TYPE Brief Research Report
PUBLISHED 27 March 2024
DOI 10.3389/frai.2024.1293084

Check for updates

OPEN ACCESS

EDITED BY
David Romero,
Monterrey Institute of Technology and Higher Education (ITESM), Mexico

REVIEWED BY
Elena Bellodi,
University of Ferrara, Italy
Anisa Rula,
University of Brescia, Italy

*CORRESPONDENCE
Samuel Kernan Freire
✉ s.kernanfreire@tudelft.nl

RECEIVED 12 September 2023
ACCEPTED 14 March 2024
PUBLISHED 27 March 2024

CITATION
Kernan Freire S, Wang C, Foosherian M, Wellsandt S, Ruiz-Arenas S and Niforatos E (2024) Knowledge sharing in manufacturing using LLM-powered tools: user study and model benchmarking.
Front. Artif. Intell. 7:1293084.
doi: 10.3389/frai.2024.1293084

Knowledge sharing in manufacturing using LLM-powered tools: user study and model benchmarking

Samuel Kernan Freire[1]*, Chaofan Wang[1], Mina Foosherian[2], Stefan Wellsandt[2], Santiago Ruiz-Arenas[3] and Evangelos Niforatos[1]

[1]Faculty of Industrial Design Engineering, Delft University of Technology, Delft, Netherlands, [2]BIBA—Bremer Institut für Produktion und Logistik GmbH, Bremen, Germany, [3]Grupo de Investigación en Ingeniería de Diseño (GRID), Universidad EAFIT - Escuela de Administración, Finanzas e Instituto Tecnológico, Medellin, Colombia

Recent advances in natural language processing enable more intelligent ways to support knowledge sharing in factories. In manufacturing, operating production lines has become increasingly knowledge-intensive, putting strain on a factory's capacity to train and support new operators. This paper introduces a Large Language Model (LLM)-based system designed to retrieve information from the extensive knowledge contained in factory documentation and knowledge shared by expert operators. The system aims to efficiently answer queries from operators and facilitate the sharing of new knowledge. We conducted a user study at a factory to assess its potential impact and adoption, eliciting several perceived benefits, namely, enabling quicker information retrieval and more efficient resolution of issues. However, the study also highlighted a preference for learning from a human expert when such an option is available. Furthermore, we benchmarked several commercial and open-sourced LLMs for this system. The current state-of-the-art model, GPT-4, consistently outperformed its counterparts, with open-source models trailing closely, presenting an attractive option given their data privacy and customization benefits. In summary, this work offers preliminary insights and a system design for factories considering using LLM tools for knowledge management.

KEYWORDS

natural language interface, benchmarking, Large Language Models, factory, industrial settings, industry 5.0, knowledge sharing, information retrieval

1 Introduction

Human-centric manufacturing seeks to harmonize the strengths of humans and machines, aiming to enhance creativity, human wellbeing, problem-solving abilities, and overall productivity within factories (May et al., 2015; Fantini et al., 2020; Alves et al., 2023). Despite these advancements, a significant challenge persists in effectively managing and utilizing the vast knowledge generated within these manufacturing environments, such as issue reports and machine documentation (Gröger et al., 2014). This knowledge is crucial for optimizing operations, yet it remains largely untapped due to the difficulties in processing and interpreting the disconnected, sometimes unstructured, technical information it contains (Edwards et al., 2008; Leoni et al., 2022).

Traditionally, leveraging this knowledge has been cumbersome, with operators choosing to use personal smartphones over official procedures (Richter et al., 2019) and AI unable to handle the complexity of the data (Edwards et al., 2008). However, recent Large Language Models (LLMs) like GPT-4 show promise in addressing these challenges. LLMs can effectively interpret, summarize, and retrieve information from vast text-based datasets (Lewis et al., 2020) while concurrently aiding the capture of new knowledge (Kernan Freire et al., 2023b). These capabilities could significantly support operators in knowledge-intensive tasks, making it easier to access relevant information, share new knowledge, and make informed decisions rapidly.

While LLMs offer promising capabilities, their application in manufacturing is not straightforward. The specific, dynamic knowledge required in this domain poses unique challenges (Feng et al., 2017). For instance, a foundational LLM may have limited utility in a factory setting without significant customization, such as fine-tuning or incorporating specific context information into its prompts (Wang Z. et al., 2023). Additionally, the practical and socio-technical risks and challenges of deploying LLMs in such environments remain largely unexplored—factors key to human-centered AI (Shneiderman, 2022). Concerns include the accuracy of the information provided, the potential for “hallucinated" answers (Zuccon et al., 2023), and the need for systems that can adapt to the highly specialized and evolving knowledge base of a specific manufacturing setting (Feng et al., 2017).

In response to these challenges, we developed an LLM-powered tool to leverage factory documents and issue analysis reports to answer operators’ queries. Furthermore, the tool facilitates the analysis and reporting of new issues. This tool demonstrates the feasibility of using LLMs to enhance knowledge management in manufacturing settings. To understand its effectiveness and potential, we conducted a user study in a factory environment, evaluating the system’s usability, user perceptions, adoption, and impact on factory operations.

Our approach also addresses the lack of specific benchmarks for evaluating LLMs in manufacturing. We benchmarked several LLMs, including both closed and open-source models, recognizing that the standard benchmarks[1] primarily focus on general knowledge and reasoning. As such, they may not adequately reflect the challenges of understanding manufacturing-specific terminology and concepts. This benchmarking focused on their ability to utilize factory-specific documents and unstructured issue reports to provide factual and complete answers to operators’ queries.

2 Background

In this section, we address the topic of industry 5.0, LLM-powered tools for knowledge management, benchmarking LLMs, and the research questions informing this work.

2.1 Human-centered manufacturing

Industry 5.0, the latest phase of industrial development, places human beings at the forefront of manufacturing processes, emphasizing their skills, creativity, and problem-solving abilities (Xu et al., 2021; Maddikunta et al., 2022; Alves et al., 2023). Human-centered manufacturing in Industry 5.0 focuses on providing a work environment that nurtures individuals’ creativity and problem-solving capabilities (Maddikunta et al., 2022). It encourages workers to think critically, innovate, and continuously learn. With machines handling repetitive and mundane tasks, human workers can dedicate their time and energy to more complex and intellectually stimulating activities. This shift could enhance job satisfaction and promote personal and professional growth, as workers could acquire new skills and engage in higher-level decision-making (Xu et al., 2021; Alves et al., 2023). Emphasis on human-machine collaboration and the continuous emergence and refinement of technology increases the need for adequate human-computer interaction (Brückner et al., 2023). One of the approaches to address this topic is using conversational AI to assist humans in manufacturing (Wellsandt et al., 2021).

2.2 LLM-powered knowledge management tools

Training Large Language Models (LLMs) on numerous, diverse texts results in the embedding of extensive knowledge (Zhao et al., 2023). LLMs can also adeptly interpret complex information (Jawahar et al., 2019), general reasoning (Wei et al., 2022a), and aiding knowledge-intensive decision-making. Consequently, researchers have been exploring applying LLM-powered tools in domain-specific tasks (Wen et al., 2023; Xie T. et al., 2023; Zhang W. et al., 2023).

Despite their potential benefits, the responses generated by LLMs may have two potential issues: (1) outdated information originating from the model’s training date, and (2) inaccuracies in factual representation, known as “hallucinations” (Bang et al., 2023; Zhao et al., 2023). To address these challenges and leverage the capabilities of LLMs in domain-specific knowledge-intensive tasks, several techniques can be used, such as chain-of-thought (Wei et al., 2022b), few-shot prompting (Brown et al., 2020; Gao et al., 2021), and retrieval augmented generation (Lewis et al., 2020).

Using few-shot prompting to retrieve information across diverse topics, Semnani et al. (2023) introduced an open-domain LLM-powered chatbot called WikiChat. WikiChat utilizes a 7-stage pipeline of few-shot prompted LLM that suggests facts verified against Wikipedia, retrieves additional up-to-date information, and generates coherent responses. They used a hybrid human-and-LLM method to evaluate the chatbot on different topics for factuality, alignment with real-worth truths and verifiable facts, and conversationality. This compound metric scores how informational, natural, non-repetitive, and temporally correct the response is. Their solution significantly outperforms GPT-3.5 in factuality, with an average improvement of 24.4% while staying on par in conversationality. Others have explored the capabilities of LLMs in domain-specific tasks such as extracting

[1] https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard (accessed February 26, 2024).

structured data from unstructured healthcare texts (Tang et al., 2023), providing medical advice (Nov et al., 2023), simplifying radiology reports (Jeblick et al., 2023), Legal Judgement Prediction from multilingual legal documents (Trautmann et al., 2022), and scientific writing (Alkaissi and McFarlane, 2023).

Several manufacturers are cautiously adopting LLMs, while seeking solutions to mitigate their associated risks. For example,[2] used AI with ChatGPT integrated through Azure OpenAI Service to enhance quality management and process optimization in vehicle production. This AI-driven approach simplifies complex evaluations for quality engineers through dialogue-based queries. Xia et al. (2023) demonstrated how using in-context learning and injecting task-specific knowledge into an LLM can streamline intelligent planning and control of production processes. Kernan Freire et al. (2023a) built a proof of concept for bridging knowledge gaps among workers by utilizing domain-specific texts and knowledge graphs. Wang X. et al. (2023) conducted a systematic test of ChatGPT's responses to 100 questions from course materials and industrial documents. They used a zero-shot method and examined the responses' correctness, relevance, clarity, and comparability. Their results suggested areas for improvement, including low scores when responding to critical analysis questions, occasional non-factual or out-of-manufacturing scope responses, and dependency on query quality. Although Wang X. et al. (2023) provides a comprehensive review of ChatGPT's abilities to answer questions related to manufacturing; it did not include the injection of task-specific knowledge into the prompts.

To improve the performance of an LLM for domain-specific tasks, relevant context information can be automatically injected along with a question prompt. This technique, known as Retrieval Augmented Generation (RAG), involves searching a corpus for information relevant to the user's query and inserting it into a query template before sending it to the LLM (Lewis et al., 2020). Using RAG also enables further transparency and explainability of the LLM's response. Namely, users can check the referenced documents to verify the LLM's response. Factories will likely have a large corpus of knowledge available in natural language, such as standard work instructions or machine manuals. Furthermore, factory workers continually add to the pool of available knowledge through (issue) reports. Until recently, these reports were considered unusable by AI natural language processing due to quality issues such as poorly structured text, inconsistent terminology, or incompleteness (Edwards et al., 2008; Müller et al., 2021). However, the leap in natural language understanding that LLMs, such as ChatGPT, have brought about can overcome these issues.

2.3 Evaluating LLMs

Large Language Model evaluation requires the definition of evaluation criteria, metrics, and datasets associated with the system's main tasks. There are two types of LLM evaluations: intrinsic and extrinsic evaluation. Intrinsic evaluation focuses on the internal properties of a Language Model (Wei et al., 2023). It means the patterns and language structures learned during the pre-training phase. Extrinsic evaluation focuses on the model's performance in downstream tasks, i.e., in the execution of specific tasks that make use of the linguistic knowledge gained upstream, like code completion (Xu et al., 2022). Despite extrinsic evaluation being computationally expensive, only conducting intrinsic evaluation is not comprehensive, as it only tests the LLMs capability for memorization (Jang et al., 2022). Here, we focus on extrinsic evaluation as we are primarily interested in the performance of LLM-based tools for specific real-world tasks.

Extrinsic evaluation implies assessing the systems's performance in tasks such as question answering, translation, reading comprehension, and text classification, among others (Kwon and Mihindukulasooriya, 2022). Existing benchmarks such as LAMBADA, HellaSwag, TriviaQA, BLOOM, Galactica, ClariQ and MMLU, among others, are widely reported in the literature for comparing language models. Likewise, domain-specific Benchmarks for tasks such as medical (Singhal et al., 2023), fairness evaluation (Zhang J. et al., 2023), finance (Xie Q. et al., 2023), robot policies (Liang et al., 2022), and 3D printing code generation (Badini et al., 2023) can also be found. Experts also evaluate the performance of large-language models (LLMs) in specific downstream tasks, such as using physicians to evaluate the output of medical specific LLMs (Singhal et al., 2023).

LLM benchmarks range from specific downstream tasks to general language tasks. However, to our knowledge, LLMs have not been benchmarked for answering questions in the manufacturing domain based on context material, a technique known as Retrieval Augmented Generation (Lewis et al., 2020). Material such as machine documentation, standard work instructions, or issue reports will contain domain jargon and technical information that LLMs may struggle to process. Furthermore, the text in an issue report may pose additional challenges due to abbreviations, poor grammar, and formatting (Edwards et al., 2008; Oruç, 2020; Müller et al., 2021). Therefore, as part of this work, we benchmarked several LLMs on their ability to answer questions based on factory manuals and unstructured issue reports. Furthermore, we conducted a user study with factory operators and managers to assess the potential benefits, risks and challenges. The following research questions informed our study:

1. *What are the perceived benefits, challenges, and risks of using Large Language Models for information retrieval and knowledge sharing for factory operators?*
2. *How do Large Language Models compare in performance when answering factory operators' queries based on factory documentation and unstructured issue reports?* We consider performance as the factuality, completeness, hallucinations, and conciseness of the generated response.

3 System

We built a fully functional system to assess the potential of using LLMs for information retrieval and knowledge sharing for factory operators. Benefiting from LLMs' in-context learning capabilities, we use this to supply an LLM with information in the form of factory manuals, and issue reports relevant to the user's question, a technique known as Retrieval Augmented Generation

[2] https://group.mercedes-benz.com/innovation/digitalisation/industry-4-0/chatgpt-in-vehicle-production.html (accessed February 26, 2024).

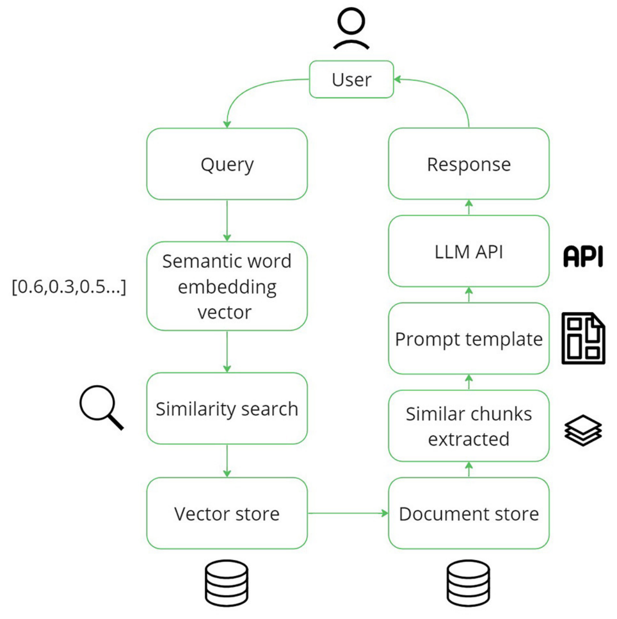


FIGURE 1
The steps of Retrieval Augmented Generation (RAG) from user query to response.

(RAG) (Lewis et al., 2020), see Figure 1. As noted by Wei et al. (2022a), training LLMs using a prompt packed with query-related information can yield substantial performance enhancement. Users can ask questions in the chat box by typing or using voice input. The response is displayed at the top of the page, and the document chunks used for the answer can be checked at the bottom (see Figure 2).

3.1 Tool dependencies

The tool was constructed utilizing two innovative technologies—Gradio and LlamaIndex. Gradio, a tool developed by Abid et al. (2019), serves as the backbone for both our front and back ends. Primarily used to simplify the development and distribution of machine learning applications, Gradio allows the quick creation of intuitive, user-friendly web interfaces for machine learning models.

Additionally, we use LlamaIndex, created by Liu (2022), for retrieving the context material in response to the user queries and handling the queries to the LLM. LlamaIndex, initially known as GPT Index, is a cutting-edge data framework designed for the efficient handling and accessibility of private or domain-specific data in LLMs applications.

Since the factory documents can be long, they may overflow the LLM's context window or result in unnecessary computational demand. To overcome this, we segment the materials into manageable chunks, each comprising ∼400 tokens. This method effectively incorporates the materials into the LLM prompt without compromising the conversation flow. Following the segmentation, each document chunk is processed through LlamaIndex using the OpenAI Embedding API.[3] Utilizing the "text-embedding-ada-002" model, LlamaIndex transforms each chunk into a corresponding embedding vector. These resulting vectors are then securely stored, ready for future retrieval and use.

3.2 Knowledge base construction

Our experiment incorporates two distinct types of domain-specific data: factory manuals and shared knowledge from factory workers. Factory manuals outline information on machine operation, safety protocols, quality assurance, and more. These resources, provided by factory management teams, initialize the knowledge base for each specific factory. The materials come in various formats, including PDF, Word, and CSV files.

In addition to the factory manuals, we integrate issue analysis reports from factory workers. This information is gathered from the production line, utilizing the five-why process, an iterative root-cause analysis technique (Serrat, 2017) (right side of Figure 2). The five-why technique probes into cause-and-effect relationships underlying specific problems by repeatedly asking "Why?" until the root cause is revealed, typically by the fifth query. This process enables us to gather real-world issues encountered on production lines, which may not be covered in the factory manuals. Upon entering all required information, including one or more "whys", the operator presses "check", triggering a prompt to the LLM that performs a logical check of the entered information and checks for inconsistencies with previously reported information. The operator can revise the entered information and submit it as is. Then, the submitted report will be added to a queue for expert operators to check before it is added to the knowledge base.

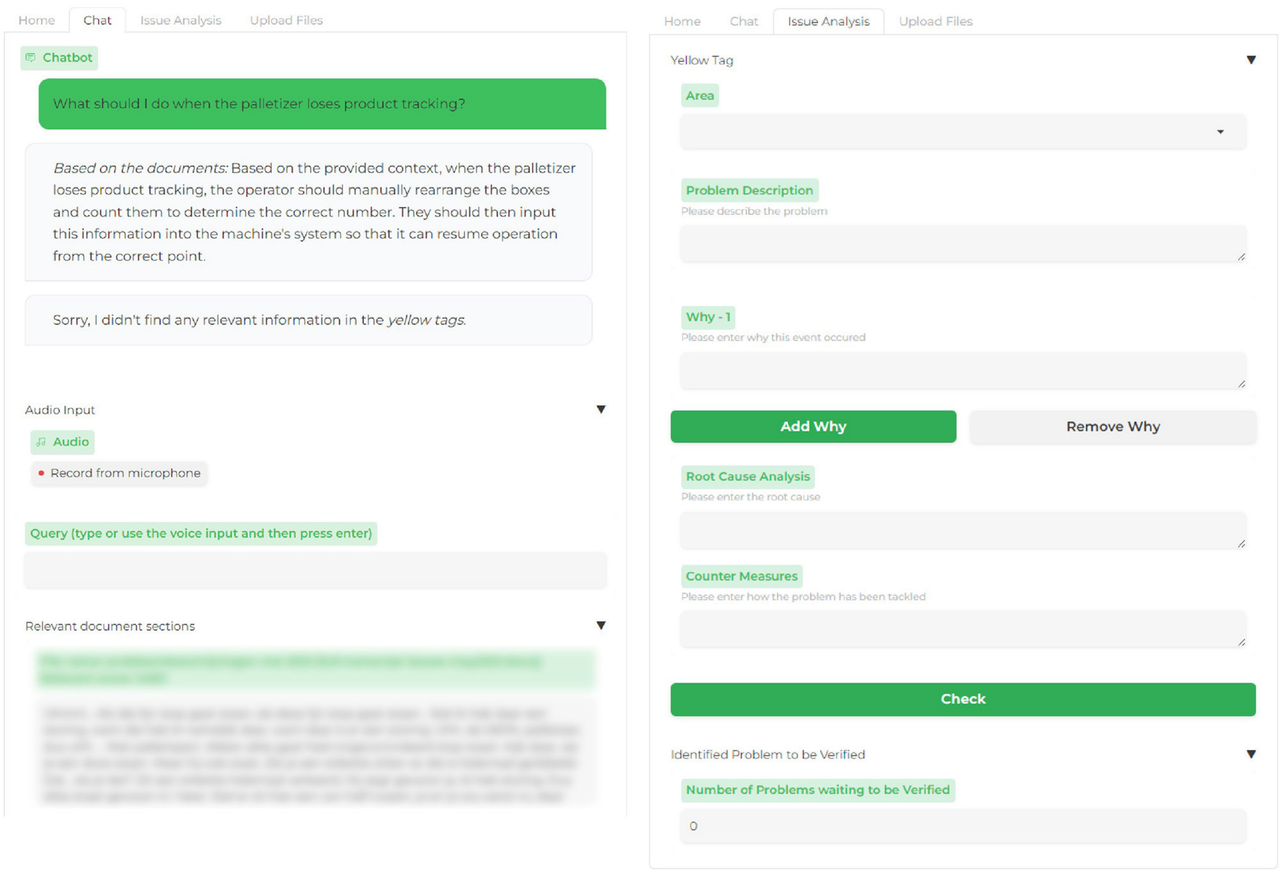
3.3 Query construction

To retrieve the document data relevant to specific user queries, we employ the same embedding model, "text-embedding-ada-002" to generate vector representations of these queries. By leveraging the similarity calculation algorithm provided by LlamaIndex, we can identify and retrieve the top-K most similar segmented document snippets related to the user query. This allows us to construct pertinent LLM queries. Once the snippets are retrieved, they are synthesized into the following query template based on the templates used by LlamaIndex[4]:

> You are an assistant that assists detergent production line operators with decision support and advice based on a knowledge base of standard operating procedures, single point lessons (SPL), etc. We have provided context information below from relevant documents and reports.
>
> ---
>
> [Retrieved Document Snippets]
>
> ---
>
> Given this information, please answer the following question: [Query]

[3] https://api.openai.com/v1/embeddings (accessed February 26, 2024).

[4] https://docs.llamaindex.ai (accessed February 26, 2024).



What should I do when the palletizer loses product tracking?

Based on the documents: Based on the provided context, when the palletizer loses product tracking, the operator should manually rearrange the boxes and count them to determine the correct number. They should then input this information into the machine's system so that it can resume operation from the correct point.

Sorry, I didn't find any relevant information in the *yellow tags*.

Please describe the problem

Please enter why this event occured

Please enter the root cause

Please enter how the problem has been tackled

0

FIGURE 2
The main screens for the tool's interface are the chat interface and issue analysis screen. The "relevant document sections" part is blurred for confidentiality as it shows the title of a company's document and its content.

> If the provided context does not include relevant information to answer the question, please do not respond.

However, considering our data originates from two distinct sources—factory manuals and shared tactical knowledge—we have decided to segregate these into two separate LLM queries. This approach is designed to prevent potential user confusion from combining data from both sources into a single query.

4 User study in the field

We conducted a user study on the system to uncover perceived benefits, usability issues, risks, and barriers to adoption. The study comprised four tasks: (1) ask the system several questions about how to solve a specific production issue and/or perform a standard procedure, (2) complete a "yellow tag" (issue analysis report) based on a recent issue, (3) request a logical check of the completed report, and finally, (4) upload new documents to the system. After each task, they were asked to provide feedback. Then, after completing all tasks, the participants were posed several open questions about the system's benefits, risks, and barriers to adoption. Finally, demographic information, such as age, gender, and role, was collected.

4.1 Participants

We recruited $N = 9$ participants from a detergent factory, of which $n = 4$ were managers (P1-4), and $n = 5$ were operators (P5-9). Of the nine participants, $n = 3$ were women, and $n = 6$ were men. Participant age was distributed over three brackets, namely $n = 2$ were 30–39, $n = 4$ were 40–49, and $n = 3$ were 50–59.

4.2 Qualitative analysis

An inductive thematic analysis (Guest et al., 2011) of the answers to the open questions resulted in six themes discussed below.

- **Usability**: the theme of usability underlines the system's ease of use and the need for clear instructions. Users mentioned the necessity for a "user-friendly" (P2) interface and highlighted the importance of having "more instructions and more details need to be loaded" (P1) to avoid confusion. This indicates a desire for intuitive navigation that could enable workers to use the system effectively without extensive training or referencing external help. The feedback suggests that the system already

works well, as reflected in statements like "Easy-to-use system" (P3) and the system "works well" (P7).
- **Access to information**: users appreciated the "ease of having information at hand" (P1), facilitating immediate access to necessary documents. However, there is a clear call for improvements, such as the ability to "Include the possibility of opening IO, SPL, etc. in pdf format for consultation" (P3). This theme is supported by requests for direct links to full documents, suggesting that while "the list of relevant documents from which information is taken is excellent" (P4), the ability to delve deeper into full documents would significantly enhance the user experience.
- **Efficiency**: users value the "greater speed in carrying out some small tasks" (P3). However, there are concerns about the system's efficiency when it does not have the answer, leading to "wasting time looking for a solution to a problem in case it is not reported in the system's history" (P3). Statements like "quick in responses" (P3) contrast with the need for questions to be "too specific to have a reliable answer" (P7), indicating tension between the desire for quick solutions and the system's limitations.
- **Adoption**: users highlight several factors affecting adopting the new system. It includes challenges such as "awareness and training of operators [might hinder adoption]" (P3) and the need for "acceptance by all employees" (P4), which indicates that the system's success is contingent on widespread user buy-in. The generational divide is also noted: "That older operators use it [on what may hinder adoption]" (P7) suggests that demographic factors may influence the acceptance of new technology.
- **Safety**: a manager expressed apprehension that "if the responses are not adequate, you risk safety" (P1), emphasizing the critical nature of reliable information in a high-risk factory setting. Beyond information being outdated or useless, the possibility of "hallucinated" responses leading to dangerous situations in a factory that processes chemicals is especially concerning.
- **Traditional vs. novel**: there is a noticeable preference for established practices among some users. For instance, "It's faster and easier to ask an expert colleague working near me rather than [the system]" (P8) captures the reliance on human expertise over the assistant system. This tension is further demonstrated by the sentiment that "Operators may benefit more from traditional information retrieval systems" (P9), suggesting a level of skepticism or comfort with the status quo that the new system needs to overcome.

5 LLM benchmarking

In our benchmarking experiment, we evaluated various commercial and open-source LLMs, including OpenAI's ChatGPT (GPT-3.5 and GPT-4 from July 20th 2023), Guanaco 65B and 35B variants (Dettmers et al., 2023) based on Meta's Llama (Large Language Model Meta AI) (Touvron et al., 2023), Mixtral 8x7b (Jiang et al., 2024), Llama 2 (Touvron et al., 2023), and one of its derivatives, StableBeluga2[5]. This selection represents the state-of-the-art closed-sourced models (e.g., GPT-4) and open-source models (e.g., Llama 2). We included the (outdated) Guanaco models to demonstrate the improvements in the open-source sphere over the past year.

We used a web UI for LLMs[6] to load and test the Mixtral 8x7B, Guanaco models, and the StableBeluga2. The models were loaded on a pair of Nvidia A6000s with NVlink and a total Video Random Access Memory (VRAM) capacity of 96 GB. The 65B model was run in 8-bit mode to fit in the available VRAM. We used the llama-precise parameter preset and fixed zero seed for reproducibility. Llama 2 was evaluated using the demo on huggingface.[7]

To rigorously assess the models, we prepared 20 questions of varying complexity based on two types of context material: half from operating manuals and half from unstructured issue reports. The operating manuals included excerpts from actual machine manuals and standard operating procedures, while the informal issue reports were free-text descriptions of issues we had previously collected from operators. The model prompt was constructed using the above template (3.3). Ultimately, the difficulty of a question is a combination of the question's complexity and the clarity of the source material. Simple questions include retrieving a single piece of information clearly stated in the context material, for example, "At what temperature is relubrication necessary for the OKS 4220 grease?". Conversely, difficult questions require more reasoning or comprise multiple parts, for example, "What should I do if the central turntable is overloaded?" which has a nuanced answer dependent on several factors not clearly articulated in the context material.

In addition to measuring response length in words, every response is manually scored on factuality, completeness, and hallucinations as defined below:

- **Factuality**: responses align with the facts in the context material.
- **Completeness**: responses contain all the information relevant to the question in the context material.
- **Hallucinations**: response appears grammatically and semantically coherent but is not based on the context material.

The following scoring protocol is applied: one is awarded for a completely factual, complete, or hallucinated response. In contrast, a score of 0.5 is awarded for a slightly nonfactual, incomplete, or hallucinated response (e.g., the response includes four out of the five correct steps). Otherwise, a score of zero is awarded. Therefore, wrong answers are penalized heavily. If the model responds by saying it cannot answer the question and does not make any attempt to do so, it is scored zero for factuality and completeness, but no score is given for hallucination. As

[5] https://huggingface.co/stabilityai/StableBeluga2 (accessed February 26, 2024).

[6] https://github.com/oobabooga/text-generation-webui/tree/main (accessed February 26, 2024).

[7] https://huggingface.co/meta-llama/Llama-2-70b-chat-hf (accessed February 26, 2024).

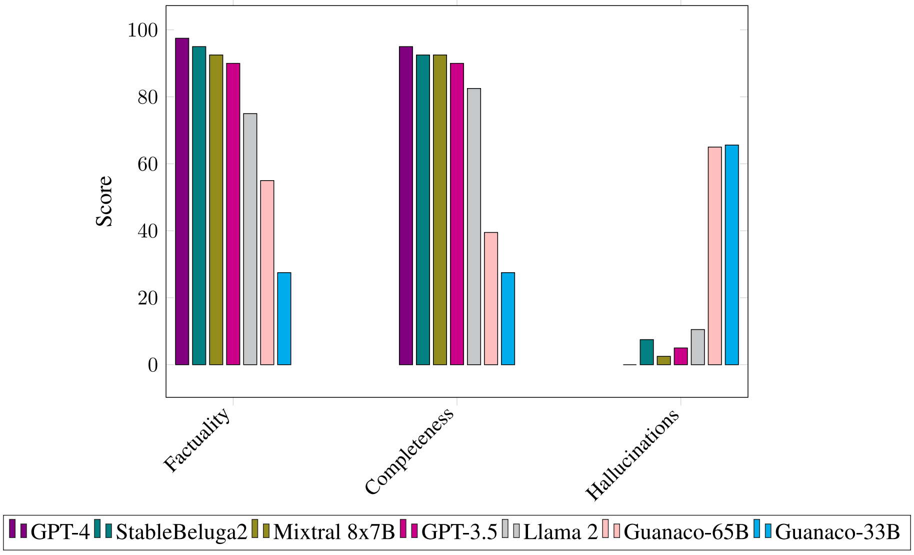


FIGURE 3
Benchmark of seven LLMs for generating answers based on factory materials.

TABLE 1 Model benchmarking scores (out of 100) and average response length.

Model	Factuality	Completeness	Hallucinations	Words
GPT-4	97.5	95	0	69
StableBeluga2	95	92.5	7.5	58
Mixtral 8x7B	92.5	92.5	2.5	66
GPT-3.5	90	90	5	89
Llama 2	77.5	82.5	13	128
Guanaco 65B	55	39.5	65	131
Guanaco 33b	27.5	27.5	65.6	190

such, the final score for hallucination is calculated as follows: corrected score $= \frac{\text{score}}{20 - \text{number of unanswered questions}} \times 100$

As shown in Figure 3 and Table 1, GPT-4 outperforms other models regarding factuality, completeness, and lack of hallucinations but is closely followed by StableBeluga2 and GPT-3.5. The Guanaco models, based on Llama 1, perform significantly worse. The conciseness of the responses showed a similar pattern, except that StableBeluga2 produced the shortest answers (58 words), followed closely by Mixtral 8x7B (66 words) and GPT-4 (69 words).

6 Discussion

6.1 GPT-4 is the best, but open-source models follow closely

GPT-4 performs best across all measures but is closely followed by StableBeluga2, Mixtral 8x7B, and GPT-3.5. Compared to GPT-4, the cost per input token for GPT-3.5 is significantly lower.[8] However, the higher costs of GPT-4 are partially counteracted by its concise yet complete responses. If longer, more detailed responses were desired (e.g., for training purposes), the prompt could be adjusted. We observed that the less powerful models, such as GPT-3.5 and Llama 2, tended to be wordier and include additional details that were not directly requested. In contrast, GPT-4, StableBeluga2, and Mixtral 8x7B generated more concise responses.

The latest generation of open-source models, such as Mixtral 8x7B and Llama 2 variants, such as StableBeluga2, demonstrates a clear jump forward relative to their predecessors based on Llama-1, which were more prone to hallucinations and exhibited poorer reasoning abilities over the context material. While open-source models like StableBeluga2 and Mixtral 8x7B do not score as high as GPT-4, they ensure better data security, privacy, and customization if hosted locally. This can be a crucial consideration for companies with sensitive data or unique needs.

6.2 The tool is beneficial but inferior to human experts

Users appreciate the system's functionality and see it as a tool for modernizing factory operations and speeding up operations. They are keen on improvements to be made for better user

[8] https://openai.com/pricing#language-models (accessed February 26, 2024).

experience and utility, especially in the areas of content, feature enhancements, and user training. However, they express concerns about potential safety risks and the efficacy of information retrieval compared to consulting expert personnel. While these concerns are understandable, the tool was not designed to replace human-human interactions; instead, it can be used when no human experts are present or when they do not know or remember how to solve a specific issue. This would come into play during the night shift at the factory where we conducted the user study as a single operator operates a production line, leaving limited options for eliciting help from others.

6.3 Limitations and future work

We used the same prompt for all LLMs; however, it is possible that some of the LLMs would perform better with a prompt template developed explicitly for it. For consistency, we matched the LLMs' hyperparameters (e.g., temperature) as closely as possible across all the tested models, except for Llama 2, as we did not have access to the presets as we did not host it locally. Our model benchmarking procedure involved 20 questions, and a singular coder assessed the responses. This introduces the potential for bias, and the limited number of questions may not cover the full spectrum of complexities in real-world scenarios. To mitigate these shortcomings, we varied query complexity and source material types.

The study's design did not include a real-world evaluation involving end users operating the production line, as this was considered too risky for our industry partner. Such an environment might present unique challenges and considerations not addressed in this research, such as time pressure. Yet, by involving operators and managers and instructing them to pose several questions based on their actual work experience, we could still evaluate the system and collect valid feedback.

These limitations suggest directions for future research, for example, longitudinal studies where operators use the tool during production line operations and more comprehensive prompt and model customization. Longitudinal studies will be key to understanding the real-world impact on production performance, operator wellbeing, and cognitive abilities.

7 Conclusion

The results demonstrated GPT-4's superior performance over other models regarding factuality, completeness, and minimal hallucinations. Interestingly, open-source models like StableBeluga2 and Mixtral 8x7B followed close behind. The user study highlighted the system's user-friendliness, speed, and logical functionality. However, improvements in the user interface and content specificity were suggested, along with potential new features. Benefits included modernizing factory operations and speeding up specific tasks, though concerns about safety, efficiency, and inferiority to asking human experts were raised.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by Human Research Ethics Committee (HREC) from TU Delft. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

SK: Writing – original draft, Visualization, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. CW: Writing – original draft, Software, Methodology, Conceptualization. MF: Writing – original draft. SW: Writing – original draft. SR-A: Writing – original draft. EN: Writing – review & editing, Supervision, Methodology, Conceptualization.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the European Union's Horizon 2020 research and innovation program via the project COALA "COgnitive Assisted agile manufacturing for a LAbor force supported by trustworthy Artificial Intelligence" (Grant agreement 957296).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abid, A., Abdalla, A., Abid, A., Khan, D., Alfozan, A., and Zou, J. (2019). *Gradio: Hassle-Free Sharing and Testing of ML Models in the Wild*. doi: 10.48550/arXiv.1906.02569

Alkaissi, H., and McFarlane, S. I. (2023). Artificial hallucinations in chatgpt: implications in scientific writing. *Cureus* 15:e35179. doi: 10.7759/cureus.35179

Alves, J., Lima, T. M., and Gaspar, P. D. (2023). Is industry 5.0 a human-centred approach? A systematic review. *Processes* 11. doi: 10.3390/pr11010193

Badini, S., Regondi, S., Frontoni, E., and Pugliese, R. (2023). Assessing the capabilities of chatgpt to improve additive manufacturing troubleshooting. *Adv. Ind. Eng. Polym. Res.* 6, 278–287. doi: 10.1016/j.aiepr.2023.03.003

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., et al. (2023). A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv*. doi: 10.18653/v1/2023.ijcnlp-main.45

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. (2020). "Language models are few-shot learners," in *Advances in Neural Information Processing Systems, Vol. 33*, eds H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin (Red Hook, NY: Curran Associates, Inc.), 1877–1901.

Brückner, A., Hein, P., Hein-Pensel, F., Mayan, J., and Wölke, M. (2023). "Human-centered hci practices leading the path to industry 5.0: a systematic literature review," in *HCI International 2023 Posters*, eds C. Stephanidis, M. Antona, S. Ntoa, and G. Salvendy (Cham: Springer Nature Switzerland), 3–15.

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. (2023). *Qlora: Efficient Finetuning of Quantized Llms*. doi: 10.48550/arXiv.2305.14314

Edwards, B., Zatorsky, M., and Nayak, R. (2008). Clustering and classification of maintenance logs using text data mining. *Data Mining Anal.* 87, 193–199.

Fantini, P., Pinzone, M., and Taisch, M. (2020). Placing the operator at the centre of industry 4.0 design: modelling and assessing human activities within cyber-physical systems. *Comp. Ind. Eng.* 139:105058. doi: 10.1016/j.cie.2018.01.025

Feng, S. C., Bernstein, W. Z., Thomas Hedberg, J., and Feeney, A. B. (2017). Toward knowledge management for smart manufacturing. *J. Comp. Inf. Sci. Eng.* 17:3. doi: 10.1115/1.4037178

Gao, T., Fisch, A., and Chen, D. (2021). "Making pre-trained language models better few-shot learners," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (Association for Computational Linguistics), 3816–3830.

Gröger, C., Schwarz, H., and Mitschang, B. (2014). "The manufacturing knowledge repository - consolidating knowledge to enable holistic process knowledge management in manufacturing," in *Proceedings of the 16th International Conference on Enterprise Information Systems* (SCITEPRESS - Science and and Technology Publications), 39–51. doi: 10.5220/0004891200390051

Guest, G., MacQueen, K. M., and Namey, E. E. (2011). *Applied* thematic analysis. Thousand Oaks, CA: Sage Publications.

Jang, J., Ye, S., Lee, C., Yang, S., Shin, J., Han, J., et al. (2022). Temporalwiki: a lifelong benchmark for training and evaluating ever-evolving language models. *arXiv*. doi: 10.18653/v1/2022.emnlp-main.418

Jawahar, G., Sagot, B., and Seddah, D. (2019). "What does BERT learn about the structure of language?," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Florence. Association for Computational Linguistics), 3651–3657.

Jeblick, K., Schachtner, B., Dexl, J., Mittermeier, A., Stüber, A. T., Topalis, J., et al. (2023). ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. *Eur. Radiol.* 1–9. doi: 10.1007/s00330-023-10213-1

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., et al. (2024). *Mixtral of Experts*. doi: 10.48550/arXiv.2401.04088

Kernan Freire, S., Foosherian, M., Wang, C., and Niforatos, E. (2023a). "Harnessing large language models for cognitive assistants in factories," in *Proceedings of the 5th International Conference on Conversational User Interfaces, CUI '23* (New York, NY: Association for Computing Machinery). doi: 10.1145/3571884.3604313

Kernan Freire, S., Wang, C., Ruiz-Arenas, S., and Niforatos, E. (2023b). "Tacit knowledge elicitation for shop-floor workers with an intelligent assistant," in *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–7.

Kwon, B. C., and Mihindukulasooriya, N. (2022). "An empirical study on pseudo-log-likelihood bias measures for masked language models using paraphrased sentences," in *TrustNLP 2022 - 2nd Workshop on Trustworthy Natural Language Processing, Proceedings of the Workshop* (New York, NY), 74–79. doi: 10.1145/3544549.3585755

Leoni, L., Ardolino, M., El Baz, J., Gueli, G., and Bacchetti, A. (2022). The mediating role of knowledge management processes in the effective use of artificial intelligence in manufacturing firms. *Int. J. Operat. Prod. Manag.* 42, 411–437. doi: 10.1108/IJOPM-05-2022-0282

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., et al. (2020). "Retrieval-augmented generation for knowledge-intensive nlp tasks," in *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS'20* (Red Hook, NY: Curran Associates Inc.).

Liang, J., Huang, W., Xia, F., Xu, P., Hausman, K., Ichter, B., et al. (2022). *Code as Policies: Language Model Programs for Embodied Control*. doi: 10.48550/arXiv.2209.07753

Liu, J. (2022). *LlamaIndex*. Available online at: https://github.com/jerryjliu/llama_index

Maddikunta, P. K. R., Pham, Q.-V., B, P., Deepa, N., Dev, K., Gadekallu, T. R., et al. (2022). Industry 5.0: a survey on enabling technologies and potential applications. *J. Ind. Inf. Integr.* 26:100257. doi: 10.1016/j.jii.2021.100257

May, G., Taisch, M., Bettoni, A., Maghazei, O., Matarazzo, A., and Stahl, B. (2015). A new human-centric factory model. *Proc CIRP* 26, 103–108. doi: 10.1016/j.procir.2014.07.112

Müller, M., Alexandi, E., and Metternich, J. (2021). Digital shop floor management enhanced by natural language processing. *Procedia CIRP* 96, 21–26. doi: 10.1016/j.procir.2021.01.046

Nov, O., Singh, N., and Mann, D. (2023). *Putting Chatgpt's Medical Advice to the (Turing) Test*. doi: 10.48550/arXiv.2301.10035

Oruç, O. (2020). A semantic question answering through heterogeneous data source in the domain of smart factory. *Int. J. Nat. Lang. Comput.* 9.

Richter, S., Waizenegger, L., Steinhueser, M., and Richter, A. (2019). Knowledge management in the dark: the role of shadow IT in practices in manufacturing. *IJKM*. 15, 1–19. doi: 10.4018/IJKM.2019040101

Semnani, S. J., Yao, V. Z., Zhang, H. C., and Lam, M. S. (2023). *Wikichat: A Few-Shot Llm-Based Chatbot Grounded With Wikipedia*. doi: 10.48550/arXiv.2305.14292

Serrat, O. (2017). *The Five Whys Technique. Knowledge Solutions: Tools, Methods, and Approaches to Drive Organizational Performance*, 307–310.

Shneiderman, B. (2022). *Human-Centered AI*. Oxford: Oxford University Press.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., et al. (2023). Large language models encode clinical knowledge. *Nature* 620, 172–180. doi: 10.1038/s41586-023-06291-2

Tang, R., Han, X., Jiang, X., and Hu, X. (2023). *Does Synthetic Data Generation of Llms Help Clinical Text Mining*? doi: 10.48550/arXiv.2303.04360

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., et al. (2023). *Llama 2: Open Foundation And fine-Tuned Chat Models*. doi: 10.48550/arXiv.2307.09288

Trautmann, D., Petrova, A., and Schilder, F. (2022). *Legal Prompt Engineering for Multilingual Legal Judgement Prediction*. doi: 10.48550/arXiv.2212.02199

Wang, X., Anwer, N., Dai, Y., and Liu, A. (2023a). Chatgpt for design, manufacturing, and education. *Proc. CIRP* 119, 7–14. doi: 10.1016/j.procir.2023.04.001

Wang, Z., Yang, F., Zhao, P., Wang, L., Zhang, J., Garg, M., et al. (2023b). Empower large language model to perform better on industrial domain-specific question answering. *arXiv*. doi: 10.18653/v1/2023.emnlp-industry.29

Wei, C., Wang, Y.-C., Wang, B., and Kuo, C. C. J. (2023). *An Overview on Language Models: Recent Developments and Outlook*. doi: 10.1561/116.00000010

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., et al. (2022a). *Emergent Abilities of Large Language Models*. doi: 10.48550/arXiv.2303.05759

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, b., Xia, F., et al. (2022b). "Chain-of-thought prompting elicits reasoning in large language models," in *Advances in Neural Information Processing Systems, Vol. 35*, eds S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Red Hook, NY: Curran Associates, Inc.), 24824–24837.

Wellsandt, S., Hribernik, K., and Thoben, K.-D. (2021). "Anatomy of a digital assistant," in *Advances in Production Management Systems. Artificial Intelligence for Sustainable and Resilient Production Systems*, eds A. Dolgui, A. Bernard, D. Lemoine, G. von Cieminski, and D. Romero (Cham: Springer International Publishing), 321–330.

Wen, C., Sun, X., Zhao, S., Fang, X., Chen, L., and Zou, W. (2023). Chathome: development and evaluation of a domain-specific language model for home renovation. *ArXiv*. doi: 10.48550/arXiv.2307.15290

Xia, Y., Shenoy, M., Jazdi, N., and Weyrich, M. (2023). Towards autonomous system: flexible modular production system enhanced with large language model agents. *arXiv*. doi: 10.1109/ETFA54631.2023.10275362

Xie, Q., Han, W., Zhang, X., Lai, Y., Peng, M., Lopez-Lira, A., et al. (2023a). Pixiu: A large language model, instruction data and evaluation benchmark for finance. *arXiv*. doi: 10.48550/arXiv.2306.05443

Xie, T., Wan, Y., Huang, W., Yin, Z., Liu, Y., Wang, S., et al. (2023b). Darwin series: domain specific large language models for natural science. *arXiv*. doi: 10.48550/arXiv.2308.13565

Xu, F. F., Alon, U., Neubig, G., Hellendoorn, V. J., and Hel, V. J. (2022). A systematic evaluation of large language models of code. *arXiv*. doi: 10.48550/arXiv.2202.13169

Xu, X., Lu, Y., Vogel-Heuser, B., and Wang, L. (2021). Industry 4.0 and industry 5.0—inception, conception and perception. *J. Manuf. Syst*. 61, 530–535. doi: 10.1016/j.jmsy.2021.10.006

Zhang, J., Chen, Y., Niu, N., and Liu, C. (2023a). A preliminary evaluation of chatgpt in requirements information retrieval. *arXiv*. doi: 10.2139/ssrn.4450322

Zhang, W., Liu, H., Du, Y., Zhu, C., Song, Y., Zhu, H., et al. (2023b). Bridging the information gap between domain-specific model and general llm for personalized recommendation. *arXiv*. doi: 10.48550/arXiv.2311.03778

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., et al. (2023). *A Survey of Large Language Models*. doi: 10.48550/arXiv.2303.18223

Zuccon, G., Koopman, B., and Shaik, R. (2023). "Chatgpt hallucinates when attributing answers," in *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region, SIGIR-AP '23* (New York, NY: Association for Computing Machinery), 46–51.