Computers in Industry 162 (2024) 104129

Contents lists available at ScienceDirect

Computers in Industry

journal homepage: www.sciencedirect.com/journal/computers-in-industry

Assessment of a large language model based digital intelligent assistant in assembly manufacturing

Silvia Colabianchi, Francesco Costantino, Nicolò Sabetta [*]

Department of computer, control, and management engineering Antonio Ruberti, Sapienza University of Rome, Via Ariosto 25, Rome 00185, Italy

ARTICLE INFO

Keywords:
Chatbot
Experimental design
Artificial intelligence
Natural language processing
Industry 4.0
Industry 5.0

ABSTRACT

The use of Digital Intelligent Assistants (DIAs) in manufacturing aims to enhance performance and reduce cognitive workload. By leveraging the advanced capabilities of Large Language Models (LLMs), the research aims to understand the impact of DIAs on assembly processes, emphasizing human-centric design and operational efficiency. The study is novel in considering the three primary objectives: evaluating the technical robustness of DIAs, assessing their effect on operators' cognitive workload and user experience, and determining the overall performance improvement of the assembly process. Methodologically, the research employs a laboratory experiment, incorporating a controlled setting to meticulously assess the DIA's performance. The experiment used a between-subjects design comparing a group of participants using the DIA against a control group relying on traditional manual methods across a series of assembly tasks. Findings reveal a significant enhancement in the operators' experience, a reduction in cognitive load, and an improvement in the quality of process outputs when the DIA is employed. The article contributes to the study of the DIA's potential and AI integration in manufacturing, offering insights into the design, development, and evaluation of DIAs in industrial settings.

1. Introduction

The impact of Industry 4.0 on the manufacturing sector is widely acknowledged for its seamless integration of various technologies. Nevertheless, excessive reliance on technological solutions and the underwhelming outcomes of some Industry 4.0 initiatives have led to a marginalization of the human element in manufacturing processes (Neumann et al., 2021). This change has prompted a reassessment of the role of human factors in technological deployments, which is now often referred to as Industry 5.0. This builds upon the aforementioned foundation as a paradigm shift that emphasizes the role of research and innovation in steering toward a sustainable, human-centric, and resilient industry (Lu et al., 2022; Ordieres-Meré et al., 2023). Similarly, today we refer to the transition from Operator 4.0 to Operator 5.0. The widespread use of Industry 4.0 technologies has generated significant sensor data, which requires constant attention and increases cognitive burden, especially in the context of agile networked manufacturing lines (Lall et al., 2017). In this context, the Operator 4.0 concept highlights the vision of the importance of trust and connection between humans and technological systems. Yet, the emphasis shifts towards prioritizing the social aspects of technology integration, compelling a revaluation of the adoption strategies to ensure they are aligned with human wellbeing (Gladysz et al., 2023). Operating agile manufacturing lines requires extensive knowledge and cognitive resources, even for experienced operators (Freire et al., 2023). Workers may be responsible for complex operations, maintenance, setups, etc. To address these activities, it is essential to implement solutions that reduce cognitive demands and provide timely feedback (Colabianchi et al., 2023). In this context, voice-enabled Digital Intelligent Assistants (DIAs) are designed to enhance human capabilities by reducing cognitive load and aligning tasks, thereby elevating the operator's role to that of a decision-maker rather than a mere task performer (Colabianchi et al., 2023). Unlike basic chatbots, DIAs are equipped with advanced functionalities that enable them to perform "intelligent" actions. These actions are designed to streamline operations that operators might find challenging, thereby enhancing efficiency and effectiveness in task execution (Wellsandt et al., 2021). His capability reflects a significant evolution from traditional chatbots or other industrial systems to more sophisticated AI-driven interfaces that can adaptively assist human operators in complex environments (Le and Wartschinski, 2018; Kernan Freire et al., 2023). DIAs, which are enabled by a range of AI functionalities, including cutting-edge Large Language Models (LLMs), are transforming

[*] Corresponding author.
E-mail addresses: silvia.colabianchi@uniroma1.it (S. Colabianchi), francesco.costantino@uniroma1.it (F. Costantino), nicolo.sabetta@uniroma1.it (N. Sabetta).

https://doi.org/10.1016/j.compind.2024.104129
Received 13 March 2024; Received in revised form 23 May 2024; Accepted 14 July 2024
Available online 31 July 2024

human-machine interactions by enabling voice-based engagement, information generation, and action execution (Colabianchi et al., 2023; Bernabei et al., 2023). The reasons outlined above are driving the increasing prevalence of DIAs in the industrial sector. Despite limited research on industrial DIAs, their contributions are significant. DIAs support production operations, training (Kernan Freire et al., 2023; Colabianchi et al., 2022), and process analysis, enhancing flexibility in automated manufacturing systems (Colabianchi et al., 2023; Kernan Freire et al., 2023; Trappey et al., 2022). Several studies show how they facilitate rapid data analysis, improve decision-making, and reduce costs and downtime while promoting knowledge transfer among workers (Kernan Freire et al., 2023; Trappey et al., 2022; Li et al., 2023; Melluso et al., 2022). DIAs are particularly useful in training for complex tasks and guiding operators through processes (Chen et al., 2021). Their hands-free and eyes-free capabilities allow workers to access information and perform tasks without losing focus on their primary activities (Ludwig et al., 2023). Additionally, DIAs simplify tasks that cause cognitive overload, such as complex assembly processes characterized by high variability of components and a significant risk of errors (Hoedt et al., 2017), by providing real-time guidance and reducing the cognitive burden (Le and Wartschinski, 2018; Carvalho et al., 2020). They offer a centralized platform for accessing information through various user-friendly interfaces, including voice interactions (Longo and Padovano, 2020; Wellsandt et al., 2023). Recently, Large Language Models, such as GPTs, have been integrated into DIAs, enhancing information retrieval, task automation, and usability (Xia et al., 2024), and supporting production control, technical design (Fan et al., 2024; Wang et al., 2023), and planning in flexible systems (Xia et al., 2023).

However, there are still several open questions in the literature. Specifically, further research is required to evaluate DIA performance from multiple perspectives, particularly in real-life conversations that differ greatly from structured, written text documents (Xia et al., 2024; Dinan et al., 2022). The variability of language presents challenges such as robustness to real-world use cases, noise perturbations, and potential adversarial attacks (Church and Yue, 2023). Additionally, a review of case studies in the manufacturing sector indicates that rule-based DIAs are predominantly utilized (Colabianchi et al., 2023) and the findings suggest a noticeable reluctance to incorporate more human-like qualities into these systems, with DIAs often lacking empathy and only facilitating brief interactions. serves to illustrate the inherent limitations of rule-based architectures, particularly concerning the characteristics of emotional intelligence and sustained engagement, which current LLMs are better equipped to address. Moreover, to the best of the authors' knowledge, no studies within the existing literature investigate the application of LLM-based DIAs specifically for an assembly process. Moreover, further studies are necessary to evaluate user experience, including usability, cognitive load, and overall process benefits within the industrial context. In this scenario, this research article aims to investigate whether integrating an LLM-based voice-enabled DIA can enhance the performance and overall experience of complex, highly cognitive load, and alienating assembly tasks compared to traditional assembly methods.

The remainder of the paper is organized as follows. Section 2 presents the objectives and the novelty of this research while Section 3 introduces the methods followed for the experimental design. Section 4 describes the experiment addressing the system's architecture, and functionalities. Section 5 presents the results and Section 6 discusses the findings. Finally, Section 7 concludes and outlines the follow-up research.

2. Objectives and novelty

The objective of this study is to assess the applicability of an LLM-based DIA within a manufacturing setting, utilizing a multi-dimensional evaluation criteria. The initial analysis will focus on the technical robustness of the DIA, examining the precision and reliability of its responses to user queries and the accuracy of the speech recognition system in the context of user variability. This preliminary assessment is designed to ascertain the technical viability of the DIA for future applications. Subsequently, the study will investigate the impact of the DIA on operators. This will include an examination of potential benefits, such as the reduction of cognitive load, and an assessment of the usability and user experience of the DIA. These aspects are of critical importance in determining whether, despite its technical capabilities, the DIA can be readily accepted by operators and thus facilitate effective human-machine collaboration. Finally, the study will analyze the main performance indicators of the production process to identify potential improvements, such as reductions in process times and errors, which could lead to enhanced quality. Significant enhancements in these areas could serve as compelling incentives for the adoption of this technology in actual industrial environments.

To address these overarching objectives, the research poses the following specific research questions:

- **RQ1: Technical robustness.** How does the technical robustness of a DIA affect its deployment in the manufacturing assembly processes?
- **RQ 2.1: Cognitive workload.** What is the impact of the DIA on the cognitive workload of the operator during assembly tasks?
- **RQ 2.2: System usability and experience.** How does the usability of the DIA influence the overall experience of the operator in assembly processes?
- **RQ 3: Assembly process performance.** What performance benefits does the introduction of a DIA offer in the assembly process?

Thus, the problem's hypotheses can be defined through RQs, which transform them from interrogative to assertive terms. For instance, RQ3 "What performance benefits does the introduction of a DIA offer in the assembly process?" corresponds to a hypothesis to be tested "The DIA improves process performance". The formulation of a testable hypothesis helps in identifying independent, dependent, and control variables. In detail, the hypotheses from RQs are the following in Table 1. In the continuation of the paper, RQs are used for clarity of exposition.

3. Methods

In this section, we describe the methods used for the experimental design to answer the RQs. The first activity identified the specific type of experiment, referring to the determinants of the experiment (Sørensen et al., 2010). The DIAs are the studied innovation, the focus is the investigation of the effects of DIA, the level of complexity considers micro-behaviours (e.g. the accuracy of every single answer of the DIA)

Table 1
Hypothesis for research questions.

Research question	Hypothesis
RQ1. Technical robustness How does the technical robustness of a DIA affect its deployment in the manufacturing assembly processes?	H1.1 DIA accuracy satisfies the process request H1.2 DIA reliability satisfies the process request H1.3 DIA speech recognition accuracy satisfies the process request
RQ 2.1: Cognitive workload What is the impact of the DIA on the cognitive workload of the operator during assembly tasks?	H2.1 DIA decrease the cognitive workload
RQ 2.2: System usability and experience How does the usability of the DIA influence the overall experience of the operator in assembly processes?	H2.2 The operator evaluates the usability and the experience positively
RQ 3: Assembly process performance What performance benefits does the introduction of a DIA offer in the assembly process?	H3.1 DIA reduces the time of the assembly process H3.1 DIA increases the quality of the assembly process

and meso-behaviours (e.g. the general level of acceptance of the DIA). Thus, the control of the experiment, abstraction and sense-making was done by the authors (the researchers), the experiment type is a qualitative laboratory experiment, and the results are an abstraction of innovation effects. Fig. 1

The experiment is defined by traditional scientific steps of experimental design. Experimental design encompasses structuring and conducting experiments to answer questions efficiently and clearly.

The research methodology relies on the well-known tasks of the scientific method and the experiment design (Montgomery, 2017; Maxwell et al., 2017; Campbell and Stanley, 2015). These tasks define the following: Problem; Hypothesis; Independent variables; Dependent variables; Control variables; Sampling; Experiment procedures; Data Collection; Data analysis. The data analysis led to results and discussions.

4. Experiment

4.1. Problem and hypothesis

To verify the hypothesis defined in Table 1 the research team defined an application context and an assembly process. The application context relies on a real company that assembles cases for electromedical instruments. The cases are assembled at operating stations, with no automation, using basic tools and highly variable component configurations. The defined assembly process comprises a series of assembly tasks that are characterized by memory-intensive, monotonous, and repetitive operations. To ensure replicability, the team proposed a simplified version of the process to be tested in a laboratory. This adaptation was designed to carefully preserve the inherent challenges of the original process, particularly the significant variability in the components involved and the potential presence of defects. The study aims to authentically replicate industrial conditions within a controlled environment, allowing for consistent and reliable testing across different settings. This is achieved by maintaining the aforementioned complexities.

The objective of the assembly problem is to assemble a box and its compartments and place all components required by the order inside. Operators are assigned specific jobs (e.g. "Italy job", "USA job"), each with a unique assembly detail and elements to be introduced into the box. During the process, the operator will also be responsible for addressing any product defects. The assembly process consists of four sequential steps (as shown in the Fig. 2):

- Step A: Selection of the box;
- Step B: Assembly of the dividers;
- Step C: Insertion of the components;
- Step D: Label placement.

The initial step requires the operator to select the box linked to the job that requires completion. Each box is identified by an alphanumeric identifier. The identifier may be affected by a defect, such as an unreadable or incomplete identifier. In such cases, a specific process has been established to report the defect and resolve the issue. The second step is to assemble and place two dividers into the box, creating four compartments. The size of each compartment will depend on the operator's assigned task. Then the components associated with the job are placed. The components are as follows: screws, dome head screws, wall stops, and nails. Each job requires a different amount of each component. Once the assembly of the box is completed, the operator must place a label on the box. The labels are identified with an alphanumeric code. The assembly process ends when the label is placed. The entire process just described was extensively detailed in a manual including all job specifications and made available for the experiment.

4.2. Independent variables

The independent variables are the differentiating factors to observe effects. In the experiment, there is a single independent variable which is the use of the DIA by the operator. Thus, the experiment uses the

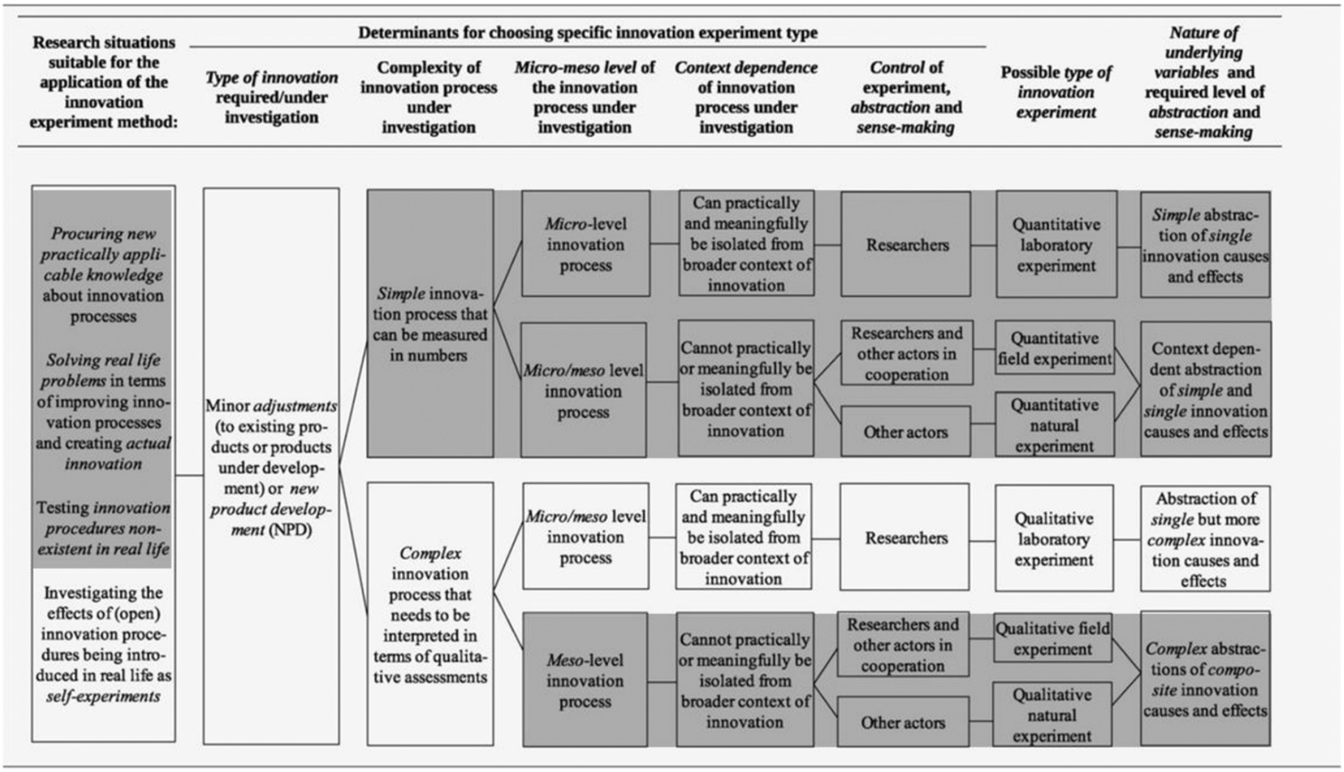
Research situations suitable for the application of the innovation experiment method:
Determinants for choosing specific innovation experiment type
Type of innovation required/under investigation
Complexity of innovation process under investigation
Micro-meso level of the innovation process under investigation
Context dependence of innovation process under investigation
Control of experiment, abstraction and sense-making
Possible type of innovation experiment
Nature of underlying variables and required level of abstraction and sense-making
Procuring new practically applicable knowledge about innovation processes
Solving real life problems in terms of improving innovation processes and creating actual innovation
Testing innovation procedures non-existent in real life
Investigating the effects of (open) innovation procedures being introduced in real life as self-experiments
Minor adjustments (to existing products or products under development) or new product development (NPD)
Simple innovation process that can be measured in numbers
Complex innovation process that needs to be interpreted in terms of qualitative assessments
Micro-level innovation process
Micro/meso level innovation process
Micro/meso level innovation process
Meso-level innovation process
Can practically and meaningfully be isolated from broader context of innovation
Cannot practically or meaningfully be isolated from broader context of innovation
Can practically and meaningfully be isolated from broader context of innovation
Cannot practically or meaningfully be isolated from broader context of innovation
Researchers
Researchers and other actors in cooperation
Other actors
Researchers
Researchers and other actors in cooperation
Other actors
Quantitative laboratory experiment
Quantitative field experiment
Quantitative natural experiment
Qualitative laboratory experiment
Qualitative field experiment
Qualitative natural experiment
Simple abstraction of single innovation causes and effects
Context dependent abstraction of simple and single innovation causes and effects
Abstraction of single but more complex innovation causes and effects
Complex abstractions of composite innovation causes and effects

Fig. 1. Selection of experiment type (applied from (Sørensen et al., 2010)).

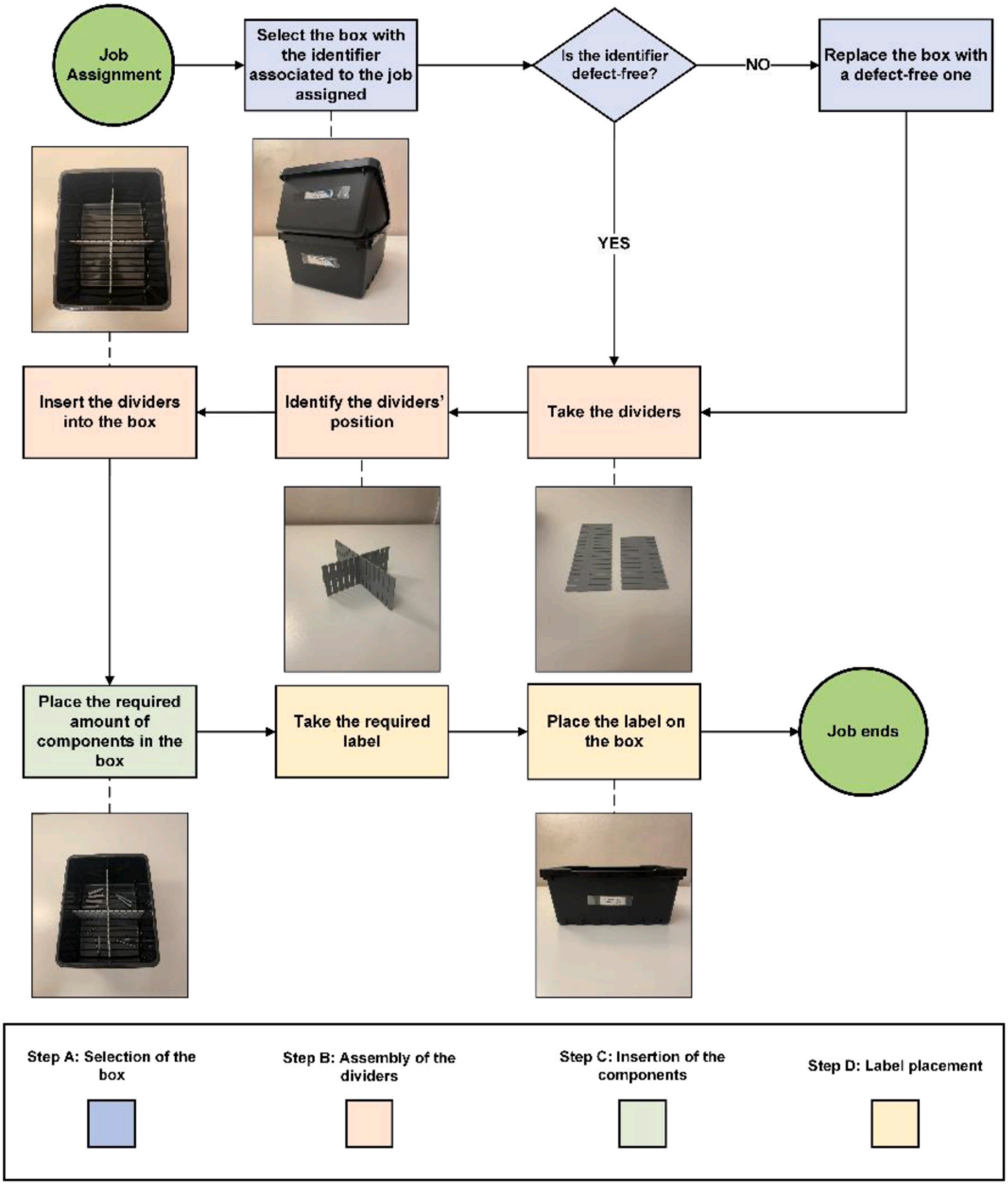
Job Assignment
Select the box with the identifier associated to the job assigned
Is the identifier defect-free?
NO
Replace the box with a defect-free one
YES
Insert the dividers into the box
Identify the dividers' position
Take the dividers
Place the required amount of components in the box
Take the required label
Place the label on the box
Job ends
Step A: Selection of the box
Step B: Assembly of the dividers
Step C: Insertion of the components
Step D: Label placement

Fig. 2. Process flow chart.

between-subjects design, where different groups are exposed to the changes in the independent variable. One group of people uses the DIA while another does not, allowing for a comparison of effects. To set the independent variable a DIA was developed. The DIA details are presented as it could be considered the independent variable. Fig. 3 shows the architecture of the DIA, which is composed of three modules:

- Speech to text system;
- Dialogue system;
- Text to speech system.

The DIA was developed using the top available technologies. The Dialogue System uses GPT-4 and the LangChian Framework. GPT-4 is OpenAI's LLM model (OpenAI, 2023a) most performative model for understanding and generating human-like text. LangChain is the most common framework (LangChain,) for building applications with large language models, enabling automated workflows and advanced natural language processing capabilities (OpenAI, 2023a). At the time of the experiment, Google Speech Recognition (Pypi., 2023) and OpenAI TTS (OpenAI, 2023b) were the most powerful technologies for converting spoken words to text and vice versa. The DIA development includes two analyses: defining the characteristics and defining the architecture.

4.3. DIA characteristics

The authors defined the DIA characteristics with an analysis of the functional specifications and application boundaries of the DIA for assembly, using the industrial conversational agent taxonomy

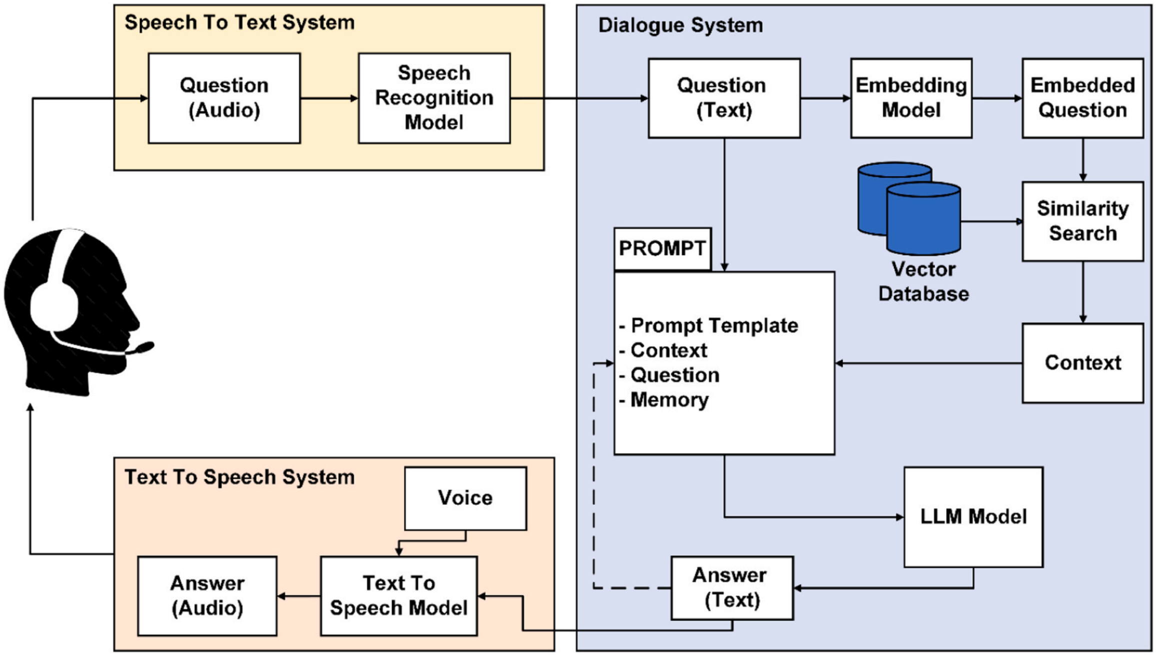

Fig. 3. Digital intelligent assistant architecture.

(Colabianchi et al., 2023). This taxonomy includes 18 dimensions organized from both the DIA and DIA-user perspectives.

The primary objective (D1) of the developed DIA is to provide user support in the assembly tasks of the box, specifically aimed at aiding operators during the five steps. This is achieved through a Specific Domain (D2) that gathers detailed process information, directly from the instruction manual. The Chatbot Intelligent Framework, identified in the third dimension (D3), is a hybrid system that combines characteristics of both AI-based and Retrieval systems, utilizing LLMs.

The DIA is designed for one user at a time and operates as an interpersonal chatbot (D7) without the ability to retain previous conversations.

The DIA proposed in this paper does not include integrated services (D4), additional human support (D5), gaming components (D6), socio-emotional behavior (D8), and interface personification (D9) as they are not strictly necessary for the purpose of the study. However, these characteristics may be implemented in a more advanced version of the system. Additionally, the DIA lacks a compelling front-end user interface (D10) as it is scripted in Python and integrated into the running device.

Regarding Chatbot-User Interaction, the DIA functions as a virtual assistant that assists operators in assembly tasks. The communication modality (D11) is exclusively voice-based, while interactions (D12) are conducted through free text, without buttons or graphical interfaces. The DIA is acknowledged as a multiturn chatbot (D13), which facilitates multiple interactions to elaborate on user queries. These interactions are typically of medium-long length, attributable to the DIA's capability to offer responses that extend beyond mere binary terms like "YES" or "NO." Instead, the DIA is equipped to provide comprehensive explanations of the assembly process. The conversation is jointly led by the chatbot and the user (D15). Initially, the DIA provides instructions to the operator, particularly the task that needs to be completed, and the operator then initiates further communication as necessary (D16). It should be noted that a single operator interacts with the chatbot (D17), serving as a facilitator in the assembly process (D18). Table 2 provides a concise summary of all the characteristics of the DIA.

Table 2
DIA characteristics.

Perspective	Design Dimension	Characteristics
DIA	D1 Primary goal	User Support
	D2 Knowledge domain	Specific Domain
	D3 Intelligence framework	Hybrid: AI-based + Retrieval
	D4 Integrated service	None
	D5 Additional human support	Not Present
	D6 Gamification	Not Present
	D7 Service provided	Interpersonal
	D8 Socio-emotional behaviour	Not Present
	D9 Interface personification	Not Present
	D10 Front-end user interface	App
DIA-User Interaction	D11 Communication modality	Only Voice
	D12 Interaction modality	Interactive
	D13 Length of conversation	Multi-turn
	D14 Duration single iteration	Medium - Long interaction
	D15 Leader of conversation	Mixed
	D16 Frequency of interaction	When required
	D17 Number of participants	Individual
	D18 Chatbot role	Facilitator

Table 3
Parameters of DIA.

Architecture's elements	Models used in the DIA
Speech Recognition Model	Google Speech Recognition
Embedding Model	Text_embedding_ada_002
Chain	Load_qa_chain + RetrievalQA
Retriever	FAISS
LLM model	GPT−4 Turbo
Memory type	Window Buffer Memory (Size = 2)
Text to Speech model	OpenAI STT ('Alloy' voice)

4.4. DIA architecture

The DIA architecture was developed based on the characteristics, following the general architecture shown in Fig. 3. Table 3 resumes all the chosen parameters for the elements in the architecture.

4.4.1. *Knowledge base definition*

The knowledge base definition is shown in Fig. 4. The instruction manual has been divided into separate text files, with each file corresponding to a specific job. Additionally, each file has been segmented into smaller text chunks before being vectorized. To avoid any loss of information during the embedding process, the authors chose to preserve the context of each phase and not divide a single phase into multiple chunks. Five distinct chunks for each job, with varying sizes were created. One chunk provides general information about the process, including the name and number of phases, while the remaining four chunks contain detailed information about each phase. Table 4 displays the chunk sizes (number of characters including spaces) for the ITALY job and their corresponding content. These chunk sizes are also indicative of the other jobs.

Finally, the text chunks of a single job were passed through the ’text_embedding_ada_002’ embedding model provided by OpenAI (as cited on the OpenAI website). This created different vector databases for each job using the Python library FAISS (as cited on the FAISS website). The databases were saved locally and retrieved each time a job was performed.

4.4.2. *LangChain framework definition*

For the proposed DIA, the authors combined two different chains:

- the Load_qa_chain (Langchain. Load_qa_chain, 2023), which includes the LLM model, the prompt, and the type of memory.
- RetrievalQA (Langchain. RetrievalQA chain,) involves the integration of the previous chain and the retriever, a vector database from which the DIA retrieves information.

Below is a description of the aforementioned elements that describe the chains:

- LLM model. GPT-4 Turbo, the latest model release from OpenAI (OpenAI, 2023a), was adopted due to its excellent capabilities and lower cost per token compared to GPT-4 (OpenAI. OpenAI Pricing, 2023).
- Retriever. The Retriever used to store the vectors is FAISS (Meta, 2017), which allows the similarity search through the KNN algorithm.
- Memory type. The utilization of this memory type may present a challenge, as the cost of elaborating and responding to a question increases with the length of the prompt. To address this concern, a ’window buffer memory’ was implemented. The buffer retains the most recent two iterations of the dialogue between the DIA and the user within the prompt. An iteration encompasses the operator’s question and the DIA’s response.
- Prompt. Here is the prompt used for the DIA:

Table 4
Chunk size and content of "Italy" job chunks.

Chunks	Chunk Size (char)	Chunk Content
Chunk 1	641	General description of the process
Chunk 2	560	Step A: Selection of the box
Chunk 3	543	Step B: Assembly of dividers
Chunk 4	412	Step C: Insertion of the components
Chunk 5	539	Step D: Label placement

Your name is Rich and you are the attentive assistant to an operator carrying out an assembly operation. Your answers must be no longer than two sentences. They must be to the point. You must help the operator by answering his questions in a timely and concise manner. Answer only what you are asked.

If you cannot answer the question or if you are in doubt, ask the operator to repeat the question.

Do not make things up and only answer based on context and chat_history.

The steps of the process are as follows, and are performed in the following order:

- *Step A: Selection of the box*
- *Step B: Assembly of the dividers*
- *Step C: Insertion of the components*
- *Step D: Label placement.*

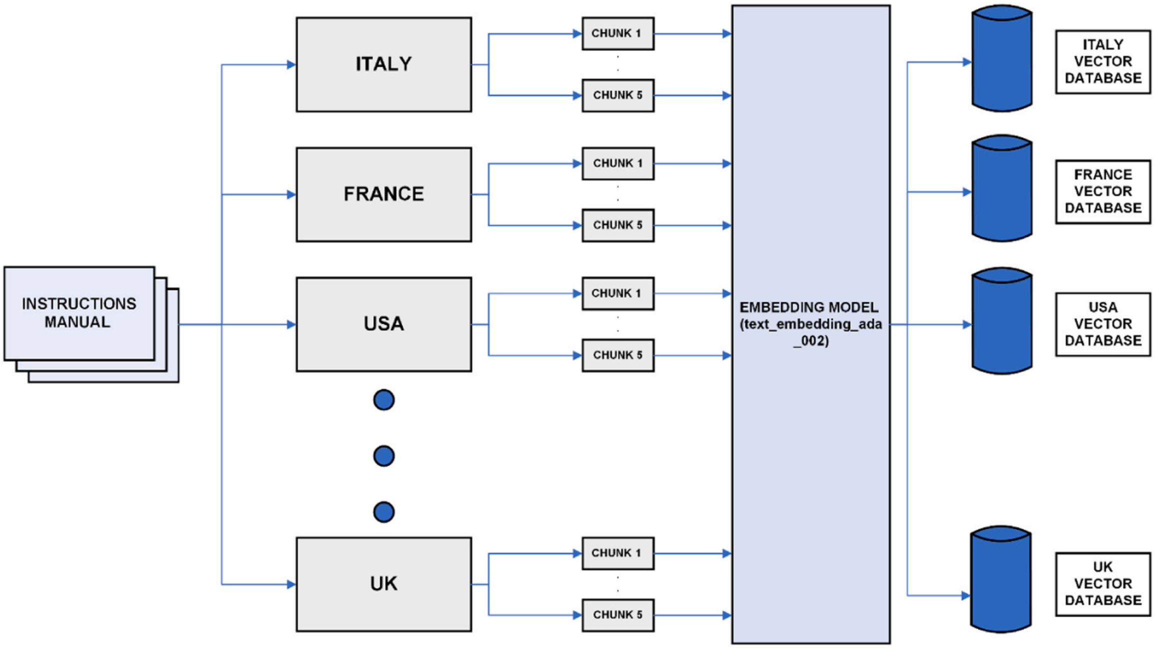

Fig. 4. Chunking and Embedding Process.

When the operator has completed a particular operation, respond by indicating the next step. If there are no next steps, state the end of the task. If the first question is of the type: how do I start, how do I proceed, what do I have to do, you must answer with the corresponding step A of the process.

{context}

Use the chat_history or the list of phases to answer follow-up questions such as What should I do next?, How do I proceed?, What next?, How do I continue?, Now?

{chat_history}

4.5. Dependent variables

As previously mentioned, the literature on DIAs reveals a significant lack of comprehensive metrics and analysis dimensions to evaluate the practicality of such solutions (Bousdekis et al., 2022). The dimensions considered for analysis are technical robustness, cognitive workload, system usability, and experience, as well as assembly process performance benefits. The dependent variables in the DIA experiment were chosen to provide a measure of the hypothesis (Table 5). Every variable is explained in the following paragraphs.

4.6. Technical robustness dimensions

To assess Technical Robustness, a structured classification of questions is used. Responses are classified as accurate (AC), not accurate (NA), or hallucinations (HALL). AC responses accurately answer the operator’s questions and can be classified as ’complete’ (COM) if they fulfill all requests or ’incomplete’ (INC) if they only provide partial information. Not accurate (NA) responses may occur when the DIA provides incorrect details, such as information about the wrong step instead of the one required. Additionally, LLMs may exhibit a recurrent behavior known as HALLs, where the DIA generates concepts about the process unrelated to the specific process knowledge base. Although several techniques exist to reduce this phenomenon, such as prompt optimization, it may still occur, leading to incorrect information and becoming a source of product defects. This classification is closely linked to system safety. Incorrect answers or hallucinations could indeed lead to incorrect actions by the operator, posing a danger in a manufacturing environment. Although NA and HALL responses appear similar in nature, they are categorized as distinct errors within the DIA. This decision is based on the intrinsic characteristics of these errors. An inaccurate response, where the DIA provides information that deviates from the user’s query but remains relevant to the process in question, can occur in systems without LLMs. Conversely, HALL errors are specific to the functioning of LLMs and thus warrant particular consideration in the performance evaluation of DIAs.

Moreover, two types of hallucinations may be identified: contextualized hallucinations (CONTs) and decontextualized hallucinations (DECONTs). CONTs provide invented concepts in the answers but contextualize them with the query. In this case, the DIA attempts to respond using its previous knowledge, contradicting the process knowledge base. In DECONTs, the system provides completely illogical answers, such as stating that it cannot reply because the information is not in its knowledge base.

Table 5
Hypothesis for research questions.

Hypothesis	Dependent variable
H1.1 DIA accuracy satisfies the process request	Answer Accuracy
H1.2 DIA reliability satisfies the process request	Reliability (Hallucinations)
H1.3 DIA speech recognition accuracy satisfies the process request	Speech Recognition Accuracy: Word Error Rate (WER)
H2.1 DIA decrease the cognitive workload	Task Load Index (NASA TLX)
H2.2 The operator evaluates the usability and the experience positively	User Experience Questionnaire (UEQ) System Usability Scale (SUS) Chatbot Usability Questionnaire (CUQ) Number of interactions Human-DIA
H3.1 DIA reduces the time of the assembly process	Lead time Cycle time
H3.1 DIA increases the quality of the assembly process	Product conformity

Moreover, the DIA’s technical robustness will be evaluated. Technical robustness is the ability to accurately capture the operator’s voice and convert it into text. This is crucial for the system’s industrial application, particularly in noisy environments with machinery and operators who have varying speech patterns. The system’s performance will be evaluated using the Word Error Rate (WER) (Popović and Ney, 2007), as previously suggested by Chen et al. (2021). The index is calculated by comparing the speech input from the system with the translated text from the speech-to-text system. This identifies correctly translated words (C), replaced words (S), deleted words (D), and inserted words (I), which are then used in the Eq. (1):

$$WER = \frac{S + D + I}{S + D + C} \tag{1}$$

A lower WER value, closer to 0, indicates a higher level of accuracy and robustness of the speech-to-text system. Furthermore, the WER can be compared with the rate of responses classified as accurate to extract further information such as the system’s ability to adapt to translation errors.

4.7. Cognitive workload

Cognition refers to the human processing of incoming information, while Cognitive Load pertains to how this information is buffered by the brain’s limited storage capacity (Schmidhuber et al., 2021). The NASA Task Load Index (NASA TLX) (Hart and Staveland, 1988) was used to measure workload, as it is the most established and widely used subjective method for detailed workload analysis (Bousdekis et al., 2022). It is a psychometric, multidimensional evaluation tool that assesses the workload perceived by users when completing specific tasks.

It consists of six dimensions: mental demand, physical demand, temporal demand, performance, effort, and frustration. Users rate each dimension on a scale from 0 to 100 with 5-point steps to indicate their perceived workload. This tool evaluates the cognitive workload perceived by users during the assembly process. Both the experimental and control groups complete the questionnaire to compare the cognitive workload of the tasks.

4.8. System usability and experience

To evaluate System Usability and Experience, three distinct questionnaires will be used: the System Usability Scale (SUS), the Chatbot Usability Questionnaire (CUQ), and the User Experience Questionnaire (UEQ). Additionally, the total number of interactions between the user and the DIA will be evaluated. This further measure will provide insight into how operators use the tool, either as a facilitator or relying on it entirely.

The System Usability Scale (SUS) is the most widely recognized tool designed to evaluate the usability of a system. Participants answer 10 questions on a 5-point Likert scale. SUS has been widely tested and can be used on small sample sizes with reliable results, effectively differentiating between usable and unusable hardware, software, mobile devices, websites, and applications (Bangor et al., 2008).

The Chatbot Usability Questionnaire (CUQ) focuses on gathering feedback related to specific aspects of the user interface, functionality, and overall satisfaction. The questionnaire was specifically designed to measure the usability of chatbots and consists of sixteen balanced questions related to different aspects of chatbot usability. Eight relate to positive aspects of chatbot usability, and eight relate to negative aspects. All sixteen questions are scored using a five-point Likert-type scale. The CUQ consists of specific questions enabling users to express their opinions on various dimensions, contributing to a nuanced understanding of usability (Holmes et al., 2023).

Additionally, the User Experience Questionnaire (UEQ) (Schrepp et al., 2014) evaluates the overall user experience of interactive products, covering six scales. Participants respond to items on a seven-point scale, providing insights into emotional and experiential facets of user interaction. These questionnaires provide a comprehensive evaluation of usability and user experience aspects, allowing for a thorough analysis of the system's performance from multiple perspectives. Aspects such as attractiveness, perspicuity, efficiency dependability, stimulation, and novelty are measured by UEQ.

4.9. Assembly process performance

To assess the advantages of implementing DIA, several measures are evaluated, including cycle time and lead time, as well as a qualitative analysis of the process output (Fig. 5).

The first two dimensions consider the average cycle times for each step of the job and the total time for completing the job of the assembly process. The objective is to identify any possible improvements or deteriorations compared to the traditional case, where the operator relies solely on the instruction manual. It should be noted that while the DIA system is innovative, it is still a basic system. As a result, it may generate delays due to non-optimized technical aspects (e.g. microphone sensitivity, Internet bandwidth speed), such as the time it takes to capture voice input and process responses. To ensure accuracy, the cycle times of the actual steps and the total job will be calculated, excluding these time inefficiencies.

The process output quality is also analyzed to identify any defects and their causes. It is important to determine whether these defects were caused by incorrect information from the DIA or human error. The aim is to determine whether the DIA reduces process defects compared to using an instruction manual.

4.10. Control variables

Some control variables were set to be kept constant across the experiment to prevent them from affecting the results. It was deemed necessary to ensure that the participants had a similar experience and skill level with regard to the process and the technology. During the participant selection phase, the researchers successfully recruited 30 master's degree students specializing in Management Engineering, aged between 22 and 28 years. While acknowledging that this demographic is not ideal for an experiment primarily aimed at industry applicability, the researchers encountered challenges in enlisting actual manufacturing workers. To minimize any biases stemming from diverse experiences, it was confirmed that none of the participants had prior familiarity with the specific assembly process under study or related tasks. This precaution was critical to maintaining the integrity and validity of the experiment's outcomes. The gender balance was granted in the selection. The environmental conditions, including lighting, noise level, and temperature, were maintained consistently throughout the experiment, in accordance with the background noise level. As the effectiveness of the DIA is contingent upon the specific language, all participants were Italian-speaking individuals.

4.11. Experimental procedures

The industrial feasibility of the DIA was evaluated based on an experiment with multiple scenarios and a control group designed to assess the interactions of the human operators with the DIA and to compare the experience and performance with traditional methods. The participants were expected to complete tasks while interacting with the system during the experiment. As mentioned before, the tasks focused on assembling a toolbox.

The participants were thirty students who were recruited through class presentations and email flyers. The participants were informed that the experiment would involve trying a DIA solution. The decision to select this specific number of participants was made in accordance with the literature regarding similar experiments (Chen et al., 2021; Roldán

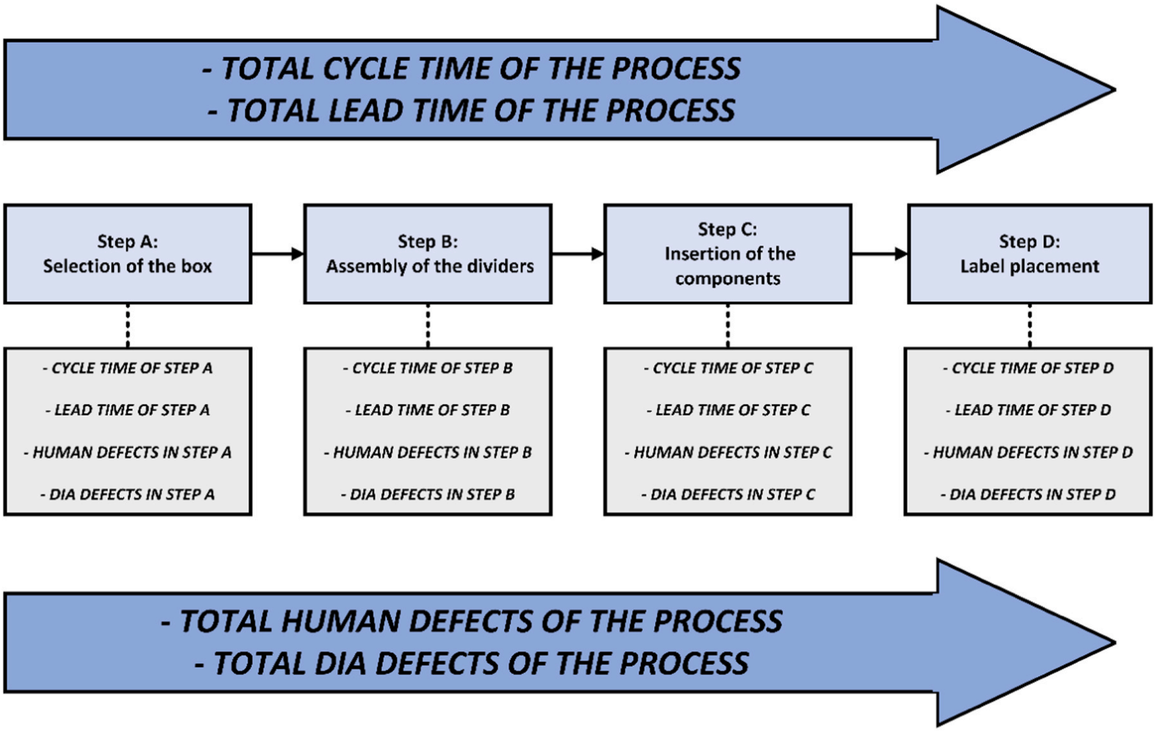
- TOTAL CYCLE TIME OF THE PROCESS
- TOTAL LEAD TIME OF THE PROCESS
Step A: Selection of the box
Step B: Assembly of the dividers
Step C: Insertion of the components
Step D: Label placement
- CYCLE TIME OF STEP A
- LEAD TIME OF STEP A
- HUMAN DEFECTS IN STEP A
- DIA DEFECTS IN STEP A
- CYCLE TIME OF STEP B
- LEAD TIME OF STEP B
- HUMAN DEFECTS IN STEP B
- DIA DEFECTS IN STEP B
- CYCLE TIME OF STEP C
- LEAD TIME OF STEP C
- HUMAN DEFECTS IN STEP C
- DIA DEFECTS IN STEP C
- CYCLE TIME OF STEP D
- LEAD TIME OF STEP D
- HUMAN DEFECTS IN STEP D
- DIA DEFECTS IN STEP D
- TOTAL HUMAN DEFECTS OF THE PROCESS
- TOTAL DIA DEFECTS OF THE PROCESS

Fig. 5. Assembly Process Indicators.

Table 6
Parameters collected during the experiment.

Parameters of evaluation		Collection method
$t_{i,j,p}$	Time to complete step *i* of job *j* by the operator *p*	Manually
$t_{j,p}$	Time to complete job *j* by the operator *p*	Manually
t_p	Time to complete the entire process by the operator *p*	Manually
$Q_{k,j,p}$	Question in the interaction k posed to the DIA to complete job i by operator p	DIA
$A_{k,j,p}$	Answer in the interaction k provided by the DIA to complete job i by operator p	DIA
$I_{j,p}$	Number of interactions in the realization of job j between the DIA and the operator p	DIA
$IT_{j,p}$	Number of tokens in input to the LLM model associated with the questions posed to the DIA to complete job i by operator p	DIA
$OT_{j,p}$	Number of output tokens from the LLM model associated with the answers to questions posed to the DIA to complete job i by operator p	DIA
$TT_{j,p}$	Number of total tokens of job j completed by operator p	DIA
TT_p	Number of total tokens of operator p	DIA
$r_{i,j,p}$	DIA Inefficiency Time spent listening to questions and processing answers to complete step i of job j by operator p	DIA
$DH_{j,p}$	Number of defects found in the job j completed by the operator p caused by a human error	Manually
$DD_{j,p}$	Number of defects found in the job j completed by the operator p caused by a DIA error	Manually
$TD_{j,p}$	Number of total defects found in the job j completed by the operator p	Manually
GA	Set of the operators who used the DIA in the realization of the jobs	Manually
GB	Set of the operators who used the instruction manual in the realization of the jobs	Manually

et al., 2019; Li et al., 2022), which typically recommends a minimum sample size of 20 individuals to ensure statistical significance. The sample consisted of 15 males and 15 females, with two participants in each time slot. The experiment was conducted over several days, with each time slot lasting 45 minutes. During this period, participants completed both the training and the activity. The data was collected between the 5th and 15th of December, 2023.

The experiment was conducted in a setting with minimal noise pollution. In this controlled setting, two researchers were present throughout the experiment of each participant. Their duties included instructing the participants at the outset of the experiment and gathering the necessary data for analysis in the subsequent phase. In more detail, the experiment is structured in two phases. In the initial phase of the experiment, all participants were trained to assemble the box in accordance with the specified time schedule. This was done in order to align the sample and to provide comprehensive knowledge about the process. The participants were instructed by the researchers and provided with a manual containing the requisite quantities and specifications for each task. The second Step involved individual participants sequentially assembling two boxes, each associated with different jobs. The participants were divided into two groups: one group, called the experimental group (Group A), completed two jobs with DIA support. A second group, called the control group (Group B), completed the jobs using the manual as in the training. In this phase of the experiment, participants were granted unlimited time to complete the tasks assigned to them. This approach was intended to remove time pressure, thereby allowing for a focus on the accuracy and quality of the work performed, and to better understand the impact of the provided support systems (the DIA or the instruction manual). This setup also facilitated a detailed observation of the participants' interaction with the support tools without the confounding factor of time constraints.

Following the initial training phase, participants completed the Nasa TLX questionnaire to assess their perception of the cognitive load of the task. They were then equally divided into two groups: the control group (n=15) and the experimental group (n=15). Upon division, participants were informed about the experimental procedure, data storage, and voice recording for the experimental group. The experimental group completed a training task to familiarize themselves with the DIA and its interactions. Subsequently, each participant was assigned two boxes to complete in row. Upon completion of the boxes, participants were asked to fill out the Nasa TLX to reassess their workload. The experimental group completed a questionnaire on their experience (CUQ and UEQ) and the usability (SUS) of the DIA. The researchers monitored the activities, measured the timing of each step, and verified that the assembly was defect-free. They also measured assembly time, errors, and the number of hints given. After completing all assembly steps, participants were asked to describe their feelings during the assembly. The statements were recorded and subsequently analyzed. Participants were not permitted to communicate with the researchers during the actual assembly process in the second phase; however, they were permitted to seek clarifications during the initial training phase and while completing the questionnaires. The structure was designed to emulate a realistic work environment while still allowing for controlled experimental conditions.

4.12. Data collection

During the experiment, a range of parameters were collected in order to evaluate the overall process performance and the DIA performance (in terms of technical robustness and safety, as well as accuracy and reliability).

The data collection for the study was conducted using both manual and automated methods. The authors gathered some of the data manually by employing a chronometer and conducting direct observations of the process and making notes. Furthermore, additional data were automatically collected through specialized software integrated with the DIA. The dual-method approach enabled a comprehensive

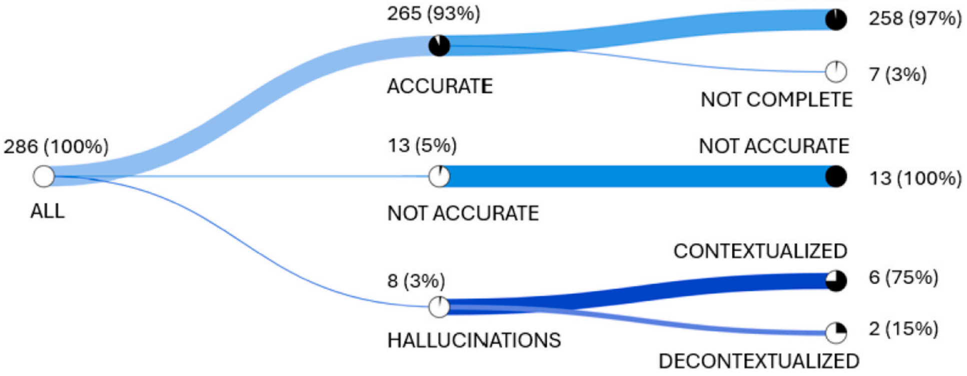

Fig. 6. Answer Classification and Sub-Classification.

Table 7
Examples of Answer Classification and Sub-Classification.

$A_{k,j,p}$	Question	DIA Answer	Correct Answer	Answer Classification	Answer Sub-Classification
$A_{7,FR,2}$	Which label should I attach?	Place the FRENCH label with code 'FR02ST' in the centre of one of the long sides of the box.	/	AC	CO
$A_{15,IT,7}$	Which label should I attach?	Place the label in the centre of one of the long sides of the box.	Place the ITALY label with code 'IT02ST' in the centre of one of the long sides of the box.	AC	INC
$A_{4,IT,13}$	Where should I insert the long divider?	Insert the nails in one of the compartments next to the identifier 'A21-A02'.	Count 8 bars from one end of the long divider and locate the assembly hole in which to insert the short divider	NA	NA
$A_{8,IT,9}$	How many nails should I put in and where should I put them?	Insert four nails, one in each corner of the compartments created by the dividers.	Insert 4 nails in one of the compartments next to the identifier.	HALL	CONT
$A_{3,FR,12}$	Tell me how many screws, dowels, and nails I have to insert	Insert 2 domed screws, 3 dowels, and 2 small screws. You did not mention nails, so I cannot give you that information.	Insert 2 domed screws, 3 dowels, 4 nails, and 2 small screws.	HALL	DECONT

$A_{k,j,p}$: Answer in the interaction k provided by the DIA to complete job i by operator p AC: accurate answer CO: complete answer INC: not complete answer NA: not accurate answer Hall: hallucination answer CONT: contextualized hallucination DECONT: decontextualized hallucination

assessment of the process under investigation, thereby ensuring the generation of a robust dataset for subsequent analysis. Table 6 presents a comprehensive overview of these findings.

5. Results

5.1. Technical robustness

The initial analysis concentrates on the technical robustness of the DIA. The accuracy and reliability of the DIA and speech recognition system were assessed by scrutinizing all 286 questions posed to the system and their corresponding answers.

In order to fully assess the accuracy and reliability of the DIA, a thorough classification of answers was conducted manually by the authors, as outlined in Section 3.6. This involved the analysis of each iteration between the users and the DIA. The classification and sub-classification of answers, along with their respective percentages of the total answers, are displayed in Fig. 6. The results show that 93 % of the answers were accurate, with 97 % of those being exhaustive for the operator's question. However, there are still several critical responses that have been categorized as not accurate (5 %) and hallucinations (3 %). Further exploration of the latter revealed that the majority of hallucinations were contextualized within the knowledge base (75 %), while only a few were decontextualized (15 %). Table 7 provides examples of the different types of answers collected during the experiment.

The speech recognition accuracy was evaluated using the Word Error Rate (WER) to assess the ability to accurately capture the operator's voice and convert it into text. To calculate the Word Error Rate (WER), the input sentences were compared to the model's real sentences and identified substituted words (S), deleted words (D), and inserted words (I). The WER is calculated based on the number of correct words (C) through Eq. 1. Out of the total 286 responses, only 62 had a WER greater than 0. However, upon further examination, it was found that only 3 of these responses were inaccurate, demonstrating the DIA's excellent ability to adapt to translation errors. Table 8 shows the mean WER values for each step of the process. These values are related to the accuracy of the answers in those steps, which is calculated by dividing the number of accurate answers by the total number of answers.

Table 8
Mean WER values for each step of the process.

Step	WER	Accuracy
A	4,13 %	89,58 %
B	3,72 %	93,88 %
C	6,81 %	94,52 %
D	5,27 %	88,37 %

5.2. Cognitive workload

The cognitive workload was assessed using the NASA TLX questionnaire. Each participant completed the questionnaire after the training session and after completing the two tasks. Table 9 reports the mean evaluation and the standard deviations for each dimension of Group A, the experimental group, and Group B, the control group, during the two phases of the experiment (training and job execution). Fig. 7 shows a comparison between the two groups. Note that Group A's assessment of the six dimensions improved between the first and second part of the experiment, resulting in a reduction in the cognitive load of using the DIA. In contrast, group B's assessment of the dimensions remained almost unchanged, except for frustration, which showed a slight increase. When comparing Group A and Group B, it is confirmed that the perceived cognitive load is lower overall when completing tasks using the DIA. Finally, it is important to note that there is an imbalance in the temporal demand between the two groups. Group A perceived a reduced time effort compared to Group B, although, DIA users took longer on average than those who used the instruction manual to complete all the tasks.

5.3. System usability and experience

The System Usability Scale (SUS) was used to quantify the overall usability, revealing a mean score of 81. Table 6 reports the score contributions and SUS score for each participant, along with their standard deviation. All participants reported a high SUS score (Fig. 8: SUS Score Distribution, Table 10), indicating satisfaction with the DIA and its ease of use (Fig. 9, Table 11).

The ease of use has also been confirmed in the User Experience Questionnaire (UEQ). Fig. 10 and Table 12 show the mean value per item after score normalization, along with the standard deviation, where +3 represents the most positive and −3 represents the most negative value. Participants found the DIA easy to use and interesting and appreciated the clear and precise information provided during the implementation process. Overall, they perceived it as an attractive and pleasant tool. However, some errors in the answers and slow processing times were reported, which could be demotivating. Additionally, comparing our average scores on the five UEQ evaluation scales (Table 13) with the benchmark data in the UEQ dataset (Fig. 11), it is clear that our DIA tool is significantly superior to the average for most dimensions, except for efficiency, which is still acceptable.

Instead, Fig. 12 and Table 14 report the mean value and standard deviation score of each item of the Chatbot Usability Questionnaire (CUQ), generating an overall score of 80. The DIA has received mostly positive evaluations, although some issues have elicited mixed opinions (Fig. 13, Table 15). Specifically, some operators have described the

Table 9
NASA TLX dimensions mean value and standard deviation.

NASA TLX Dimensions	TRAIN GROUP A		GROUP A		TRAIN GROUP B		GROUP B	
	Mean score	Std. Dev	Mean score	Std. Dev	Mean score	Std. Dev	Mean score	Std. Dev
Mental Demand	52,67	19,07	32,00	23,96	46,67	15,43	42,00	15,68
Physical Demand	23,33	18,39	23,33	17,99	16,67	8,16	20,00	11,95
Temporal Demand	36,67	19,52	23,33	11,13	44,67	18,46	47,33	12,80
Performance	68,67	24,75	75,33	22,64	73,33	22,57	74,00	20,98
Effort	47,33	21,54	27,33	17,92	38,00	20,07	38,00	18,97
Frustration	30,00	22,04	18,00	12,07	26,00	24,14	36,00	25,58

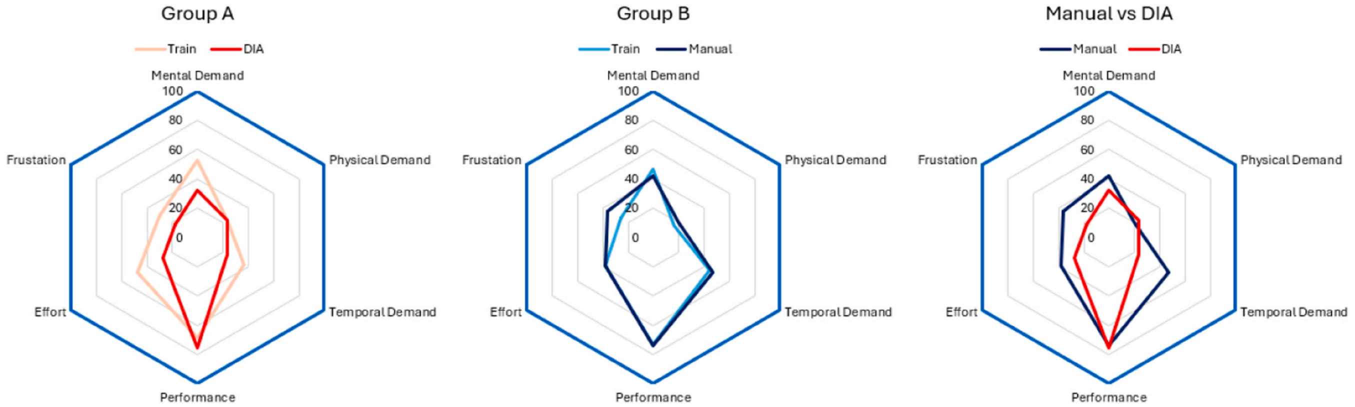

Fig. 7. Operators' evaluation of NASA TLX dimensions.

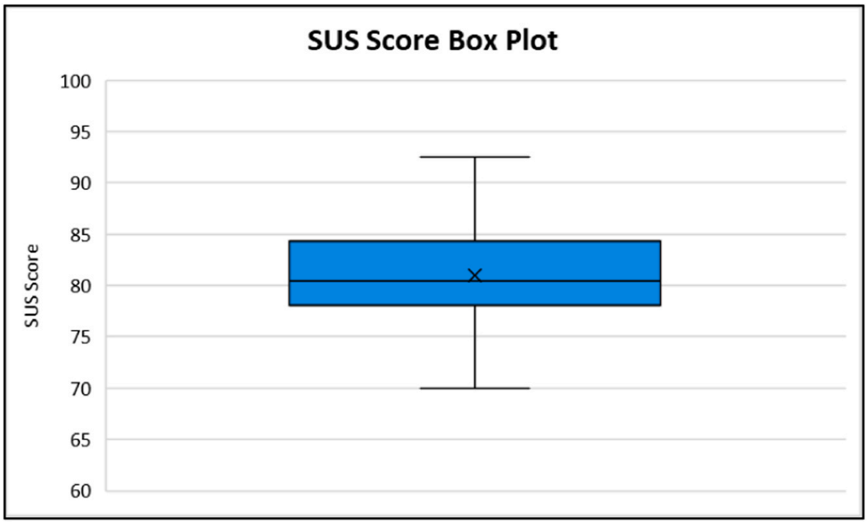

Fig. 8. SUS Score Distribution.

DIA's voice as robotic, while others have experienced difficulty understanding the DIA in certain cases.

The number of interactions of each operator is also reported in Table 16. It can be observed that the number of interactions between DIA and the operator generally decreases from the job "Italy" to the job "France". This implies a learning process with the system. Further data related to the number of interactions can be found in the Appendix.

5.4. Assembly process performance

Considering the last dimension of analysis, the text reports the mean cycle time of each step for both Group A and Group B. It is important to note that for Group A, the time without buffers is also considered, where buffers refer to the lead times caused by DIA in the listening and answer elaboration process. It is important to acknowledge that although the DIA system represents a technological advancement, it is currently in a prototype testing version. Consequently, it is susceptible to delays attributed to unrefined technical elements, such as microphone sensitivity and internet bandwidth speed, which can affect the efficiency with which it captures voice inputs and processes responses. Fig. 14 shows that, for most cases with DIA, the average completion times are higher than with the traditional method. However, once the lead time of listening and processing the question is removed, completion times become similar, if not inferior, to those of realization with the instruction manual. This is also evident in the graphs of the cumulative average times of the two jobs (Fig. 15), whose trends are almost identical. The appendix provides information on the timing of each phase.

Finally, Table 17 displays the total defects resulting in the process,

Table 10
SUS Score.

Operator	SUS Score
1	85
2	75
3	80
4	77,5
5	80
6	92,5
7	70
8	92,5
9	80
10	80
11	82,5
12	70
13	82,5
14	85
15	82,5
Mean Evaluation	81,00
Std. Dev.	6,53

categorized by group and typology. Group A reported fewer defects than Group B. Half of the defects in Group A were due to DIA errors, which provided incorrect information leading to operator errors. The other half of the errors were due to operators' errors. It is important to note that the defects do not refer to a single box, but to any incorrect operation carried out during the process. Therefore, multiple errors may have occurred within a single job, resulting in a significant number of defects in the assembled box. The incidence of human error has decreased across all steps of the process. Additionally, there are no steps in which the number of defects has increased in comparison to traditional assembly methods. Overall, the table shows that the operators who performed the DIA-led tasks achieved better results in terms of product quality.

6. Discussion

In the following section, a discussion of the results has been conducted in an attempt to answer the initial RQs.

RQ1: How does the technical robustness of the DIA impact its application in the manufacturing assembly processes?

The experiment results indicate that the DIA demonstrated a high level of robustness by adapting to different operators and understanding their speech patterns. The low average WER value demonstrated this. Additionally, when analyzing questions with a WER value greater than zero, only a few responses were inaccurate, highlighting the system's excellent ability to interpret the operator's intent, even when it is not explicitly stated. However, it should be noted that the experiment was conducted in a noise-free environment, which may not be representative of real-world industrial settings (Casas et al., 2014). Therefore, while the results are promising for testing in an industrial environment, it is important to evaluate the DIA's adaptability to noisy environments. This can be achieved by improving the Speech to Text system with better instrumentation and conducting further industrial tests.

DIA achieved a 93 % accuracy rate in responses, demonstrating the LLMs' capabilities in comprehension and text generation indicating a good performance and flexibility of the system. However, incomplete responses remain an issue that could be addressed by investing in prompt and chain improvements (Korzynski et al., 2023; Lo, 2023).

A further issue is that of hallucinations, which is a well-known phenomenon in LLMs. This also manifested in the non-experimental group. It is notable that instances of hallucinations occurring in the absence of context were almost non-existent, indicating a high level of robustness in the system. Although hallucinations in the responses of DIA are relatively rare, particularly those that lack contextual relevance, they present a potential risk of causing confusion among operators in industrial settings. In the context of industrial mass customization, operators typically possess extensive experience with the processes in

Table 11
SUS item mean time and standard deviation.

Item	Mean Value	Std. Dev.
1	4,07	1,03
2	1,20	0,41
3	4,33	0,90
4	2,20	1,15
5	4,07	1,10
6	2,07	1,03
7	4,53	1,06
8	1,67	0,90
9	4,33	0,72
10	1,67	1,11

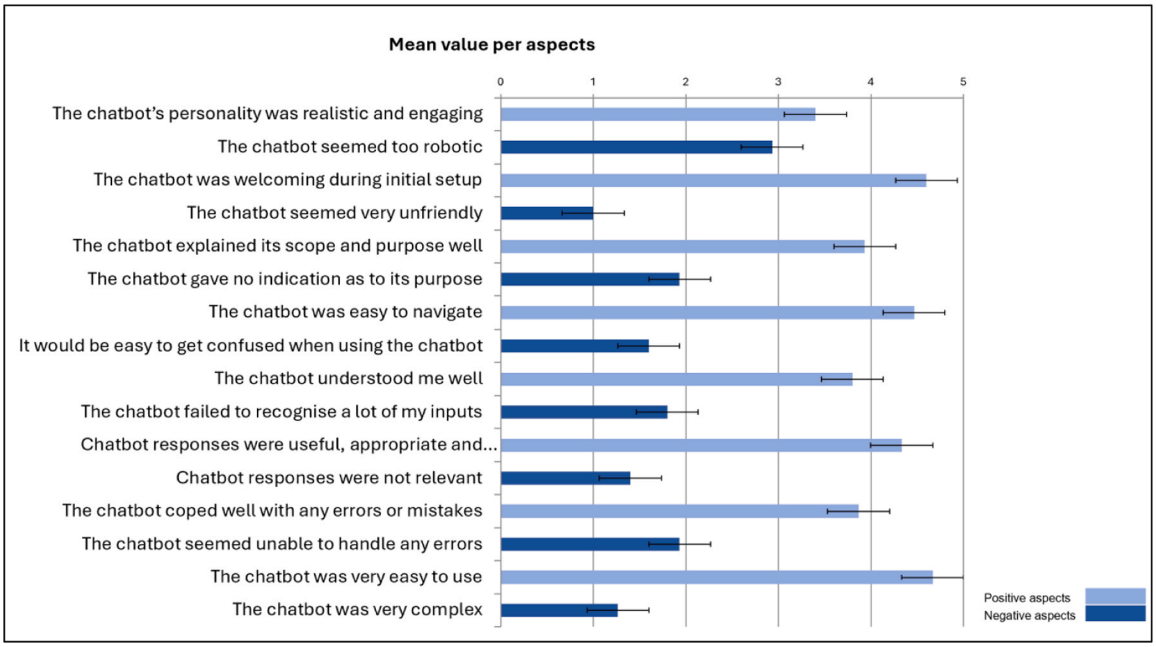

Fig. 9. SUS Mean value per aspects.

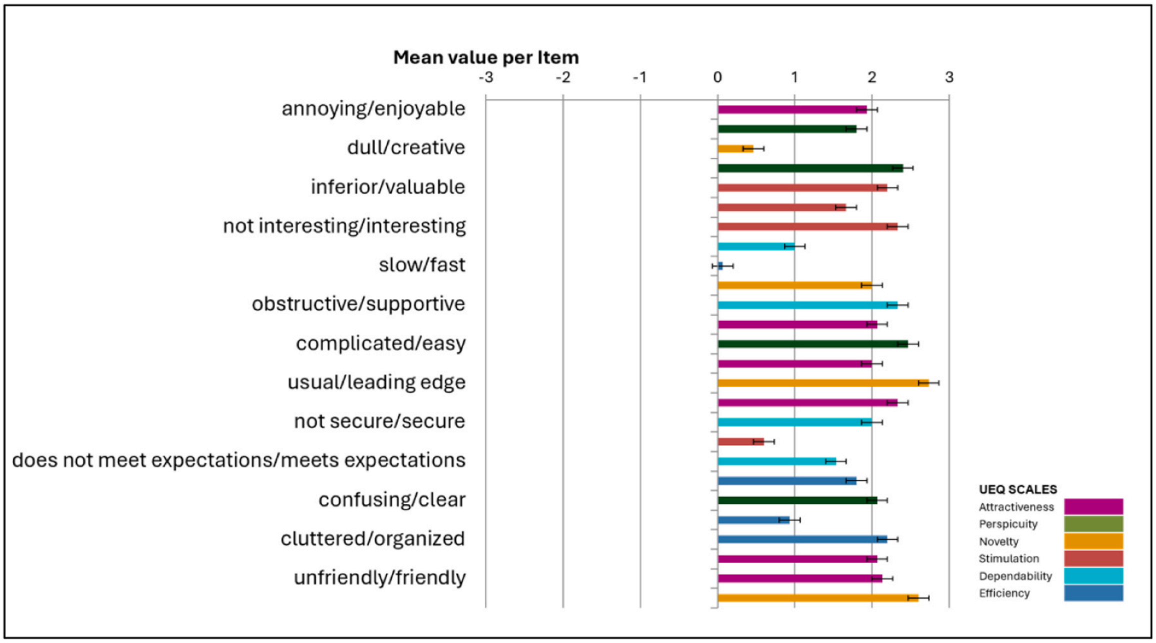

Fig. 10. UEQ Mean value per Item.

Table 12
UEQ mean value and standard deviation per Item.

Item	Mean	Std. Dev.
1	1,93	0,88
2	1,80	0,86
3	0,47	2,13
4	2,40	1,30
5	2,20	0,77
6	1,67	1,05
7	2,33	0,72
8	1,00	1,56
9	0,07	1,79
10	2,00	1,13
11	2,33	0,72
12	2,07	1,22
13	2,47	0,64
14	2,00	1,07
15	2,73	0,46
16	2,33	0,62
17	2,00	0,85
18	0,60	1,68
19	1,53	1,25
20	1,80	0,94
21	2,07	0,80
22	0,93	1,28
23	2,20	0,77
24	2,07	0,70
25	2,13	0,64
26	2,60	0,51

Table 13
UEQ scales mean value and standard deviation.

Scale	Mean	Std. Dev.
Attractiveness	2,09	0,65
Perspicuity	2,18	0,58
Efficiency	1,25	0,65
Dependability	1,72	0,62
Stimulation	1,70	0,77
Novelty	1,95	0,84

question, having executed them on numerous occasions. This familiarity allows them to mitigate the effects of any hallucinatory responses by leveraging their existing knowledge. Consequently, the disruptive potential of hallucinations is amplified when the operator is less experienced. Consequently, it is of paramount importance to emphasize the necessity of comprehensive training for operators. Such training should equip operators with the skills necessary to distinguish between accurate responses and those provided by the DIA that deviate from the established knowledge base.

Despite the excellent results, the system is not yet error-free enough to ensure operator safety in any industrial environment. If the DIA is used for complex tasks and integrated with machines and robots, even a small incorrect response could cause problems, not only in terms of product and process defects but, more importantly, in terms of operator safety (Polak-Sopinska et al., 2019; Costantino et al., 2021). However, it is believed that this system has significant room for improvement and can be easily integrated with other solutions, such as sensors, to monitor dangerous situations.

RQ 2.1: What is the impact of the DIA on the cognitive workload of the operator during assembly tasks?

The NASA TLX questionnaires showed that using the DIA reduced cognitive load compared to the traditional method, confirming what has already been observed in several studies (Schmidhuber et al., 2021; Lee et al., 2019). The innovative tool increased operator involvement and reduced physical and mental effort. The level of involvement was high, as confirmed by the participants' perception of time in Group A compared to Group B. Although Group A took longer to complete the process overall, they perceived it as shorter due to the absence of alienation in their operations. In contrast, group B experienced a growing sense of frustration throughout the experiment. *This growing frustration can be attributed to the prolonged engagement with the process specifications outlined in the manual, which became increasingly alienating over time. Furthermore, the frequent necessity to consult the manual in order to verify the accuracy of their actions serves to exacerbate their sense of alienation. This observation suggests that the manual's complexity and the constant requirement for reference may hinder operational efficiency and affect worker satisfaction negatively.* Overall, it can be concluded that the DIA reduced the cognitive load on the operator, who viewed the system as a tool to support their work. However, it is important to note that the results may be biased due to the wow effect that these technologies can

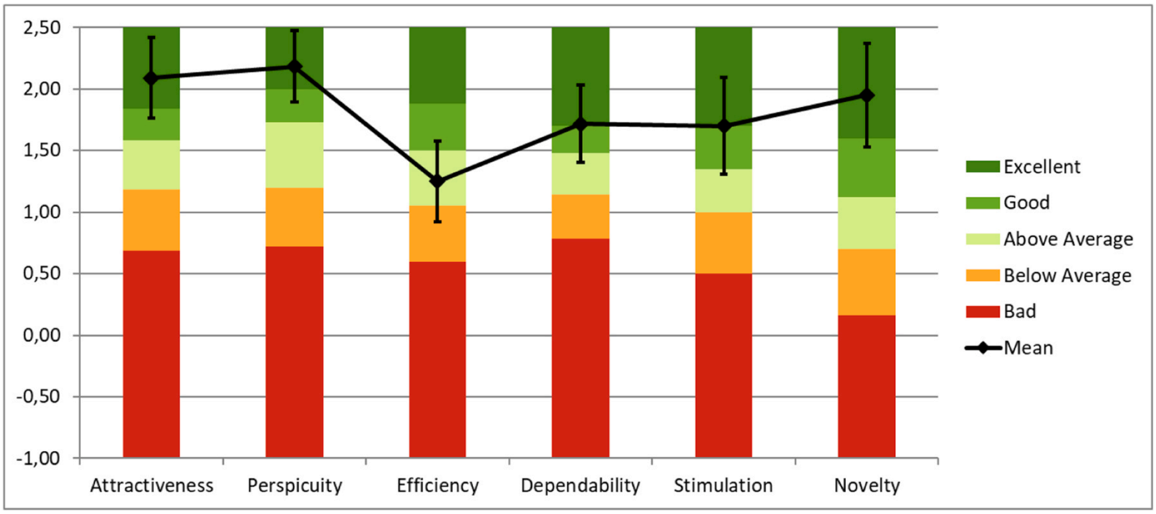

Fig. 11. Comparison between our UEQ scales and benchmark scales.

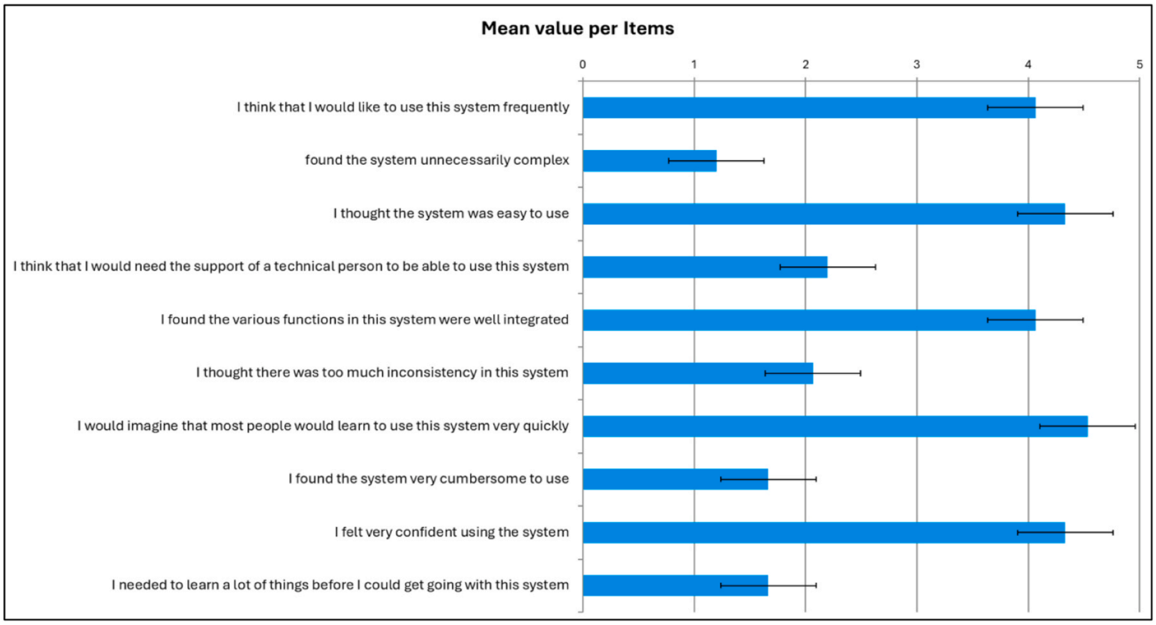

Fig. 12. CUQ mean value per Items.

Table 14
CUQ questions' mean score and standard deviation.

Question	Mean score	Std. Dev.
1	3,40	0,99
2	2,93	1,22
3	4,60	0,74
4	1,00	0,00
5	3,93	1,22
6	1,93	1,28
7	4,47	0,52
8	1,60	0,63
9	3,80	0,77
10	1,80	0,77
11	4,33	0,62
12	1,40	1,06
13	3,87	0,83
14	1,93	1,16
15	4,67	0,49
16	1,27	0,46

generate (Gong et al., 2021).

RQ 2.2: How does the usability of the DIA influence the overall experience of the operator in assembly processes?

The SUS, UEQ, and CUQ questionnaires confirm that the operator found the DIA to be highly acceptable (Ruiz et al., 2023). Across all three questionnaires, the system's ease of use was identified as its strongest attribute. Furthermore, the tool's information clarity contributed to a positive evaluation. The ease of learning to use the system was also noted. This was confirmed by the number of interactions between DIA and humans, which decreased between the first and second jobs. This indicates that the operator, once he or she understands how it works, tends to be able to understand how best to use it, thus improving overall performance (Holmes et al., 2019). This is important as it could facilitate the tool's acceptance in an industrial setting. The questionnaires identified areas for improvement, particularly the DIA's efficiency, which was perceived as slow and robotic.

RQ 3: What are the benefits of introducing a DIA into the assembly process in terms of performance, such as time and quality?

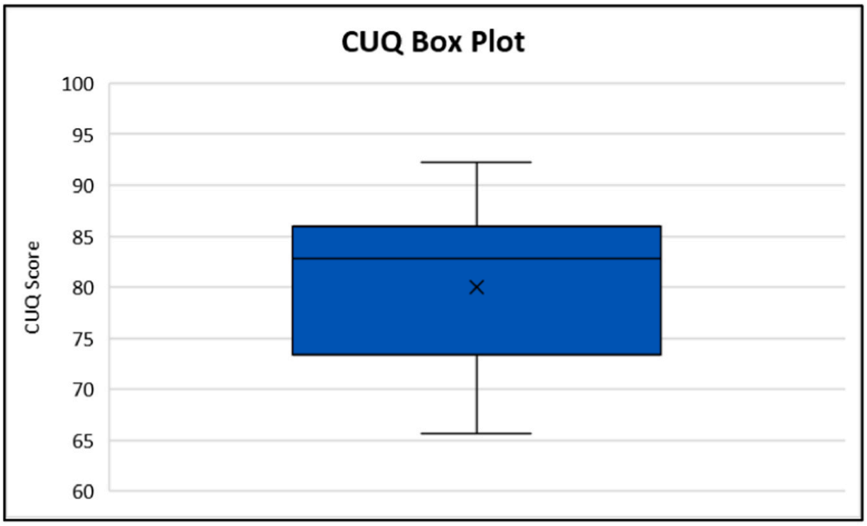

Fig. 13. CUQ Score distribution.

Table 15
CUQ scores.

Operator	CUQ Score
1	75,00
2	85,94
3	87,50
4	84,38
5	85,94
6	70,31
7	87,50
8	71,88
9	65,63
10	76,56
11	82,81
12	75,00
13	85,94
14	73,44
15	92,19
Mean Score	80
Std. Dev.	7,86

Table 16
Number of interactions between operator and DIA.

Operator	ITALY	FRANCE
1	9	7
2	10	7
3	7	8
4	10	15
5	15	6
6	8	7
7	16	8
8	8	7
9	17	9
10	13	9
11	14	8
12	6	6
13	9	7
14	12	10
15	10	8

The analysis examined the DIA’s support in the assembly process with regard to process time and product output quality. The DIA recorded longer job realization times than the traditional process. However, if one assumes times without the inefficiencies caused by listening to questions and processing answers, the times become almost identical. This situation can be considered because, as previously stated, the present DIA is a prototype tool with significant potential for technical improvement. Although the desired results for time performance were not achieved, there was a significant improvement in the quality of process output. Specifically, there was a notable reduction in errors compared to the traditional method. This finding is consistent with previous studies (Fan et al., 2024; Bousdekis et al., 2021), although they are still in their early stages. In this study, we conducted additional analyses by combining both experiential and process parameters. The use of the DIA reduced the difficulty of carrying out complex operations, which may be connected to the reduction in cognitive load experienced by the operators in group A. They felt less fatigued and more focused, enabling them to avoid many errors. Furthermore, it is noteworthy that the superior quality of the output may be attributed to the ease of verifying the accuracy of the work using the DIA, in contrast to the instruction manual. In addition, the effectiveness of the DIA is greatly enhanced when used in complex and varied operations where individual expertise cannot be relied upon (Freire et al., 2023).

7. Conclusion

The article presents a novel exploration of integrating a voice-enabled Digital Intelligent Assistant (DIA) utilizing advanced Large Language Models (LLMs) within manufacturing assembly processes. A key innovation of this study is its experimental design, which introduces explicit analytical dimensions and well-defined indicators to address the lack of clear evaluation parameters for DIAs in manufacturing—a noted gap in current literature. The study provides a comprehensive analysis of the DIA’s applicability, assessing technical robustness, cognitive workload, process performance, usability, and the overall experience of operators involved in complex assembly tasks. Additionally, the simplicity of the experimental implementation enhances the study’s replicability. This research represents also a significant improvement as it is the first application of LLMs for a smart assembly process. advancement as it is the first to apply LLMs in a smart assembly process. The system’s development follows a meticulous approach, incorporating a well-defined taxonomy and architecture. Each step of its creation has been critically discussed, ensuring a robust foundation for implementation. The study assesses the system from multiple perspectives, considering both technical and social aspects, contributing valuable insights to the discourse on user-centric design and the integration of advanced AI technologies in manufacturing.

From a practical perspective, this research serves as an essential prototype, providing promising results in terms of operator experience, cognitive load, and product output quality. However, it is crucial to

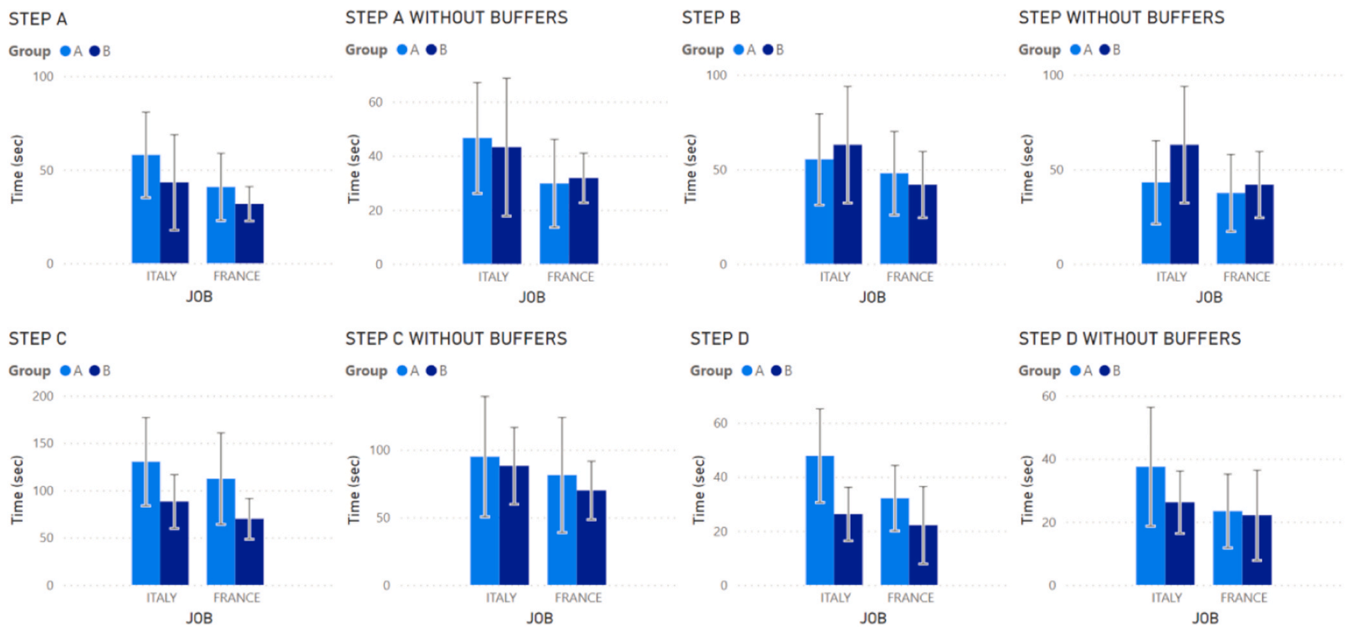

Fig. 14. Average completion time of each step of the process.

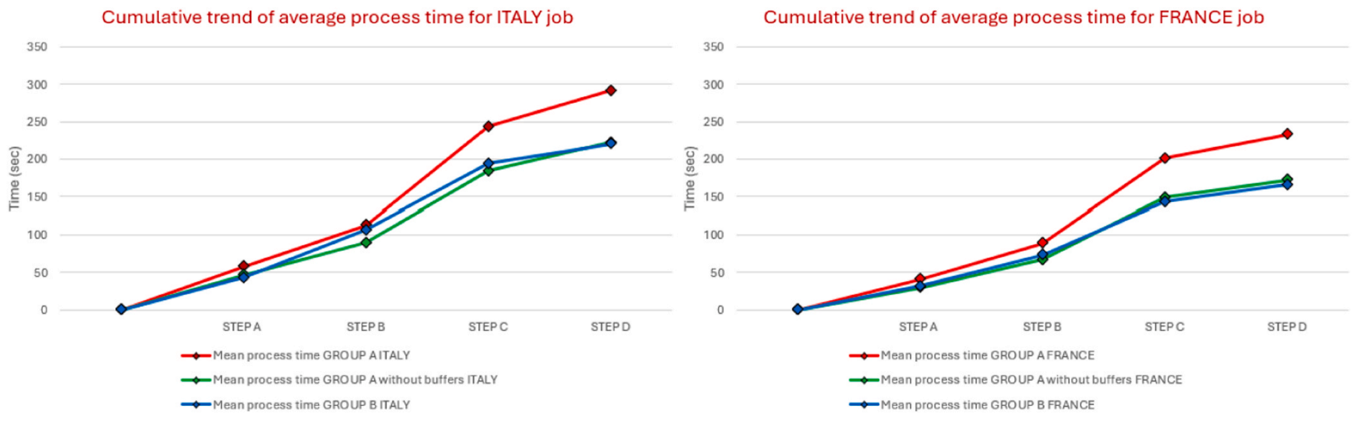

Fig. 15. Cumulative trend of average process time for "Italy" and "France" jobs.

Table 17
Total number of defects.

Steps of the process	Group A		Group B	
	DIA Defects	*Human Defects*	*DIA Defects*	*Human Defects*
Step A: Selection of the box	2	4	/	6
Step B: Assembly of the dividers	1	1	/	7
Step C: Insertion of the components	3	0	/	3
Step D: Label placement	0	0	/	3
TOTAL DEFECTS	11		19	

acknowledge that this is a preliminary step, and further testing in an industrial context is necessary.

Future developments should focus on organizing an industrial test environment, expanding the scope to more complex and varied tasks. Technical enhancements, such as implementing a fact-checking system for hallucinations within the chain and refining the noise reduction in the speech-to-text system, are suggested for continued progress.

From a user experience design perspective, future improvements should focus on enhancing the adaptability of the DIA to different operators' profiles, skills, and backgrounds. This could potentially be achieved by incorporating personas. Finally, the integration of a visual component can enhance the overall usability and effectiveness of the DIA, providing a more multimodal experience.

CRediT authorship contribution statement

Nicolò Sabetta: Writing – original draft, Visualization, Software, Methodology, Data curation, Conceptualization. **Francesco Costantino:** Writing – review & editing, Validation, Supervision, Methodology. **Silvia Colabianchi:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Conceptualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Appendix

A. Number of Interactions

Table 18
Number of interactions between human and DIA.

Ij,p	j="Italy"	j = "France"
p= 1	9	7
P=2	10	7
p= 3	7	8
p= 4	10	15
p= 5	15	6
p= 6	8	7
p= 7	16	8
p= 8	8	7
p= 9	17	9
p= 10	13	9
p= 11	14	8
p= 12	6	6
p= 13	9	7
p= 14	12	10
p= 15	10	8

B. Time details for each step of the process

Table 19
Step time.

$t_{i,j,p}$		j="Italy"				j = "France"			
		i= "STEP A"	i= "STEP B"	i= "STEP C"	i= "STEP D"	i= "STEP A"	i= "STEP B"	i= "STEP C"	i= "STEP D"
GROUP A	p= 1	35,30	46,15	124,32	42,60	36,30	40,71	96,57	34,58
	P=2	116,42	52,34	136,66	54,76	75,18	36,12	124,05	34,90
	p= 3	50,56	71,95	73,06	31,01	79,55	99,99	93,98	51,69
	p= 4	34,74	30,20	151,14	63,99	43,98	53,75	233,23	27,10
	p= 5	92,02	112,87	208,15	75,69	15,04	50,68	88,27	31,47
	p= 6	49,24	40,19	112,46	30,58	31,62	34,25	100,50	36,04
	p= 7	50,43	41,44	228,79	84,87	37,20	25,01	112,01	21,87
	p= 8	62,60	27,05	136,00	42,64	39,22	47,87	79,31	38,43
	p= 9	54,91	91,02	182,36	16,41	29,60	83,92	101,67	19,49
	p= 10	59,67	77,90	92,79	49,40	63,40	28,15	97,62	27,63
	p= 11	44,14	69,02	133,14	48,25	20,18	40,82	225,53	17,92
	p= 12	28,12	22,87	77,04	32,32	30,27	18,86	91,63	20,58
	p= 13	73,39	50,15	60,50	37,50	43,75	36,04	58,98	27,51
	p= 14	76,72	43,29	138,68	55,96	39,09	44,52	107,76	64,88
	p= 15	40,38	52,45	99,16	51,16	27,60	78,28	73,96	27,50
GROUP B	p= 1	14,89	37,45	56,38	10,19	22,17	24,58	47,56	9,44
	P=2	20,47	91,73	143,72	21,30	20,02	68,10	72,71	17,67
	p= 3	50,69	24,70	62,21	26,87	44,84	36,63	62,78	11,87
	p= 4	38,89	38,71	77,19	10,65	40,78	33,95	47,86	20,78
	p= 5	13,82	75,40	94,81	28,86	24,61	69,35	99,75	11,74
	p= 6	31,02	55,39	130,53	9,37	28,42	49,74	78,31	29,26
	p= 7	31,11	44,93	100,44	43,14	37,34	38,31	48,98	9,92
	p= 8	24,19	41,37	85,38	31,18	26,84	37,54	117,47	68,84
	p= 9	81,02	98,31	59,96	32,41	28,52	45,65	63,63	19,15
	p= 10	72,62	63,33	91,38	31,57	38,78	29,56	78,02	22,75
	p= 11	31,30	36,42	39,35	28,37	26,23	19,08	50,36	15,95
	p= 12	76,02	53,73	94,55	22,30	42,64	19,92	77,98	20,00
	p= 13	33,10	47,56	60,22	22,74	16,25	24,27	35,39	19,93
	p= 14	30,36	91,19	109,33	34,60	32,68	60,33	74,46	16,94
	p= 15	98,59	144,49	116,26	39,65	46,42	71,30	93,05	36,86

Table 20
Job Time and Process Time.

$t_{j,p}$		j="Italy"	j = "France"	t_p
GROUP A	p= 1	248,37	208,16	456,53
	P=2	360,19	270,25	630,44
	p= 3	226,58	325,21	551,79
	p= 4	280,07	358,05	638,12
	p= 5	488,73	185,45	674,17
	p= 6	232,46	202,40	434,86
	p= 7	405,54	196,09	601,63
	p= 8	268,29	204,83	473,12
	p= 9	344,70	234,67	579,37
	p= 10	279,75	216,80	496,56
	p= 11	294,56	304,46	599,02
	p= 12	160,35	161,35	321,70
	p= 13	221,54	166,29	387,83
	p= 14	314,65	256,25	570,90
	p= 15	243,15	207,34	450,49
GROUP B	p= 1	118,91	103,75	222,66
	P=2	277,22	178,50	455,72
	p= 3	164,47	156,12	320,59
	p= 4	165,44	143,37	308,81
	p= 5	212,89	205,45	418,34
	p= 6	226,31	185,73	412,04
	p= 7	219,62	134,55	354,17
	p= 8	182,12	250,69	432,81
	p= 9	271,70	156,95	428,65
	p= 10	258,90	169,11	428,01
	p= 11	135,44	111,62	247,06
	p= 12	246,60	160,54	407,14
	p= 13	163,62	95,84	259,46
	p= 14	265,48	184,41	449,89
	p= 15	398,99	247,63	646,62

Table 21
Inefficiency time.

ri,j,p	j="Italy"				j = "France"			
	i= "STEP A"	i= "STEP B"	i= "STEP C"	i= "STEP D"	i= "STEP A"	i= "STEP B"	i= "STEP C"	i= "STEP D"
p= 1	7,05	13,13	28,78	6,09	6,92	6,13	30,54	6,68
P=2	13,79	6,22	29,57	9,75	19,22	10,43	41,11	8,92
p= 3	8,83	23,02	12,35	19,00	16,04	15,40	14,70	8,44
p= 4	6,84	19,54	48,66	13,43	7,80	6,75	52,96	11,06
p= 5	15,52	12,10	26,81	5,72	14,26	10,97	22,96	10,51
p= 6	10,22	6,39	23,43	11,33	9,72	5,91	24,27	6,66
p= 7	14,65	8,14	48,99	4,94	16,40	8,41	47,74	6,40
p= 8	9,62	6,37	28,20	5,32	9,92	7,63	27,39	6,15
p= 9	5,77	20,19	47,11	9,37	8,83	19,03	39,60	6,93
p= 10	11,30	13,04	25,03	22,09	11,32	14,04	21,13	20,69
p= 11	9,08	21,12	121,70	6,93	10,99	20,98	33,17	6,88
p= 12	9,62	7,31	17,45	7,49	7,25	6,53	26,51	5,26
p= 13	17,54	8,35	16,27	5,58	12,36	7,46	18,06	5,00
p= 14	20,61	5,73	34,25	19,26	7,93	6,90	41,19	13,53
p= 15	9,27	12,05	21,96	8,27	6,12	10,37	23,11	6,90

C. Recorded Defects Details

Table 22
Defects of the process.

Defects		j="Italy"			j = "France"		
		$DH_{j,p}$	$DD_{j,p}$	$TD_{j,p}$	$DH_{j,p}$	$DD_{j,p}$	$TD_{j,p}$
GROUP A	p= 1	0	0	0	0	0	0
	P=2	0	0	0	0	0	0
	p= 3	0	0	0	0	0	0
	p= 4	1	1	2	1	1	2
	p= 5	0	0	0	0	0	0
	p= 6	0	0	0	0	0	0

(*continued on next page*)

Table 22 (*continued*)

Defects		*j="Italy"*			*j = "France"*		
		$DH_{j,p}$	$DD_{j,p}$	$TD_{j,p}$	$DH_{j,p}$	$DD_{j,p}$	$TD_{j,p}$
	p= 7	0	0	0	0	0	0
	p= 8	1	0	1	0	0	0
	p= 9	2	1	3	0	0	0
	p= 10	0	0	0	0	1	1
	p= 11	0	0	0	0	0	0
	p= 12	1	0	1	0	0	0
	p= 13	0	0	0	0	1	1
	p= 14	0	0	0	0	0	0
	p= 15	0	0	0	0	0	0
GROUP B	*p= 1*	2	/	2	0	/	0
	P=2	0	/	0	0	/	0
	p= 3	2	/	2	1	/	1
	p= 4	0	/	0	0	/	0
	p= 5	3	/	3	3	/	3
	p= 6	0	/	0	0	/	0
	p= 7	4	/	4	1	/	1
	p= 8	0	/	0	0	/	0
	p= 9	0	/	0	0	/	0
	p= 10	0	/	0	0	/	0
	p= 11	2	/	2	0	/	0
	p= 12	0	/	0	0	/	0
	p= 13	0	/	0	0	/	0
	p= 14	1	/	1	0	/	0
	p= 15	0	/	0	0	/	0

D. LLM Tokens' count and cost

Table 23
Tokens' count LLM.

GA	*j="Italy"*			*j = "France"*			TT_p
	$IT_{j,p}$	$OT_{j,p}$	$TT_{j,p}$	$IT_{j,p}$	$OT_{j,p}$	$TT_{j,p}$	
p= 1	6840	202	7042	2934	116	3050	10092
P=2	7410	298	7708	5192	222	5414	13122
p= 3	5235	187	5422	5943	212	6155	11577
p= 4	7364	234	7598	10437	299	10736	18334
p= 5	11437	342	11779	4544	142	4686	16465
p= 6	6121	219	6340	5290	205	5495	11835
p= 7	11990	342	12332	5699	194	5893	18225
p= 8	6129	200	6329	5233	161	5394	11723
p= 9	12412	362	12774	6444	240	6684	19458
p= 10	10035	288	10323	6693	209	6902	17225
p= 11	10399	314	10713	5593	162	5755	16468
p= 12	4736	163	4899	4487	152	4639	9538
p= 13	6682	214	6896	5028	189	5217	12113
p= 14	9497	468	9965	7693	302	7995	17960
p= 15	7168	227	7395	5634	188	5822	13217

Table 24
Cost of the experiment.

Operator	Job	Job Cost	Operator Cost (€)
1,00	ITALY	0,10	0,17
	FRANCE	0,08	
2,00	ITALY	0,11	0,19
	FRANCE	0,08	
3,00	ITALY	0,08	0,16
	FRANCE	0,09	
4,00	ITALY	0,10	0,24
	FRANCE	0,14	
5,00	ITALY	0,15	0,22
	FRANCE	0,07	
6,00	ITALY	0,09	0,17
	FRANCE	0,08	
7,00	ITALY	0,15	0,24
	FRANCE	0,08	

(*continued on next page*)

Table 24 (*continued*)

Operator	Job	Job Cost	Operator Cost (€)
8,00	ITALY	0,09	0,16
	FRANCE	0,08	
9,00	ITALY	0,16	0,25
	FRANCE	0,09	
10,00	ITALY	0,13	0,23
	FRANCE	0,09	
11,00	ITALY	0,14	0,22
	FRANCE	0,08	
12,00	ITALY	0,07	0,14
	FRANCE	0,07	
13,00	ITALY	0,09	0,17
	FRANCE	0,08	
14,00	ITALY	0,14	0,25
	FRANCE	0,11	
15,00	ITALY	0,10	0,18
	FRANCE	0,08	
Total cost (€)			2,99

References

Bangor, A., Kortum, P.T., Miller, J.T., 2008. An Empirical Evaluation of the System Usability Scale. Int J. Hum. -Comput. Inter. 24, 574–594. https://doi.org/10.1080/10447310802205776.

Bernabei, M., Colabianchi, S., Falegnami, A., Costantino, F., 2023. Students' use of large language models in engineering education: a case study on technology acceptance, perceptions, efficacy, and detection chances. Comput. Educ. Artif. Intell. 5 https://doi.org/10.1016/j.caeai.2023.100172.

Bousdekis, A., Mentzas, G., Apostolou, D., Wellsandt, S., 2022. Evaluation of AI-based digital assistants in smart manufacturing. IFIP Adv. Inf. Commun. Technol. 664 IFIP, 503–510. https://doi.org/10.1007/978-3-031-16411-8_58.

Bousdekis, A., Wellsandt, S., Bosani, E., Lepenioti, K., Apostolou, D., Hribernik, K., et al., 2021. Human-AI collaboration in quality control with augmented manufacturing analytics. In: Dolgui, A., Bernard, A., Lemoine, D., von Cieminski, G., Romero, D. (Eds.), Adv. Prod. Manag. Syst. Artif. Intell. Sustain. Resilient Prod. Syst. Springer International Publishing, Cham, pp. 303–310. https://doi.org/10.1007/978-3-030-85910-7_32.

Campbell, D.T., Stanley, J.C., 2015. Experimental and Quasi-Experimental Designs for Research. Ravenio Books.

Carvalho, A.V., Chouchene, A., Lima, T.M., Charrua-Santos, F., 2020. Cognitive manufacturing in industry 4.0 toward cognitive load reduction: a conceptual framework. Appl. Syst. Innov. 3, 55. https://doi.org/10.3390/asi3040055.

Casas, W.J.P., Cordeiro, E.P., Mello, T.C., Zannin, P.H.T., 2014. Noise mapping as a tool for controlling industrial noise pollution. J. Sci. Ind. Res 73, 262–266.

Chen, T.Y., Chiu, Y.C., Bi, N., Tsai, R.T.H., 2021. Multi-modal chatbot in intelligent manufacturing. IEEE Access. https://doi.org/10.1109/ACCESS.2021.3083518.

Church, K.W., Yue, R., 2023. Emerging trends: smooth-talking machines. Nat. Lang. Eng. 29, 1402–1410. https://doi.org/10.1017/S1351324923000463.

Colabianchi, S., Bernabei, M., Costantino, F., 2022. Chatbot for training and assisting operators in inspecting containers in seaports, vol. 64, 6–13. https://doi.org/10.1016/j.trpro.2022.09.002.

Colabianchi, S., Tedeschi, A., Costantino, F., 2023. Human-technology integration with industrial conversational agents: a conceptual architecture and a taxonomy for manufacturing. J. Ind. Inf. Integr. 35 https://doi.org/10.1016/j.jii.2023.100510.

Costantino, F., Falegnami, A., Fedele, L., Bernabei, M., Stabile, S., Bentivenga, R., 2021. New and emerging hazards for health and safety within digitalized manufacturing systems. Sustainability 13, 10948. https://doi.org/10.3390/su131910948.

Dinan E., Abercrombie G., Bergman A., Spruit S., Hovy D., Boureau Y.-L., et al. SafetyKit: First Aid for Measuring Safety in Open-domain Conversational Systems. In: Muresan S, Nakov P, Villavicencio A, editors. Proc. 60th Annu. Meet. Assoc. Comput. Linguist. Vol. 1 Long Pap., Dublin, Ireland: Association for Computational Linguistics; 2022, p. 4113–33. https://doi.org/10.18653/v1/2022.acl-long.284.

Fan, H., Liu, X., Fuh, J.Y.H., Lu, W.F., Li, B., 2024. Embodied intelligence in manufacturing: leveraging large language models for autonomous industrial robotics. J. Intell. Manuf. https://doi.org/10.1007/s10845-023-02294-y.

Freire, S.K., Panicker, S.S., Ruiz-Arenas, S., Rusak, Z., Niforatos, E., 2023. A Cognitive assistant for operators: AI-powered knowledge sharing on complex systems. IEEE Pervasive Comput. 22, 50–58. https://doi.org/10.1109/MPRV.2022.3218600.

Gladysz, B., Tran, T., Romero, D., van Erp, T., Abonyi, J., Ruppert, T., 2023. Current development on the Operator 4.0 and transition towards the Operator 5.0: A systematic literature review in light of Industry 5.0. J. Manuf. Syst. 70, 160–185. https://doi.org/10.1016/j.jmsy.2023.07.008.

Gong, L., Fast-Berglund, Å., Johansson, B., 2021. A framework for extended reality system development in manufacturing. IEEE Access 9, 24796–24813. https://doi.org/10.1109/ACCESS.2021.3056752.

Hart, S.G., Staveland, L.E., 1988. Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. Adv. Psychol. 52, 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9.

Hoedt, S., Claeys, A., Van Landeghem, H., Cottyn, J., 2017. The evaluation of an elementary virtual training system for manual assembly. Int J. Prod. Res 55, 7496–7508. https://doi.org/10.1080/00207543.2017.1374572

Holmes, S., Bond, R., Moorhead, A., Zheng, J., Coates, V., McTear, M., 2023. Towards validating a chatbot usability scale. In: Marcus, A., Rosenzweig, E., Soares, M.M. (Eds.), Des. User Exp. Usability, vol. 14033. Springer Nature Switzerland, Cham, pp. 321–339. https://doi.org/10.1007/978-3-031-35708-4_24.

Holmes, S., Moorhead, A., Bond, R., Zheng, H., Coates, V., Mctear, M., 2019. Usability testing of a healthcare chatbot: can we use conventional methods to assess conversational user interfaces? Proc. 31st Eur. Conf. Cogn. Ergon. BELFAST United. ACM, Kingdom, pp. 207–214. https://doi.org/10.1145/3335082.3335094.

Kernan Freire, S., Niforatos, E., Wang, C., Ruiz-Arenas, S., Foosherian, M., Wellsandt, S., et al., 2023. Lessons learned from designing and evaluating CLAICA: A continuously learning AI cognitive assistant. Proc. 28th Int. Conf. Intell. User Interfaces, Sydney NSW. ACM, Australia, pp. 553–568. https://doi.org/10.1145/3581641.3584042.

Korzynski, P., Mazurek, G., Krzypkowska, P., Kurasinski, A., 2023. Artificial intelligence prompt engineering as a new digital competence: Analysis of generative AI technologies such as ChatGPT. Entrep. Bus. Econ. Rev. 11, 25–37. https://doi.org/10.15678/EBER.2023.110302.

Lall, M., Torvatn, H., Seim, E.A., 2017. Towards industry 4.0: Increased need for situational awareness on the shop floor. IFIP Adv. Inf. Commun. Technol. 513, 322–329. https://doi.org/10.1007/978-3-319-66923-6_38.

LangChain n.d. https://python.langchain.com/docs/get_started/introduction.

Langchain. Load_qa_chain 2023. https://api.python.langchain.com/en/latest/chains/langchain.chains.qa_with_sources.loading.load_qa_with_sources_chain.html#langchain.chains.qa_with_sources.loading.load_qa_with_sources_chain.

Langchain. RetrievalQA chain n.d. https://api.python.langchain.com/en/latest/chains/langchain.chains.retrieval_qa.base.RetrievalQA.html#.

Le, N.-T., Wartschinski, L., 2018. A Cognitive Assistant for improving human reasoning skills. Int J. Hum. -Comput. Stud. 117, 45–54. https://doi.org/10.1016/j.ijhcs.2018.02.005.

Lee K., Jo J., Kim J., Kang Y. Can Chatbots Help Reduce the Workload of Administrative Officers? - Implementing and Deploying FAQ Chatbot Service in a University. In: Stephanidis C, editor. HCI Int. 2019 - Posters, Cham: Springer International Publishing; 2019, p. 348–54. https://doi.org/10.1007/978-3-030-23522-2_45.

Li, C., Chrysostomou, D., Pinto, D., Hansen, A.K., Bøgh, S., Madsen, O., 2023. Hey max, can you help me? An intuitive virtual assistant for industrial robots. Appl. Sci. Switz. 13 https://doi.org/10.3390/app13010205.

Li, C., Zhang, X., Chrysostomou, D., Yang, H., 2022. ToD4IR: a humanised task-oriented dialogue system for industrial robots. IEEE Access 10, 91631–91649. https://doi.org/10.1109/ACCESS.2022.3202554.

Lo, L.S., 2023. The art and science of prompt engineering: a new literacy in the information age. Internet Ref. Serv. Q 27, 203–210. https://doi.org/10.1080/10875301.2023.2227621.

Longo, F., Padovano, A., 2020. Voice-enabled assistants of the operator 4.0 in the social smart factory: prospective role and challenges for an advanced human–machine interaction. Manuf. Lett. 26, 12–16. https://doi.org/10.1016/j.mfglet.2020.09.001.

Lu, Y., Zheng, H., Chand, S., Xia, W., Liu, Z., Xu, X., et al., 2022. Outlook on human-centric manufacturing towards Industry 5.0. J. Manuf. Syst. 62, 612–627. https://doi.org/10.1016/j.jmsy.2022.02.001.

Ludwig, H., Schmidt, T., Kühn, M., 2023. Voice user interfaces in manufacturing logistics: a literature review. Int J. Speech Technol. 26, 627–639. https://doi.org/10.1007/s10772-023-10036-x.

Maxwell, S.E., Delaney, H.D., Kelley, K., 2017. Designing Experiments and Analyzing Data: A Model Comparison Perspective, Third Edition. Routledge.

Melluso, N., Grangel-González, I., Fantoni, G., 2022. Enhancing Industry 4.0 standards interoperability via knowledge graphs with natural language processing. Comput. Ind. 140, 103676 https://doi.org/10.1016/j.compind.2022.103676.

Meta. FAISS 2017.

Montgomery, D.C., 2017. Design and Analysis of Experiments. John Wiley & Sons.

Neumann, W.P., Winkelhaus, S., Grosse, E.H., Glock, C.H., 2021. Industry 4.0 and the human factor – a systems framework and analysis methodology for successful development. Int J. Prod. Econ. 233, 107992 https://doi.org/10.1016/J.IJPE.2020.107992.

OpenAI. GPT-4 Technical Report 2023a.

OpenAI. G.P.T. models 2023a.

OpenAI. OpenAI Text to Speech 2023b. https://platform.openai.com/docs/guides/speech-to-text.

OpenAI. OpenAI Pricing 2023. https://openai.com/pricing.

Ordieres-Meré, J., Gutierrez, M., Villalba-Díez, J., 2023. Toward the industry 5.0 paradigm: Increasing value creation through the robust integration of humans and machines. Comput. Ind. 150, 103947 https://doi.org/10.1016/j.compind.2023.103947.

Polak-Sopinska, A., Wisniewski, Z., Walaszczyk, A., Maczewska, A., Sopinski, P., 2019. Impact of industry 4.0 on occupational health and safety. Adv. Intell. Syst. Comput. 971, 40–52. https://doi.org/10.1007/978-3-030-20494-5_4.

Popović, M., Ney, H., 2007. Word error rates: decomposition over Pos classes and applications for error analysis. Proc. Second Workshop Stat. Mach. Transl. - StatMT 07, Prague, Czech Repub.: Assoc. Comput. Linguist. 48–55. https://doi.org/10.3115/1626355.1626362.

Pypi. Python Speech Recognition 2023. https://pypi.org/project/SpeechRecognition/.

Roldán, J.J., Crespo, E., Martín-Barrio, A., Peña-Tapia, E., Barrientos, A., 2019. A training system for Industry 4.0 operators in complex assemblies based on virtual reality and process mining. Robot Comput.-Integr. Manuf. 59, 305–316. https://doi.org/10.1016/j.rcim.2019.05.004.

Ruiz, E., Torres, M.I., del Pozo, A., 2023. Question answering models for human–machine interaction in the manufacturing industry. Comput. Ind. 151, 103988 https://doi.org/10.1016/j.compind.2023.103988.

Schmidhuber J., Schlögl S., Ploder C. Cognitive Load and Productivity Implications in Human-Chatbot Interaction. 2021 IEEE 2nd Int. Conf. Hum.-Mach. Syst. ICHMS, 2021, p. 1–6. https://doi.org/10.1109/ICHMS53169.2021.9582445.

Schrepp, M., Hinderks, A., Thomaschewski, J., 2014. Applying the user experience questionnaire (UEQ) in different evaluation scenarios. In: Marcus, A. (Ed.), Des. User Exp. Usability Theor. Methods Tools Des. User Exp., vol. 8517. Springer International Publishing, Cham, pp. 383–392. https://doi.org/10.1007/978-3-319-07668-3_37.

Sørensen, F., Mattsson, J., Sundbo, J., 2010. Experimental methods in innovation research. Res Policy 39, 313–322. https://doi.org/10.1016/j.respol.2010.01.006.

Trappey, A.J.C., Trappey, C.V., Chao, M.-H., Wu, C.-T., 2022. VR-enabled engineering consultation chatbot for integrated and intelligent manufacturing services. J. Ind. Inf. Integr. 26 https://doi.org/10.1016/j.jii.2022.100331.

Wang, X., Anwer, N., Dai, Y., Liu, A., 2023. ChatGPT for design, manufacturing, and education, vol. 119, 7–14. https://doi.org/10.1016/j.procir.2023.04.001.

Wellsandt, S., Foosherian, M., Bousdekis, A., Lutzer, B., Paraskevopoulos, F., Verginadis, Y., et al., 2023. Fostering Human-AI collaboration with digital intelligent assistance in manufacturing SMEs. IFIP Adv. Inf. Commun. Technol. 689 AICT, 649–661. https://doi.org/10.1007/978-3-031-43662-8_46.

Wellsandt, S., Klein, K., Hribernik, K., Lewandowski, M., Bousdekis, A., Mentzas, G., et al., 2021. Towards using digital intelligent assistants to put humans in the loop of predictive maintenance systems, vol. 54, 49–54. https://doi.org/10.1016/j.ifacol.2021.08.005.

Xia, L., Li, C., Zhang, C., Liu, S., Zheng, P., 2024. Leveraging error-assisted fine-tuning large language models for manufacturing excellence. Robot Comput.-Integr. Manuf. 88 https://doi.org/10.1016/j.rcim.2024.102728.

Xia Y., Shenoy M., Jazdi N., Weyrich M. Towards autonomous system: Flexible modular production system enhanced with large language model agents. vol. 2023-September, 2023. https://doi.org/10.1109/ETFA54631.2023.10275362.