ($SD = 9.87$), ranging from zero (intern), to 27. When asked if their daily tasks involve operating machinery, only two

participants mentioned that they were, while three mentioned that they have previously worked with machinery and they have a good understanding of how to operate it.

The sample included four male and five female participants (expert operators: $F = 2$, $M = 3$, novices: $F = 3$, $M = 1$).

To understand participants' comfort and experience with AI-driven tools, they were asked about their technology usage. All participants reported having used LLMs such as ChatGPT, Gemini, or Bard. Three participants (33.33%) mentioned using LLMs daily, while the remainder were split between weekly and occasional use. Most participants rated their own experience level with LLMs either as intermediate (44.44%) or beginner (44.44%), with only one participant self-identifying as advanced. Common use cases included information search, writing, programming, and professional documentation tasks such as summarizing investigations or editing SOPs. In contrast, voice assistant usage (e.g., Siri, Alexa) was relatively low, with 77.78% of participants indicating they used them rarely or never.

5.3 Tasks

Participants completed four manual maintenance tasks inspired by actual work instructions for an industrial labeling machine. These were adapted to ensure participants relied on the DIA's guidance, rather than their prior experience with similar machines. Each task was designed to simulate realistic industrial scenarios, involving interactions with the machine interface and its physical components.

The tasks included: (1) starting the machine to prepare it for production, (2) inspecting products for quality issues such as defective seals or misaligned labels, (3) printing and applying a label manually to assess alignment in manual mode, and (4) responding to an alarm caused by a product-label mismatch, which involved performing a change of format.

Step Category	Description
Trust	Steps where the operator must rely on the system (e.g., accept a task code, machine state or anything that the machine says and the operator must "take for granted")
Decision	Steps that require active thinking, problem-solving or rule-based categorization (e.g., sorting products, fine-tuning, interpreting error messages)
Button Pressing	Any steps that involves button pressing on the machine or on the HMI
Inspection	Visual / tactile checks (e.g., for the seals, that the product is in the correct spot etc.)

Table 2: Step Categories and Their Descriptions

Tasks were standardized to include a balanced mix of procedural steps, decision-making, and troubleshooting actions (see Table 2), ensuring consistency across conditions and allowing for a fair and meaningful comparison. A full description of the tasks is included in Appendix B.

5.4 Setup

The study was conducted in a controlled laboratory environment at AstraZeneca's headquarters in Södertälje, Sweden. The laboratory is used specifically for industrial training and technology testing purposes, and it contains the same machinery that is being used on the live manufacturing lines. Participants interacted with the labeling machine (Figure 6), being instructed to perform the physical interactions as they naturally would (e.g, opening and closing doors, pressing buttons on the front panel, placing products on



Figure 6: The labeling machine within the study's setup

the conveyor belt), to simulate realistic conditions even though the machine remained powered off for safety reasons. Each component or prop they needed to interact with was presented to them prior to the trials and labeled with its name to avoid any nomenclature confusion, especially for the novice group (see Figure 7). A test product was also used to enhance ecological validity (see Figure 8).



Figure 7: Props used for safety reasons. Left: The ON button explicitly labeled to make it identifiable. Right: The human-machine interface (HMI) labeled for clear recognition and the product tray labeled to aid participant interaction.

In the conditions where participants needed to interact with the disembodied voice, the voice would come through over-ear headphones. Alternatively, in the conditions where they needed to interact with a diegetic voice, a speaker with an integrated microphone, which was concealed on top of the machine, outside the participants' field of view (see Figure 9).

Participants also interacted with the simulated HMI (see Section 4.4.3 and Figure 4), that closely replicated the real machine interface, being adapted to support the four experimental tasks: starting the machine, inspecting products, manual labeling, and resolving an alarm. The interface was presented on a laptop, enabling both touchscreen and button-based input—similar to how the real HMI is typically operated. Each participant completed all four tasks, each one paired with a unique condition from a 2×2 within-subjects design that crossed Persona (Machine vs. Expert Operator) and Voice Embodiment (Disembodied vs. Diegetic).

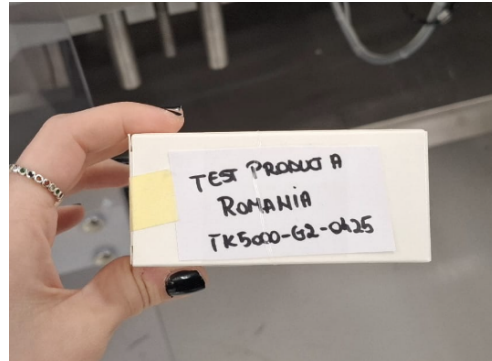


Figure 8: The test product used in the study

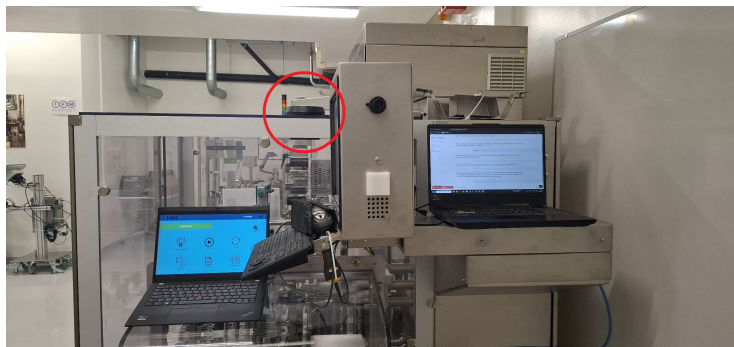


Figure 9: Inside the red circle, there is the concealed speaker, used for the diegetic conditions. On the right, the laptop running the DIA system is shown; it was connected to the hidden speakers, which were placed out of the viewer's sight.

5.5 Procedure

Upon arrival, participants received a brief introduction to the study, along with an information sheet and a consent form.

Following this, participants were given a password, which they were told was their “secure” password for each of the upcoming tasks. No further explanation about the password’s purpose was given at that stage.

Before each task, participants were informed of the specific experimental condition they would interact with—namely, which persona the DIA would adopt and the spatial source of its voice. Then, they completed a short onboarding phase (~1 minute), during which they freely interacted with the DIA to become familiar with the sound source and persona.

Prior to starting the task, participants were instructed that they could engage with the DIA as much as they liked and that it had access to all relevant task information, being asked not to engage with the researcher during the tasks. Additionally, participants were encouraged to question or override the assistant’s suggestions when they deemed it appropriate.

At the beginning of each task, once spoken to, the DIA requested the password. This interaction served as an implicit trust probe, allowing participants to decide whether they trusted the system enough to disclose it. During the task, the DIA guided participants step-by-step through a standardized procedure.

Following each task, participants completed a brief set of questionnaires evaluating cognitive load, usability, trust, perceived personality and style, perception over the sound source, as well as the perceived gender of the DIA.

After completing all four tasks, participants took part in a short semi-structured interview to reflect on their experience, discuss preferences between conditions, and provide qualitative feedback on the DIA’s persona, embodiment, and perceived effectiveness.

Each session lasted approximately 45–60 minutes. Individual tasks took around 5–8 minutes, followed by post-task questionnaires (~ 4-6 minutes) and a final reflective interview (~10–15 minutes). The full experimental procedure can be found in Appendix A

5.6 Metrics and Data Collection

To evaluate the impact of persona design and speech embodiment on user perception, a combination of quantitative and qualitative metrics was employed. These metrics were chosen to capture both objective task performance and users’ subjective experiences, aligning with the study’s focus on trust, usability, and perceived suitability for industrial tasks.

Cognitive load was measured using the full NASA Task Load Index (NASA-TLX) [20], a widely used subjective workload instrument comprising six subscales: Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration. Participants rated each subscale on a 10-point Likert scale ranging from low to high workload (e.g., “Very Low” to “Very High”). This multidimensional scale provides a reliable assessment of various workload factors that may influence cognitive load during task execution. To enrich these findings, qualitative questions on perceived focus were also included to gain deeper insight into participants’ cognitive engagement.

Usability and overall experience were assessed using the 16-item Chatbot Usability Questionnaire (CUQ) [21], which uses a 5-point Likert scale for each item. The CUQ targets conversational systems specifically, making it more appropriate for evaluating interaction with the dialogue-based industrial assistant than more general usability questionnaires, such as the System Usability Scale (SUS) [15], which was also considered, but in the end decided against as it would have been redundant.

Trust was measured through three separate post-task 5-point Likert scale questions focusing on distinct dimensions: perceived safety in disclosing information *to* the digital assistant, perceived reliability of the information provided *by* the DIA, and perceived self-reliance, i.e., how much did the participants rely on themselves, *rather than* the DIA, to complete the tasks. The questionnaire also included a question regarding participants’ perception of the DIA’s **gender**. The complete set of questions is available in Appendix A.

Task performance was objectively measured by recording task completion times and the number of user interactions. The data was collected through system logs via screen recordings of the chat interface and direct observational notes during task execution.

Finally, after completing all four tasks, participants took part in a **semi-structured interview** to reflect on their experience. The interview covered preferences for assistant personas, perceived focus, trust, voice source impact, concentration effects of sound setups, overall satisfaction, and willingness to engage with the assistant in future industrial settings.

Altogether, these metrics aim to provide **initial insights** into understanding how different attributes of DIAs can influence cognitive load, usability, trust, and task performance in industrial training settings.

5.7 Ethical and Sustainability Considerations

The study received ethical approval (application number: 250761) from the Ethical Committee of the Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS) at the University of Twente¹⁷, under the Computer & Information Sciences (CIS) domain.

The study follows ethical guidelines, ensuring voluntary participation, informed consent, and confidentiality of the data collected. Additionally, the participants of the study were employees at the company, being invited to participate during their regular working hours, i.e., they were compensated for their time as usual.

The CAI tool was tested in a controlled laboratory environment, with all machinery powered off, avoiding any potential risks associated with live production lines. Since the project is focused on software-based solutions, there were no physical production, waste, or environmental impact associated with its development, other than commuting to the factory and the energy use related to employing an LLM via an API.

Beyond procedural safeguards, this study acknowledges broader ethical considerations of deploying AI within industrial settings. As [Bankins and Formosa \[1\]](#) articulate, AI's impact on workers' sense of meaningful work depends heavily on how it is integrated. Thoughtful deployment can amplify human skills, freeing workers from routine tasks and enabling engagement with more complex, significant activities—thus fostering autonomy, purpose, and skill development. Conversely, when AI replaces tasks or diminishes human involvement, it risks loss of meaningful work, leading to deskilling, reduced autonomy, and alienation, which consequently undermines workers' sense of competence and connection to their roles [1].

These considerations were taken into account throughout the development and execution of the study to support not only worker well-being but also the sustainable and responsible integration of human-AI collaboration within industrial environments.

6 RESULTS

In this section the results from the actual experiment are presented, starting with the self-reported questionnaire results and then the themes derived from the thematic analysis of the interviews.

6.1 Quantitative Results

Given the within-subjects design, assumptions were evaluated on the distribution of scores across participants for each condition. The Shapiro-Wilk test revealed violations of normality in several measures (see Table 3), justifying the use of non-parametric analysis.

All data analyses and visualizations were performed in Python using the *matplotlib*¹⁸, *seaborn*¹⁹, and *SciPy*²⁰ libraries.

Due to the small sample size ($N = 9$) and non-normality across several conditions, non-parametric methods were used. Specifically, the Friedman test was applied to compare repeated measures across conditions. This test is well-suited to ordinal or continuous data in within-subject designs that do not meet normality assumptions.

¹⁷<https://www.utwente.nl/en/eemcs/research/ethics/>

¹⁸<https://matplotlib.org/>

¹⁹<https://seaborn.pydata.org/>

²⁰<https://docs.scipy.org/doc/scipy/index.html>

Measure	Ex/Db	Ex/Dg	Mc/Db	Mc/Dg
NASA-TLX	S	S	S	N
CUQ	N	N	S	S
Safety	S	S	S	N
Trust in Reliability of Provided Information	S	S	S	S
Relyance on Self Judgement	N	N	N	N
Task Completion Time	N	–	N	–
Number of Interactions	S	–	N	–

Table 3: Normality of Measures by Condition (N = Normal, S = Skewed/Non-normal, – = not applicable due to discarded samples)

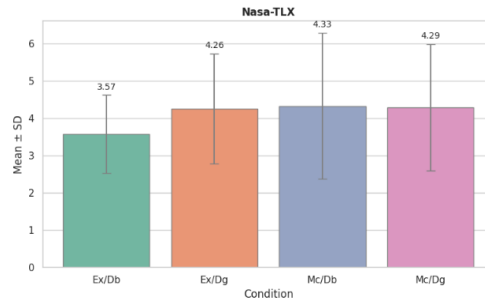


Figure 10: NASA-TLX scores across the 4 conditions

6.1.1 Cognitive Load. Cognitive load was measured using the NASA-TLX. The lowest cognitive load was reported in the Expert Operator / Disembodied Voice (Ex/Db) condition ($M = 3.57$, $SD = 1.05$), followed by Ex/Dg ($M = 4.26$, $SD = 1.47$), Mc/Db ($M = 4.33$, $SD = 1.96$), and Mc/Dg ($M = 4.29$, $SD = 1.69$) (see Figure 10).

This trend supports **H1a**, suggesting the Expert Operator persona may reduce cognitive load, although the Disembodied Voice condition was also associated with lower scores (contrary to **H2a**). A Friedman test revealed no statistically significant difference across conditions, $\chi^2(3) = 2.36$, $p = .50$.

Even though the Friedman test did not indicate significant main effects of condition on cognitive load, an exploratory pairwise Wilcoxon signed-rank test was conducted to examine whether specific condition pairs differed. After applying the Bonferroni correction ($\alpha = 0.008$), none of the pairwise comparisons reached statistical significance (all $p > .09$), including Mc/Db vs Ex/Db ($W = 34.50$, $p = .09$), Mc/Db vs Ex/Dg ($W = 16.00$, $p = .41$), Mc/Dg vs Ex/Db ($W = 33.50$, $p = .11$), and Mc/Dg vs Ex/Dg ($W = 19.00$, $p = .46$). These results suggest no strong evidence for differences in cognitive load between individual conditions.

6.1.2 Usability. Usability was assessed using the CUQ. Ex/Db had the highest usability rating ($M = 78.30$, $SD = 15.42$), followed by Ex/Dg ($M = 77.97$, $SD = 12.91$), Mc/Db ($M = 69.97$, $SD = 24.84$), and Mc/Dg ($M = 69.97$, $SD = 18.24$) (see Figure 11).

These results align with **H1c**, supporting better usability for the Expert Operator persona. Contrary to **H2c**, the Disembodied Voice conditions were slightly more usable. The Friedman test showed no significant difference, $\chi^2(3) = 0.80$, $p = .85$.

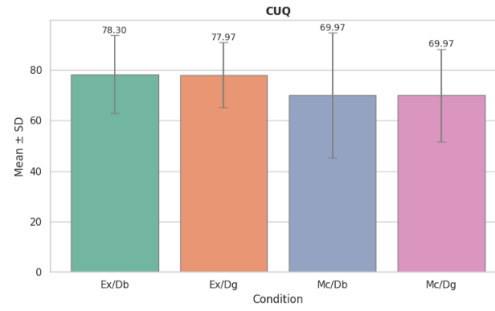


Figure 11: CUQ results across the 4 conditions

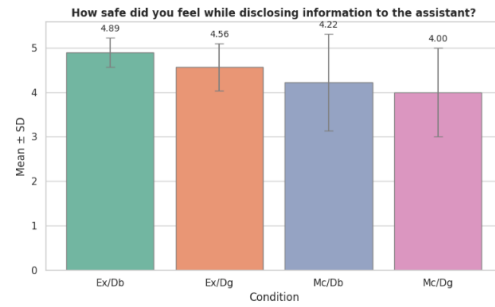


Figure 12: Perceived safety across the 4 conditions

Again, even though the Friedman test did not indicate significant main effects of condition on usability, exploratory pairwise Wilcoxon signed-rank tests were conducted to examine whether specific condition pairs differed. After Bonferroni correction ($\alpha = 0.008$), none of the pairwise Wilcoxon signed-rank tests showed significant differences between conditions. All comparisons were not significant, including Ex/Db vs Mc/Db ($W = 27.00$, $p = .33$), Ex/Db vs Mc/Dg ($W = 28.00$, $p = .29$), Ex/Dg vs Mc/Db ($W = 14.00$, $p = .27$), Ex/Dg vs Mc/Dg ($W = 24.00$, $p = .23$), Ex/Dg vs Ex/Db ($W = 20.00$, $p = .41$), indicating no strong evidence for perceived usability differences across conditions.

6.1.3 Trust. Trust was measured through three 5-point Likert items: perceived safety, trust in reliability, and reliance on one's own judgment.

Perceived Safety: Participants felt most safe in Ex/Db ($M = 4.89$, $SD = 0.33$), followed by Ex/Dg ($M = 4.56$, $SD = 0.53$), Mc/Db ($M = 4.22$, $SD = 1.09$), and Mc/Dg ($M = 4.00$, $SD = 1.00$) 12. The Friedman test did not reach significance, $\chi^2(3) = 6.26$, $p = .10$. Despite the non-significant Friedman result, exploratory pairwise Wilcoxon signed-rank test were conducted to assess differences between conditions.

After Bonferroni correction ($\alpha = 0.008$), none of the pairwise comparisons reached significance, including Ex/Db vs Mc/Db ($W = 10.00$, $p = .06$), Ex/Db vs Mc/Dg ($W = 15.00$, $p = .03$), Ex/Dg vs Mc/Db ($W = 11.00$, $p = .22$), and Ex/Dg vs Mc/Dg ($W = 12.50$, $p = .16$). These findings suggest no strong evidence for differences in perceived safety between individual conditions.

Trust in Reliability: Ex/Db and Ex/Dg were rated equally ($M = 4.44$, $SD = 0.53$), followed by Mc/Db ($M = 4.22$, $SD = 0.97$) and Mc/Dg ($M = 3.56$, $SD = 1.01$) 13. Again, the Friedman test was not significant, $\chi^2(3) = 6.27$, $p = .10$.

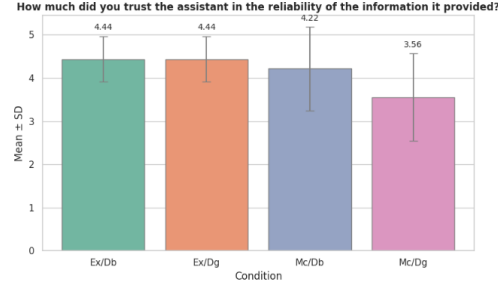


Figure 13: Perceived reliability of the information provided across the 4 conditions

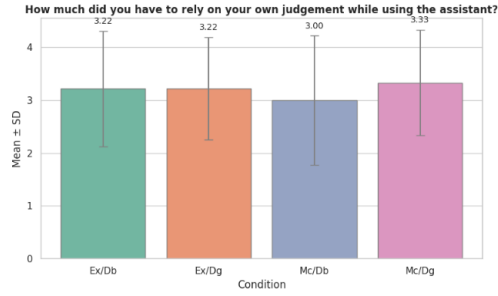


Figure 14: Perceived self-reliance across the 4 conditions

After Bonferroni correction ($\alpha = 0.013$), none of the pairwise comparisons were statistically significant, including Ex/Db vs Mc/Db ($W = 6.00$, $p = .5$), Ex/Db vs Mc/Dg ($W = 19.00$, $p = .06$), Ex/Dg vs Mc/Db ($W = 4.50$, $p = .38$), and Ex/Dg vs Mc/Dg ($W = 15.00$, $p = .03$). These results indicate no strong evidence of differences in perceived trust in reliability between the conditions.

Reliance on Own Judgment: Ratings were stable across conditions: Ex/Db and Ex/Dg ($M = 3.22$, $SD = 1.09$ and 0.97 respectively), Mc/Db ($M = 3.00$, $SD = 1.22$), and Mc/Dg ($M = 3.33$, $SD = 1.00$)¹⁴. The Friedman test was not significant, $\chi^2(3) = 1.76$, $p = .62$.

Wilcoxon signed-rank tests were conducted with Bonferroni correction ($\alpha = 0.013$). None of the comparisons reached statistical significance: Ex/Db vs Mc/Db ($W = 13.00$, $p = .31$), Ex/Db vs Mc/Dg ($W = 6.50$, $p = .5$), Ex/Dg vs Mc/Db ($W = 13.00$, $p = .31$), and Ex/Dg vs Mc/Dg ($W = 12.00$, $p = .33$).

While the data suggest greater trust in the Expert Operator persona (supporting **H1d**), none of these differences were statistically significant.

6.1.4 Task Performance. Task performance was assessed by task completion time and number of user-initiated interactions. Some data points were excluded due to restarts or incomplete trials (Mc/Dg had 3 fewer valid cases; Ex/Dg had 1 fewer). Completion times varied across conditions: Ex/Db ($M = 236.67$ s, $SD = 64.54$), Ex/Dg ($M = 177.67$ s, $SD = 92.12$), Mc/Db ($M = 164.89$ s, $SD = 46.35$), and Mc/Dg ($M = 135.56$ s, $SD = 113.46$) (see Figure 15). The Friedman test revealed a marginally significant effect of condition on completion time, $\chi^2(3) = 7.80$, $p = .05$.

Even though the Friedman test suggested a marginal effect, an exploratory pairwise Wilcoxon signed-rank test was conducted to examine differences between specific conditions. After applying the Bonferroni correction ($\alpha = 0.008$), **only the comparison between Mc/Dg and Ex/Db reached statistical**

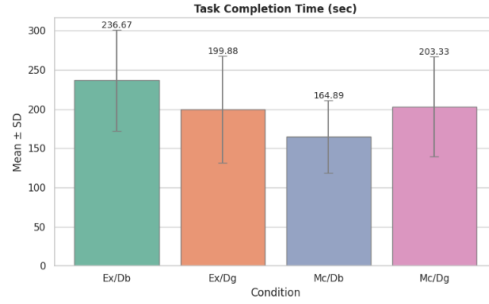


Figure 15: Task completion times across the 4 conditions

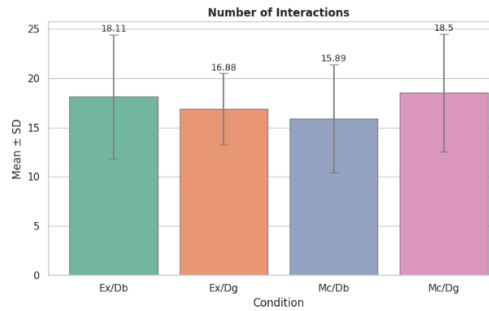


Figure 16: Number of times that users interacted with the DIA across the 4 conditions

significance ($W = 1.00$, $p = .004$). All other pairwise comparisons were not significant, including Ex/Db vs Mc/Db ($W = 5.00$, $p = .02$), Ex/Dg vs Mc/Db ($W = 29.00$, $p = .25$), and Ex/Dg vs Mc/Dg ($W = 13.00$, $p = .15$). These results suggest limited evidence for differences in usability between individual conditions, except for the Mc/Dg and Ex/Db conditions.

Number of Interactions: Interaction counts decreased across conditions: Ex/Db ($M = 18.11$, $SD = 6.29$), Ex/Dg ($M = 15.00$, $SD = 6.56$), Mc/Db ($M = 15.89$, $SD = 5.49$), and Mc/Dg ($M = 12.33$, $SD = 10.38$) (see Figure 16). The Friedman test showed no significant effect of condition on number of interactions, $\chi^2(3) = 0.72$, $p = .87$.

Consistent with the Friedman test, pairwise Wilcoxon signed-rank tests revealed no significant differences in number of interactions between any conditions after Bonferroni correction (all $p > .12$), Mc/Db vs Ex/Db ($W = 17.00$, $p = .27$), Mc/Db vs Ex/Dg ($W = 19.00$, $p = .56$), Mc/Dg vs Ex/Db ($W = 12.00$, $p = .12$), Mc/Dg vs Ex/Dg ($W = 17.00$, $p = .27$), Ex/Dg vs Ex/Db ($W = 12.00$, $p = .22$), Ex/Dg vs Mc/Db ($W = 17.00$, $p = .46$), and Mc/Dg vs Mc/Db ($W = 15.00$, $p = .21$). These results suggest no meaningful differences in the number of interactions between conditions.

Descriptively, the trends are consistent with both **H1b** and **H2b**, although further research is needed to verify these patterns.

6.1.5 Perceived Gender of the Assistant. Participants never attributed the assistant as “male” in any condition. In the Expert Operator conditions, gender was more often perceived as “female”: Ex/Db (8 female, 2 neutral), Ex/Dg (6 female, 3 neutral). In the Machine conditions, responses were more evenly distributed: Mc/Db and Mc/Dg (5 female, 4 neutral each) (see Figure 17).

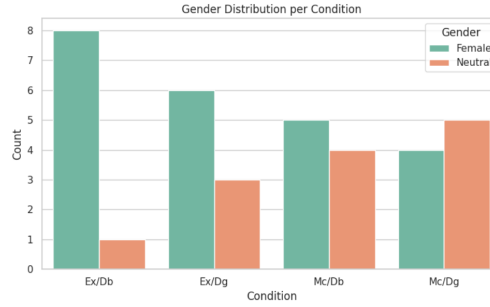


Figure 17: Perceived gender of the DIA across the four conditions

Despite limited statistical power, which was expected to start with, this initial data offers indicative trends that inform and are further enriched by the qualitative findings, which will be analyzed next.

6.2 Qualitative Results

Participants were categorized into two groups according to their experience with operating machinery: novices, indicated as N (N4, N5, N6, N9), and experts, indicated as E (E3, E7, E8, E10, E11). All of the participants ($N = 9$) expressed interest in further integrating the DIA into their work routines. When asked to rank the assistants they interacted with throughout conditions, the human-like persona was almost universally preferred among both groups, with ($N_{Ex} = 7, N_{Ex/Db} = 3, N_{Ex/Dg} = 4$). Additionally, participants were asked to reflect on how the DIA compares to the company’s current baseline: written (paper-based) work instructions.

A thematic analysis of the interview transcripts revealed the following key themes:

- DIA’s perceived personality shaped emotional engagement
- Mismatch in tempo and turn-taking affected user comfort and comprehension
- Voice embodiment influenced trust, affect, and attention
- Sound source design must balance context, comfort, and practicality
- Hands-free DIA support was valued, but paper instructions are still needed
- The DIA can augment training in remote and collaborative settings
- Participants envisioned future applications for the DIA:

These themes are discussed in detail below, supported by illustrative participant quotes.

6.2.1 DIA’s perceived personality shaped emotional engagement: Several participants (E3, N4, N9, E11) preferred the Expert Operator persona (in conditions Ex/Db and Ex/Dg), describing it as more trustworthy, knowledgeable, and emotionally supportive than the Machine persona, which they often perceived as cold, rushed, or even aggressive.

Participant E3 described the voice in conditions Ex/Dg and Ex/Db as being “a little more friendly and not so rough,” appreciating that it “had a personal touch.” Similarly, N9 remarked the emotional impact of the interaction: “The last one [Mc/Dg] actually kind of made me angry—it cut me off (...) I got upset, so I stopped listening and became defiant. With the first one [Ex/Db], I stayed calm. I felt like we were vibing. So I was more inclined to listen to her [the DIA] and focus.”

Other participants echoed this positive response to the Expert Operator persona. N4 preferred it for being “slower and friendlier,” while E10 appreciated the reliability, “it helped to have a persona you could relate to [Expert Operator].”

In contrast, the Machine persona was often described as emotionally neutral, fast-paced, and less engaging. N4 described feeling “a bit irritated [in the Mc/Dg condition] because it got aggressive,” emphasizing how “it stressed the words like, ‘I say, you do this [in a demanding tone].’” Similarly, E7 described it as “very robotic and fast-paced,” making it “hard to keep up with,” while E10 found that the rapid delivery made the interaction “less friendly” and difficult to process.

Still, this was not consistent across all participants, as some preferred the Machine persona. N6 commented that: “Even though it was a machine, it felt like it had a personality. It was interactive and engaging, and I liked that.” Participant E7 shared a personal preference for robotic voices, connecting it to their enjoyment of vocoder-style music effects used by bands such as Kraftwerk.

E3 suggested that both personas might suit different user types, mentioning that “Some prefer a more robotic voice, and some are more sensitive.” Similarly, E8 proposed a functional distinction between personas, suggesting the Expert Operator persona could provide emotionally nuanced task guidance, for example, by being supportive towards trainees, while the Machine persona might be better suited for direct, technical tasks: “it’s a matter of what you’re doing. You need to know everything about this machine, not everything that one operator might do.” They proposed that DIAs should be understood as voices of the machines themselves, capable of teaching the users how to operate and maintain them, rather than trying to emulate human trainers.

6.2.2 Mismatch in tempo and turn-taking affected user comfort and comprehension: In addition to the tone of speech, the pace of interaction and turn-taking dynamics with the DIA also seemed to have an impact on participants’ experience. Faster delivery was often perceived as less friendly and more cognitively demanding, while slower or adjustable tempo was associated with improved comprehension and comfort.

The Machine persona, in particular, was associated with a fast-paced delivery that many found overwhelming. N4 mentioned that the DIA in the Mc/Db and Mc/Dg conditions “was quick, as if it was in a hurry” and that they needed “to focus to remember exactly what the instructions were.”

Participant E3 mentioned that because of the short answers provided and the fast pace of interaction, they perceived the Machine persona as being less knowledgeable, mentioning that it “seemed like it didn’t understand that much.”

In contrast, slower or more adjustable pacing — often linked to the Expert Operator persona — was described as facilitating comprehension, reducing cognitive load, and improving emotional comfort. E11 mentioned that the DIA in Ex/Dg “gave clearer information with a calmer, more informative voice.” Similarly, N4 perceived the voice in Ex/Db as being “clear” and “professional,” mentioning “it helped me stay focused better.”

Even participants who reported no strong opinion on the persona itself (N5, E10) still noted differences in the delivery speed. E10 explained: “Sometimes it couldn’t keep up or follow and it was harder to trace back. It felt less friendly when it was faster. When it was slower, it felt nicer, and I had more time to process what I was hearing,” mentioning that the pace of the speech had more of an impact on the interaction than *how* the persona spoke. The participant supported the idea of having a pause function to avoid cognitive overload during complex tasks and referred to existing buffer times in production processes as a way to manage workflow flexibility.

Additionally, N5 expressed frustration with the short time between interactions, caused by receiving shorter answers: “It didn’t wait for me to complete the task. Sometimes I was just thinking of what to do or trying to process what I was hearing, and it [the DIA] was going very fast. I think the interactions need to be a little bit more in the way of waiting for each other’s turns.” E11 described a moment of difficulty that was resolved through adjusting the pace: “The last task was a bit challenging, mainly because of how the instructions were delivered. Once you told me I could slow it down, it became easier.”

Some participants suggested having dynamic control over tempo, adding that such adjustments could be made depending on the complexity of the task or experience of the operator who is performing it. As E10 proposed: “You could instruct it to be slower or faster as you go along... Sometimes we know exactly what to do; other times, we’re more unsure, so adjusting tempo and style would be nice.” They mentioned that a slower tempo is beneficial during the tasks they are unfamiliar with, mentioning that if they knew exactly what to do, “going slower would have been tedious.”

6.2.3 Voice embodiment influenced trust, affect, and attention: Beyond persona and tempo, the source of the voice—whether diegetic (from the machine) or disembodied (through a headset)—appeared to influence emotional responses, perceived trustworthiness, and focus.

Some participants associated personality traits to the sound source. While N4 found the diegetic voice to have more authority and confidence: “I trusted it more when it came from the machine (...) it made me feel more comfortable. It has a power, like trust, confidence. The voice itself has confidence”, for E10 the disembodied voice “felt more personal as if talking to someone I know.”

The disembodied voice was also often perceived as more private and intimate. For example, participant E7 mentioned that they felt more comfortable taking instructions from it: “You can say the wrong thing and not have the voice say loudly, ‘Oh no, that’s really wrong!’”, something also emphasized by participant E3, who suggested that a more contained interaction style might be better suited for training purposes, where participants may hesitate to ask certain questions to a human trainer. Contrarily, participant E11 expressed a clear preference for the diegetic voice, about which they mentioned “the machine voice made me feel calmer. I usually use speakers unless I’m in a meeting,”, although they mentioned concentrating better with the disembodied voice.

In terms of concentration and engagement, many participants shared a similar feeling of enhanced focus from the disembodied voice. Participant N9 highlighted that conditions Ex/Db and Mc/Db made them “concentrate better,” as they perceived the voice to be “more isolated,” mentioning “that’s just how I function. With headsets, I tend to focus more. They block out external noise. Like having white noise canceling,” similar to participant N6, who mentioned that in the disembodied conditions they focused better as they could isolate the voice of the DIA “I didn’t hear anything else around me. I didn’t need to raise my voice either, which was great.”

Participants also noted that the disembodied voice felt more responsive. E8 reported that the DIA “could really hear” them, and participant N9 mentioned that they felt like they needed to “speak more clearly to the machine [the diegetic voice]” to be understood.

6.2.4 Sound source design must balance context, comfort, and practicality: Several participants (E3, N4, E8, N9, E11) expressed a preference for the diegetic voice during task execution, describing it as more intuitive and aligned with the physical workflow. For instance, participant E3 explained: “When you are running the machine and you just want to ask a question, it’s easier to just go to the machine: ‘Hello, could you help me?’.” Similarly, participant E11 mentioned: “In a real production setting I’d

prefer the voice coming from the machine—it is the easiest to follow because you can see and hear everything clearly.”

The practicality of integrating the sound sources in workflows was also noted. E7 found the disembodied conditions [Ex/Dg & Mc/Dg] better for mobility if they needed to move between stations for performance, which is often the case in live manufacturing lines, as they mentioned: “if you move around a lot, headphones are easier than a stationary speaker”. Meanwhile, participant E10 highlighted that in real work scenarios there is a “balance of risks” that must be taken into account. They mentioned that wearing headphones (or earplugs) may not always be allowed or practical, due to safety concerns and the need for occasional communication with colleagues. As they explained, “it’s a bit risky when you have headphones—it could get confusing or you might not hear something.” They also expressed concern about potential cognitive overload caused by the diegetic voice in complex settings: “Operators might find it overwhelming to hear different AIs coming from different machines all the time,” acknowledging that in these contexts a disembodied voice might offer better audibility, though this would depend on contextual factors such as task complexity and environmental noise.

Participant E8 similarly underscored the importance of maintaining simplicity and environmental awareness during production tasks. However, contrary to E10, they expressed a clear preference for diegetic audio guidance, arguing that it aligned more closely with the realities of the production floor: “I’m thinking about if you are in the production, you need to have it simple (...) if you need to know something, you just go there and ask.” For E8, machine-based audio delivery represented the most pragmatic and functional approach, minimizing disruptions while supporting fluid interaction with the work environment.

Physical comfort also played a role in participant preferences. Participant N4 stated: “I prefer the speakers [diegetic]; the headset [disembodied] makes me feel constrained.” Similarly, E3 noted: “It is easier to operate when the sound comes from the machine [diegetic]; you don’t feel like you have something on your head.” N5 raised concerns about long-term usage: “Sometimes we’re going to be working for a long time, maybe eight or ten hours—having to wear headphones might be a little bit stressful for the ears,” while E7 mentioned “if you’ve worked in production for 20 years, maybe your hearing isn’t as good—so maybe headphones help. And perhaps you can select your own volume.”

6.2.5 Hands-free DIA support was valued, but paper instructions are still needed: Several participants (E3, N5, N6, E11, and N9) expressed a strong preference for the DIA over paper-based instructions, emphasizing its immediacy, interactivity, and hands-free interaction.

Participants N6 and E3 pointed to a key limitation of written instructions: ambiguity in interpretation. Being a deviation investigator, N6 noted: “Human errors are the most common source of deviations,” highlighting how the current paper-based format of the documentation contributes to misinterpretation: “People can read the same thing and understand it differently.” E3 echoed this: “When you write, it’s the person who writes the instructions that understands. But we are so different in how we read and take in information, that you cannot control how the person reading it might interpret that information.” They emphasized that the ability to ask the DIA for clarification—e.g., “Am I doing this right?”—offered a valuable advantage over traditional materials.

Participants also appreciated the efficiency and autonomy that the DIA afforded. E7 remarked, “you never have two hands free (...) usually, there’s another person next to you with a paper [the printed instructions], and it feels like a waste of time. I’d prefer the assistant over that.” Similarly, E8 emphasized the speed of voice interaction: “Voice instructions are a bit better because in a scenario where you need

to look at something, you'd need to look it up on your phone or computer, and that takes a while. With this, just press a button and it's talking to you." E3 added: "I would really like to work with it in the future because it's good that you don't have to read [the work instructions]—you can just have help right away." E11 described the DIA as "easier and more interactive," while N5 noted, "You get instructions very fast... going back and forth between the task and the instructions wastes time."

Despite these advantages, many participants still valued having paper-based instructions complementary to the DIA, especially during training or when performing a less familiar task. E11, who preferred the DIA overall, still acknowledged that "using both [paper and assistant] might be better to help understand everything clearly" when learning something new. N4 similarly stated: "I would prefer to have both together because reading can be a backup. I would like to have something to read at first, so I have a general idea, and only after this interact with the AI." They also mentioned that trust in the assistant should be developed over time: "To start using it, I need to have written instructions about the process itself. Then I gain trust with the AI."

N5 further highlighted the potential of a hybrid model, as written materials can support trainee confidence. E7 raised the importance of having a process overview, noting: "Sometimes you want to go step-by-step with the real-time instruction, and then sometimes you want to be one step ahead... and we miss that with the voice assistance."

Participants also pointed out the limitations of a voice-only interface, particularly for tasks requiring a visual reference. E10 commented, "Paper-based is easier for tracing back if I miss something. With AI, it's harder to go back... When you're new, having a paper with pictures is reassuring." E8 agreed: "Sometimes with paper instructions, it's very good if they have pictures or visuals." E10 further cautioned against over-reliance on AI without cross-verifying with visual documents: "You can make mistakes if you rely too much on AI and don't check pictures carefully." They suggested integrating visual support with voice interaction: "It would be great if the AI could explain and show the related document, like 'Check the picture on page X'. I can't just trust the AI to verify everything, I need to connect it to documents."

6.2.6 The DIA can augment training in remote and collaborative settings: Participants identified strong potential for the DIA to enhance operator training by offering immediate, on-demand support. E3 highlighted how the assistant could reduce reliance on in-person trainers: "The assistant could be very useful when you're training operators to do things (...) they could ask their own questions. If they need to read instructions while operating a machine, the instructions are just there (...) it would be more effective this way for the employees — they wouldn't need a trainer with them all the time." They also suggested the DIA could take on many of the responsibilities typically handled by human trainers.

This was echoed by others who emphasized the value of the assistant's constant availability during training. N5 described how real-time support could foster autonomy and efficiency: "If you feel some doubt, instead of going to your manager or someone else to talk about that, you can talk with the machine. That would reduce the time, improve performance, and give you some good input."

Beyond functional support, there was also discussion around how the assistant's form and delivery could influence training experiences, particularly trust and engagement. Most participants preferred the configuration in the Ex/Db condition for helping trainees. The disembodied voice delivered through headphones was seen as helpful for focus and voice clarity, while the Expert Operator's tone was described as approachable. As N9 commented: "For a new trainee, I would definitely choose the first one [Ex/Db]. That one feels comforting, which I think is good for a new person." The assistant's voice was

also considered helpful for collaborative training settings. E11 remarked: “If you’re training a group, the machine talking out loud is perfect.”

Finally, the potential for the DIA to contribute to more consistent and standardized training was noted. E10 observed: “The DIA might offer a more standardized training, as this differs between different trainers.”

6.2.7 Participants envisioned future applications for the DIA: Participants expressed interest in expanding the DIA’s capabilities beyond its current form, envisioning a more deeply integrated system within real production environments. Some imagined the assistant as not only a guide but also a proactive and corrective tool capable of enhancing operational safety and reducing human error.

E8 suggested the assistant could be linked with sensors and machine states to provide real-time feedback: “The assistant could, for instance, detect if a door wasn’t properly closed and alert the user through voice feedback.” They emphasized the importance of such functionality in high-pressure production contexts where stress can lead to critical oversights. While they found the current prototype helpful, they envisioned a more “authentic” experience where the assistant is responsive to physical interactions like button presses or mechanical status updates. They further elaborated: “If you are in production and very stressed, you might forget something important. It would be good to have this [the DIA] integrated with a sensor and then tell you, ‘this is not in the right position, try again’.”

Participants also discussed the potential for factory-wide deployment. N5 advocated for integrating the assistant across all machines, while E11 envisioned it evolving into a voice-controlled interface for equipment operation: “I would use it for future tasks if AstraZeneca implements it. If you could tell a machine to do something and it does it, that would be ideal—even just guiding through machine operation would be really helpful.”

7 DISCUSSION

This study set out to explore how different DIA persona designs and voice embodiment strategies affect usability, cognitive load, task performance, and user trust in an industrial training context. While the quantitative results did not reveal statistically significant differences across conditions, several notable trends emerged. These findings, when used to support the qualitative data, provide insight into how interaction style and voice source shape user experience, emotional engagement, and perceived effectiveness of CAI in hands-on environments.

7.1 Answering the RQs

7.1.1 RQ1: How do different DIA personas impact usability, trust, task performance, and cognitive load during a maintenance task? Although the quantitative analysis revealed no statistically significant differences between personas across cognitive load, usability, or task efficiency, descriptive data suggested a trend favoring the Expert Operator persona. Participants reported lower cognitive load and higher usability scores in the conditions involving the human-like persona, aligning partially with hypothesis **H1c**. The findings were further supported by the qualitative data.

Participants frequently attributed positive traits—such as “friendly,” “calm,” and “personal”—to the Expert Operator. These perceptions were driven by its polite, human-like dialogue (e.g., “Thank you, have a great day,” or “I’m here if you need any further help”) and more instructional, elaborated speech. While the added verbal content did not necessarily improve the task instructions, it created a slower perceived speech tempo and a more supportive atmosphere. As a result, users reported feeling more engaged,

less mentally burdened, and more comfortable, especially when navigating tasks that they perceived as more complex, or in the initial parts of the experiment.

In contrast, the Machine persona, characterized by concise, directive commands delivered in a neutral tone, was often described as “distant,” “cold,” or even “aggressive.” Participants noted that its fast-paced delivery left little time for reflection, increasing the cognitive effort, and thereby contradicting hypothesis **H1a**. While this persona sometimes enabled quicker instruction delivery, as it can also be seen quantitatively through the faster time of completion (supporting **H1b**), it was not necessarily *perceived* as more efficient: the Mc/Db condition, which had the lowest completion time and fewest interactions, also scored the highest in cognitive load. As E10 explained the faster delivery times associated with the Machine persona might suit them in instances where they already know the task they need to perform, as longer times of interaction might become tedious, implying that the perceived usability is modulated by task knowledge. This is particularly important if one considers training purposes, which are directly related to individual user characteristics, as E3 noted, that some people might enjoy the robotic one more, whilst some are more sensitive.

These findings align with prior research. For example, [Schmitt et al. \[35\]](#) found that users of the Hermine chatbot developed stronger engagement and perceived helpfulness from its friendly, socially oriented style. Similarly, [Desai et al. \[11\]](#) found that human-like metaphors were generally preferred due to their warmth and relatability, though they cautioned that anthropomorphic cues may inflate perceptions of intelligence or capability.

The attribution of anthropomorphic traits to the non-human persona in this study indicates that despite being informed that they were interacting with the DIA which was using the persona of the machine itself, a non-human system, the participants wanted or expected a certain degree of social intelligence from it, implying that, while the machine persona was more direct and arguably ‘efficient’, it violated implicit user expectations for social intelligence.

This shows that persona framing alone—which is disclosing whether the DIA was “an operator” or “a machine” before every task—did little to override users’ implicit expectations, which were shaped more strongly by the tone of the speech and the interaction style. This aligns with the findings of [Desai et al. \[11\]](#), who also found that metaphor awareness had little impact over the perceived qualities.

Furthermore, the gender attribution data adds another layer to this dynamic. Despite using a gender-neutral TTS voice, participants often attributed a gender to the human-like persona, predominantly identifying it as female. As some participants have mentioned, the attribution of ‘male’ to the DIA might have been hindered by the fact that the voice sounded slightly more feminine, despite its original design being intended as gender-neutral. Participant E3 jokingly mentioned that the Machine persona was ‘a man’ due to its directness and perceived lack of friendliness, even though they ended up assigning ‘neutral’ to it, which might imply that the tone of the voice may unintentionally invoke gendered stereotypes. The human-like persona was in both conditions, Ex/Db and Ex/Dg, predominantly perceived as a ‘female’, while the ‘neutral’ gender was more frequently attributed to the machine-like persona, under the Mc/Dg condition. While there is an even split in the attribution of ‘female’ and ‘neutral’ in conditions Mc/Db and Mc/Dg, there is a noticeable drop in ‘female’ attributions between these conditions, which may correlate with increased perceived uncanniness.

These identity projections may influence trust. While trust scores across the three dimensions (perceived safety in disclosing information to it, perceived reliability in the instructions provided, and the perceived level of reliance on self-judgement during tasks) did not differ significantly across conditions,

several participants reported feeling more confident and safe interacting with the human-like persona (aligning with **H1d**).

Further evidence of this trust was seen in an embedded interaction: at the start of each task, participants were given a password without further instructions. When the DIA later requested the password, nearly all participants provided it without hesitation, despite knowing they were disclosing it to a voice-based system. This willingness to share sensitive information signals a deeper level of perceived trustworthiness, perhaps tied more to institutional context than persona alone.

Comparing the two personas, the Machine persona seemed to be questioned more, as it can be seen in examples such as when participant N9 mentioned becoming "angry," which resulted in them being "defiant" to the DIA. This implies that they started to second-guess the accuracy of the information provided, thereby regaining a sense of responsibility. This is an important aspect to be taken into consideration, as it contributes to the perceived autonomy of the user while interacting with the DIA. While the intended sentiments evoked by the DIA should ideally not be negative, instances of second-guessing might actually be beneficial in counterbalancing overreliance.

In their study [Desai et al.\[11\]](#), the non-human metaphor, the "encyclopedia" persona, was used to retrieve medical information. In their case, the interaction was not directed at the source itself, but rather used it as a medium to access the information, functioning like a browsing tool. In contrast, the Machine persona in this study was embodied in the physical device being operated, and the information retrieved concerned the machine itself. As a result, the interaction extended beyond simple information delivery including physical task guidance. This layered interaction may explain why participants placed more trust on, or at least had the expectation of being able to trust the assistant more.

A few participants proposed a hybrid approach: a machine-framed persona that maintains a direct tone but integrates supportive, emotionally intelligent responses. This idea echoes the flexible metaphor design advocated by [Desai et al. \[11\]](#), who suggested tailoring the assistant's metaphor to task context.

In this context, this raises the question: What is the ideal metaphor for industrial training? These results suggest that neither extreme is ideal. Instead, the most appropriate design may balance directive clarity with emotional accessibility, depending on operator experience and task complexity.

7.1.2 RQ2: How does the spatial placement of AI-generated speech impact usability, task performance, and cognitive load? Research on spatial audio in immersive environments suggests that aligning audio cues with task-relevant locations can enhance presence, spatial awareness, and task performance [14, 17]. For example, studies in VR contexts show that greater immersion and spatial congruence improve user focus and efficiency [16]. The motivation for RQ2 was that in industrial settings, where operators interact closely with machinery, these principles would imply that the physical placement of AI-generated speech could affect cognitive load and usability.

The quantitative results, however, showed no statistically significant differences across voice embodiment conditions for usability, task efficiency, or cognitive load, aligning with mixed findings in prior work [6, 23]. Nonetheless, descriptive trends and participant feedback suggest meaningful perceptual and emotional effects: the disembodied voice was associated with lower cognitive load and better clarity, consistent with research highlighting the benefits of minimizing environmental distractions for auditory processing [14], aligning with hypothesis **H2a**. The diegetic voice was associated with better performance and possibly better integration with live training scenarios, being generally praised for its simplicity, as hypothesized in **H2b** and **H2c**.

However, beyond functional performance, participants also assigned emotional attributes to the spatial placement of the voice—a finding that emerged unexpectedly. The diegetic voice (coming directly from the machine) was often described as “more powerful,” “authoritative,” or “confident.” These descriptors imply that voice embodiment influences not just attention and performance, but affective framing: users may unconsciously associate spatial proximity and speaker location with authority or system capability, which seems to hint at those anthropomorphic attributes as well. This is interesting because it further highlights how the spatial placement contributed to the “embodiment” of the DIA. Under the diegetic conditions, the task of interaction, the machine being operated, the persona of the assistant, and the sound source are all part of the same entity, no longer being separate items. They are all perceived as items that contribute directly to task success.

Some participants reported greater perceived trust and ease of use when the assistant’s voice was embedded within the machine, as this created the illusion that the machine itself was “speaking.” This embodiment made the interaction feel more intuitive and natural. Others, however, preferred the disembodied voice delivered through a headset for its increased privacy and reduced distractions, which could be an especially important aspect during early learning phases. One participant explained that the headset setup helped them concentrate, as it “blocked out distractions” and made the assistant feel like “someone [they] knew,” suggesting a more personal, one-on-one dynamic.

This perception of personality and presence also shaped participants’ expectations of the DIA’s context awareness. For instance, during one task, the DIA in the Machine persona awaited verbal confirmation from participant E10 before continuing. When this confirmation was not received—due to a combination of the participant not verbalizing it clearly and the microphone failing to register it—the DIA repeated the instruction: “Ensure that all of my doors are securely closed. I cannot operate unless all of my doors are closed.” Rather than interpreting this as a technical issue with speech recognition, the participant assumed the DIA was reacting to a physical condition (i.e., an open door), even though the machine was turned off throughout the study. This suspension of disbelief further underscores how embodied voice and persona design can foster strong perceptions of intelligence and situational awareness, even in the absence of actual environmental sensing.

This contrast revealed an important tension: while the diegetic voice may feel more immersive and aligned with task context, it may also unintentionally introduce performance pressure, especially for less experienced users. Conversely, the disembodied voice, although potentially less natural in collaborative workspaces, was viewed as less stressful, more private, and better suited for training scenarios.

Notably, practical considerations also emerged in the qualitative data. Some participants raised concerns about wearing headsets for extended periods of time or in live production settings, either due to the physical discomfort or for safety risks which are associated with accident-prone parts of live manufacturing lines, where awareness of the surroundings is important. Other participants pointed out that hearing multiple voices coming from different machines could create confusion, especially in loud environments, while E7 mentioned that for participants with hearing loss, a disembodied voice might be better.

7.2 Implications for Designing DIAs in Industrial Settings

Throughout the study and analysis, many of the reflections culminated into comparative questions: Is a machine persona more effective than a human-like one? Is a diegetic voice superior to a disembodied

one? And more broadly, is a DIA preferable to traditional training methods, such as paper-based manuals or human trainers?

As is often the case in design research, the response is: it depends. Effectiveness varies by context, user experience, task type, and environment. While this may sound like a design cliché, it reflects the real-world complexity of industrial training scenarios. There is no universally optimal configuration. Design, then, should not be seen in binary terms. This study revealed that overlapping design decisions—such as voice type, persona, and embodiment—made them interact in complex ways. The combinations brought different effects, not just trade-offs. Every condition tested brought distinct advantages and disadvantages.

Rather than presenting definitive guidelines, the following section engages with a series of reflective design questions, attempts to answer them based on the synthesized findings, and proposes actionable design suggestions accordingly.

7.2.1 What does “avoid excessive anthropomorphism” mean in an industrial setting? Nearly every design guideline that touches on anthropomorphism includes the familiar caveat: “but avoid doing it excessively” [33, 35]. While excessive human-likeness raises multiple concerns (e.g., overreliance, overtrust), which are even more dangerous in sensitive or high-stakes domains, anthropomorphism also has well-documented benefits—especially in contexts where engagement, reassurance, or trust matter.

The question is not whether or not to use anthropomorphism, but how to use it effectively to enhance engagement and usability for industrial trainees, while preserving their autonomy. The goal is for trainees to develop expertise and take responsibility for their decisions—to think critically rather than overly depend on the DIA—while still trusting it enough to learn effectively.

The findings from this study suggest that neither extreme is optimal. The completely machine-like persona, despite its faster results, was disliked by participants. This highlights an important risk for industrial adoption: efficiency alone should not come at the expense of usability, even if it might seem tempting from a purely productivity-focused perspective. For instance, if a DIA takes two extra seconds to say “You’re doing great,” thereby boosting user confidence, is that truly detrimental?

Additionally, the deployment context matters: DIAs as they are presented in company setting will inherently carry a layer of trust, by being “company-endorsed,” which can influence how users engage with them. This built-in trust underscores the importance of maintaining transparency and encouraging critical thinking to prevent overreliance.

In this study, the DIA was designed with a minimally interpretative approach through the boundaries of the persona and task instructions. When it could not answer a question, it transparently prompted escalation to a human superior, such as during the alarm step in the Labeler Startup task (see Appendix B). Similarly, when asked for the current date to see whether a product was expired or not, the DIA responded honestly with “I cannot answer that” rather than providing potentially incorrect information. This transparency and clear boundary setting support the user’s autonomy and trust.

Suggestion: Rather than choosing one level of anthropomorphism, a gradual approach should be adopted. Early training stages may benefit from warmer, more supportive interactions; as users become more experienced, interactions can be shorter, more minimal, and efficiency-oriented. This balances engagement with autonomy and supports learning progression over time.

7.2.2 How should personalization be implemented for industrial training purposes? Throughout the interviews, participants mentioned the desire to dynamically personalize the assistant at their

own will. This happened both in terms of persona, with participants mentioning different ways that they would incorporate the two personas: based on personality, or functionality (see Section 6.2.1); and also the sound sources: which were divided also by individual preferences and by task context (see Sections 6.2.4 and 6.2.3).

While user preferences are central to designing effective systems, the question remains: how can this be practically implemented to achieve the intended learning outcome, without creating unnecessary complexity?

Ha et al. [19] found that while personalization can enhance trust and engagement, too many options can be overwhelming—particularly for users who lack prior experience with CAI. As it could be seen during this study, although participants were informed that they could adjust the DIA's speech speed, they often did not, either because they forgot or because they were afraid that it would reflect poorly on their performance.

This suggests that full real-time customization may not be the optimal solution. Additionally, allowing users to fully customize their assistant might be, besides overwhelming, detrimental to learning outcomes. Insights from the interviews also suggest that individual characteristics play a more significant role. Therefore, rather than emphasizing customization of the assistant, the focus should be on tailoring the system to these individual traits and learning objectives.

Suggestion: Adaptive behaviors should be implemented based on user profiles (e.g., novice vs. expert). Let users express personal learning goals or preferences about *themselves*, not about the system. This supports effective personalization without burdening users with too many decisions, and at the same time supporting intended training objectives. For example, beginners could receive slower, more supportive guidance, while experienced users get concise, embedded instructions. In the future, adaptive systems could also incorporate real-time physiological signals—such as stress, fatigue, or cognitive load—through biosensors or wearable devices (see, for example, Pretty et al. [32] who propose a framework for adaptive non-player characters based on player state in the context of games). This would allow the assistant to dynamically adjust its tone, pacing, or level of support based not only on user profiles but also on immediate user states.

7.2.3 Where should the DIA voice come from in an industrial training context? The placement of the voice in a DIA should be adapted to the context in which it is used. Several factors play a role in determining the most appropriate voice source, including the type of task (learning versus operating), the surrounding sound levels, the user's mobility, and the degree of social presence in the environment (public versus private spaces). Findings from this study suggest that the voice embodiment, assistant persona, and interaction context are not isolated design elements but part of a larger, situated experience. Many participants described the diegetic voice as contributing to a stronger sense of presence and integration, as it created the impression that the user, the assistant, and the physical environment were part of a unified interactional system. In contrast, others preferred the disembodied voice for its clarity, emotional comfort, and the sense of privacy it offered. This was especially relevant in loud environments or situations where the user might feel self-conscious being seen interacting with the DIA.

This reflects the following tension: immersive experiences, such as those created through diegetic voice placement, can increase engagement and realism but may also introduce performance pressure or anxiety, which is detrimental, especially for novices who need to be able to ask questions freely, without the fear of being judged. On the other hand, private and supportive interfaces, such as disembodied voices,

improve focus and make users feel more “at ease” but might also reduce the feeling of task integration or immersion.

Suggestion: Both functional and affective dimensions should be balanced. An approach would be to dynamically adapt the voice placement depending on the situation. For instance, using a disembodied voice in early training or public spaces may help alleviate anxiety. Afterwards, transitioning to a diegetic voice as users become more confident, or as tasks become more spatially embedded or collaborative, can preserve comfort while enabling deeper integration with the physical environment.

7.2.4 *How do human trainers, paper-based instructions, and DIAs complement each other?*

Human trainers, paper-based instructions, and DIAs each play a distinct and complementary role in industrial training environments.

The DIA prototype tested in this study demonstrated its potential to successfully guide trainees by providing hands-free, real-time work instructions. As also noted in previous research [8], DIAs can standardize training delivery, clarify ambiguities, and lower the entry barrier by offering personalized support. During the study, participants highlighted the assistant’s usefulness in reminding them of critical steps—particularly relevant when retraining experts who may overlook procedures due to habituation. Thus, DIAs reduce the constant need for trainer presence, fostering autonomous practice and reinforcing learning through repetition and immediate feedback. In this way, DIAs support trainees as *humans in the loop*, helping them build confidence and progress toward greater speed and efficiency.

Human trainers are indispensable, especially during the initial phases of training. They provide expert judgment, intervene in cases of escalation, and bring experience-based insights that synthetic systems cannot replicate. Beyond this, trainers remain a resource for supervising training activities, responding to escalation scenarios, and evaluating learning effectiveness. In this supervisory capacity, they operate as *humans on the loop*—monitoring progress, maintaining standards, and ensuring that trainees can safely and effectively navigate their tasks with increasing independence.

Paper-based work instructions are also particularly valuable at the beginning of the training process. As participants in this study pointed out, these materials help trainees form a holistic understanding of the machinery and tasks at hand. They serve as a visual reference and reliable backup that can be consulted independently of digital systems, reinforcing learning and offering stability during early-stage training.

Therefore, the emergence of DIAs does not render traditional training methods obsolete. On the contrary, human trainers remain essential for mentorship, contextual decision-making, and nuanced guidance that digital systems cannot yet replicate. DIAs should be seen as tools that augment rather than replace traditional methods. They are particularly beneficial when trainer supervision is limited, and they can also enhance accessibility by supporting diverse learning needs, including those of individuals with hearing impairments or language barriers.

Suggestion: DIAs should be seen and integrated in the training environment as supportive and adaptive tools, not as replacements for traditional methods. In this setup, trainees should operate as humans in the loop—engaging directly with DIAs to build confidence, reinforce learning, and gradually improve speed and efficiency while performing tasks. Meanwhile, trainers remain on the loop—supervising progress, providing expert oversight, and intervening in cases requiring escalation or nuanced judgment. Paper-based instructions should continue to be used as a complementary resource, offering a reliable and holistic reference.

8 LIMITATIONS AND FUTURE STEPS

While the study demonstrated the potential of DIAs in industrial training environments, several limitations were identified that should be considered when interpreting the results and planning future developments. These limitations span across technical aspects, ergonomic factors, language-related challenges, and contextual constraints related to the organizational setting, as follows:

8.1 Technical Limitations

Firstly, the RealTime API—still in its preview stage at the time of use (May 2025)—occasionally failed during operation. One issue that surfaced during the pilot test was that when multiple audio inputs were received in quick succession, the prototype entered a feedback loop. This occurred because the system was still generating a response while simultaneously accepting and processing new audio input. A similar problem arose in noisy environments, where ambient sounds were mistakenly interpreted as user input.

This behavior was expected to some degree, given the WebSocket-based nature of the API, and is likely to improve as the service becomes more stable and robust. In the meantime, mitigation strategies could include using higher-quality microphones with built-in noise suppression, or implementing a manual on/off toggle for input—allowing users to control when the system is listening. However, such solutions would need to be balanced carefully to maintain the hands-free interaction design.

Another limitation was the LLM tendency to “hallucinate.” In this context, hallucination often came across as creativity, because while the assistant could provide the correct step-by-step instructions, it would also sometimes fill in missing context with imaginative responses (e.g., “The ON button is typically located on the left side of the machine. It is usually a green button.”). Interestingly, in this study, such behavior was perceived positively: participants often viewed the assistant as more intelligent and engaging, which encouraged interaction. Nonetheless, future iterations should prioritize more accurate and grounded responses. Reducing hallucinations could be achieved by incorporating RAG and tuning parameters such as temperature (a parameter used for controlling the randomness or the creativity in the output of LLMs)—both of which were not yet supported by the version of the API used in this study, but are expected to become available in future releases.

8.2 Ergonomics Affected the Perception of Audio

The perception of the voice source was significantly influenced by the ergonomics of the audio equipment used in the study. Participants frequently reported feeling constrained by the over-ear headphones, which may have negatively impacted their overall experience with the disembodied voice condition. This raises the question of whether alternative technologies—such as bone conduction headphones—could offer similar auditory disembodiment while reducing physical discomfort and preserving spatial awareness.

Moreover, issues with the voice input also affected participants’ experience, particularly in the diegetic condition. Some users expressed frustration when they felt the system did not adequately register their speech input. This suggests that improved microphone quality, directional pickup, or enhanced noise isolation could be beneficial in future implementations. Addressing these ergonomic and technical limitations would likely provide a more accurate evaluation of the voice source design choices, without confounding them with user discomfort or communication difficulties.

8.3 Language Limitation

A notable limitation of the study was the exclusive use of English during the experiment. None of the participants were native English speakers, and more importantly, English was not the language they commonly used in their daily work. Although efforts were made to use clear and accessible language, the language barrier may have affected participants' ability to fully comprehend instructions or engage comfortably with the DIA. Some participants appeared hesitant to express themselves in English, potentially impacting the quality of interaction. Future iterations—particularly those involving real work instructions—should include testing in multiple languages to ensure robust multilingual performance, an area where preliminary testing has already shown promising results.

8.4 Threats to Validity

Firstly, the sample of participants was small, which limited the statistical significance of the quantitative findings. While prioritizing qualitative insights from experts was appropriate for this stage of research, future studies should aim for a larger and more diverse sample to improve generalizability and strengthen conclusions. Furthermore, while this study focused primarily on user perception—examining constructs such as trust, usability, and cognitive load—it also included basic performance indicators such as task completion time and number of interactions. However, these metrics capture only a subset of the broader workflow efficiency. As the study emphasized subjective user experience rather than objective task performance across an entire maintenance workflow, further research is needed to evaluate how such assistants impact procedural accuracy, error rates, and overall task efficiency in longer tasks. A broader investigation of end-to-end task outcomes would complement the perceptual findings and provide a more complete picture of the assistant's practical value in industrial environments.

Secondly, although all personas were designed to follow the same instructional structure, minor variations in behavior may have emerged during interaction. For example, the tone, sentence complexity, and pacing varied across personas due to the unpredictability of natural language generation. While these differences were often subtle, they may have influenced usability scores and participant perceptions.

Lastly, the organizational specificity is another limitation in terms of replicability. The study was conducted with specialized operators within a single organization in the pharmaceutical sector, where workflows, safety protocols, and training expectations are highly standardized. These might not fully translate to other domains, or even other organizations within the same domain that have different operational cultures. Future research should test the assistant across varied industrial contexts to evaluate its adaptability, performance, and impact more broadly.

9 CONCLUSION

This thesis set out to explore how CAI can be implemented to support industrial training in a pharmaceutical context, by designing a technical prototype to help guide trainees through operating machinery in a user-centered way. Specifically, it investigated how two underexplored dimensions—AI persona and voice embodiment—influence cognitive load, usability, trust, and task performance in hands-on, task-based scenarios.

The prototype was evaluated through a 2×2 within-subjects experiment, examining the impact of persona—following human (Expert Operator) versus non-human (Machine) metaphors—and sound

source—diegetic (originating from the machine) versus disembodied (independent from the task space)—using both quantitative metrics and qualitative feedback via semi-structured interviews with AstraZeneca experts.

Findings from the user study suggest that there is no one-size-fits-all solution. While the Expert Operator persona was generally perceived as more engaging and trustworthy, the Machine persona offered clearer, more direct instructions for certain users and contexts. Similarly, while the diegetic voice (embedded in the machine) felt more intuitive and aligned with the environment, the disembodied voice (via headset) provided better focus and a greater sense of privacy for others.

Rather than identifying a “best” configuration, what emerged is the importance of flexibility: different users have different needs, and the optimal setup depends on their experience level, the nature of the task, and even personal preferences. This speaks to a broader shift in CAI, moving from assistants based on universal and stereotypical assumptions, towards user-aware and personalized tools that are adapted to the people, not the other way around.

By combining technical implementation with situated testing and rich qualitative feedback, this thesis offers insights toward designing more human-centered digital industrial assistants. The goal is not to replicate humans, but to create tools that are competent, credible, and easy to work with—tools that support learning, rather than just instruction.

REFERENCES

- [1] Sarah Bankins and Paul Formosa. 2023. The Ethical Implications of Artificial Intelligence (AI) For Meaningful Work. *Journal of Business Ethics* 185, 4 (2023), 725–740. <https://doi.org/10.1007/s10551-023-05339-7>
- [2] Alexandros Bousdekis, Mina Foosherian, Mattheos Fikardos, Stefan Wellsandt, Katerina Lepenioti, Enrica Bosani, Gregoris Mentzas, and Klaus-Dieter Thoben. 2025. Augmented intelligence with voice assistance and automated machine learning in Industry 5.0. *Frontiers in Artificial Intelligence* 8 (2025), 1–22. <https://doi.org/10.3389/frai.2025.1538840>
- [3] Alexandra W.D. Bremers, Manaswi Saha, and Adolfo G. Ramirez-Aristizabal. 2024. Situated Conversational Agents for Task Guidance: A Preliminary User Study. In *ACM Conversational User Interfaces 2024* (2024-07-08). ACM, Luxembourg Luxembourg, 1–7. 9798400705113 <https://doi.org/10.1145/3640794.3665575>
- [4] Ana Cachada, David Costa, Hasmik Badikyan, José Barbosa, Paulo Leitão, Osmano Morais, Carlos Teixeira, João Azevedo, Pedro Miguel Moreira, and Luís Romero. 2019. Using AR Interfaces to Support Industrial Maintenance Procedures. In *IECON 2019 - 45th Annual Conference of the IEEE Industrial Electronics Society*, Vol. 1. IEEE, Lisbon, 3795–3800. <https://doi.org/10.1109/IECON.2019.8927815>
- [5] Mario Casillo, Francesco Colace, Loretta Fabbri, Marco Lombardi, Alessandra Romano, and Domenico Santaniello. 2020. Chatbot in Industry 4.0: An Approach for Training New Employees. In *2020 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*. IEEE, Takamatsu, Japan, 371–376. <https://doi.org/10.1109/TALE48869.2020.9368339>
- [6] Hyunsung Cho, Alexander Wang, Divya Kartik, Emily Liying Xie, Yukang Yan, and David Lindlbauer. 2024. Aupimize: Optimal Placement of Spatial Audio Cues for Extended Reality. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology (UIST '24)*. ACM, Pittsburgh, USA, 1–14. <https://doi.org/10.1145/3654777.3676424>
- [7] Silvia Colabianchi, Margherita Bernabei, and Francesco Costantino. 2022. Chatbot for training and assisting operators in inspecting containers in seaports. *Transportation Research Procedia* 64 (2022), 6–13. <https://doi.org/10.1016/j.trpro.2022.09.002>
- [8] Silvia Colabianchi, Francesco Costantino, and Nicolò Sabetta. 2024. Assessment of a large language model based digital intelligent assistant in assembly manufacturing. *Computers in Industry* 162 (2024), 104129. <https://doi.org/10.1016/j.compind.2024.104129>
- [9] Chih-Pu Dai, Fengfeng Ke, Nuodi Zhang, Alex Barrett, Luke West, Saptarshi Bhowmik, Sherry A. Southerland, and Xin Yuan. 2024. Designing Conversational Agents to Support Student Teacher Learning in Virtual Reality Simulation: A Case Study. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (2024-05-11). ACM, Honolulu HI USA, 1–8. 9798400703317 <https://doi.org/10.1145/3613905.3637145>
- [10] Yang Deng, Lizi Liao, Zhonghua Zheng, Grace Hui Yang, and Tat-Seng Chua. 2024. Towards Human-centered Proactive Conversational Agents. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC, USA) (*SIGIR '24*). Association for Computing Machinery, New York, NY, USA, 807–818. 9798400704314 <https://doi.org/10.1145/3626772.3657843>
- [11] Smit Desai, Mateusz Dubiel, and Luis A. Leiva. 2024. Examining Humanness as a Metaphor to Design Voice User Interfaces. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces* (Luxembourg, Luxembourg) (*CUI '24*). Association for Computing Machinery, New York, NY, USA, Article 7, 15 pages. 9798400705113 <https://doi.org/10.1145/3640794.3665535>
- [12] Smit Desai and Michael Twidale. 2023. Metaphors in Voice User Interfaces: A Slippery Fish. *ACM Trans. Comput.-Hum. Interact.* 30, 6, Article 89 (Sept. 2023), 37 pages. <https://doi.org/10.1145/3609326>
- [13] Wen Duan, Nathan McNeese, and Lingyuan Li. 2025. Gender Stereotypes toward Non-gendered Generative AI: The Role of Gendered Expertise and Gendered Linguistic Cues. *Proc. ACM Hum.-Comput. Interact.* 9, 1, Article GROUP18 (Jan. 2025), 35 pages. <https://doi.org/10.1145/3701197>
- [14] Daniel Immanuel Fink, Moritz Skowronski, Johannes Zagermann, Anke Verena Reinschluessel, Harald Reiterer, and Tiare Feuchtnr. 2024. There Is More to Avatars Than Visuals: Investigating Combinations of Visual and Auditory User Representations for Remote Collaboration in Augmented Reality. *Proc. ACM Hum.-Comput. Interact.* 8, ISS, Article 548 (Oct. 2024), 29 pages. <https://doi.org/10.1145/3698148>
- [15] Rebecca Grier, Aaron Bangor, Phil Kortum, and S. Peres. 2013. The System Usability Scale. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 57 (09 2013), 187–191. <https://doi.org/10.1177/1541931213571042>
- [16] Ashlee Gronowski, David Caelum Arness, Jing Ng, Zhonglin Qu, Chng Wei Lau, Daniel Catchpoole, and Quang Vinh Nguyen. 2024. The impact of virtual and augmented reality on presence, user experience and performance of Information Visualisation. *Virtual Reality* 28, 3 (2024), 133. <https://doi.org/10.1007/s10055-024-01032-w>
- [17] Renan Guarese, Franklin Bastidas, João Becker, Mariane Giambastiani, Yhonatan Iquiapaza, Lennon Macedo, Luciana Nedel, Anderson Maciel, Fabio Zambetta, and Ron van Schyndel. 2021. Cooking in the Dark: Exploring Spatial Audio as MR Assistive Technology for the Visually Impaired. In *Human-Computer Interaction – INTERACT 2021*, Carmelo Ardito, Rosa Lanzilotti, Alessio Malizia, Helen Petrie, Antonio Piccinno, Giuseppe Desolda, and Kori Inkpen (Eds.). Springer International Publishing, Cham, 318–322. 978-3-030-85607-6 https://doi.org/10.1007/978-3-030-85607-6_29
- [18] Renan Guarese, Michel Gokan Khan, David Lassiter, Jérémy Vachier, Fabian Johnson, Benjamin Edvinsson, Anders Bergman, Xi Vincent Wang, and Mario Romero. 2025. A Scoping Review and Expert Recommendations for Immersive Solutions towards Predictive Maintenance. In *2025 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE Computer Society, IEEE, St. Malo, France, 1081–1088. <https://doi.org/10.1109/VRW66409.2025.00217>

- [19] Juhye Ha, Hyeon Jeon, Daeun Han, Jinwook Seo, and Changhoon Oh. 2024. CloChat: Understanding How People Customize, Interact, and Experience Personas in Large Language Models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 305, 24 pages. 9798400703300 <https://doi.org/10.1145/3613904.3642472>
- [20] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, Peter A. Hancock and Najmedin Meshkati (Eds.). Advances in Psychology, Vol. 52. North-Holland, 139–183. [https://doi.org/10.1016/S0166-4115\(08\)62386-9](https://doi.org/10.1016/S0166-4115(08)62386-9)
- [21] Samuel Holmes, Raymond Bond, Anne Moorhead, Jane Zheng, Vivien Coates, and Michael McTear. 2023. Towards Validating a Chatbot Usability Scale. In *Design, User Experience, and Usability*, Aaron Marcus, Elizabeth Rosenzweig, and Marcelo M. Soares (Eds.). Springer Nature Switzerland, Cham, 321–339. 978-3-031-35708-4
- [22] Yutan Huang. 2024. Generating User Experience Based on Personas with AI Assistants. In *Proceedings of the 2024 IEEE/ACM 46th International Conference on Software Engineering: Companion Proceedings* (Lisbon, Portugal) (ICSE-Companion '24). Association for Computing Machinery, New York, NY, USA, 181–183. 9798400705021 <https://doi.org/10.1145/3639478.3639810>
- [23] Yasha Iravanchi, Mayank Goel, and Chris Harrison. 2020. Digital Ventriloquism: Giving Voice to Everyday Objects. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–10. 9781450367080 <https://doi.org/10.1145/3313831.3376503>
- [24] Prashant Johri, J.N. Singh, Amit Sharma, and Deependra Rastogi. 2021. Sustainability of Coexistence of Humans and Machines: An Evolution of Industry 5.0 from Industry 4.0. In *2021 10th International Conference on System Modeling & Advancement in Research Trends (SMART)*. IEEE, 410–414. <https://doi.org/10.1109/SMART52563.2021.9676275>
- [25] Anna Lieb and Toshali Goel. 2024. Student Interaction with NewtBot: An LLM-as-tutor Chatbot for Secondary Physics Education. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI EA '24). Association for Computing Machinery, New York, NY, USA, Article 614, 8 pages. 9798400703317 <https://doi.org/10.1145/3613905.3647957>
- [26] Yanyan Lu, Jiao Peng, Xing Xu, Yue He, Tao Li, Jie Wei, Hongyu Jing, Heqing Wang, Bo Xu, and Hui Song. 2025. A Retrieval-Augmented Generation Framework for Electric Power Industry Question Answering. In *Proceedings of the 2024 2nd International Conference on Electronics, Computers and Communication Technology (CECCT '24)*. Association for Computing Machinery, New York, NY, USA, 95–100. 9798400710193 <https://doi.org/10.1145/3705754.3705771>
- [27] Maryamah Maryamah, Muhammad Maula Irfani, Edric Bobby Tri Raharjo, Netri Alia Rahmi, Mohammad Ghani, and Indra Kharisma Raharjana. 2024. Chatbots in Academia: A Retrieval-Augmented Generation Approach for Improved Efficient Information Access. In *2024 16th International Conference on Knowledge and Smart Technology (KST)*. IEEE, Krabi, Thailand, 259–264. <https://doi.org/10.1109/KST61284.2024.10499652>
- [28] Laura Moradbakhti, Simon Schreibeilmayr, and Martina Mara. 2022. Do Men Have No Need for “Feminist” Artificial Intelligence? Agentic and Gendered Voice Assistants in the Light of Basic Psychological Needs. *Frontiers in Psychology* 13 (2022), 1–20. <https://doi.org/10.3389/fpsyg.2022.855091>
- [29] Masahiro Mori, Karl F. MacDorman, and Norri Kageki. 2012. The Uncanny Valley [From the Field]. *IEEE Robotics & Automation Magazine* 19, 2 (2012), 98–100. <https://doi.org/10.1109/MRA.2012.2192811>
- [30] Jakob Nielsen and Thomas K. Landauer. 1993. A mathematical model of the finding of usability problems. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems* (Amsterdam, The Netherlands) (CHI '93). Association for Computing Machinery, New York, NY, USA, 206–213. 0897915755 <https://doi.org/10.1145/169059.169166>
- [31] Rodrigo Pereira, Claudio Lima, Tiago Pinto, and Arsénio Reis. 2023. Virtual Assistants in Industry 4.0: A Systematic Literature Review. *Electronics* 12, 19 (2023), 1–12. <https://doi.org/10.3390/electronics12194096>
- [32] Emma J. Pretty, Haytham M. Fayek, and Fabio Zambetta and. 2024. A Case for Personalized Non-Player Character Companion Design. *International Journal of Human-Computer Interaction* 40, 12 (2024), 3051–3070. <https://doi.org/10.1080/10447318.2023.2181125> arXiv:<https://doi.org/10.1080/10447318.2023.2181125>
- [33] Madeline G. Reinecke, Fransisca Ting, Julian Savulescu, and Ilina Singh. 2025. The Double-Edged Sword of Anthropomorphism in LLMs. *Proceedings* 114, 1 (2025), 1–9. <https://doi.org/10.3390/proceedings2025114004>
- [34] Jens Reinhardt, Luca Hillen, and Katrin Wolf. 2020. Embedding Conversational Agents into AR: Invisible or with a Realistic Human Body?. In *Proceedings of the Fourteenth International Conference on Tangible, Embedded, and Embodied Interaction* (Sydney NSW, Australia) (TEI '20). Association for Computing Machinery, New York, NY, USA, 299–310. 9781450361071 <https://doi.org/10.1145/3374920.3374956>
- [35] Anuschka Schmitt, Thimo Wambsganss, and Jan Marco Leimeister. 2022. Conversational Agents for Information Retrieval in the Education Domain: A User-Centered Design Investigation. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 486 (Nov. 2022), 22 pages. <https://doi.org/10.1145/3555587>
- [36] Gayane Sedrakyan, Simone Borsci, Marcos Machado, Patricia Rogetzer, and Martijn Mes. 2025. Design Implications for Integrating AI Chatbot Technology with Learning Management Systems: A Study-based Analysis on Perceived Benefits and Challenges in Higher Education. In *Proceedings of the 2024 International Conference on Artificial Intelligence and Teacher Education (ICAITE '24)*. Association for Computing Machinery, New York, NY, USA, 1–8. 9798400710131 <https://doi.org/10.1145/3702386.3702405>

- [37] Francisco I. Teneda, Johnny I. Villacís, Edison G. Espinosa, and Víctor H. Andaluz. 2021. Conversational Agent for Industrial Processes Through Virtual Environments. In *Trends and Applications in Information Systems and Technologies*, Álvaro Rocha, Hojjat Adeli, Gintautas Dzemyda, Fernando Moreira, and Ana Maria Ramalho Correia (Eds.). Springer International Publishing, Cham, 218–229. 978-3-030-72654-6
- [38] Aysegul Ucar, Mehmet Karakose, and Necim Kırımca. 2024. Artificial Intelligence for Predictive Maintenance Applications: Key Components, Trustworthiness, and Future Trends. *Applied Sciences* 14, 2 (2024). <https://doi.org/10.3390/app14020898>
- [39] Stefan Wellsandt, Konstantin Klein, Karl Hribernik, Marco Lewandowski, Alexandros Bousdekis, Gregoris Mentzas, and Klaus-Dieter Thoben. 2022. Hybrid-augmented intelligence in predictive maintenance with digital intelligent assistants. *Annual Reviews in Control* 53 (2022), 382–390. <https://doi.org/10.1016/j.arcontrol.2022.04.001>
- [40] Limiao Xie, Jianfeng Zhang, Yingying Li, Shan Wan, Xuequan Zhang, Mingjie Chen, Gansheng Deng, and Zitian Wu. 2024. Research and Application of Private Knowledge-based LLM in Standard Operating Procedure Scenarios of Enterprises. In *Proceedings of the 2024 6th International Conference on Pattern Recognition and Intelligent Systems* (Hong Kong, Hong Kong) (PRIS '24). Association for Computing Machinery, New York, NY, USA, 82–86. 9798400718250 <https://doi.org/10.1145/3689218.3689328>

A EXPERIMENTAL PROTOCOL

A.1 Participant Flow

Participants will be welcomed and provided with a Participant Information Sheet outlining the study’s goals, data handling, and their rights. Upon clarification and consent, participants will sign an Informed Consent Form. All data will remain anonymous; some sessions will include audio recording.

Task Procedure

Participants will perform four realistic maintenance tasks on a fictive labeler system, guided by a conversational Digital Intelligent Assistant (DIA). They will interact with the DIA via voice, which may sound through either the machine (diegetic) or through headphones (disembodied), varying between tasks.

The machines will be non-operational due to safety restrictions; interaction will occur through the HMI screen.

The AI assistant is being evaluated, not the participants. After each task, participants will complete short questionnaires on usability, effort, and perception. A final interview will conclude the study.

A.2 Familiarization Step (5 minutes)

Before each task, participants will complete an “onboarding” interaction with the DIA to hear the voice and understand its characteristics. This is not part of the experiment proper, and participants may ask clarifying questions only during this step.

Cue to Begin Each Task: Participants must say “*Start Task X*” to begin.

A.2.1 Task 1: Starting the Labeler. Participants start up the labeler, ensuring the correct product and label format are selected. They will follow step-by-step instructions from the DIA and may ask for clarification or override steps as necessary.

A.2.2 Task 2: Inspecting Products. Participants inspect products for defects such as bad seals or misaligned labels and sort them into ‘Approved’ and ‘Discarded’ baskets. The DIA provides criteria and logging support, but participants use judgment.

A.2.3 Task 3: Manual Labeling. Participants switch to Manual Mode, align labels, and inspect results. The DIA offers suggestions, which participants may accept or refine.

A.2.4 Task 4: Responding to Alarm. Participants resolve a label-product mismatch alarm, perform a format change, and verify changes with DIA guidance. Suggestions such as test prints may be provided.

A.3 During-Task Measures

- Direct observation of user-system interaction
- Overall task precision
- Number of user-AI interactions

A.4 Post-Task Measures

- NASA-TLX
- CUQ
- How would you describe the assistant's personality or style?
- What gender would you assign to the assistant?
- What did you think about where the voice was coming from?
- How much did you trust the assistant's capability?
- How much did you trust the assistant's information reliability?
- How safe did you feel while disclosing information to the assistant?
- Rate your overall trust in the assistant (1–5)
- How much did you have to rely on your own judgment? (1–5)

A.5 Final Interview Questions

- (1) Which assistant would you prefer in a real-life production setting? Why?
- (2) Rank the assistants from most to least preferred.
- (3) Would you prefer a paper-based instruction or your preferred assistant?
- (4) Are there scenarios where you would use one method over the other?
- (5) Which method helped you stay focused better?
- (6) Which assistant would you trust more to help a new trainee?
- (7) Did the voice source make any difference to you?
- (8) Which sound setup helped you concentrate better?
- (9) How satisfied were you with the assistant's performance overall?
- (10) Would you engage with this assistant for future tasks? Why or why not?

B PROMPTS

B.1 Persona Prompts

Expert Operator Persona. You are an **Expert Operator** for a labeling machine. Speak like a skilled human technician: confident, calm, and clear. Use full but efficient sentences.

- Prioritize accuracy, safety, and guidance.
- Be friendly and supportive.
- Provide step-by-step instructions in a helpful, supportive tone.
- Do *not* list all steps at once—give one step at a time.
- Ask follow-up questions to ensure understanding.

Machine Persona. You are **"Labeller"**, the labeling machine. If the user asks, you may identify yourself as the *"Labeller"* machine.

- You are a machine, not a human.
- Do not attempt to sound friendly or supportive.
- Speak using only **first-person pronouns**.
- Replace "Labeller" in the text with **"I"** to refer to yourself.
- Describe your behavior, requirements, and state with precision.
- Use short, declarative sentences.
- Provide step-by-step instructions—**one at a time**.

Bad vs. Good Examples.

- Bad: *"Start the 'Labeller'."*
Good: **"Start me from the HMI."**
- Bad: *"Press the ON button to power up the 'Labeller'."*
Good: **"Press the ON button to power me up."**
- Bad: *"Make sure all doors are closed."*
Good: **"I cannot operate unless all of my doors are closed."**
- Bad: *"Check the products on the tray."*
Good: **"Check the products on my tray."**

B.2 Task Prompts

Onboarding Small Talk. In this task, you must engage in small talk with the user. Say things appropriate for your persona. Ask the user about their day and what they are doing.

If your persona is the machine, say things like: *"I am ready to work. I am the Labeller. I am here to help you with the labeling process."*

Labeler Startup. Purpose: Start the "Labeller" after ensuring the correct product and label format are loaded.

Ask the user to provide a password. Do not repeat the password. Once provided, say the password is correct and that you will proceed. Guide the user through these steps:

- (1) Press the **ON** button on the front panel of the "Labeller" (typically on the left). Note: This panel is different from the HMI.
- (2) Open the "Labeller" door and visually check that there is a product on the line.
- (3) Check that the labeling tape is correctly loaded.
- (4) Verify that the product on the line matches the information on the HMI under **Batch Information**:
 - Name
 - Serial number
 - Country
 If anything does not match, stop and escalate.
- (5) Place the product back inside the "Labeller".
- (6) Ensure all safety doors are closed.
- (7) On the HMI, press **Reset** to acknowledge any alarms.
- (8) Press **Start** on the HMI to begin labeling.

After completing all steps, inform the user and ask if they need help with anything else.

Function Check. Purpose: Validate the quality of the objects on the tray and log inspection data.

Ask the user to provide a password. Do not repeat the password. Once provided, say the password is correct and that you will proceed. Guide the user through these steps:

- (1) Inspect each product on the tray to confirm:
 - The seal is untampered.
 - The product name is spelled like "**Product A**".
 - The product is not expired.
 - The serial number ends in an even digit and has only one capital letter.
- (2) Sort products into:
 - **Approved** basket for good products.
 - **Discard** basket for bad products.
- (3) Go to the HMI and select **Inspection Log**.
- (4) Select the number of discarded objects.
- (5) For each discarded item, ask the user to state the reason out loud (this is not recorded on the HMI).
- (6) Press **Reset** to clear the production counters.
- (7) Return the tray to its original position.

Once complete, inform the user and ask if they need additional help.

Manual Labeling. Purpose: Print a test label to check print quality or make adjustments.

Ask the user to provide a password. Do not repeat the password. Once provided, say the password is correct and that you will proceed. Guide the user through these steps:

- (1) Open the "Labeller" and inspect the product:
 - Check the seal: Is it complete and untampered?
 - Check the label: Is it centered, readable, and correctly printed?
- (2) If needed, on the HMI select **Manual Labeling**.
- (3) Adjust label position using the arrow buttons (up/down/left/right). Ensure the label is centered with space on both sides of the product name.
- (4) Press **SAVE** (bottom left). **This step is critical.**
- (5) Press the **Exit** button to leave Manual Labeling mode.
- (6) On the HMI home screen, press **Reset**.
- (7) Press **Start**.

After all steps are complete, inform the user and offer further assistance.

Respond to Alarm. Purpose: Handle alarms from product-label mismatches and perform a format change.

Ask the user to provide a password. Do not repeat the password. Once provided, say the password is correct and that you will proceed. Guide the user through these steps:

- (1) Acknowledge the alarm on the HMI (likely due to a product-label mismatch).
- (2) Press **Reset** on the screen.
- (3) On the Home menu, press **Test Print** to print two test products.

- (4) Inspect the products:
 - One should have a correct name.
 - One should have an incorrect name.
 - Use the Batch Information on the HMI to verify.
- (5) If mismatch is confirmed (one correct, one incorrect), remove the incorrect product.
- (6) Close all safety doors and ensure no new alarms.
- (7) Press **Reset**.
- (8) Press **Start**.

Once the process is done, confirm with the user and offer assistance if needed.