Evaluating the Performance of ChatGPT in the Automation of Maintenance Recommendations for Prognostics and Health Management

Sarah Lukens[1], Asma Ali[2],

[1] *LMI, Tysons, VA, 22102, USA*
sarah.lukens@lmi.org

[2] *GE Vernova, Chicago, IL, 60661, USA*
asma.ali1@ge.com

Abstract

Until now, automation of maintenance recommendations for Prognostics and Health Management (PHM) has been a domain-specific technical language processing (TLP) task applied to historical case data. ChatGPT, Bard, GPT-4 and Sydney are a few examples of generative large language models (LLMs) that have received significant media attention for their proficiency in natural language tasks across a variety of domains. Preliminary exploration of ChatGPT as a tool for generating maintenance recommendations has shown promise in its ability to generate and explain engineering concepts and procedures, but the precise scope of its capabilities and limitations remains uncertain. Currently we know of no performance criteria related to formally measuring how well ChatGPT performs as a tool for industrial use cases. In this paper, we propose a methodology for the evaluation of the performance of LLMs such as ChatGPT for the task of automation of maintenance recommendations. Our methodology identifies various performance criteria relevant for PHM such as engineering criteria, risk elements, human factors, cost considerations and corrections. We examine how well ChatGPT performs when tasked with generating recommendations from PHM model alerts and report our findings. We discuss the various strengths and limitations to consider in the adoption of LLM's as a computational support tool for prescriptive PHM as well as the different risks and business case considerations.

1. Introduction

The rapid advancements by the artificial intelligence (AI) community in the development in the area of natural language processing (NLP) are creating a widening gap between their transformative impact on everyday life and their abilities to address the technical language needs in industry. The engineering approach which leverages NLP tools intelligently and selectively for technical language data has been dubbed Technical Language Processing (TLP) (Brundage, Sexton, Hodkiewicz, Dima, & Lukens, 2021). Although out-of-the-box NLP models trained on general language may not directly apply to industrial text, promising advancements have been made in adapting NLP techniques to technical domains, generating substantial interest within the Prognostics and Health Management (PHM) community (Dima, Lukens, Hodkiewicz, Sexton, & Brundage, 2021; Nandyala, Lukens, Rathod, & Agarwal, 2021).

One such area of promising NLP development which has emerged in the past year are large language models (LLMs) such as ChatGPT, which are a subset of the broader domain of Generative AI. Generative AI encompasses AI algorithms that generate new data or content, distinct from discriminative AI that categorizes and classifies existing data. Through deep learning, NLP, and computer vision, generative AI models analyze patterns in existing data to produce statistically similar output. Applications span from generating realistic images to composing music and creating art, showcasing the vast possibilities enabled by generative AI technologies. LLMs specifically employ massive amounts of text data to generate human-like language, enabling tasks such as conversational responses, text completion, and document generation.

PHM involves a strategic decision-making process based on diagnostics or prognostics information, resource availability, and operational demands. It encompasses crucial elements such as data collection, predictive modeling and the critical task of initiating appropriate actions based on extracted information, while continuously validating the accuracy of predictions. One specific task in PHM, which may be well-suited for LLMs, is generating recommendations for troubleshooting actions in response to alerts from PHM models. Currently, this process involves multiple stakeholders, where an analyst or reliability engineer identifies an alert and collaborates with others for troubleshooting.

This work aims to propose a methodology for evaluating the effectiveness of LLMs in generating troubleshooting recommendations in response to PHM model alerts. Application of Generative AI to this task introduces a paradigm shift away from relying on using historical data from past alerts for training AI models and towards the utilization of pre-trained models which contain non-technical data in their training corpus. This transition carries inherent risks as incorrect suggestions or actions can result in significant financial, safety and environmental consequences. For this reason, we have developed a systematic approach to evaluate LLMs, which start with general maintenance and reliability knowledge, focus on PHM specific knowledge, and lastly focus on specific recommendations.

The following questions were used as guidance in the development on our approach for assessing the performance and feasibility of an LLM (such as ChatGPT) in the industrial domain.

- How well can the LLM grasp central concepts within the industrial domain?
- How well does the LLM transfer to the technical language domain?
- How often will the LLM return inconsistent responses and how will these affect the scores?
- Is the LLM appropriately specific or too general in its responses?
- How does the hallucinogenic effect manifest and to what extent?
- Are the LLM responses physically sound?
- Are there safety risks in the LLM responses?
- Can the LLM make deductions to return a response?

The results of our study are targeted at addressing some of the larger, more broad questions: How well do these chat models perform in generating recommendations? Should we consider shifting towards LLMs as a preferred approach over traditional methods that heavily rely on historical cases? In this paper, we first develop a rubric for scoring LLM responses and then take a step by step approach for the evaluation using the rubric. First, a general knowledge assessment of general maintenance and reliability knowledge is performed. We then perform a more specific knowledge assessment using proficiency questions for an analyst monitoring predictive health models. Lastly, once the extent of LLMs grasp of fundamental knowledge and concepts is assessed, we specifically evaluate LLM as a recommendation engine in response to PHM alerts.

The remainder of this paper is organized as follows. Section 2 provides a background on the PHM recommendation task and Generative AI models. In Section 3, we present the methodology for evaluating the performance of the models, including details on the test design, utilized data, experimental setup, scoring rubric, and evaluation process. Section 4 presents the results obtained from the maintenance and reliability knowledge exam, while Section 5 focuses on the results of the PHM knowledge exam. Section 6 provides the results specific to PHM troubleshooting recommendations. The study's conclusions can be found in Section 7. Finally, in Section 8, we engage in a detailed discussion and present recommendations for the PHM community to consider.

2. BACKGROUND

In a Monitoring and Diagnostics (M&D) center, analysts play a critical role in real-time monitoring of industrial assets from the outputs of sensor-based PHM models. When an alert is triggered, these analysts are responsible for assessing the situation and deciding whether to escalate the issue to the reliability and maintenance organization at the plant level. Drawing from their deep technical experience, the analysts provide recommendations for initiating appropriate actions such as recommended steps for troubleshooting the cause of the fault. At this stage, a person in the reliability organization at the plant identifies the potential root cause of the fault and creates a work order notification to initiate the work management process (ex: order parts, schedule maintenance, etc). The ultimate vision is for a system to automatically generate the work orders tied to actionable alerts that would direct the plant reliability teams as to what to do. However, the complex relationship between alerts and potential failure modes necessitates an intermediary troubleshooting step. Accurately pinpointing the root cause in remote monitoring space remains a challenge, requiring domain expertise and comprehensive analysis to determine the appropriate course of action.

The development of automatic recommendations in response to alerts from sensor-based PHM models serves two primary purposes: first, to capture the expertise of analysts and provide support to junior professionals, addressing the challenges posed by an aging workforce; and second, to standardize recommendations, ensuring consistent outcomes for identical inputs measured on equipment, independent of the analyst involved. Many M&D centers have databases of historical cases which have grown in size over the years, and several recent studies have explored applying a TLP approach to this historical case data for knowledge extraction in maintenance and troubleshooting.

2.1. Related Work: Suggestion Systems in Industry

Real-time suggestion systems have been under development in the industrial domain and can be viewed as a class of TLP tasks related to chat tasks through the deliverable of an actionable recommendation. Real time-suggestion systems are algorithms aimed at suggesting information of interest and can provide prescriptive decision support in many applications but require populated knowledge frameworks (Lepenioti, Bousdekis, Apostolou, & Mentzas, 2020; Lepenioti, Pertselakis, et al., 2020). Knowledge systems are systems which can access and extend a collection of knowledge stored in a representation language, such as a standard ontology or schema. Ontologies developed specific to maintenance data and fault diagnosis have been developed (Rajpathak, 2013; Karray, Ameri, Hodkiewicz, & Louge, 2019; Hodkiewicz, Klüwer, Woods, Smoker, & Low, 2021) as well as maintenance actions (Woods, Selway, Bikaun, Stumptner, & Hodkiewicz, 2023) and maintenance procedures (Woods, French, Hodkiewicz, & Bikaun, 2023). TLP can be used to formally structure unstructured text for populating such systems, and there has been work developing larger knowledge frameworks for industrial equipment in manufacturing where creating knowledge databases which integrate tools for structuring the unstructured data fields (Ansari, 2020; Ansari, Glawar, & Nemeth, 2019).

Unified frameworks which standardize and merge siloed data sources as a knowledge base for operational systems can be useful for many applications such as maintenance planning, performance benchmarking, root cause analysis, predictions on future performance. Real time suggestion systems have been proposed which utilize structured databases of historical failures and actions taken to make recommendations for maintenance actions in real time (Bastos, Pedro and Lopes, Isabel and Pires, LCM, 2012) (Bokinsky et al., 2013). Other similar proposed real-time suggestion systems estimate workload based on the nature of the failure mode (Usuga Cadavid, Grabot, Lamouri, Pellerin, & Fortin, 2020) and suggest where to route a work order based on information from the text (Bouabdallaoui, Lafhaj, Yim, Ducoulombier, & Bennadji, 2020). Real time suggestion systems in knowledge frameworks have also been suggested for root cause analysis (Brundage, Kulvatunyou, Ademujimi, & Rakshith, 2017) and for finding similar solutions to past issues (Alfeo, Cimino, & Vaglini, 2021).

Real time suggestion systems specifically designed for the task of offering troubleshooting recommendations in response to a PHM model alert have been under development as well. Pau, Tarquini, Iannitelli, and Allegorico (2021) (Pau, Tarquini, Iannitelli, & Allegorico, 2021) utilized NLP techniques for consistent troubleshooting insights in an M&D center, while Peshave et al. (Peshave et al., 2022) evaluated approaches for vectorization of short-text case titles. Trilla, Mijatovic and Vilasis-Cardona (2022) (Trilla, Mijatovic, & Vilasis-Cardona, 2022) used TLP for troubleshooting in PHM and developed a failure ontology and a data-driven quality strategy. Pires, Leitao, Moreira and Ahmad (2023) (Pires, Leitão, Moreira, & Ahmad, 2023) compared different recommendation systems, including their own discrete event simulation model, and demonstrated improved user ratings over state-of-the-art recommendation systems. Addepalli, Weyde, Namoano, Ayodeji Oyedeji, Wang, Erkoyuncu and Roy developed a knowledge extraction framework providing information in response to a degradation where historical degradation information was extracted from full text papers (Addepalli et al., 2023).

2.2. Generative AI

Generative AI technologies are rapidly evolving and are being utilized across various fields such as finance, healthcare, and entertainment, among others. One of the most popular generative AI technologies is ChatGPT. ChatGPT is a natural language processing model that was first introduced as a research preview prototype in November 2022 (OpenAI, 2022). It is a variant of the GPT-3.5 LLM that has been pretrained on a large dataset of text obtained from various online sources, such as books, websites, and articles published until 2021. The exact size of the dataset has not been disclosed. ChatGPT has been fine-tuned using Reinforcement Learning from Human Feedback or Reinforcement Learning from Human Preference (RLHF/RLHP) techniques to improve its responses' coherence and contextual appropriateness, by incorporating feedback and correction from human inputs.

Other LLMs include BARD (Building Automated Reasoning and Decision-making), which uses deep reinforcement learning to generate models that can make decisions based on real-world situations (Google, 2023). AWS Titan is a generative AI technology that utilizes GANs (Generative Adversarial Networks) to create high-quality synthetic data, which is useful for training machine learning models (Amazon, 2023). Bedrock is another powerful generative AI technology that can create highly realistic synthetic data, and it has been used in various industries such as healthcare and finance. Finally, GPT4 is the latest iteration of the GPT (Generative Pretrained Transformer) series, and it is expected to further enhance the capabilities of language generation and other natural language processing tasks (OpenAI, 2023). These five generative AI technologies are just a few examples of the vast and rapidly evolving landscape of AI technology, and their potential applications are vast and varied.

Benefits and Risks.

Adoption of LLMs as industry tools is a nascent area and formal identification of their benefits and risks specifically in the PHM space and mapping these risks to mitigation strategies and recommendations is still an open research topic for the PHM community. Since the inception of this article in February 2023 and making final revision in August 2023, there have been many works published across different domains, including domains of relevance to industrial applications in this area. Generic benefits and risks are well reviewed in (Ray, 2023). More specific to the industrial domain, are publications such as which suggest many relevant tasks such as demand forecasting, logistics, inventory management and supply chain risk management (Bahrini et al., 2023; Chowdhury et al., 2023), uses for engineering education (Qadir, 2023) and manufacturing (Wang, Anwer, Dai, & Liu, 2023; Rathore, 2023).

We provide a non-exhaustive list of benefits and risks here as an overview to provide a sample of the various aspects which need consideration. In terms of benefits, LLMs excel at formulaic writing tasks, earning ChatGPT the reputation of a brainstorming tool. LLMs have proven to be a useful writing aid that can assist users in crafting more coherent and appropriate responses. More specifically to maintenance recommendations, LLMs have the potential for use in recommendation systems, automating communication of insights from analysis from large amounts of data and data mining data from different sources.

While LLMs have shown great potential in providing human-like responses to a wide range of queries, there are many risks associated with its use for decision making. One concern is lack of data diversity, which may result in cultural bias and lead to inaccurate or inappropriate responses. Data security is another issue, as any data input into LLMs becomes part of its training dataset and is effectively in the public domain. Other significant risks include lack of interpretability and safety concerns, such as spreading misinformation, are also significant risks. Furthermore, LLMs currently lack regulation, and scaling up its use could prove difficult. Limited understanding of the LLMs capabilities and limitations among users could lead to unrealistic expectations and potential misuse. It is essential to carefully consider these risks and develop appropriate safeguards when using LLMs for decision making.

The term "hallucination" is frequently used in reference to the "hallucinogenic effect" that these models can have. While these models are designed to generate coherent and contextually relevant responses, their underlying architecture and training methodologies can give rise to unexpected outcomes. The hallucinogenic effect observed in many responses across the study are due to the model's ability to simulate human-like response while lacking truth or subjective experiences. LLMs hallucinate because they "lack the understanding of the cause and effect of their actions" (Ortega et al., 2021). While some generative AI tools have shown promise in laboratory settings, concerns remain that they may produce coherent-sounding yet inaccurate information (Peng et al., 2023).

When it comes to industrial assets, safety is paramount. Dependence on LLMs for critical information related to the operation, maintenance, or troubleshooting of industrial assets carries inherent risks. LLMs are trained on vast amounts of text data, including diverse sources that may not always be reliable or up to date. As a result, the responses generated by LLMs may contain inaccuracies, outdated guidelines, or conflicting information, posing safety hazards if followed without proper verification.

The internet has seen numerous instances of LLMs such as ChatGPT generating content that is demonstrably inaccurate. Doug Hubbard's *exsupero ursus* fallacy, which posits that algorithms must be perfect to be preferred by experts regardless of the performance of the alternative, is an interesting point to consider in this context (Hubbard, 2020). Historically, critics have tended to view anything less than perfect performance as proof of shortcomings in established methods. However, it is worth noting that single anecdotes cannot be used to compare the relative performance of one model against another and motivates the need to formalize model evaluation rubrics as well as safeguards around when and how to appropriately use such models to support decision making.

2.3. Assessing ChatGPT performance in other domains

Various studies across different domains such as the medical community have been reported for the assessment of ChatGPT and other LLMs. An extensive review of applications across the domains of healthcare and medicine, business and finance, banking, law and legal services, creative writing and content generation, education and training, programming and debugging, media and entertainment, sales and marketing and public outreach are found in (Ray, 2023). For general evaluation of models abilities to perceive, understand, judge and reason, (Bian et al., 2023) evaluated ChatGPT's reasoning ability using 11 open-source commonsense QA datasets across 8 domains. ChatGPT was able to accurately answer commonsense questions, particularly in science and prototypical questions. However, its problem-solving skills and ability to distinguish relevant and irrelevant knowledge were limited.

(Kung et al., 2023) evaluated ChatGPT's performance on the United States Medical Licensing Examination (USMLE) using 376 test questions using an accuracy, concordance, and insight (ACI) scoring system. ChatGPT achieved a 94.6% concordance across all questions, outperforming other LLM models and showing potential in augmenting medical education. In (Jalil, Rafi, LaToza, Moran, & Lam, 2023), ChatGPT's ability to answer questions related to software testing was evaluated using a popular textbook. Results showed that

ChatGPT was correct or partially correct in about 44% of cases and provided correct or partially correct explanations in 57% of cases.

In (Rao et al., 2023), ChatGPT's performance on radiologic decision-making for breast cancer screening prompts was evaluated and achieved moderate accuracy. (Yeo et al., 2023) assessed ChatGPT's accuracy, completeness, and reproducibility in answering frequently asked questions about managing patients with cirrhosis and hepatocellular carcinoma. ChatGPT performed better in basic knowledge, lifestyle, and treatment compared to the domains of diagnosis and preventive medicine.

3. METHODOLOGY

We describe the evaluation tests used to assess the LLM, the scoring rubric used for evaluating the LLMs and the experiment and grading process.

3.1. Data

Our experimental approach for assessing LLM performance is in three stages of evaluation: first we evaluate the LLMs general knowledge of maintenance and reliability in order to get a general gauge of its knowledge across the maintenance domain. Second, we evaluate the LLMs general knowledge specific to PHM with knowledge required for an M&D analyst. Lastly, we test out the LLMs capability for the recommendation task specifically.

Maintenance & Reliability Knowledge Assessment. We used a 76- question multiple-choice exam adapted from Ramesh Gulati's "Maintenance and Reliability Best Practices" (R. Gulati and R. Smith, 2021) as the test for basic maintenance and reliability knowledge. The book is a commonly used resource for professional certification exams, such as the Society of Maintenance and Reliability (SMRP)'s Certified Maintenance and Reliability Profession (CMRP) certification. The book contains the multiple choice questions along with their solutions and explanation, which were used as an answer key for grading the model performance. We assume a score of 60% is "passing".

The questions cover various aspects of maintenance and reliability knowledge which cover the five pillars in the SMRP Body of Knowledge and Best Practices (Society of Maintenance & Reliability Professionals, 2017): Business Management, Manufacturing Process Reliability, Equipment Reliability, Organization and Leadership and Work Management. Due to class imbalance in topics across the 5 Pillars (the Work Management pillar contains 45 out of the 68 (66%) of the metrics as well as concepts ranging from planning, scheduling, work execution, different work types, stockroom and inventory management and predictive maintenance), we map the questions to topics more balanced and relevant to our analysis. The different categories are shown and explained in Table 1.

PHM Industrial Domain Knowledge Assessment. The test used for assessment of PHM knowledge was a 63 question exam adapted from the GE Vernova's commercial PHM solution (APM SmartSignal) and this exam is used by the Industrial Managed Service (IMS) team as a knowledge test for new Industrial Subject Matter Expert (SME) hires in the M&D Center. Overall passing requirement is 80%, where a score 80% or below is deemed unsuitable for an applicant applying for the position. Due to proprietary reasons of the nature of the test, the test itself cannot be shared.

The exam is a test of in-depth knowledge base questions that are designed to test an individual's knowledge of process, industrial equipment and failure modes and mechanisms as well as testing for a strong ability to diagnose and troubleshoot problems in a remote monitoring environment. The exam covers questions that cover in-depth industrial verticals such as Power, O&G and Mining & Metals, knowledge of the equipment and process and overall domain knowledge. The different categories are shown and explained in Table 2.

PHM Troubleshooting Assessment The specific test for assessing a LLMs capability for a PHM recommendation task was performed by asking specific questions based on historical cases, such as those made publically available by GE Vernova (GE, 2023). The focus for this assessment was on the ability to make a troubleshooting assessment, and for reporting, the test was focused on evaluating the response to one question asked multiple ways.

3.2. Experiment

For both knowledge examinations, each question was input to the LLM in a standard format. If the question was True or False, the prompt was formatted as: "Answer the following true or false question: [Question]. Why?". If the question was multiple choice, the prompt was formatted as: "Answer the following multiple choice question: [Question]. Why?". Two examples are shown below.

- **True or False question:** "Answer the following true or false question: Vibration monitoring can detect uniform impeller wear. Why?"
- **Multiple choice question:** "Answer the following multiple choice question: What percentage of maintenance work should be proactive? (a) 100%; (b) 85% or more; (c) 50%. Why?"

The approach developed in this work is generic across LLM. At the time of writing, there have been 5 Generative AI language model open to use, but we focus on one primarily with a second for illustrative purposes, both of which will certainly have been updated since the publication of this manuscript.

Table 1. Different categories covered by the maintenance and reliability knowledge examination.

Category	SMRP Pillar(s)	Definition	Example
Work Management	Work Management	Related to Work Planning, Scheduling and Execution as well as different types of work	Question on the primary purpose of scheduling work
Storeroom & Inventory Management	Work Management		Question about Economic Order Quantity (EOQ)
Operator-Driven Reliability	Manufacturing Process Reliability		Question about Overall Equipment Effectiveness (OEE)
Strategy Development & Management	Work Management	Around the development and management of maintenance strategies	Question about Reliability-Centered Maintenance (RCM) or Failure mode and effects analysis (FMEA)
Metrics	Equipment Reliability	About particular metrics, such as for reliability, availability and maintainability	Question about calculating Failure Rate or lagging indicators
Workforce Management & Leadership	Business Management; Organization and Leadership	Around managing the workforce and programs to support and contribute to business results	Questions about workforce skills, reliability culture or team development
PHM Technologies	Work Management	Related to specific condition assessment technologies	Question about vibration analysis

Table 2. Different categories covered by the PHM knowledge examination

Category	Definition	Example
Fundamental Definition	Regarding equipment or process or root causes	What is SAC and DLE in Combustion Turbines?
SME Knowledge & Expertise	Knowledge of verticals versus Equipment versus Process	If Hot Start occurs, can we attempt the startup? State the reasons.
Fundamental Equipment Knowledge	Knowledge of industrial equipment	How many stages on axial compressor in LM2500 and how does it differ with LM6000?

The LLMs used in the study were ChatGPT (based on the GPT-3.5 architecture, developed by OpenAI, with a knowledge cutoff of September 2021; accessed March 2023) and Google's Bard (based on the BERT language model architecture; accessed April 2023). As the main purpose of this paper is to propose how to evaluate LLMs for PHM rather than compare specific models (which will most definitely change or have been updated at time of publication), the ChatGPT responses will be referred to as "AI1" and the Bard responses as "AI2" for the rest of this report in order to make the analysis more generic to any LLM model comparison.

3.3. Scoring Rubric

Responses from the LLM were broken down into two parts, *answer* (Overall Correctness) and *explanation*. Overall Correctness is the straightforward score corresponding to how an exam would be graded for anyone through matching correct multiple choice answers. The *explanation* part consists of more in-depth scoring for analyzing strengths and weaknesses of the LLMs' responses and addressing our guiding research questions. For this work, we adapted the Accuracy, Concordance, Insight (ACI) scoring system from (Kung et al., 2023) to the industrial domain, using the below definitions:

- **Accuracy**. A response is *accurate* if it identifies the central concept being tested, is specific, is physically sound and answers the question correctly.
- **Concordance**. A response is *discordant* if any part of the explanation contradicts itself, otherwise it is concordant.
- **Insight**. A response contains *insight* if it demonstrates knowledge and intuition. For our purposes, we focused on if the answer was general (could apply to anything), defines a term in the question and demonstrates some capacity to reason.

Two additional bookkeeping scoring questions were added to the Accuracy category. The observed tendency of LLMs to hallucinate motivated the development of a scoring question to track the wrong claims made by the AI across the responses. The second scoring question was based on the additional observed tendency of LLMs is to contain extraneous

Table 3. Scoring rubric for evaluating ChatGPT response adapted from the ACI scoring system

Category	Sub-category	Question	Possible Answer
Answer	Overall Score	Was the response correct?	Yes/No
Accuracy	Accuracy Rollup Score	If the following 4 measures are true:	Yes/No
	Central Concept	AI response identified the central concept being tested	Yes/No/(Partially)
	Correct	AI response answers the query correctly	Yes/No/(Partially)
	Specific	AI response is specific	Yes/No/(Partially)
	Physically Sound	AI response is physically sound	Yes/No/(Partially)
Accuracy	Bookkeeping	AI response contains an unrelated concept	Yes/No
	Bookkeeping	AI response hallucinates	Yes/No
Concordance		Is the explanation concordant?	Yes/No
Insight	Generality	Measure of Generality	Least (1) to Most Specific (5)
	Definitional	AI Response defines a term in the input question	Yes/No/(Partially)
	Nonobvious	AI Response demonstrates external knowledge or deduction to the question input	Yes/No/(Partially)
PHM Recommendation specific	People elements	AI response provides instructions a person can follow	Yes/No/(Partially)
	Risk/Safety elements	AI response does not contain safety issues in the recommendations	Yes/No/(Partially)

content, and tracks responses containing unrelated extraneous content. Two additional measures specific to the task of PHM recommendations were also added:

- **People elements**: Is it readable? Does the response provide instructions that someone can follow?
- **Risk (Safety) elements**: Are there any safety issues in the recommendation?

The full scoring rubric is summarized in Table 3.

3.4. Response labeling

The *explanation* part was scored using responses of "Yes", "No" with an option of "Partially" to a series of questions. We note that in knowledge examinations, "Partially" may be an acceptable answer, but in many cases which map to physical systems, if an answer is not correct then it must be wrong. In the cases when "Partially" was used as an acceptable answer, two scores per scenario were compared: those censoring "Partially" (treat as a no), and those that included "Partially" (treat as a yes) to compare the spread.

The grading of the Maintenance and Reliability exam was completed by the paper authors, who (1) have CMRP certification and (2) have over ten years of experience in the development and support of commercial software and services for maintenance and reliability. The PHM Industrial knowledge exam was assessed by GE Vernova's Data Science & Analytics team and grading was completed by four of GE Vernova's Industrial Managed Services subject matter experts, all with 25+ years of industrial experience. For both exams, reasoning for "No" or "Partially" responses were recorded when grading and later used by the paper authors when auditing contradictory responses. Final scores used were agreed upon by consensus.

4. RESULTS 1: MAINTENANCE & RELIABILITY KNOWLEDGE EXAM

For Overall Correctness, AI1 got 55 questions correct with a score of 72% and AI2 got 49 questions correct with a score of 64%. A summary plot of the accuracy scores is shown in Figure 1 with a 60% passing score on the overall correctness to the questions. Overall, both models tended to perform better in grasping the central concept and by being specific more so than getting the correct answer.

The measures of concordance were 82.2% for AI1 and 77.6% for AI2, both reflecting generally high consistency in the responses. This is supported by the correctness-concordance contingency table (counts between correct-concordant, correct-discordant, incorrect-concordant and incorrect-discordant), where AI1 was correct-concordant 72% of the time and AI2 was correct-concordant 61% of the time, showing high agreement with correct responses and concordance. Both had mean insight scores around 3 (AI1; 3.1 and AI2; 3.5), but AI1 had significantly more spread.

LLM responses often contain extraneous information. AI2 had significant higher scores for containing unrelated concepts (51%) compared to AI1 which has unrelated content in the responses about 13% of the time, meaning that about half the time the responses from AI2 contained extraneous information. AI1 demonstrated a definitional ("AI response defines a term in the input question") rate of 45%/49% (with/without censoring), whereas AI2 exhibited a higher definitional rate of 62%. The increased definitional rate of AI2 aligns with its tendency to include unrelated concepts in

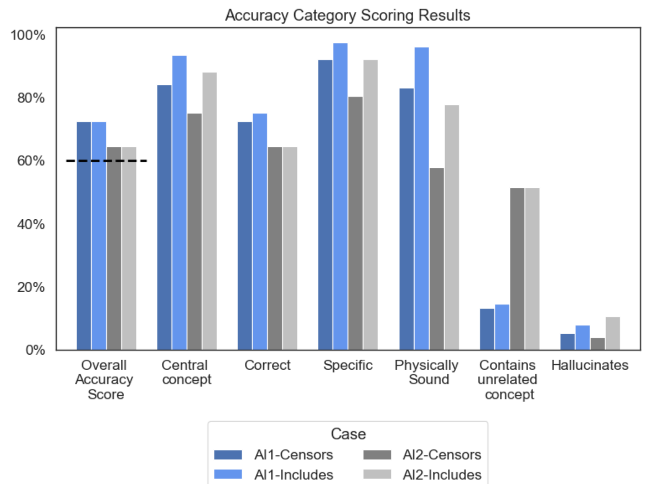

Figure 1. Maintenance and Reliability Knowledge Exam: Summary of scoring in the "Accuracy" category across the two models (AI1 and AI2) in both the cases where scores of "Partial" are censored and included. The dashed lines across the overall accuracy score denote 60% which is a passing score.

its responses. In general, AI2 provided a surplus of information, regardless of its relevance, with 39% of its responses containing both definitions and unrelated concepts, compared to 8% for AI1. Both AI models had limited deductive capabilities and instead relied on providing extensive information.

LLMs may (but not always) hallucinate concepts in maintenance and reliability. AI1 responses contained incorrect hallucinations 5%/8% of the time (censored/included), while AI2's hallucinations ranged from 4%/10% of the time. Hallucinations were observed through the AI creating definitions or technology purpose. An example of a made up definition was: "The 10% Rule of [Preventative Maintenance (PM)] is a guideline that suggests that maintenance spending should be around 10% of the replacement value of an asset". In reality, the 10% Rule of PM is a PM compliance rule stating that a PM work orders must be completed within 10% of its time frequency. Under technology, the AI responded: "Karl Fischer's coulometric titration method is an effective technique to determine the metallic content (in [parts per million (PPM)] in an oil sample". In actuality, Karl Fischer's titration method quantifies water content in a sample.

Hallucinations were also observed through valid logic applied to wrong numbers. In one example, the AI responded: "Typically, it is recommended that 30-50% of assets should be ranked as critical based on the risk to the business." The logic is correct as the purpose of asset criticality is to prioritize which assets to focus on in a reliability initiative. However, in a plant with tens of thousands of assets, focusing on 30-50% is not practical.

Accurate and industry-aligned definitions play a crucial role in the field of maintenance and reliability. However, it was observed the responses often returned definitions that do not align well with industry standards and best practices established through years of consensus within the maintenance and reliability community which could be potentially problematic. For example, Reliability is the probability that an asset or component will perform its intended function for a specified time period under specified conditions (citation), but AI2 defined: "Reliability is a measure of how often an asset is available to produce good parts" (AI2). Similarly, Maintainability is the measure of the ability of an item to be retained in or restored to specified condition when maintenance is performed (citation), but AI2 defined: "Maintainability is a measure of how easy it is to maintain an asset" (AI2).

Additionally, both AI models tended to struggle with understanding the concept of Reliability-Centered Maintenance (RCM). While both models had the knowledge of the general textbook structure for implementing RCM (alignment with SAE's JA1011 (JA1011, SAE, 2009)), neither approach made the deduction that RCM was a risk-based approach to developing a maintenance strategy, and instead both suggested use of a risk-based approach over RCM.

LLM performance varied across the different knowledge categories. The number of correct responses for each model against each of the category groupings are shown in Figure 2. In general, both models performed higher across Workforce Management and Leadership questions (which were more generic and less domain specific) and questions around manufacturing operations. The lower performing areas across both models were Work Management and maintenance strategy development (driven by the general observation that both models struggled with understanding RCM despite having knowledge of RCM). AI2 performed lower at PHM technologies, inventory management and metrics. Much of the lower performance by metrics was observed by its non-ability to make deductions about relationships, even when knowing their definitions.

The Accuracy scores (Table 1) showed the highest variation in scoring was across the Physically Sound measurement. The Physically Sound scores across both models are shown in Figure 3, showing the largest discrepancies between the models in the PHM Technologies and Metrics categories. For example, AI2 did not deduce that the metric Mean Time Between Failures (MTBF) is the inverse of Failure Rate, even though it knew the definitions of both metrics which are inverses of each other.

Table 4 presents an example within the PHM technologies domain comparing two AI responses to a prompt regarding vibration monitoring (True or False: Vibration monitoring can detect uniform wear). AI1 is correct in responding that visual inspections and wear measurement are the best methodology to track wear measurement due to the reasons stated. How-

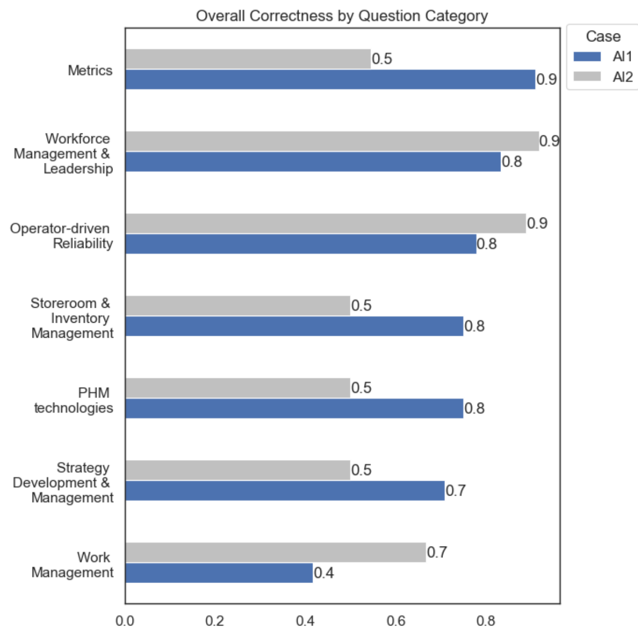

Figure 2. Maintenance and Reliability Knowledge Exam: Overall correctness by question category for Maintenance and Reliability Knowledge examination between the two models (AI1 and AI2)

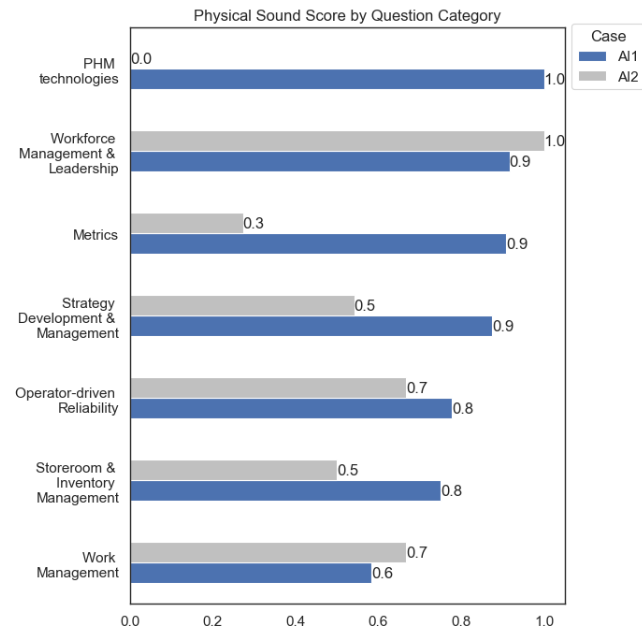

Figure 3. Maintenance and Reliability Knowledge Exam: Comparison of the accuracy scores for "AI response is physically sound" across the different question categories for both LLLMs.

ever, in practice, it is also possible to detect wear through vibration monitoring if its not uniform; uniform wear will create less imbalance. By the time the manifestation of profile change in vibration will occur, the wear may be significant enough or may get convoluted with other operational issues. The responses indicate an example of extremely complicated phenomena to diagnose in the world of remote monitoring and diagnostics.

5. RESULTS 2: PHM KNOWLEDGE EXAM

For Overall Correctness, AI1 got 43 questions correct with a score of 67% and AI2 got 35 questions correct with a score of 56%. A summary plot of the accuracy scores is shown below in Figure 4. Similar to the general knowledge exam, both models overall tended to perform better in grasping the central concept and by being specific more so than getting the correct answer.

The measures of concordance were 65% for AI1 and 49% for AI2, reflecting high amount of discordance in response, particularly in comparison to the high concordance of the maintenance and reliability knowledge examination. There was high correlation between accuracy and concordance for this set of questions, shown by the correctness-concordance contingency table shown in Figure 5. AI1 was either correct-concordant or incorrect-discordant 98% of the time, while AI2 was 93%. For this set of questions, if the answer was correct, it was nearly always concordant and vice versa.

LLM responses tended to be either very specific or too general. The Generality average was 3.7 for AI1 and 2.8 for AI2. While the averages have central values, the generality distribution across the questions were bimodal for both LLMs, shown in Figure 6, an indication of the response tendency to fall in either one extreme or the other. Generally, the degree of specificity varies. For example, in response to a question regarding Bearing failures, AI2 provided a long list of answers surrounded around how to address a bearing failure for a car, ignoring that bearings can be applied in many different applications and generality around it.

LLM responses tended to not be as strong at highly specific technical questions or acronyms. Both models had lower response accuracy for questions under the "Fundamental Definition" category as under the other two categories, as summarized in Figure 7. More specifically, AI1 scored 58% on fundamental definition, where as AI2 scored 37%, AI1 scored 69% on the SME Knowledge & Expertise area, AI2 scored 65% for the same category and finally, AI1 scored 72% on equipment knowledge and AI2 scored 61%.

An example prompt with responses are shown in Table 6. This example requires understanding of acronyms with domain knowledge specific to gas turbines. AI2 missed the context and hallucinated meanings for the acronyms which may sound viable to someone outside of the domain. However, in the context of Combustion Turbine, this answer is incorrect. A second example prompt with responses are shown

Table 4. Example of LLM responses for a PHM technology knowledge question. In this example, both answers are interesting. AI1 is theoretically correct, but in practice, it is posible (though not ideal) to detect wear through vibration monitoring.

Prompt	True or False: Vibration monitoring can detect uniform wear?
AI1	False. Vibration monitoring is not an effective method for detecting uniform impeller wear... Vibration monitoring is a useful tool for detecting changes in the vibration signature of rotating machinery. ... However, vibration monitoring is not well-suited to detecting uniform wear, including uniform impeller wear. Uniform wear typically results in a gradual change in the overall vibration signature of the equipment, which may be difficult to detect using vibration monitoring alone. Instead, other methods, such as regular visual inspections or wear measurements, may be more effective for detecting uniform wear in equipment components such as impellers.
AI2	True. Vibration monitoring can detect uniform impeller wear. This is because uniform impeller wear will cause the impeller to become unbalanced, which will result in increased vibration levels. The increased vibration levels can be detected by vibration monitoring equipment.

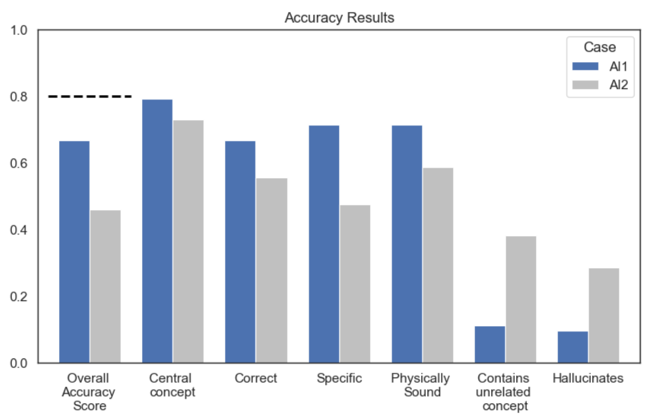

Figure 4. PHM Knowledge Exam: Comparison of the accuracy scores for "Overall Correctness" across the different question categories for both LLLMs. The dashed lines across the overall accuracy score denote 80% which is considered a passing score for an M&D analyst position.

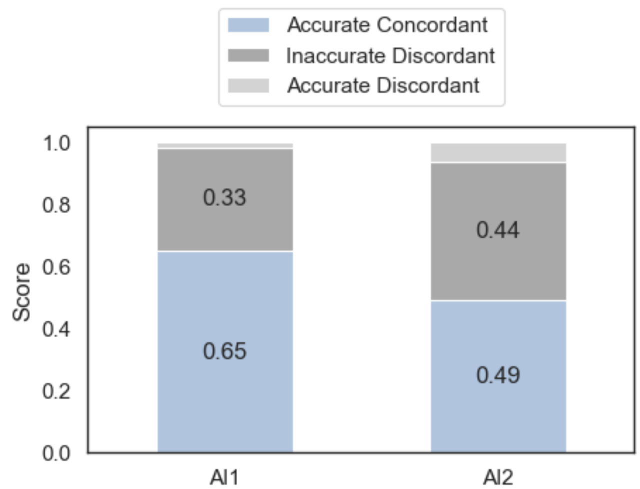

Figure 5. PHM Knowledge Exam: Correctness-concordance contingency table for PHM knowledge exam results. In general, if the AI was correct, then it was concordant and vice versa.

in Table 5, illustrating an instance where the central concept of the question was missed by AI2, but the always respond with some response, whether or not the response addresses the central concept or is correct.

Another example is in response to a question about being able to predict AGB failure (AGB stands for Accessory Gearbox in Gas Turbines). AI2 missed the context of the question and provided an answer that talked about star's luminosity, effect temperatures, rotation rate and so one. This highlights the AI's difficulty with specific acronyms, resulting in a misunderstanding of the central concept and generating incorrect or irrelevant responses.

6. RESULTS 3: PHM TROUBLESHOOTING GUIDELINES

Samples of AI1's responses for one troubleshooting task prompted in three different ways are summarized in Table 7. AI1 was prompted with the same question three times, each question was promoted slightly differently with minor grammatical changes however the central concept of the question remained the same: step change on compressor bearing on an Alstom GT11 gas turbine. The three prompts differed through asking (1) What can be the cause, (2) How to troubleshoot and (3) How to mitigate and address.

Each prompt had different responses by AI1 and the full responses are made available in the supplemental material. In all three cases, AI1 grasped the central concept of the query and was concordant. However, all three responses were extremely generic. From an accuracy perspective, these recommendations would be deemed partially correct as they do not contain the level of required specificity.

Overwhelmingly, all AI responses for PHM recommendations do not meet the "AI provides instructions a person can follow" criteria due to the length of the responses. The nature of output being too lengthy would make it unattractive for an analyst in an M&D center to consume. It may help a person early in the career to get the understanding of the concept, help diagnose and troubleshoot in a remote manner but definitely not ready to be consumed at a plant level to write out the investigative or repair work orders.

Table 5. Example of LLM responses from the PHM knowledge examination.

Prompt	What are the components of bleed system for a Combustion Turbine?
AI1	The bleed system in a combustion turbine is responsible for extracting and controlling the flow of air or gases from various stages of the turbine for specific purposes. The specific components of a bleed system can vary depending on the turbine design and application. However, here are some common components found in a typical bleed system: Bleed Valves: These are the primary components of the bleed system and are used to control the flow of air or gases. Bleed valves are typically located at strategic points within the turbine and are designed to open or close as needed to regulate the amount of extracted air or gas. Bleed Lines: These are the pipes or conduits that carry the extracted air or gases from the turbine stages to their intended destinations. Bleed lines are designed to handle the high-temperature and high-pressure conditions of the turbine environment. Intercoolers: In some combustion turbines, intercoolers are used in the bleed system to reduce the temperature of the extracted air or gases. Intercoolers are heat exchangers that remove excess heat, improving the efficiency and performance of downstream processes or systems that use the extracted air or gases. Air/Fuel Control System: The bleed system may be integrated with the turbine's air/fuel control system. This system monitors and regulates the flow of air and fuel to maintain optimal combustion conditions. The bleed system can be used to adjust the air-to-fuel ratio by extracting excess air from the combustion process, ensuring efficient and stable operation. Bleed Air Utilization: The extracted air or gases from the bleed system can be utilized for various purposes, depending on the turbine's configuration. Common uses include providing air for cooling turbine components, supplying air for auxiliary systems, or being utilized in other industrial processes. It's important to note that the specific configuration and components of a bleed system can vary significantly depending on the turbine model, manufacturer, and application. Detailed information about the bleed system can typically be found in the turbine's technical documentation and specifications provided by the manufacturer.
AI2	In printing, bleed is the area of a design that extends beyond the trim size of the final piece. Bleed is required to ensure that the design is not cut off when the paper is trimmed. For example, if you are printing a business card with a red background, you will need to extend the red background by 1/8 inch on all sides of the business card. This will ensure that the red background is not cut off when the business card is trimmed. Bleed is also required for designs that have intricate details that could be lost when the paper is trimmed. For example, if you are printing a brochure with a photo of a flower, you will need to extend the photo by 1/8 inch on all sides of the brochure. This will ensure that the photo is not cut off when the brochure is trimmed. Bleed is an important part of the printing process and should always be considered when designing a piece. By adding bleed to your design, you can ensure that your final piece will look exactly the way you intended. Here are some of the benefits of using bleed in printing: Prevents white space from showing around the edges of your design. Creates a seamless look and feel to your design. Allows for more creative freedom when designing your piece. Ensures that your design is printed accurately. If you are planning to print a piece, be sure to add bleed to your design. It is a small detail that can make a big difference in the final product.

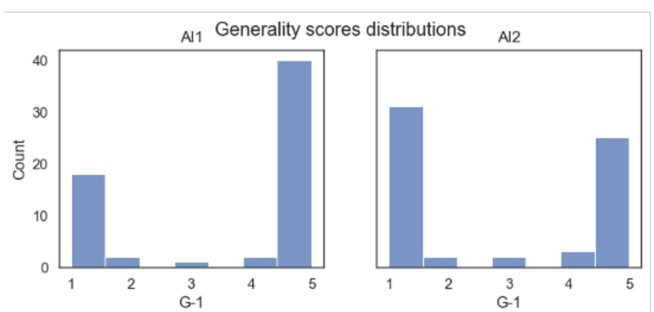

Figure 6. PHM Knowledge Exam: Distributions of generality scores across responses to the PHM knowledge exam questions. The bi-modal distributions for both AI1 and AI2 indicate the responses were either too general or very specific without many responses in the middle.

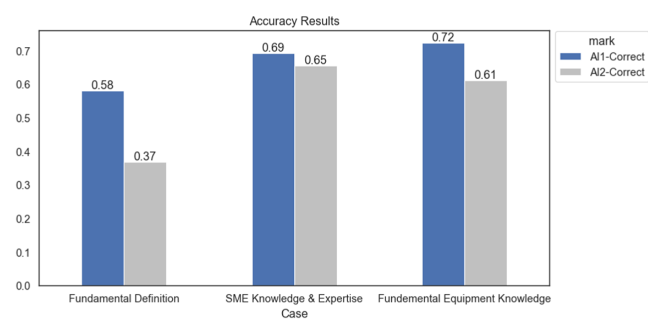

Figure 7. PHM Knowledge Exam: Correctness scores by different test question categories.

Further, while it was generally observed that the AI responses were overly conservative in their recommendations (for example, recommending to first turn off critical systems due to a sensed increase bearing temperature), for the specific response shown, only one of the three prompts (How to mitigate and address) recommended addressing safety protocols as an immediate response. For a highly critical asset (such as a gas turbine), safety protocols are important. In other words, while the AI was generally too conservative it was not always conservative and it is possible for safety risks to be present in the response.

Table 6. Example of LLM responses from the PHM knowledge examination.

Prompt	What is SAC and DLE in a Combustion Turbine?
AI1	In the context of a combustion turbine, SAC and DLE refer to different types of combustion systems: SAC: SAC stands for Single Annular Combustor. It is a type of combustion system used in gas turbines. In a SAC, the combustion process takes place in a single annular chamber where the fuel and air mixture is burned. The combustion chamber typically consists of a series of fuel nozzles arranged in a ring around the turbine's central axis. This design allows for efficient mixing of fuel and air and promotes stable combustion. DLE: DLE stands for Dry Low Emissions. It is another type of combustion system used in gas turbines. DLE technology is employed to reduce the emission of nitrogen oxides (NOx) from the exhaust gases of the turbine. In a DLE system, the combustion process is carefully controlled to minimize the formation of NOx. This is achieved by using advanced burner designs, precise fuel and air mixing, and optimized operating parameters. DLE systems often incorporate advanced sensors and controls to monitor and adjust the combustion process in real-time for maximum efficiency and low emissions. Both SAC and DLE combustion systems are used in modern gas turbines, but their specific design and operation can vary between different turbine manufacturers and models. The choice of system depends on factors such as emissions regulations, efficiency requirements, and the turbine's intended application.
AI2	SAC stands for Shaft Alignment Check and DLE stands for Drive End Level. These are two important parameters that are used to ensure that the gas turbine is properly aligned and that the drive end is level. Shaft Alignment Check: The shaft alignment check is a procedure that is used to ensure that the gas turbine shaft is properly aligned with the drive shaft. This is important to ensure that the gas turbine operates smoothly and efficiently. Drive End Level: The drive end level is a procedure that is used to ensure that the drive end of the gas turbine is level. This is important to ensure that the gas turbine is properly supported and that it does not vibrate excessively. Both SAC and DLE are important parameters that should be checked regularly to ensure that the gas turbine is operating properly. By checking these parameters, you can help to prevent costly repairs and downtime. Here are some of the benefits of checking SAC and DLE regularly: Prevents costly repairs: By checking SAC and DLE regularly, you can identify potential problems before they cause costly repairs. Prevents downtime: By checking SAC and DLE regularly, you can prevent downtime caused by unexpected repairs. Increases safety: By checking SAC and DLE regularly, you can increase the safety of the gas turbine operation. If you are the owner or operator of a gas turbine, it is important to check SAC and DLE regularly. By doing so, you can help to ensure that your gas turbine operates safely and efficiently.

7. CONCLUSIONS

A summary of our findings from the three exams is presented in Table 8. We observed that LLMs exhibit a surprising ability to comprehend central concepts and contain a vast amount of knowledge and information in the industrial domain. However, it is important to note that LLMs are not always accurate, and distinguishing between correct and hallucinated responses is challenging without prior knowledge of the answers. Another obstacle in their general utility is the verbosity of the generated responses. Frequently, LLMs provide extraneous information alongside the required details, resulting in lengthy and impractical recommendations. Providing concise information remains a significant challenge for LLMs. Their output often tends to be generic and high-level, occasionally irrelevant, which can overwhelm users seeking quick and precise answers, particularly in industrial use-cases where analysts at M&D centers may struggle to navigate through the excess information provided by these LLMs.

The LLMs exhibited varying performance across different areas despite their strong grasp of central concepts and industrial knowledge. In the maintenance and reliability knowledge exam, the LLMs demonstrated higher accuracy on questions that were less domain-specific and those related to manufacturing operations. This observation could potentially be attributed to the comparatively greater emphasis on operations rather than maintenance in the manufacturing world, which may have influenced the training data. However, further investigation is needed to confirm this speculation.

Both AI models used in the PHM knowledge experiments performed well in the Fundamental Equipment Knowledge category, achieving accuracy results exceeding 50%. This could be attributed to the inclusion of numerous manuals available on the internet in the training data, which allowed the generative AI to respond effectively to questions regarding equipment knowledge.

7.1. Limitations

Several limitations should be considered in this study. The knowledge exam experiments took place in March 2023, and the PHM troubleshooting tests were conducted in May 2023. Since the experiments, both LLMs utilized in this paper have undergone updates, resulting in observed improvements in certain responses. For instance, areas such as deducing maintenance metric definition relationships have shown improvement, warranting ongoing monitoring. Moreover, comparing results from May with those from March may introduce bias. Nevertheless, the authors believe that the major conclusions remain relevant.

Another consideration pertains to the models' learning process. Although our tests may or may not have included unseen questions initially, conducting the tests again could introduce bias since the LLMs would have already encoun-

Table 7. Three different prompts for a PHM troubleshooting task. The LLM responded differently depending on how the query was worded.

Prompt	No. steps in AI response	Response sample
1. **What can be the cause for** step change on vibrations on compressor bearing for an Alstom GT11 combustion turbine?	8 steps	A step change in vibrations on a compressor bearing for an Alstom GT11 combustion turbine can have various causes. Here are some potential reasons to consider: 1. Mechanical Damage: The compressor bearing may have experienced mechanical damage due to factors such as excessive loads, misalignment, or foreign object damage. This can lead to sudden changes in vibration levels. 2. Bearing Wear: Over time, the compressor bearing can wear out, leading to increased vibrations. If the wear suddenly accelerates or if the bearing experiences a failure, it can result in a step change in vibrations. 3. ...
2. **How to troubleshoot** a step change in vibrations on a compressor bearing for an Alstom GT11 combustion turbine	10 steps	To troubleshoot a step change in vibrations on a compressor bearing for an Alstom GT11 combustion turbine, you can follow these troubleshooting steps: 1. Review Historical Data: Analyze historical vibration data to determine when the step change occurred. Look for any patterns or events that coincide with the change in vibrations. This can help identify potential triggers or causes. 2. Conduct Visual Inspection: Perform a thorough visual inspection of the compressor bearing and surrounding components. Look for signs of damage, wear, misalignment, or rubbing. Check for loose fasteners, oil leaks, or any other abnormalities that could contribute to the step change in vibrations. 3. ...
3. **How to mitigate and address** a step change in vibrations on a compressor bearing for an Alstom GT11 combustion turbine	7 steps	Mitigating and addressing a step change in vibrations on a compressor bearing for an Alstom GT11 combustion turbine involves taking appropriate corrective actions. Here are some steps to mitigate and address the issue: 1. Safety: Ensure that all safety procedures are followed when working on the turbine. Adhere to lockout/tagout procedures and obtain necessary permits to work on the equipment. 2. Immediate Response: If the step change in vibrations is significant and poses an immediate risk to the turbine or personnel, **consider shutting down the turbine until the issue can be addressed.** 3. ...

tered the questions. Furthermore, there were limitations in our study design. Firstly, the experiment did not involve asking the same question multiple times. This decision was influenced by the laborious nature of grading the exams. Additionally, during the experiment, the authors began to discover the models' ability to perform better when posed with follow-up or clarifying questions. While the test was not designed to guide the models deductively through questioning, the authors informally re-asked questions during the grading process, which could be an interesting aspect to account for in future studies.

8. DISCUSSION

In this paper, we developed and implemented a methodology to evaluate the performance of LLMs as a tool for industrial use cases, specifically in the context of PHM. Through this approach, we identified both strengths and weaknesses of the models, both in terms of their ability to generate PHM related recommendations and their general knowledge in the field. While the paper highlighted instances where the models fell short, it is important to remember the *exsupero ursus* fallacy, as these examples do not define the overall success or failure of the AI. In fact, our results demonstrate that the models possess some knowledge of many highly technical concepts. However, it is crucial for readers to discern the appropriate areas of application, exercise caution, and recognize the need for further development before using these models for PHM applications. We provide specific recommendations for the consideration of the PHM community in addressing the identified areas below.

Need to integrate technical information with LLM capabilities. The use of LLMs shows promise for troubleshooting recommendations based on PHM model alerts. However, there is still work to be done to ensure proper guidelines are in place. When it comes to industrial assets in sectors such as power generation, oil & gas, mining, and metals, the domain-specific expertise of out-of-the-box LLMs may pose challenges. These sectors involve complex operations, equipment, and processes that require specialized knowledge. LLMs, despite their impressive language capabilities, often lack exposure to the intricacies of these industries during their training. Their training data usually consists of publicly available information, which may not provide sufficient depth. As a result, important details, unique challenges, safety considerations, and regulatory frameworks may be missing from LLM knowledge. Therefore, it is crucial to augment LLMs with domain-specific knowledge, curated data sets, and expert input to enhance their subject matter expertise in the context of industrial asset management.

For future research development, LLMs could benefit from techniques that emphasize brevity and refinement in their generated responses. Approaches like summarization, abstraction, and information extraction could help distill the relevant information from extensive texts, enabling LLMs to provide more concise and easily digestible responses tailored to the specific information needs of users in industrial domains.

Need for integration of PHM specific safeguards. A specific extension to integration of technical information is towards addressing safety risks. There is risk that the model could provide incorrect or outdated guidance regarding the operation, maintenance, or troubleshooting of industrial assets, which could lead to unsafe practices or equipment malfunctions. Additionally, LLMs may lack the contextual understanding of specific industrial processes, safety protocols, or regulatory requirements, making it prone to offering suggestions that may violate safety guidelines. Furthermore, the model's inability to verify the accuracy or completeness of user inputs increases the risk of misinterpretation, leading to potentially hazardous actions. To mitigate these safety risks, it is crucial to implement robust validation mechanisms, domain-specific training, and human supervision to ensure the reliability and safety of LLM's responses. Additionally, employing stringent quality control measures, regular model updates, and user feedback loops can enhance the accuracy and safety of language model outputs, thus reducing the likelihood of safety incidents or accidents in industrial settings. To mitigate these safety risks, it is crucial to exercise caution when relying on LLM responses for industrial assets. Employing human oversight and verification processes can help ensure that the generated information is accurate, up to date, and aligned with industry-specific safety standards. Domain experts should be involved in the development and validation of LLM models for industrial applications to ensure that safety considerations are properly addressed. Additionally, integrating real-time monitoring, feedback loops, and quality control mechanisms can help detect and rectify any potential safety issues in LLM responses, enhancing the overall safety of industrial operations.

Furthermore, LLMs lack contextual understanding specific to industrial processes, safety protocols, and industry-specific regulations. They may not be aware of the unique risks and safety considerations associated with industrial assets. Consequently, the model's responses may not align with established safety guidelines or regulatory requirements, potentially leading to unsafe practices or equipment failures.

Need for requirements around AI responses. The verbosity of LLM models poses a practical challenge for their application in the industrial community. During the knowledge exam, LLMs often provided extraneous information and generated recommendations that were excessively wordy, making it impractical for humans to read and act upon them in real-world scenarios. One possible reason for this verbosity could be associated with pricing models that may reward longer responses based on the number of words. Additionally, the inclination towards providing extensive information might stem from a safeguarding response, where the inclusion of more details increases the likelihood of containing the correct information. Nevertheless, there is a pressing need to explicitly define requirements for AI responses and fine-tune language models specifically for targeted technical tasks. By establishing precise guidelines, the issue of verbosity can be effectively addressed, enabling LLMs to deliver concise and actionable responses in the industrial context.

Need for prompt engineering specific to PHM tasks. In this study, the authors observed that LLM responses are highly influenced by the wording of the prompt, highlighting the significance of prompt engineering. Prompt engineering, a rapidly growing discipline, involves applying engineering principles to fine-tune language models like ChatGPT, enhancing their ability to generate high-quality responses. By crafting well-designed prompts that include clear instructions and domain context (e.g., Industrial Knowledge, SME knowledge, equipment knowledge), prompt engineering can increase the model's competence in generating accurate answers while avoiding pitfalls such as verbosity, ambiguity, or irrelevant information/hallucination.

Through prompt engineering, the model gains a better understanding of user inquiries, leading to more accurate and coherent responses. Carefully constructed prompts with strong context and domain knowledge encourage critical thinking and help the model extract relevant information from its knowledge base, resulting in improved factual accuracy and logical consistency. Prompt engineering also allows for the incorporation of specific guidelines, biases, or preferences, enabling customization of ChatGPT's behavior to meet specific objectives.

One related concept is chain-of-thought reasoning (CoT)

(Wei et al., 2022), which involves utilizing a series of intermediate reasoning steps to enhance the ability of large language models to perform complex reasoning tasks. In this study, the authors found that re-asking a question often led to the model providing correct answers, demonstrating the potential of prompt refinement. As a recommendation, templatizing questions for the industrial domain can be explored as a means to streamline and optimize the prompt engineering process, further enhancing the model's performance.

General recommendations. The above recommendations are all aimed towards incorporation of this technology in the development of tools for PHM. However, there are larger outstanding considerations that the PHM community also needs to consider. Such considerations span social and economic considerations, such as around the processes and personas in development teams, business case considerations, security considerations and many more. There is a lot of ongoing and future work in the PHM space in this area.

When interpreting the results of this study, it is crucial to acknowledge that the evaluation focused on a specific chat task, particularly related to making recommendations. It is important to recognize that within the realm of PHM, there exist numerous other NLP tasks, such as work order classification, where LLMs may have distinct roles. In these tasks, concepts like word representations become more relevant, while factors like hallucinations and prompt engineering may have a relatively lesser impact. It is worth considering the broader landscape of AI tasks and the specific requirements and nuances of each task when drawing conclusions from this study.

ACKNOWLEDGMENT

The authors wish to acknowledge and thank GE Vernova's Industrial Managed Services SME team (Theresa Cardia, David Saad, David Roe, Venkat Bala) and Christopher Schreiner, Sr Manager from Analytics & Data Science from GE Vernova The authors also wish to thank ChatGPT for its assistance in writing this article.

NOMENCLATURE

PHM	Prognostics and Health Management
TLP	Technical Language Processing
AI	Artificial Intelligence
LLM	Large Language Model
RLHF	Reinforcement Learning from Human Feedback
RLHP	Reinforcement Learning from Human Preference
M&D	Monitoring and Diagnostics
SME	Subject Matter Expert
SMRP	Society of Maintenance and Reliability Professionals
CMRP	Certified Maintenance and Reliability Professional
PM	Preventative Maintenance
RCM	Reliability-Centered Maintenance
PPM	Parts Per Million
CoT	Chain-of-thought reasoning
O&G	Oil and Gas

REFERENCES

Addepalli, S., Weyde, T., Namoano, B., Oyedeji, O. A., Wang, T., Erkoyuncu, J. A., & Roy, R. (2023). Automation of knowledge extraction for degradation analysis. *CIRP Annals*, *72*(1), 33–36. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0007850623000070` doi: https://doi.org/10.1016/j.cirp.2023.03.013

Alfeo, A. L., Cimino, M. G., & Vaglini, G. (2021). Technological troubleshooting based on sentence embedding with deep transformers. *Journal of Intelligent Manufacturing*, *32*(6), 1699–1710.

Amazon. (2023). *Amazon titan.* `https://aws.amazon.com/bedrock/titan/`. (Accessed: June 16, 2023)

Ansari, F. (2020). Cost-based text understanding to improve maintenance knowledge intelligence in manufacturing enterprises. *Computers & Industrial Engineering*, *141*, 106319. Retrieved from `https://www.sciencedirect.com/science/article/pii/S036083522030053X` doi: https://doi.org/10.1016/j.cie.2020.106319

Ansari, F., Glawar, R., & Nemeth, T. (2019). PriMa: a prescriptive maintenance model for cyber-physical production systems. *International Journal of Computer Integrated Manufacturing*, *32*(4-5), 482–503.

Bahrini, A., Khamoshifar, M., Abbasimehr, H., Riggs, R. J., Esmaeili, M., Majdabadkohne, R. M., & Pasehvar, M. (2023). Chatgpt: Applications, opportunities, and threats. In *2023 Systems and Information Engineering Design Symposium (SIEDS)* (pp. 274–279).

Bastos, Pedro and Lopes, Isabel and Pires, LCM. (2012). A maintenance prediction system using data mining techniques. In *World congress on engineering 2012* (Vol. 3, pp. 1448–1453).

Bian, N., Han, X., Sun, L., Lin, H., Lu, Y., & He, B. (2023). ChatGPT is a Knowledgeable but Inexperienced Solver: An Investigation of Commonsense Problem in Large Language Models. *arXiv preprint arXiv:2303.16421*.

Bokinsky, H., McKenzie, A., Bayoumi, A., McCaslin, R., Patterson, A., Matthews, M., ... Eisner, L. (2013). Application of natural language processing techniques to marine V-22 maintenance data for populating a CBM-Oriented database. pdfs.semanticscholar.org.

Bouabdallaoui, Y., Lafhaj, Z., Yim, P., Ducoulombier, L., & Bennadji, B. (2020). Natural language processing model for managing maintenance requests in buildings. *Buildings*, *10*(9), 160.

Brundage, M. P., Kulvatunyou, B., Ademujimi, T., & Rakshith, B. (2017). Smart manufacturing through a framework for a knowledge-based diagnosis system. In *International Manufacturing Science and Engineering Conference* (Vol. 50749, p. V003T04A012).

Brundage, M. P., Sexton, T., Hodkiewicz, M., Dima, A., & Lukens, S. (2021). Technical language processing: Unlocking maintenance knowledge. *Manufacturing Letters*, *27*, 42–46.

Chowdhury, M., Rifat, N., Latif, S., Ahsan, M., Rahman, M. S., & Gomes, R. (2023). ChatGPT: The Curious Case of Attack Vectors' Supply Chain Management Improvement. In *2023 IEEE International Conference on Electro Information Technology (eIT)* (pp. 499–504).

Dima, A., Lukens, S., Hodkiewicz, M., Sexton, T., & Brundage, M. P. (2021). Adapting natural language processing for technical text. *Applied AI Letters*, *2*(3), e33.

GE. (2023). *Remote monitoring powered by digital twins.* `https://www.ge.com/digital/industrial-managed-services-remote-monitoring-for-iiot/`. (Accessed: June 8, 2023)

Google. (2023). *Bard.* `https://bard.google.com/`. (Accessed: June 16, 2023)

Hodkiewicz, M., Klüwer, J. W., Woods, C., Smoker, T., & Low, E. (2021). An ontology for reasoning over engineering textual data stored in fmea spreadsheet tables. *Computers in Industry*, *131*, 103496. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0166361521001032` doi: https://doi.org/10.1016/j.compind.2021.103496

Hubbard, D. W. (2020). *The failure of risk management: Why it's broken and how to fix it.* John Wiley & Sons.

JA1011, SAE. (2009). A guide to the reliability-centered maintenance (rcm) standard (2009). *Society of Automotive Engineers, 2da Rev.*

Jalil, S., Rafi, S., LaToza, T. D., Moran, K., & Lam, W. (2023). ChatGPT and software testing education: Promises & Perils. *arXiv preprint arXiv:2302.03287*.

Karray, M. H., Ameri, F., Hodkiewicz, M., & Louge, T. (2019). ROMAIN: Towards a BFO compliant reference ontology for industrial maintenance. *Applied Ontology*, *14*(2), 155–177.

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., ... others (2023). Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models. *PLoS digital health*, *2*(2), e0000198.

Lepenioti, K., Bousdekis, A., Apostolou, D., & Mentzas, G. (2020). Prescriptive analytics: Literature review and research challenges. *International Journal of Information Management*, *50*, 57–70.

Lepenioti, K., Pertselakis, M., Bousdekis, A., Louca, A., Lampathaki, F., Apostolou, D., ... Anastasiou, S. (2020). Machine learning for predictive and prescriptive analytics of operational data in smart manufacturing. In *Advanced Information Systems Engineering Workshops: CAiSE 2020 International Workshops, Grenoble, France, June 8–12, 2020, Proceedings 32* (pp. 5–16).

Nandyala, A. V., Lukens, S., Rathod, S., & Agarwal, P. (2021). Evaluating word representations in a technical language processing pipeline. In *PHM sSciety European Conference* (Vol. 6, pp. 17–17).

OpenAI. (2022). *Chatgpt.* `https://chat.openai.com/`. (Accessed: June 16, 2023)

OpenAI. (2023). *Gpt4.* `https://openai.com/gpt-4`. (Accessed: June 16, 2023)

Ortega, P. A., Kunesch, M., Delétang, G., Genewein, T., Grau-Moya, J., Veness, J., ... others (2021). Shaking the foundations: delusions in sequence models for interaction and control. *arXiv preprint arXiv:2110.10819*.

Pau, D., Tarquini, I., Iannitelli, M., & Allegorico, C. (2021). Algorithmically exploiting the knowledge accumulated in textual domains for technical support. In *PHM Society European Conference* (Vol. 6, pp. 12–12).

Peng, B., Galley, M., He, P., Cheng, H., Xie, Y., Hu, Y., ... Gao, J. (2023). Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*.

Peshave, A., Aggour, K., Ali, A., Mulwad, V., Dixit, S., & Saxena, A. (2022). Evaluating vector representations of short text data for automating recommendations of maintenance cases. In *Annual Conference of the PHM Society* (Vol. 14).

Pires, F., Leitão, P., Moreira, A. P., & Ahmad, B. (2023). Reinforcement learning based trustworthy recommendation model for digital twin-driven decision-support in manufacturing systems. *Computers in Industry*, *148*, 103884.

Qadir, J. (2023). Engineering education in the era of chatgpt: Promise and pitfalls of generative ai for education. In *2023 IEEE Global Engineering Education Conference (EDUCON)* (pp. 1–9).

R. Gulati and R. Smith. (2021). *Maintenance and Reliability Best Practices* (3rd ed.). Industrial Press Inc.

Rajpathak, D. G. (2013). An ontology based text mining system for knowledge discovery from the diagnosis data in the automotive domain. *Computers in Industry*, *64*(5), 565-580. Retrieved from https://www.sciencedirect.com/science/article/pii/S0166361513000456 doi: https://doi.org/10.1016/j.compind.2013.03.001

Rao, A., Kim, J., Kamineni, M., Pang, M., Lie, W., & Succi, M. D. (2023). Evaluating ChatGPT as an adjunct for radiologic decision-making. *medRxiv*, 2023–02.

Rathore, B. (2023). Future of textile: Sustainable manufacturing & prediction via chatGPT. *Eduzone: International Peer Reviewed/Refereed Multidisciplinary Journal*, *12*(1), 52–62.

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, *3*, 121-154. Retrieved from https://www.sciencedirect.com/science/article/pii/S266734522300024X doi: https://doi.org/10.1016/j.iotcps.2023.04.003

Society of Maintenance & Reliability Professionals. (2017). SMRP Best Practices 6th Edition.

Trilla, A., Mijatovic, N., & Vilasis-Cardona, X. (2022). Towards learning causal representations of technical word embeddings for smart troubleshooting. *International Journal of Prognostics and Health Management*, *13*(2).

Usuga Cadavid, J. P., Grabot, B., Lamouri, S., Pellerin, R., & Fortin, A. (2020). Valuing free-form text data from maintenance logs through transfer learning with camembert. *Enterprise Information Systems*, 1–29.

Wang, X., Anwer, N., Dai, Y., & Liu, A. (2023). ChatGPT for design, manufacturing, and education. *Procedia CIRP*, *119*, 7–14.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., & Zhou, D. (2022). Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Woods, C., French, T., Hodkiewicz, M., & Bikaun, T. (2023). An ontology for maintenance procedure documentation. *Applied Ontology*, 1–38.

Woods, C., Selway, M., Bikaun, T., Stumptner, M., & Hodkiewicz, M. (2023). An ontology for maintenance activities and its application to data quality. *Applied Ontology*, 1–34. doi: 10.3233/SW-233299

Yeo, Y. H., Samaan, J. S., Ng, W. H., Ting, P.-S., Trivedi, H., Vipani, A., … others (2023). Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. *medRxiv*, 2023–02.

BIOGRAPHIES

Sarah Lukens is a Data Scientist at LMI. Her interests are focused on data-driven modeling for reliability applications by combining modern data science techniques with current industry performance data. This work involves analyzing asset maintenance data and creating statistical models that support asset performance management (APM) work processes using components from natural language processing, machine learning, and reliability engineering. Sarah completed her Ph.D. in mathematics in 2010 from Tulane University with focus on scientific computing and numerical analysis. Sarah is a Certified Maintenance and Reliability Professional (CMRP).

Asma Ali a Senior Staff Analytics Engineer with GE Vernova where she is a technical team lead for the Analytics & Data Science team. Asma earned a Bachelors Degree in Biomedical Engineering from University of Connecticut and a Masters Degree in Mechanical Engineering from University of IL while working full-time. Asma has 15+ years of industry experience focusing on software solution design, new product development and deploying process improvements, etc. Over the years, she worked closely with industrial domains like Power Generation, Oil & Gas, Mining & Metals, Aviation, Chemicals, Automotive, and more. Asma developed several dozen Digital Twins for Industrial Assets and Analytics for the APM product line. Currently, Asma is leading the efforts on Work Order Automation utilizing technical language processing for GE Vernova and providing GE Research with Power domain-related industrial expertise.

Table 8. Executive summary of observations against the key research questions across the three assessments.

Guiding Research Question	Maintenance & Reliability Knowledge Task	PHM Knowledge Task	PHM Troubleshooting Task	Summary/ Conclusion
How well can the LLM grasp central concepts within the industrial domain?	Pretty good, but better in some areas and worse in others	Pretty good. Lower rate of grasping central concept associated with wrong answer than any question category.	Very good as grasping central concepts.	Surprisingly good as grasping central concepts in general.
Domain adaptation: How well does the LLM understand the industrial domain?	Better in the areas of workforce management, leadership and operations. Poorer in areas of PHM technologies, Work Management and RCM/FMEA.	Not as good on highly specific technical questions and acronyms. Would hallucinate if did not have answer.	Very good at understanding industrial domain	Surprisingly good, but better in some areas than others
How often will the LLM return inconsistent responses and how will these affect the scores?	Surprisingly consistent responses	Consistent when correct, but when incorrect, the response can contain anything.	Consistent when correct	Observed that consistency in response can be dependent on prompt intelligence
Is the LLM appropriately specific or too general in its responses?	Inconclusive for this test.	Both: either extremely specific or extremely general.	Not specific enough and too verbose at the same time. Key remedy actions are lost in the verbosity of the responses.	Specificity is a requirement for PHM recommendations. LLM's need development in this area to contribute value in an M&D center.
How does the hallucinogenic effect manifest and to what extent?	It is present and should be a caution.	Will hallucinate incorrect response over saying it does not know.	Prompt dependent.	Is present and should be a caution.
Are the LLM responses physically sound? Are there safety risks in the LLM responses? NA	Should be a caution NA	Should be a caution General tendency to be overly conservative, except not always.	Prompt dependent; Should be a caution. Tends to be overly conservative, but safety risks can still get through guardrails.	Should be a caution.
Can the LLM make deductions to return a response?	Ability to deduct is a clear weakness	Ability to deduct is a clear weakness, and will make stuff up before it reasons.	Can deduce if you help it along through a chain of prompts	Prompt engineering and chain of thought are needed to assist LLM.
As a "ChatBot", can the LLM make practical recommendations to a person?	NA	NA	Too wordy. Key remedy actions lost in the verbosity of the responses.	Too wordy to be useful as a recommendation engine in an M&D Center.