Large language models in requirements engineering for digital twins

Nico Blasek[1], Karl Eichenmüller[1], Bastian Ernst[1], Niklas Götz[1], Benjamin Nast[1,*] and Kurt Sandkuhl[1,2]

[1]Rostock University, Albert-Einstein-Str. 22, 18059 Rostock, Germany
[2]Jönköping University, Gjuterigatan 5, 55111 Jönköping, Sweden

Abstract

Can large language models (LLMs) be used for digital twin engineering (DTE)? Engineering digital twins (DTs) is a complex process consisting of several phases and involving different disciplines. We argue that an investigation of LLM use in DTE has to define what kinds of DTs are in focus and what DTE phases shall be supported. In our work, we concentrate on the early phases of DTE, with a particular focus on requirements engineering (RE), and we focus on supervisory and operational DTs. This paper investigates the quality of LLM output for defining requirements for DTs. The main contributions of our work are results from an experiment comparing requirements to a DT of an air conditioning facility of a domain expert and ChatGPT and conclusions for prompt engineering resulting from this experiment.

Keywords

Digital Twin Engineering, Large Language Model, ChatGPT, Requirements Engineering, Air Conditioning

1. Introduction

Can large language models (LLMs) be used for digital twin engineering (DTE)? If so, can the contribution of LLMs be considered substantial enough to motivate further research? These two questions were the starting point for the research on LLM use in DTE presented in this paper. The motivation behind the questions is that both areas, DTE and LLM, receive a lot of attention in enterprises and academic research as promising technology areas with high potential for industrial application, which makes the intersection between both areas particularly interesting.

Digital twin (DT) can be defined as "a dynamic virtual representation of a physical object or system across its lifecycle, using real-time data to enable understanding, learning, and reasoning" [1]. Depending on the capability of the DT, different kinds of DTs can be distinguished, starting from supervisory DTs, which only allow for monitoring the situation of a physical object, to autonomous DTs that are self-contextualizing and self-optimizing (see section 2.2). Furthermore, engineering DTs is a complex process consisting of several phases and involving different disciplines (see section 2.2). We argue that an investigation of LLM use in DTE has to define what kinds of DTs are in focus and what DTE phases shall be supported. In our work, we

Companion Proceedings of the 16th IFIP WG 8.1 Working Conference on the Practice of Enterprise Modeling and the 13th Enterprise Design and Engineering Working Conference, November 28 – December 1, 2023, Vienna, Austria
*Corresponding author.
benjamin.nast@uni-rostock.de (B. Nast); kurt.sandkuhl@uni-rostock.de (K. Sandkuhl)
0000-0003-4659-9840 (B. Nast); 0000-0002-7431-8412 (K. Sandkuhl)

concentrate on the early phases of DTE, with a particular focus on requirements engineering (RE), and we focus on supervisory and operational DTs.

More concretely, we aim to investigate the quality of LLM output for defining requirements for DTs. For this purpose, we used one of the currently most popular LLMs, OpenAI's ChatGPT-4, to elicit requirements for DTs. As our ambition is to explore issues relevant to future research, we limited our study in a first step to only one application domain. The application domain selected is air conditioning and ventilation systems. We operationalize quality as completeness and correctness from the perspective of a domain expert.

The main contributions of our work are results from an experiment comparing requirements for a DT of an air conditioning facility of a domain expert and ChatGPT and conclusions for prompt engineering resulting from this experiment.

The paper is structured as follows: section 2 describes the background for our work from digital twin engineering and large language models. Section 3 summarizes the research method, and section 4 results of a literature study on related work. Section 5 describes the requirement elicitation for DT of air conditioning systems and compares ChatGPT output with the view of a domain expert. Section 6 discusses results and addresses the issue of prompt design. A conclusion and implications for future work are given in section 7.

2. Theoretical background

2.1. Large language models

LLMs belong to the broader category of deep learning models, address the area of natural language processing, and are designed to interpret and generate human-like text. Essential concepts of LLMs and their evolution have been widely documented, for example, in the publication by Brown et al. [2]. The most influential architecture in recent times for building LLMs is the Transformer. It uses attention mechanisms [3] to weigh the importance of different words or tokens in a sequence when producing an output. LLMs are trained on vast amounts of text data to be able to generate coherent and contextually relevant text across a wide range of topics. For instance, models like OpenAI's GPT series have been trained on books, articles, and web pages. The pre-training of LLMs is supposed to be task-agnostic [4].

Capabilities of LLMs include tasks such as translation, question-answering, summarization, and text generation without needing task-specific training data. One of the most important features of models like GPT (Generative Pre-trained Transformer) is their ability to generate coherent, diverse, and contextually relevant text over long passages. One of the currently most popular LLMs, OpenAI's GPT-4 with its Chatbot frontend - ChatGPT[1] can also be used for translation, grammar correction, or email composition [5]. While LLMs are powerful, they can sometimes produce incorrect or nonsensical answers, which often are termed "hallucinations". The use of LLMs starts from inputs (called prompts) stating the task to be completed by the LLM. LLMs are sensitive to the input phrasing. Thus, prompt engineering and prompting methods [6] have developed into a critical topic of study for LLMs as they investigate the techniques by which end-users can use LLMs to perform tasks.

[1] https://chat.openai.com

2.2. Digital twin engineering

An analysis of research on DTE showed various procedural and method approaches, for example, discussed in [7], [8] or [9]. While the specific procedures in DTE can vary depending on the application domain, industry, and specific objectives, the following general steps in DTE are visible in the literature:

- **Scoping and Goal Definition:** Setting the scope of DT development includes deciding on the physical system, process, or entity that is the subject of the DT, i.e., the scope of the project. Furthermore, there should be a clear understanding of what objective to achieve, e.g., support of predictive maintenance, process optimization, simulation, or other objectives. Defining clear goals is the first step.
- **Requirements Engineering:** In the RE phase, the way of data collection has to be defined. This includes instrumentation, i.e., what sensors are in the physical entity to collect relevant data. This could include sensors for temperature or pressure, cameras, accelerometers, and more. If historical data must be considered, this should be defined. Furthermore, the functionality required to reach the goals has to be specified.
- **Develop the Digital Representation:** Modeling the physical entity might involve CAD models, process diagrams, or other types of digital schematics. How to integrate the model with real-time data feeds from sensors and other sources has to be part of this step.
- **Calibration and Validation:** The DT requires validation by comparing its outputs and predictions to actual outcomes in the physical world. This typically includes refinement of the model until its performance meets acceptable accuracy levels.
- **Integration with Other Systems:** DTs can be more effective when integrated with other systems, such as enterprise resource planning (ERP) systems, manufacturing execution systems (MES), or building management systems.
- **Iterative Refinement:** Over time, as more data becomes available and the physical entity evolves or changes, the DT should be updated and refined to reflect these changes. This requires the implementation of a feedback mechanism that continuously updates the DT based on real-world data. This ensures the twin remains an accurate representation of its physical counterpart.

When engineering a DT, it's essential to maintain a close alignment between the digital model and the real-world entity it represents. Collaboration across disciplines, including domain experts, IT professionals, data scientists, and engineers, is often crucial for successful DT implementations. Furthermore, some scholars propose to distinguish different kinds of DT according to their capabilities [10]:

- **Supervisory DTs**, sometimes also called digital shadows, allow the real-time monitoring of the status and events of a physical object. For this purpose, the parts of the object to be monitored, necessary information, and adequate sensing equipment must be known.
- **Operational DTs** add the possibility to perform certain operations on the physical object, which can be used to control the operations. This requires actuators to initiate functional changes and the knowledge about possible and required operations.

- **Simulation DTs** have the ability to simulate the physical object based on the developed models. This type of DT can also carry out predictions to support design or operational decisions.
- **Intelligent DTs** are supposed also to show learning abilities. They are supposed to learn from operational data, for example, using machine learning (ML), which also realizes some abilities of decision support and scenario planning.
- **Autonomous DTs** additionally have the ability to perform all decision-making based on predefined parameters and manage the physical object concerned with minimal human intervention.

In this paper, we focus on RE for supervisory and operational DTs.

3. Research method

The starting point of our work is the questions presented in the introduction and the decision to focus on the phase of RE. With this background, two research questions (RQ) were defined for this paper:

- RQ 1: In the context of RE for DT, how consistent and complete is the output of ChatGPT with the information provided by domain experts?
- RQ 2: How to design the prompt chain to improve the output of ChatGPT?

The overall research strategy for work presented in this paper is of an explorative nature, i.e., we aim to gather new knowledge by exploring the potential of ChatGPT use in RE of DTs. More concretely, the work combines literature studies with quasi-experiments and argumentative-deductive work.

The literature search aimed to identify related work and results from other scholars to be taken into account when investigating the potential of LLMs for DTE. For this step, we used Kitchenham's approach for systematic literature reviews (SLR). Kitchenham [11] suggests six steps, which we briefly introduce in the following and document in detail in section 4. The first step is to develop research questions (RQ) to be answered by the SLR. The process of paper identification starts with defining the overall search space (step 2), which basically consists of determining the literature sources to take into account in the light of the research questions. Paper identification continues with the population phase (step 3). In this step, the search string is developed and applied by searching the literature sources. Afterward, the step "paper selection" follows by defining inclusion and exclusion criteria and a manual selection of relevant papers found in the population phase (step 4). The data collection phase (step 5) focuses on extracting the information relevant to answering the research question from the set of identified relevant papers. The last step is the analysis of data and interpretation, i.e., to answer the research question defined in step 1 by using collected data from relevant papers.

As the SLR returned no previous work on using LLM for DTE, we structured the field of DTE along with the tasks to perform during an engineering project (see section 2.2). This is the argumentative-deductive part of our work.

In the next step, RE for a DT in a defined application area was the subject of a quasi-experiment. A controlled experiment in software engineering and information systems development is "a

randomized or quasi-experiment in which individuals or teams (the study units) conduct one or more [...] tasks for the sake of comparing different populations, processes, methods, techniques, languages or tools (the treatments)" [12]. In our work, we perform a quasi-experiment; the study units are ChatGPT and domain experts, and the treatment is the task of eliciting requirements for a DT of an air conditioning system. A quasi-experiment is "an experiment in which units are not assigned to conditions randomly" [13]. The experiment does not aim at testing a specific hypothesis but is exploratory research to answer the research questions defined. The experiment design is described in detail in section 5.

4. Related work

Related work was identified by performing a systematic literature study based on the six steps proposed in the method of Kitchenham (see section 3). The research questions (step 1) were already introduced in section 3. The search space (step 2) consisted of the literature databases Scopus, IEEExplore, and AISeL. The search string (step 3) used in these databases combines the term "Digital Twin" in combination with "large language model" and synonyms for this term ("LLM", "neural text", "ChatGPT"). Thus, the final search string used was *"Digital Twin" AND ("large language model" OR "LLM" OR "neural text" OR "ChatGPT").* The search in title, abstract, and keywords resulted in a total of 19 papers. The inclusion criterion (step 4) was that the papers had to discuss LLM use in the context of DT.

Of the 8 hits in Scopus, 5 were excluded because they were conference proceedings that included DT papers and LLM papers, but no paper with both topics in the same paper. The remaining three papers used DT and LLM on the abstract to position the work but did not address DT development. In AISeL, the query interface only allowed for search in all meta-data. All 9 papers did not mention DT and LLM in title, abstract, or keywords; none of the nine papers concerned DT development or LLM use in a DT. In IEEE Xplore, two papers were found, and one of them was relevant to our work: [14] discuss a DT for an industrial automation system and use LLM-agents to interpret descriptive information in the DTs and to control the physical system through service interfaces. Thus, this paper is not exactly about engineering the DT but about using LLM as a component for a specific task in a DT.

Table 1 summarizes the number of papers found in the different databases and the relevant ones. In conclusion, the SLR did not return any paper addressing LLM use for DTE. Thus, the identified literature does not provide any information to aid in investigating the potential of LLM use for DTE or to answer our RQs.

Table 1
Results of the SLR

Literature Database	No. of Hits	Relevant Papers
Scopus	8	none
IEEE Xplore	2	[14]
AISeL	9	none

5. Experiment on ChatGPT use in requirements engineering

5.1. Experiment design

Based on our experience from previous work [15] in the field of air conditioning systems and an internet research, we prepared a questionnaire for the interview consisting of 17 questions. These can be divided into *General Structure and Functions of Air Conditioning Systems*, *Energy Optimizations*, and *Special Case of a Facility in a Swimming Pool Hall* (see Table 2). This list of questions not only served as a guide during the interview with a domain expert but was also used to create prompts for ChatGPT. By means of prompt engineering, in particular role prompting, it was thus possible to simulate an interview situation with ChatGPT as the interview partner. This offers the possibility of comparing statements of the domain expert with the answers of ChatGPT to have them evaluated later and thus generate further expertise and set requirements for a DT.

Table 2
Questionnaire

General Structure and Functions of Air Conditioning Systems
What is/How does an air conditioning system work?
What are the main processes in an air conditioning system?
What parts does an air conditioning system consist of?
What possibilities exist for modeling air conditioning systems?
Are there essential connections between sensors and actuators?
Which parts are monitored by sensors?
Which parts can be controlled externally?
Which environmental influences affect the air conditioning system?
Which factors affect the durability of an air conditioning system?
Which parts are maintenance-intensive?
Energy Optimizations
What are the possibilities for saving energy in air conditioning systems?
How does the heat recovery work?
Are there peaks in the possibilities for energy savings?
Special Case of a Facility in a Swimming Pool Hall
Are there any parts and conditions in the air conditioning system for the swimming pool hall that differ from conventional systems?
What special environmental influences affect the ventilation system in the swimming pool hall?
In what form is the sensor data collected, and how does the storage work?
Are there components that are particularly stressed by the special environmental influences in a swimming pool hall?

5.1.1. Interview with domain expert

An expert in the field of air conditioning systems was available as a participant in the interview. He has been working for several years in a medium-sized company in this domain, where he checks existing facilities, evaluates measurement data, and advises operators on energy optimization. In addition, the expert has basic knowledge and initial experience with DTs and

digital shadows of those systems.

The interview was conducted with the domain expert as the participant and four people who prepared the interview. One person acted as the interviewer and guided the participant through the interview on the basis of the questionnaire, while the remaining three people took notes, asked follow-up questions, or checked the course of the interview guide. For later transcription and analysis, the interview was recorded with the consent of all persons in the room.

Before the interview began, the project and the project group were briefly introduced to the expert. This was followed by an introduction of the expert, in which he described his professional career and experience with air conditioning systems and the associated energy optimization. The expert was able to answer each question from the list of questions, as well as intermediate questions, in detail, and in some cases, also with case studies, experience reports, and further explanations. In addition, the expert was able to interpret illustrations and circuit diagrams of an air conditioning system, as well as to name and explain the components and their functions.

5.1.2. Interview simulation with ChatGPT

Using ChatGPT-4, role prompting was applied with an initial prompt to assign a role to the Artificial Intelligence (AI) as an expert in the field of DTs for air conditioning systems. Here, the assignment of the role was primarily done to ensure that experience in the domain and professional background matched the domain expert. The prompts were entered in German.

"Hello, thank you for finding the time for an interview with us as a domain expert on digital twins of air conditioning systems. Thanks to your expertise and years of experience related to air conditioning systems, you are the ideal candidate for our interview. You have studied mechanical engineering and are a project engineer in a renowned East German company in the air conditioning and refrigeration sector, and your job is to look after, maintain, and improve/optimize existing air conditioning systems. Below, we have a few questions for you. The questions should be answered in as much detail as possible and be easy to understand, as we have no expertise in mechanical engineering or in air conditioning and refrigeration technology."

With the simulation of the interview, each question was asked as a single prompt to ChatGPT, where no follow-up questions were asked. It could be observed that the AI's answers follow a similar pattern, which starts with an introductory sentence. This is always followed by a list of examples with a short explanation and further information until the answer ends with a concluding paragraph. Unlike the domain expert, ChatGPT did not use specific case studies or testimonials to reflect the information better. In addition, the version of ChatGPT used does not offer the possibility to interpret illustrations and describe them if necessary.

5.1.3. Evaluation

After conducting both interviews, the recordings and texts had to be processed and evaluated accordingly for later analysis. The interview with the domain expert was transcribed, and the transcript was reviewed for errors and corrections. Afterward, a qualitative content analysis was conducted according to [16] (see section 5.2.1). Supercategories and subcategories were

inductively formed independently based on individual markers. The formed categories were then compared within the group. The suitability of the categories discussed was often measured within the groups based on the goal and task. The supercategories were unanimously defined as follows:

- **General Information About Air Conditioning Systems:** Questions concerning the construction and operation
- **Energy Optimization:** Questions concerning the possibilities for energy optimization of an air conditioning system, both structural and process-related changes
- **Use Case Swimming Pool Hall:** Questions concerning specifics of an air conditioning system in an indoor swimming pool

Once the broad topics of the questions were identified, subcategories had to be created, which helped to classify the answers to the corresponding questions and the evaluation by the domain expert in a meaningful way. Therefore, the following categories were unanimously agreed upon for the subcategories:

- **Missing Information:** Information from ChatGPT, which should have been included in the generated answer according to the domain expert.
- **Additional Helpful Information:** Information from ChatGPT, which, according to the domain expert, is outside the scope of the posed question and represents additional, non-trivial, useful background knowledge.
- **Additional Unhelpful Information:** Information from ChatGPT, which, according to the domain expert, is outside the scope of the question asked and does not provide additional useful information to the subject matter.
- **Mistake:** Marked as misinformation by the domain expert.

The following subcategories were also added after a discussion with the agreement of the majority of the working group. They were added because they allow for a more in-depth evaluation of the interview in relation to the research questions.

- **Software:** Information on software that enables digital realization of air conditioning systems.
- **Components:** Details of parts that a working air conditioning system requires.
- **Sensor Data:** Data that can be collected via sensors and is relevant for system functionality. Only metrics that can be assigned to physical quantities are to be recorded.

The following subcategories were rejected by the majority after discussion and were not adopted: *Wear Factors*, *Processes in a System*, *Factors Influencing Durability of Parts*, and *Factors Influencing Comfort/Well-Being*. These subcategories were rejected either because they lacked reference to the task or overlapped with other already defined categories and were, therefore, difficult to distinguish from one another.

Using the definitions, all four group members were then tasked with marking all text passages individually according to the definitions they had developed. These markings were compared

with each other again afterward so that the coding could be completed. The preceding definition was decisive for the coding. The coding was done in great unanimity, and there was no incident of explicit disagreement in the assignment of text passages to the respective categories. Differences in markings were limited to careless errors where consensus was quickly reached.

5.2. Results

In this chapter, the results of the qualitative content analysis are presented first. Based on this, the requirements of a DT for an air conditioning system are defined, and the preceding interview with ChatGPT is examined with regard to these requirements. Subsequently, newly created prompts are presented to provide a detailed analysis of the feasibility of using ChatGPT as a substitute for a domain expert.

5.2.1. Qualitative content analysis

Figure 1 shows the assessment of the domain expert, sorted by supercategories and subcategories. The 17 questions from the questionnaire (see Table 2) are divided into eleven questions on *General Information* about air conditioning systems, three questions about *Energy Optimization*, and three questions on information relevant to air conditioning systems *Case Swimming Pool*. For each of these questions, it was checked whether a text passage was defined, which could be assigned to one of the subcategories. The subcategories, which deal with the evaluation of ChatGPT statements by the domain expert, are counted out here. Figure 1 shows the evaluation of the statements of ChatGPT by the domain expert.

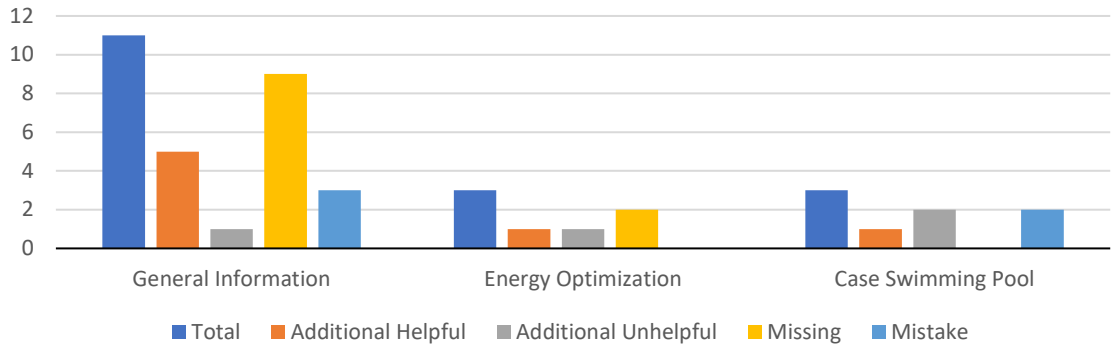

Figure 1: Evaluation of ChatGPT's answers by the domain expert.

Of the eleven *General Information* questions, five responses obtained Additional Helpful Information. One response included Additional Unhelpful Information. Information was Missing in nine responses, and three contained Mistakes.

Of three questions on *Energy Optimization*, one response provided Additional Helpful Information, and one provided Additional Unhelpful Information. In this block of questions, no Missing information or Mistakes were identified.

In three questions about the *Case Swimming Pool*, Additional Helpful Information was identified in one response and Additional Unhelpful Information in two responses. Additionally, two Mistakes were found, but no Missing information in this block of questions.

5.2.2. Identified requirements for a digital twin

Based on the interviews and our research, it was possible to identify a total of 28 elements that are important for the realization of an air conditioning system as a DT and for it to function well. These 28 elements can be divided into 23 general (*Component/Factor*) and five specific factors (*Type*). The specific factors are specifications of individual components whose differentiation is important for the successful design of an air conditioning system as a DT.

The interview with ChatGPT was examined for these parts and components listed in Table 3 to explore whether a structured interview with the AI application was sufficient to realize a DT

Table 3
Components and Factors for the Realization of a DT

Category	Component/Factor	Type	Mentioned?
Actuator	Silencer		x
	Fan		x
	Valve		x
		Volume Flow Controller	
		Throttle Valve	
		Multileaf Damper	
	Heating Register		x
	Cooling Register		x
	Pumps		
	Frost Protection Monitor		
Sensor	Temperature		x
	Humidity		x
	Air Pressure		x
	Air Quality		x
	Air Flow		x
	Utilization of the Facility		x
	Energy Sensor		x
Passive Component	Supply Air System		
	Exhaust Air System		
	Filter		x
		Dust Filter	(x)
		HEPA Filter	(x)
	Heat Exchanger		x
	Air Ducts/Shaft		x
	Outlet/Return		x
Passive Factor	Room Size		x
	Number of Persons		x
	Room Specific Factors		x
Total	23	5	19 (2)

of a technical system. Of the 23 general parts, ChatGPT correctly named 19 during the interview, representing about 83% coverage. The correct and complete naming of all relevant sensors is to be emphasized here. However, the important actuator pump was not named, nor was the frost protection monitor, which protects the air conditioning system from frost damage and thus ensures its functionality. The fact that the supply and exhaust systems are two separate systems within a facility was also not apparent from the interview. For specific components (*Type*), coverage is two out of five correctly named items (40%), more than 40% less than for general parts. The different types of valves were missing, and it was not highlighted that there are different types of airflow controllers, which would have been necessary in a DT due to the different operations and impact on airflow.

5.3. Subsequent prompt engineering

The following section deals with three new prompts that were created to get a more detailed picture of the extent to which ChatGPT can be used as a substitute for a domain expert. The main goal was to reproduce the list introduced in Table 3 and to extend it for the implementation as a DT with information not mentioned before.

Prompt 1: To model a DT, each modeling tool needs the exact mechanical units and sizes of the individual components. These range from the number of channels in millimeters to the power consumption of the motors in watts. Since none of these units were apparent during the structured interview with ChatGPT, a customized prompt was created that named the categories introduced in Table 3 and asked for the associated units.

"Please name all actuators, sensors, and passive components, as well as other passive factors that are needed from an air conditioning system to realize it as a digital twin that should have a focus on energy optimization. For all relevant components, please name their required physical quantities."

The quality of the answers can be described as very good. However, the initial question lacked the physical units for the respective variables, which were supplemented on request. It should be emphasized that by directly addressing passive factors, a higher output was achieved than in the structured interview. Thus, outdoor environmental conditions in the form of outdoor temperature and outdoor humidity, as well as building characteristics and operating hours, were mentioned again. However, from an energy optimization standpoint, the lack of motors on the valves should be considered negative. In addition, ChatGPT also provided the point that special software and algorithms for data analysis, ML, and optimization algorithms are needed for energy savings in addition to the physical components. Another negative assessment is that these are vague statements that could have provided more insight by providing more detailed information, such as the use of Message Queuing Telemetry Transport (MQTT) as an energy-saving protocol for transmitting sensor data, and do not contribute to the facts of the case by providing additional information that is not really helpful.

Prompt 2: In a further attempt, we entered the results from prompt 1 and asked ChatGPT whether it was sufficient.

"I have realized the following actuators, sensors, components and passive factors as well as their physical parameters of a ventilation system in a digital twin: (...) Please tell me whether these components are sufficient to optimize the ventilation system in terms of energy and, if not, name the missing parts and the corresponding physical values.".

On the positive side, new detailed information has again emerged, such as the status of air duct insulation, which can make a positive contribution to energy efficiency and, thus, energy optimization. ChatGPT also adds further facts for passive factors such as solar radiation or wind speed, which are negligible for an air conditioning system DT. Otherwise, the quality of the responses must be rated rather negatively. Both the consistency of the answers and the actual task of realizing a functional DT must be rated as poor. The previously mentioned parts, valves, and heat exchanger, which are both elementary for a DT, especially for energy optimization, were not mentioned. The humidity sensor was correctly detected. However, ChatGPT also provides rather poor answers for the sensors. The accuracy of the temperature sensors cannot really be determined, and the addition of noise level sensors to the sensors would help detect fan and other mechanical component failures or malfunctions, but ChatGPT sees these more in the function of noise level for a more comfortable environment, including fan speed. Building features have also been significantly reduced.

Prompt 3: This prompt is an extension of the second prompt to test the effects of role prompting on the quality of responses. The only adaptation of this prompt is to tell the language model to imagine being an expert in the air conditioning domain.

"Imagine you are an expert in the ventilation system industry and have to assess whether these components are sufficient to optimize the ventilation system in terms of energy. If they are not sufficient, please name the missing components with the corresponding physical values."

The output here is better judged than for the second prompt. The actuators were expanded to include a previously unmentioned humidifier and the previously missing heat exchanger was described more generally in terms of heat recovery systems but was therefore listed as an actuator. The missing valves were again not mentioned, but the list of passive components was at least expanded to include adjustable air diffusers. Questionable components such as noise sensors were no longer listed.

6. Discussion and evaluation

6.1. Limitations

While our research has led to a number of results, it also has many limitations that concern various aspects of the experiment and the process of DTE. The three main categories of our questionnaire were not represented with the same number of questions (evaluation is less meaningful regarding the more detailed facts about air conditioning systems). ChatGPT-4 only works with data through the end of 2021. More recent requirements, findings, and frameworks cannot be considered. In addition, the topic of DT, as well as air conditioning, has proven to be highly complex in the course of this work. Especially with regard to the physical conditions and

the mechanical as well as thermodynamic processes, a sound knowledge base is indispensable. The consistency in queries at different times is mostly not given, and ChatGPT often provides a lot of information without any real classification, and often, there are only more detailed answers for very detailed questions that require prior knowledge.

The experimental design itself also imposes limitations on the work. Due to the framework condition that no prior knowledge is allowed, it was determined that prompting in ChatGPT must not be a fixed part of the experimental setup. For this reason, the usual response pattern with respect to prompt engineering was omitted.

It is not possible to conduct an unstructured interview with ChatGPT, which is why structuring the interview remains the interviewer's task. Further limitations are that ChatGPT, in its current version, cannot interpret or analyze visual material, making it impossible to let ChatGPT evaluate information in a visual structure.

6.2. Interpretation of the results

The statements that can be derived from this work are manifold. This section first discusses the qualitative content analysis to answer the research questions; thus, the findings can be taken from the domain experts' evaluations.

RQ 1: In the context of RE for DT, how consistent and complete is the output of ChatGPT with the information provided by domain experts? There is missing information or mistakes in each category. Therefore, it is important to carefully evaluate and use ChatGPT's generated responses with discretion. Specific inquiries often include superfluous information, leading to confusion. In contrast, ChatGPT presents valuable supplementary information for more general topics. The domain expert confirms that most of the answers here are simple but correct in content but do not go beyond normal textbooks. At one point, he would have liked to swap his answer for that of ChatGPT. It can be concluded that additional unhelpful information, missing information, and errors were identified in many of the responses generated.

Given the presented data limitations, creating a DT through the assistance of ChatGPT in this experiment was not feasible. The absence of data on facility power consumption, actuator status data, 3D models of the facility, and real-time data directly fed from a network only allowed for the development of a dashboard that solely processes the available data. Nonetheless, an attempt was made to program a DT in Modelica using ChatGPT as the sole tool. However, ChatGPT is unable to perform or guide the development of a DT independently. This is because ChatGPT releases the required code in an outdated and unsupported version. Consequently, it is necessary to engage a domain expert with the required expertise to successfully implement a DT. ChatGPT could not determine, through inquiries, that implementing a DT of an air conditioning facility is feasible with the provided resources, and the proposed solution is not viable.

RQ 2: How to design the prompt chain to improve the output of ChatGPT?

Based on this work, we can initially derive a few things to answer RQ 2, but this can only be seen as a starting point and must be continued in future work. In our first attempts, we found that the use of role prompting leads to more nuanced and detailed responses. By trying out different approaches, we found that clarification and specificity, as well as encouraging ChatGPT for more detailed responses, are important aspects to consider. For example, adding the fact that we also need physical quantities of the components in prompt 1 helped improve

the answers. Using these results in prompt 2 (iterative prompting) again helped to refine and clarify the gained information. We also added a feedback loop there, which leads ChatGPT to evaluate and refine the answers.

These results are promising and give reason to go further in this direction in future work.

7. Conclusions and future work

The work presented in this paper addressed the questions of how consistent and complete the output of ChatGPT can be in the context of RE for DT compared to the knowledge of a domain expert and how to design the prompt chain to improve the output of ChatGPT. To answer these questions, we conducted an experiment in which we used ChatGPT to elicit requirements in order to compare them with the knowledge of a domain expert. We then used the knowledge gained to try to change the prompt in order to improve the output.

A first implication is that the AI tool ChatGPT should not be seen as a substitute but as a useful building block to better design DTs. ChatGPT can compensate for human errors through a synthetic approach with the domain expert and can be used to supplement the expert's remarks at many points due to its constant availability, as long as ChatGPT's statements are always critically scrutinized afterward. The advantage here is also a more effective use of the domain expert's time. The most effective way to get all the data in ChatGPT to create a DT is to create a list of components and have ChatGPT check them for completeness. With a larger input set, ChatGPT returns a more diverse output set, which can lead to the expansion of some components and the avoidance of errors due to the addition of forgotten components. Some open questions for future work derived from our results are:

(1) Does working out a topic with ChatGPT and domain experts save time compared to working it out without ChatGPT? Since ChatGPT qualitatively supports the early phases of DTE, the question arises whether a quantitative benefit (e.g., time) can be measured.

(2) How can the evaluation scale be adapted to the existing knowledge of the domain experts? On the conceptual level, the evaluation of the effort of ChatGPT by the domain expert depends on the domain expert's existing knowledge. The comparability of the statements of ChatGPT suffers from this since their evaluation depends directly on the person evaluating the statements.

(3) How does ChatGPT perform among several domain experts? As already mentioned, the evaluation of the statements of ChatGPT by the domain expert depends on the domain expert. For a more informed evaluation, it would make sense to use several domain experts for the evaluation in order to check whether the core statements differ among the various experts.

(4) How does ChatGPT perform in other disciplines? Besides DTs and air conditioning systems, a whole range of topics could be considered for processing with ChatGPT. It would be interesting to know if there are significant differences in the level of knowledge and expressiveness of ChatGPT.

In conclusion, a more comprehensive investigation of the benefits of ChatGPT for more and different use cases seems relevant and necessary to better understand its potential and limitations. It is essential that a diverse range of domain experts play a role in this research. The contribution of this paper serves to confirm the importance and promise of continued research in this area.

References

[1] B. Sniderman, M. Mahto, M. J. Cotteleer, Industry 4.0 and manufacturing ecosystems: Exploring the world of connected enterprises, Deloitte Consulting 1 (2016) 3–14.

[2] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[4] W. Huang, P. Abbeel, D. Pathak, I. Mordatch, Language models as zero-shot planners: Extracting actionable knowledge for embodied agents, in: International Conference on Machine Learning, PMLR, 2022, pp. 9118–9147.

[5] L. Floridi, M. Chiriatti, Gpt-3: Its nature, scope, limits, and consequences, Minds and Machines 30 (2020) 681–694.

[6] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig, Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, ACM Computing Surveys 55 (2023) 1–35.

[7] G. N. Schroeder, C. Steinmetz, R. N. Rodrigues, R. V. B. Henriques, A. Rettberg, C. E. Pereira, A methodology for digital twin modeling and deployment for industry 4.0, Proceedings of the IEEE 109 (2020) 556–567.

[8] Y. Qamsane, J. Moyne, M. Toothman, I. Kovalenko, E. C. Balta, J. Faris, D. M. Tilbury, K. Barton, A methodology to develop and implement digital twin solutions for manufacturing systems, IEEE Access 9 (2021) 44247–44265.

[9] J. Liu, J. Liu, C. Zhuang, Z. Liu, T. Miao, Construction method of shop-floor digital twin based on mbse, Journal of Manufacturing Systems 60 (2021) 93–118.

[10] D. J. Wagg, K. Worden, R. J. Barthorpe, P. Gardner, Digital Twins: State-of-the-Art and Future Directions for Modeling and Simulation in Engineering Dynamics Applications, ASCE-ASME J Risk and Uncert in Engrg Sys Part B Mech Engrg 6 (2020) 030901.

[11] B. Kitchenham, Procedures for performing systematic reviews, Keele, UK, Keele University 33 (2004) 1–26.

[12] D. I. Sjøberg, J. E. Hannay, O. Hansen, V. B. Kampenes, A. Karahasanovic, N.-K. Liborg, A. C. Rekdal, A survey of controlled experiments in software engineering, IEEE transactions on software engineering 31 (2005) 733–753.

[13] T. D. Cook, D. T. Campbell, W. Shadish, Experimental and quasi-experimental designs for generalized causal inference, Houghton Mifflin Boston, MA, 2002.

[14] Y. Xia, M. Shenoy, N. Jazdi, M. Weyrich, Towards autonomous system: flexible modular production system enhanced with large language model agents, in: IEEE 28th International Conference on Emerging Technologies and Factory Automation (ETFA), 2023, pp. 1–8.

[15] N. Ivanovic, B. Nast, A. Reiz, K. Sandkuhl, Technologies for a diagnostic technique for hvac systems using iot and cloud-based architecture, in: 2023 International Interdisciplinary PhD Workshop (IIPhDW), 2023, pp. 1–6.

[16] P. Mayring, Qualitative inhaltsanalyse: Grundlagen und techniken, Dt. Studien-Verlag, 1997.