IEEE Open Journal of the
Industrial Electronics Society

Received 30 March 2024; revised 21 May 2024; accepted 6 June 2024. Date of publication 11 June 2024;
date of current version 2 August 2024. The review of this article was arranged by Associate Editor D. de Silva.

Digital Object Identifier 10.1109/OJIES.2024.3412222

The Relevance of Large Language Models for Project Management

STAMATIS KARNOUSKOS (Fellow, IEEE)

SAP, 69190 Walldorf, Germany

(e-mail: karnouskos@ieee.org).

ABSTRACT The rise of artificial intelligence, particularly the emergence of large language models (LLMs) like ChatGPT, continuously reveals numerous advantages across various domains. However, the area of project management has not yet been sufficiently explored. This study fills the research gap by conducting an empirical evaluation of three well-known LLMs: OpenAI's ChatGPT-3.5 and ChatGPT-4, as well as Google's Bard. The evaluation involves subjecting these LLMs to tests designed to prepare professionals for project management certification by the Project Management Institute. The findings cast a positive light on all three LLMs, with each model achieving scores exceeding 82%. Key insights acquired include: LLMs demonstrate the ability to effectively answer project management certification exam questions; LLMs and project managers should be viewed as a dynamic and complementary partnership; and project management certification should evolve to include an assessment of how project managers collaborate with LLMs to enhance project management.

INDEX TERMS Bard, ChatGPT, Generative artificial intelligence (AI), large language models (LLMs), project management.

I. INTRODUCTION

High-visibility engineering projects often fail due to a lack of proper project management and tools. The Berlin Brandenburg airport started construction in 2006, with a delivery day in 2011, but was only delivered with a delay of 9 years in 2020, an overrun of more than €6.2 bn compared to the original budget, due to problems such as lack of a proper information management system, poor construction planning, execution, management, and corruption [1], [2]. Even if information management systems are available, information technology (IT) projects may still fail due to poor project management practices. The Airbus A380 had production facilities (16 different sites with 41 000 employees) around the world with different teams using incompatible computer-aided design tools designing parts (electrical wiring harnesses) that did not fit together (discovered only during physical assembly), eventually delaying the project for 2 years and costing $6 bn to the company [3]. U.K.'s National Health Service (NHS) started one of the largest IT projects in order to introduce new technologies in the health sector, which, however, failed and was stopped after costing approx. £11.4 bn, due to (among other reasons) the absence of phased change management, lack of sufficient end-user engagement, and underestimation of the project's scale [4]. Countless other IT project failures exist due to management reasons, among other factors [5], [6], [7]. While such project failures may result in financial and reputation-loss risks, sometimes they also result in fatal accidents, as recently the two Boeing 737-MAX crashes were linked to the design problems in sensors, technical components, and practice failures [7].

The role of the project manager is critical to achieving the project's objectives [8], and certifications exist in order to provide some degree of assurance about sufficient knowledge of the area both in theory as well as best practices. Despite the existence of best practices as well as their evolution in all angles that comprise it over time, nowadays still a large number of projects are not well managed and fail [6]. Typical pitfalls include a lack of efficient leadership within the project, insufficient planning, poor communication, ineffective change management, team conflicts, unclear roles and accountability, risk mismanagement, etc. In addition, there is a shortage of project managers in several domains, e.g., for IT projects, and their existence does not guarantee the quality of project design and execution, as often there is a lack of

time, experience, or resources. The utilization of emerging artificial intelligence (AI) technologies in the scope of project management is of interest as they have demonstrated sophisticated competencies [9], [10], which could be relevant for project management.

Over the last years, increased penetration of digital tools and digitalization of processes had a significant impact on project management and its processes [11]. As technology advances, new technological competencies can be utilized in order to increase the efficiency, transparency, and effectiveness of efforts undertaken. One technology on the rise is AI, and more specifically, Generative AI, which has the capability to use deep learning to generate, e.g., text as a response to prompts from the user. At the end of 2022, the world was taken by storm by OpenAI's ChatGPT, a chatbot that was made public, and users could interact with it in natural language in a wide range of aspects. ChatGPT is capable of generating text as a reaction to prompts, which means essays, answering questions, writing poetry, etc. In less than 2 months after its release, it reached 100 million active monthly users, which made it the fastest growing user base in history, while other viral technologies such as TikTok took 9 months and Instagram took 2.5 years [12].

The success of ChatGPT kick-started business wars on the usage of AI chatbots, especially in the field of Internet search, where technology giants such as Microsoft, Google, and others compete. ChatGPT is an example of a generative pretrained transformer (GPT), which is a type of large language model (LLM) and has been fine tuned for conversational applications. LLMs exhibit a wide range of competencies [9], [10], [13] that need to be better understood and utilized productively as part of systems, applications, and services. There is currently an ongoing discussion about what the LLMs can do, how well they can address challenges in different domains, what their limitations might be, how they can be made affordable, and even if they are subject to adversarial attacks and what implications their usage in real-world systems, applications, and services might imply for the society [9], [14], [15]. Project management could be one such domain where LLMs could be utilized, and if successful, this could have a significant impact on the way projects are managed in the future, as well as their potential success, risk mitigation, and qualitative outcomes.

The research question (RQ) investigated is: What potential do LLMs have in relation to project management practices? To grasp the potential of LLM, consider a project where the project manager, the project team, and the project stakeholders could take advice on project management issues 24/7 and drill down to their queries to understand the motivation of decisions and potential actions, something that would be practically impossible at large scale within a project today. This kind of democratization of project management competencies, availability to the stakeholders, coupled with sophisticated decision and argumentation, as well as addressing hypothetical challenges, could have a significant impact on the domain and reshape the world of project management.

Hence, the question arises if the LLMs can learn the necessary qualifications of good project management and can utilize theories and best practices toward addressing project management challenges [16]. In industry, the certifications provided by the not-for-profit professional organization Project Management Institute (PMI) are the standard for certification. Therefore, it would be of interest to assess the LLM competencies under the scope of the exams that project managers undertake in order to be certified. It is hypothesized that LLMs are suitable in a specific context, i.e., for addressing project management challenges. Therefore, it is investigated if the LLMs can possess the competencies of a specific group, i.e., certified project managers, and to evidence this hypothesis, some selected LLMs are experimented with, by exposing them to the preparation exam questions of the PMI that enables project managers to be certified.

II. COMPETENCIES AND CERTIFICATION

Business and industry have utilized competency models in order to select employees [17]. The focus is put on specific characteristics and skills that can provide some confidence with respect to the knowledge, ability, skill, experience, etc., that someone needs to possess in order to effectively perform the task(s) associated with a specific role [18], [19]. The matching of organizational needs with specific competencies in employees leads to better products and services [20]. The confidence in the competencies of an individual can be acquired via various means, e.g., university degrees and professional certifications [21], which act as a credible source that their holder possesses these skills, knowledge, experience, etc.

Certifications can be versatile, but overall, they also involve tests that need to be passed, and often, the grading signifies how well the individual has understood the material or is able to utilize it in the context of the test questions. While it would be a logical fallacy to equate the capability of passing such a theoretical test with the capability and competence of carrying out complex actions, in general, e.g., solving situated problems and challenges, such certifications and tests are the current way several competencies can be assessed, and they do offer a degree of confidence and can serve as an indicator of the expected future performance of the individual that holds such a certification [22]; hence, it should be considered as an indicator and not as a guarantee of that future performance.

Project management is a complex area that comprises of management practices, principles, processes, tools, and techniques that enable leaders and managers to design and execute projects successfully [8]. To capture the qualifications needed for good project managers, the industry standard certification offered is via the PMI, as they have a comprehensive training and certification program in place that utilizes best practices, something that is now part of the international standard ISO 21500. This certification can be awarded via an exam, and preparatory courses exist for it.

Certification of project managers is seen as beneficial [23] as it gives a certain assurance on the knowledge of best practices that can be applied in the context of planning and

execution of a project. In addition, via certification, there are additional beneficial influences on education, quality, and professionalization with respect to project management [24]. The certification assumes a mastering of the body of knowledge in the domain of project management, and passing that certification exam indicates that the holder has the necessary knowledge and knows how to apply it professionally [25]. Familiarity with the industry and technical expertise are required by project managers in order to be successful [26], and this is challenging. While best practices are outlined [8], the certified project managers should not be considered as people who follow only a specific procedure like trained technicians, but as reflective practitioners who can adapt to the needs of the situation based on their project management knowledge and best practices [27]. While best practices are outlined, e.g., in [8], the certified project managers should not be considered as people who follow mechanically a specific procedure like trained technicians, but as experienced practitioners who can adjust to the needs of the situation and utilize effectively the best approaches. Therefore, certification has grown rapidly in the last years, and certified individuals are considered in the industry as a signal of future performance [22], even if this is not guaranteed.

III. LLMS AND EXAMS

Generative AI can learn patterns and structures from training data and generate similar multimodal outputs such as text, images, sound, etc. AI overall has been suggested as a potential helping hand that could augment or automate human intelligence in complex and unstructured tasks such as those relevant to open innovation [28]. LLMs are examples of Generative AI that exhibit skills for a wide range of tasks (as opposed to task-specific neural networks). LLMs have made significant advances [13], and some of the latest ones, such as GPT-4, are considered to exhibit good inductive reasoning [29], [30] and general intelligence and competencies such as image generation, coding, mathematical abilities, interaction with the world, interactions with humans, and discriminative competencies, while their performance is in many cases very close to human level performance [9]. Today, several LLMs exist, both proprietary and open-source ones [31].

Several tests exist to evidence the performance of the LLMs on a variety of data and situations, and the results are promising in several domains [9], [10], [13], [32], [33]. For instance, ChatGPT was found to exhibit skills that allow it to perform at par with humans in algorithmic tasks, attend multiple questions in a query and answer them, create good summaries of text, code very well, is more ethical and truthful than previous models but also is worse than potential single-task fine-tuned models, might provide different results for the same prompt depending on its version, may underperform in underrepresented languages and sometimes consider only utilitarian morality to ethical dilemmas [32]. The ongoing research in LLMs will shed some more light on their competencies [33], [34].

Among all tests, however, what has mostly captured the public eye and surprised people is the capability of the LLMs to perform well on exams that were directed toward humans [34], and up to now, were considered too complex to be addressed from machines, e.g., due to their nature of open-ended questions, or the needed understanding of context. In the law domain, GPT-3.5 was able to achieve a passing rate on two categories of the bar test in the U.S. in the multistate multiple choice section of the Bar Exam known as the Multistate Bar Examination and achieve parity with the human test takers [35]. Other evaluations on GPT-4 put it in the top 10% of test takes in law exams [34]. In the medical domain, LLMs were subjected to 60K+ multiple-choice questions sourced from the Chinese National Medical Licensing Examination, where GPT-4 had the best accuracy of 61.6% but worse than the average human exam taker, which is at 71.6% [36].

Similarly, in Japan, ChatGPT-4 has been able to pass the Japanese national medical licensing examinations of the last 6 years, albeit sometimes its solutions reflect choices that should be avoided, e.g., proposing euthanasia [37]. ChatGPT-4 passed the US Medical Licensing Examination and was also capable of explaining medical reasoning and even crafting new counterfactual scenarios around a medical case [38]. In the physics domain, an exam of 100 radiation oncology physics questions at Mayo Clinic was developed, and it was passed while several models outperformed humans [39]. In university degree level physics, several course exams of 2018–2021 were passed, while overall, the performance was below that of the average student [40]. LLMs were also evaluated toward a computer science degree program [41], and while, e.g., ChatGPT-3.5 averaged 79.9% in ten modules and passed 9 out of 10, it would not pass the degree program due to limitations in mathematical calculations, which led to its failure to one of the ten modules. In the business administration domain, ChatGPT could achieve a passing grade of B to B- on the final exam of an MBA core course on operations management [42], which, however, does not provide a detailed view of the other MBA courses that could also be passed. Apart from the evaluations in several domains, the utilization of LLMs in management education could also be beneficial [43]. A set of comprehensive exams in different domains was carried out [34], which showed several LLMs passing various exams in a multitude of domains such as the uniform bar exam, LSAT, GRE, exams in art history, biology, calculus, chemistry, micro-/macroeconomics, physics, statistics, and psychology, etc.

LLMs do exhibit some competencies as they attempt to mimic complex human activities, but they differ from human cognitive structures in the way they interpret and organize knowledge [44]. LLMs evolve with the interactions they carry out and get more capable while exhibiting emergent behavior that does not express the values of their creators and it is challenging to interpret their inner workings [45]. Further fine-tuning with human feedback has the potential to better align the LLM with the human intent [46]. Their constant evolution can also be seen by the continuous roll-out and

updates on the publicly available models, e.g., ChatGPT [47] and Bard [48] (rebranded as Gemini in 2024) are updated several times in a month, something that also impacts their behavior. Apart from the evaluations in several domains, the utilization of LLMs in management education could also be beneficial [43]. Finally, efforts are being carried out toward reducing the cost of LLMs while also improving their performance [49], which may enable the much wider utilization of LLMs, while also considering a balanced view of the challenges they face and the contributions they can make [50]. It is estimated that for a significant part of the US workforce, LLMs are expected to impact their tasks independent of wage levels, jobs, and industry, and therefore, the potential implications of LLMs in the labor market could be significant [51].

IV. LLMS IN PROJECT MANAGEMENT

While several exams and utilization of LLMs in various domains are evidenced in the literature, when it comes to the domain of management, only a handful of efforts exist [16]. Some general aspects of the management domain overall could benefit from the exhibited LLM competencies [52]. However, when it comes to exams in management, limited investigations have been carried out so far, e.g., the assessment of ChatGPT on the MBA exam [42], but no other works seem to have addressed the angle pursued in this work. In literature, works exist that highlight the need and potential benefits of AI to project management [53], which point that AI is expected to significantly enhance tasks that are currently carried out by project managers and their teams, e.g., in creating project schedules, analyzing implications, tracking project progress, identify deviations, prioritize tasks, etc. However, while AI expectations are evident, empirical assessments to see how AI approaches, such as LLMs can fulfill them are scarce. For instance, recently, GPT-4 was assessed and compared to project managers with respect to project planning, concluding that while AI can help, especially as efficient starting points, human expertise is still necessary to validate and improve the AI-generated project plans [54]. This work, therefore, fits well in proving how well these identified needs can be fulfilled with LLMs.

LLMs have started expanding their competencies via the usage of tools [55], but no explicit assessment of project management tools is yet sufficiently analyzed in the literature. Such a development is important as then LLMs can expand their competencies and use existing task-specific tools or techniques that could have an impact on their competencies, e.g., they could use specific AI techniques for developing a reliable parametric cost model in the concept stage of a project [56]. This is considered important, especially since, due to the multimodal complexities, such aspects have been excluded. Because there is a lack of investigations when it comes to the utilization of LLMs in project management, this work can make some novel contributions.

This is the first study of its kind that reports on the assessment of LLMs in the domain of project management, and specifically on the exam that is used by the PMI to certify the project managers. In addition, it provides a comparative analysis of cutting-edge LLMs, i.e., ChatGPT-3.5, ChatGPT-4, and Bard, which can enable a better understanding of their competencies and evolution (e.g., from ChatGPT-3.5 to ChatGPT-4). Finally, via the critical discussion, not only does it cover a white spot in the literature but also enables the positioning with respect to the competencies of LLMs in other fields, where also exams have been realized [34].

V. RESEARCH METHODOLOGY

The aims of this work in assessing the competencies of LLMs require a systematic approach in order to be able to acquire the needed data and answer the research question. While several LLMs are available, including some as open source, for many, their instantiation requires significant computational and storage resources. Therefore, it was decided to use already hosted LLMs where access can be obtained, i.e., OpenAI's ChatGPT-3.5 (GPT version 3—first released on 30 November 2022), OpenAI's ChatGPT-4 (GPT version 4—first released on 14 March 2023), and Google's Bard (first released on 21 March 2023). The research carried out is categorized as empirical, it is grounded on the belief that direct observation is an appropriate way to measure reality and generate truth about the world [57]. The focus is on experiments with LLMs, and capturing the interactions with them in order to analyze and understand their competencies.

The selection of both ChatGPT-3.5 and ChatGPT-4 is seen as necessary as these are two different generation models, differing in the number of parameters and capabilities among them [58] as well as with Bard. However, one should not consider that because ChatGPT-4 is a more advanced model, it is also always a better one, as the empirical tests show, e.g., in Tables 3 and 6, ChatGPT-3.5 scores better than ChatGPT-4. So, both are needed for the comparative analysis because the assumption that a later model always performs better than the previous one does not hold.

A series of questions are asked to the selected LLMs and the responses are evaluated for correctness. The questions asked stem from the PMI's PMP Exam Prep, which contains questions aimed at candidates who want to acquire the project management certificate of the PMI [59]. The questions were input via the web interface of the models, and no parameters were adjusted for the model (e.g., temperature). A zero-shot approach is followed, without prompt engineering or other efforts in order to achieve the best performance from the LLM. Also, other methods such as prompt engineering, chain-of-thought, and retrieval augmented generation were not utilized, but are seen as future work to investigate the performance limits of LLMs, rather than their out-of-the-box zero-shot capabilities, which is the core focus of this research. The method is similar to the one carried out by Terwiesch [42] in the management domain as well as all the other examinations [34] and aims to rely on existing assessment tests for its objectives.

The collected data, i.e., the responses to the questions posed to each LLM, were scored along the criteria defined in the PMI exam preparation book [59], which provided the answers

TABLE 1. Overview of the Dataset Areas and Question Representation

Topic	Subactivities	Questions	Selected	Representation
Activity 1: Creating a high-performing team	7	23	23	100%
Activity 2: Starting the project	11	56	34	61%
Activity 3: Doing the work	8	37	36	97%
Activity 4: Keeping the team on track	8	31	31	100%
Activity 5: Keeping the business in mind	5	20	20	100%
Mastery 1: Reviewing creating a high-performing team		21	21	100%
Mastery 2: Reviewing starting the project		27	27	100%
Mastery 3: Reviewing doing the work		24	24	100%
Mastery 4: Reviewing keeping the team on track		21	21	100%
Mastery 5: Reviewing keeping the business in mind		15	15	100%
Total addressed in this research: Five activities and Five masteries		275	252	92%

to the asked questions or guidelines on how to assess free-text answers. The researcher compared the answers from each one of the LLMs to the exam's question, and then, scored each answer independently as if it were an answer coming from a human examinee. Points were awarded for the correct answers to multiple-choice questions. For the questions that required a (unstructured) text reply, the answer of the LLM was assessed according to the guidelines given in the exam preparation book [59], and the scores reflect the degree of correctness achieved. Each answer was normalized in the range of [0...1], denoting the percentage of correctness in the answer, with 1 denoting a 100% correct answer.

Several actions have been undertaken in order to address validity and reliability. The existing questions that are in the PMP exam preparation [59] were used and not developed by the researcher; hence, industry-standard questions and assessment practices from the PMI, which provides widely accepted certification for project management, are used. The questions correspond to a wide area of project management topics and stem from theories in the fields and measure to a degree what they are supposed to measure (validity), all of which is reused and relied upon by the experience of the certification process of PMI. To address the internal validity, the same exact questions were asked to all assessed LLMs (that exist in the PMP exam preparation [59]), and the scoring of the answers is objective as it follows the answers and evaluation considerations for each question as given in the PMP exam preparation book [59]. Finally, the exact training and testing data that were used for the LLMs are undisclosed, and therefore, it cannot be excluded that the prompts posed in the carried-out experiments do not already exist in the data (data contamination) [60], [61]. If this is the case, then the LLM is merely retrieving the solutions to the tests rather than actually creating a solution. Based on the scores achieved per criterion, it is considered unlikely (higher scores would be expected, especially in some multiple-choice questions), but it also cannot be excluded, and as such, this constitutes a threat to the validity of this work.

VI. EMPIRICAL FINDINGS

The questions from the PMP exam preparation book [59] cover a wide range of project management issues along the key areas of expertise [8] that every certified project manager should possess. The dataset collected follows the same structure as the PMP exam preparation test, which can be seen in Table 1. There are five main *Activities*, each of which has a number of subactivities for which questions are asked. Each activity is coupled with a *Mastery Builders*, which evaluates additional aspects in the same area as the respective Activity (e.g., Mastery 1 has questions in the area of Activity 1). While the questions in the Mastery tests are multiple-choice questions, the questions in the subactivities are mixed as they contain both multiple-choice questions as well as questions that need to be answered in the free-flow text. In that sense, the scoring of mastery tests is explicit, while for the questions of the subactivities that require free-flow text, guidelines are given on what the answers should contain, and the examiner (in this case, the researcher) scores them accordingly to the degree of fulfillment.

From Activity 2 (Starting the Project), the Activity 2–3 (Estimating Project Costs), which had nine questions, and Activity 2–4 (Estimating the Cost Baseline), which had six questions, were excluded. This was done because these two subactivities wanted the examinees to work with specific data files and figures, which was impossible to uniformly use along the three models under investigation in this work (only ChatGPT-4 has such capabilities); hence, it was decided to exclude them. This also explains the low representation in Activity 2, which is due to the exclusion of 15 questions from these two subactivities. In addition, it can be seen in Table 1, from the available questions in each section, some are omitted because it was asked to carry out tasks that were either of no interest for this evaluation or could not be given as a task to all three LLMs under investigation. This is clearly reflected in the final number of questions selected as well as the representation columns in Table 1.

Overall, the three LLMs were examined on a total of 252 questions each, representing 92% of the exam preparation questions in [59]. This gives confidence that the results achieved are representative, even with some categories excluded. This practice, which focuses only on text-based interaction and utilization of the exam parts that can be described in text, is also common to other exams where LLMs were assessed.

A. ACTIVITY 1: CREATING A HIGH-PERFORMING TEAM

The success of a project highly depends on the involved people, and therefore, it is vital that efforts should be taken under

TABLE 2. LLM Performance Results on Activity 1 and Mastery 1 Tests

Topic	ChatGPT-3.5	ChatGPT-4	Bard
Activity 1-1: Build a team	100%	100%	100%
Activity 1-2: Defining team ground rules	100%	100%	100%
Activity 1-3: Negotiating project agreements	50%	100%	50%
Activity 1-4: Empowering team members and stakeholders	87.50%	87.50%	87.50%
Activity 1-5: Training team members and stakeholders	75%	75%	75%
Activity 1-6: Engaging and supporting virtual teams	75%	83.33%	75%
Activity 1-7: Building shared understanding about a project	60%	100%	70%
Activity 1 (composite)	78.21%	**92.26%**	79.64%
Mastery 1: Reviewing creating a high-performing team	85.71%	**95.24%**	**95.24%**

The bold values denote the highest achieved score.

the umbrella of project management to enable the project team to carry out their tasks in an efficient and effective manner. The project manager plays a pivotal role not only in putting the team together to fit the project's needs but also in enabling a shared understanding of the project and its goals, empowering the team, e.g., with training and tools to be effective and collaborative, as well as to be able to manage the relation and interaction with the various stakeholders involved in the project [8]. The questions in this section aim to test how the LLMs can deal with these aspects.

Table 2 shows that all LLMs are effective and can address such challenges in this project area. All of them achieved high scores with the exception of Activity 1–3 where negotiation aspects were partially addressed by ChatGPT-3.5 and Bard, as well as Activity 1–7 where again these two LLMs unveiled weaknesses in areas of applying techniques for achieving agreement in agile environments and establishing a shared vision within the project. On the mastery test for this topic, ChatGPT-3.5 was less successful as it partially addressed issues that the other two models had addressed, e.g., aspects of effective team charter, while it failed on others, e.g., excluding reasons from the PMI motivation for publishing its code of ethics or what constitutes a training cost estimate. Overall ChatGPT-4 significantly outperformed the others in the Activity-1 tests and was at par with Bard on the Mastery test, while ChatGPT-3.5 lagged behind.

B. ACTIVITY 2: STARTING THE PROJECT

With the team engaged and empowered, the next step is to properly plan the project, which includes different aspects such as the budget, schedule, scope, quality, activities, procurement, and closure, following well-established methodologies and best practices [8]. The questions under Activity-2 aim to capture these factors and assess to what degree the LLMs are not only in a position to know the right actions but also apply them in hypothetical scenarios.

Table 3 shows a compelling view of the performance achieved in the different categories. While some practical applications in specific scenarios are excluded (Activities 2–3 and 2–4), it can be seen that most areas could be very well to excellently addressed by all LLMs. However, weaknesses are evident in Activities 2–9 and 2–10 where it seems that LLMs could not figure out the right course of action for a hypothetical situation, as well as when the verification of assumptions made for a phase should happen. Interestingly enough, all three LLMs had the same weaknesses in some questions. Similarly, in the mastery test, a common weakness

TABLE 3. LLM Performance Results on Activity 2 and Mastery 2 Tests

Topic	ChatGPT-3.5	ChatGPT-4	Bard
Activity 2-1: Determining the appropriate project methodology	100%	100%	80%
Activity 2-2: Planning and managing scope	82.14%	100%	82.14%
Activity 2-3: Estimating project costs	∅	∅	∅
Activity 2-4: Estimating the cost baseline	∅	∅	∅
Activity 2-5: Creating an activity list and a milestone list	100%	100%	100%
Activity 2-6: Sequencing activities	100%	100%	100%
Activity 2-7: Planning and managing quality	100%	100%	100%
Activity 2-8: Integrating project planning activities	100%	100%	100%
Activity 2-9: Managing suppliers and contracts	62.5%	75%	62.5%
Activity 2-10: Determining governance	66.67%	66.67%	66.67%
Activity 2-11: Closing a project or phase	87.5%	100%	87.5%
Activity 2 (composite)	88.76%	**93.52%**	86.53%
Mastery 2: Reviewing starting the project	**94.44%**	92.59%	87.04%

The bold values denote the highest achieved score.

TABLE 4. LLM Performance Results on Activity 3 and Mastery 3 Tests

Topic	ChatGPT-3.5	ChatGPT-4	Bard
Activity 3-1: Assessing and managing risks	50%	50%	66.67%
Activity 3-2: Executing a project with urgency	100%	100%	100%
Activity 3-3: Planning communications management	90%	70%	70%
Activity 3-4: Creating a stakeholder register	56.25%	87.50%	87.50%
Activity 3-5: Creating a stakeholder engagement assessment matrix	100%	100%	87.50%
Activity 3-6: Creating project artifacts	66.67%	75%	41.67%
Activity 3-7: Managing project changes	87.50%	87.50%	87.50%
Activity 3-8: Managing project issues	100%	100%	88.33%
Activity 3-9: Ensuring knowledge transfer for project continuity	80%	80%	100%
Activity 3 (composite)	81.16%	**83.33%**	81.02%
Mastery 3: Reviewing doing the work	72.92%	**83.33%**	64.58%

The bold values denote the highest achieved score.

was the correct process steps to create a project budget. Bard shows lower performance for both the activity and mastery tests when compared to the other two LLMs, while in the mastery test, ChatGPT-3.5 surprisingly slightly outperforms ChatGPT-4, which validates the reasoning for including both versions of ChatGPT in this test.

C. ACTIVITY 3: DOING THE WORK

The next step consists of the project execution aspects. Having a successful plan, as discussed in Section VI-B, is a promising start, which, however, needs to be coupled with effective execution in order to deliver the business value promised via the project. Key tasks include risk management and mitigation, communication management, stakeholder interaction management, change management, knowledge transfer, creation of project artifacts, etc., all of which are challenging and need to be effectively addressed [8].

Table 4 shows an overall good performance, with some weak spots for all models. For instance, in the activity tests, weaknesses were detected in how the LLMs address the positive risk (opportunity) or negative risk (threat) as well as how these risks could be triggered. Issues pertaining to artifact management and tools to use also emerged as weak spots. Similarly in the mastery test, all models faced issues identifying risks when presented with statements in a hypothetical scenario, as well as managing issues and how to proceed effectively with effective change control systems. Overall, while ChatGPT-4 performed technically best, the difference with the other models in the activity tests was marginal. However, in the mastery test, ChatGPT-4 outperformed by a wide margin both its predecessor, as well as Bard that had its worse test performance on all tests carried out in this topic.

D. ACTIVITY 4: KEEPING THE TEAM ON TRACK

Considering that the previous phases were successful, the project has started and is executing effectively. It is now more

TABLE 5. LLM Performance Results on Activity 4 and Mastery 4 Tests

Topic	ChatGPT-3.5	ChatGPT-4	Bard
Activity 4-1: Leading a team	75%	100%	75%
Activity 4-2: Using earned value management	90.62%	90.62%	84.38%
Activity 4-3: Supporting team performance	100%	100%	100%
Activity 4-4: Addressing impediments	66.67%	88.89%	66.67%
Activity 4-5: Managing conflict	100%	100%	75%
Activity 4-6: Collaborating with stakeholders	100%	100%	100%
Activity 4-7: Mentoring stakeholders	100%	100%	100%
Activity 4-8: Applying emotional intelligence	100%	100%	100%
Activity 4 (composite)	91.54%	**97.44%**	87.63%
Mastery 4: Reviewing keeping the team on track	90.48%	**95.24%**	90.48%

The bold values denote the highest achieved score.

TABLE 6. LLM Performance Results on Activity 5 and Mastery 5 Tests

Topic	ChatGPT-3.5	ChatGPT-4	Bard
Activity 5-1: Managing project compliance	50%	75%	50%
Activity 5-2: Using decision tree analysis	80%	100%	80%
Activity 5-3: Addressing internal and external business environment changes	83.33%	83.33%	83.33%
Activity 5-4: Discussing organizational influences on projects	83.33%	83.33%	83.33%
Activity 5-5: Employing continuous process improvements	83.33%	83.33%	83.33%
Activity 5 (composite)	76%	**85%**	76%
Mastery 5: Reviewing keeping the business in mind	**90%**	83.33%	83.33%

The bold values denote the highest achieved score.

important than ever to ensure that the team stays on track and it is up to the project manager to lead by guaranteeing that conflict can be addressed, obstacles can be removed, and stakeholders can purposefully interact with the team. To carry out such tasks, the project manager must be in a position to inspire, motivate, and influence all parties involved (team members, other stakeholders), to identify and address impediments, obstacles, and blockers, to be able to find appropriate resolutions to conflicts that may derail the project, and generally boost the team by also capitalizing on emotional intelligence [8].

Table 5 shows that all LLMs are successful in addressing aspects that keep the team on track. As seen from the scoring in the activity topics, most aspects are well addressed, with the potential partial exception of impediments, where the LLMs had some difficulty in properly identifying and distinguishing among impediments, blockers, and obstacles in hypothetical scenarios. ChatGPT-3.5 and Bard also had some weaknesses when it came to the leadership style that best represents the project manager. Similarly, in the Mastery, the scores were high, but all three LLMs failed to identify the best example of a blocker, something which is a similar weakness as identified in the Activity test. Overall ChatGPT-4 achieved the highest score, which was also the best overall score in the carried out tests, and addressed almost all aspects of this area excellently.

E. ACTIVITY 5: KEEPING THE BUSINESS IN MIND

All projects are influenced by internal or external factors, which can impact the project and affect its success. It is up to the project manager and her/his skills to anticipate and address the changes while ensuring compliance with regulatory requirements and the organizational structure, alternative delivery options, as well as employing continuous improvement processes [8]. All the above should be done with the aim of project success and the delivery of the business value the project targets to its beneficiaries.

Table 6 shows that most aspects can be well addressed by the LLMs; however, there are weaknesses in almost all criteria by all LLMs. In the Activity tests, some issues were detected in regulatory compliance risks and how these can be mitigated. Other common weaknesses were detected in aspects that need to be considered in a change management process of managing agile projects, as well as how the organizational culture, style, communication, and structure are reflected. On the Mastery test, all LLMs had issues in properly identifying some organizational structure in a hypothetical scenario, while some other issues were also reflected, e.g., the proper identification of tools for quality management issue identification, etc. Overall, as can be seen from the composite activity score, the competencies of all LLMs are good, with ChatGPT-4 being ahead of the competition, while surprisingly, in the Mastery test, ChatGPT3 outperformed both other LLMs, which were equally performing. This result validates the reasoning behind including both versions of ChatGPT in this test.

TABLE 7. Overall LLM Performance (Composite Scoring)

Topic	ChatGPT-3.5	ChatGPT-4	Bard
Activities 1–5 (composite)	83.13%	**90.31%**	82.16%
Masteries 1–5 (composite)	86.71%	**89.95%**	84.13%

F. OVERALL LLM PERFORMANCE

The previous sections have presented a detailed view and insights into the empirical results acquired per test. Overall, 252 questions were submitted to each model and evaluated. Considering a composition of all Activities and Masteries based on the scoring of their subparts (average), the metrics presented in Table 7 are derived.

As can be seen in the composite assessment, ChatGPT-4 achieves the highest score both in the Activity as well as in the Mastery tests. The obtained results are in line with an online LLM-based automatic evaluation [62] that is based on the AlpacaFarm evaluation set [63], which tests the ability of models to follow general user instructions. ChatGPT-4 presents a clear enhancement over its predecessor, ChatGPT-3.5, and at least in the context of the carried-out tests, also outperforms Bard. In this composite score, it is seen that ChatGPT-3.5 also outperforms Bard, albeit with less advantage than that of ChatGPT-4. While these composite scores may serve as an indication, they should not be generalized for other cases, as even in the carried-out tests, both ChatGPT-3.5 and Bard were outperforming ChatGPT-4 for specific tests. Therefore, it is suggested that the competencies of each model are always assessed in the specific tasks the model is intended to assist with.

As can be seen from the results, all the LLMs provided wrong answers in some of the tests, and none achieved a perfect score. While it would be interesting to understand why they would answer wrongly, an analysis of the wrong answers across all categories did not reveal an identifiable pattern of why specific questions were provided with wholly or partially false answers. Understanding where and why LLMs fail can be an interesting area of future research.

VII. DISCUSSION

Overall, the empirical results put the three LLMs that were experimented with in a very positive light. Considering that 92% of the questions that are typically used to prepare human project management professionals for their exam in order to be certified from PMI, and scores above 82%, means that all three LLMs are well-prepared to handle the different diverse project management challenges, at least in theory. Overall, the LLMs seem to possess remarkable competencies that could also benefit the domain of project management, and as attested in the literature, go toward exhibiting general intelligence [9]. Therefore, the research question posed with respect to the potential of LLMs in relation to project management practices, e.g., in being certified as project managers, can be answered with a degree of confidence that, at least for the tested parts, these LLMs would probably pass the exams. However, it has to be considered that some multimodal aspects were excluded from the tests; hence, formally, the LLMs would need to address that part in an official certification exam. Currently, only ChatGPT-4 has multimodal competencies and could potentially be a candidate for these parts of the PMI test.

LLMs, especially newer ones such as GPT-4 seem to have impressive capabilities in passing exams [34]. Such concerns, especially with tests designed for humans and utilizing them to evaluate AI, are seen critically, as the way knowledge is ingested may have led to memorization patterns [64], and aspects such as even slightly changing the way a question is phrased may lead to different model performances (and answers), something that may not hold for humans. Therefore, one has to be skeptical about such LLM competencies, if they reply on memorization patterns or they are a result of contaminated data (since for several LLMs the training material is not fully clear), or if the answers are highly volatile, depending even on how something is asked.

In this work, although the experimental results indicate that the LLMs could potentially pass the PMI exam and get certified as project managers, it would not imply that they can be effective project managers. In an analogy, while an LLM may pass the theoretical exam of a driving test, it does not necessarily mean that it has the competence to be a good car driver. This is considered a weakness when evaluating LLMs, as the specific exam may not capture additional angles and all the facets of daily work, e.g., other qualities that make a good project manager beyond theoretical expertise and potentially problem-solving competencies, e.g., interpersonal skills. Overall, it is known that exams such as licensing exams put a stronger emphasis on knowledge rather than how this may be applied in the real world. That the LLMs are exactly mastering such knowledge ingestion, could provide the false sense that knowing something (or deriving it from a pattern) may also imply competence in the application of that knowledge in real-world scenarios. Passing an exam does serve as a strong indicator that the theoretical parts are considered, and the justification provided shows some reasonable logic behind choosing specific answers, actions, or strategies, but this does not mean that it can be generalized that they would perform excellently as project managers.

Considering the multiple-choice questions in the tests, one could argue that based on the training of similar texts, such answers may be easily obtained. However, when presented with more complex scenarios (as is the case for several free-text questions posed in the test and shortly discussed in Section VI-F, bring into light the competence of LLMs to handle complex aspects. Several of such complexities were well-tackled in the project management domain, and this seems to be at par with similar results achieved in the other domains reported in the literature.

The results show that while all models perform reasonably well when it comes to very specific aspects of project management, they are challenged. This is in line with observations in other fields, e.g., in medicine, where general-purpose LLMs trained predominantly on general text data may have inadequate exposure to domain-specific texts (e.g., business processes and industrial electronics), and further fine-tuning could be beneficial. This would imply that LLMs could potentially improve if fine-tuned on additional project management-specific texts.

Fine-tuning, apart from enhancing the domain-specific competencies of the LLM, can also avoid potential issues such as the generation of outputs that can be perceived as hallucinations, untruthful, toxic, or even not helpful to the user. In the empirical results evaluated, such answers existed and were merely perceived as wrong answers that did not get any scoring points. However, in the overall project management context, these might be disastrous; the provision of bad project management advice may lead to unjustified resource usage and be linked to financial, organizational, and reputation losses. While this happens in project management also from reputable consulting experts, people tend to consider that machines do not make mistakes, and as such, impose higher standards on them. This is similar to evaluations from other domains, e.g., in medicine where ChatGPT confidently proposes immoral decisions such as euthanasia of patients [37].

The empirical results shown in Section VI show a very good understanding and application of the theoretical concepts, even in complex scenarios. It could, therefore, be that LLMs could act as advisors to project managers and enable them with views on how to carry out daily tasks and hypothetical situations, even if the way LLMs interpret and organize knowledge differs from humans [44]. In addition, because from the empirical results, a mastery of the theoretical aspects is attested, new roles for LLMs could emerge, such as their inclusion in management education [43]. In this case, the certification process may need to be altered and not consider the humans as a knowing-all entity but rather as a collaborative constellation between the project manager and the LLM who advises her/him. Such collaborations between humans and Generative AI may be beneficial if their competencies and limitations are well understood [14], [52]. Others also indicate that technical skills can be replaced by AI, while conceptual skills and human skills can be enhanced by AI [53].

In the future, the capability of the project manager to utilize LLMs, e.g., via prompt engineering [65], and critically assess the competencies of the LLM in specific complex situations should be included in the certification tests. In addition, the focus on human training could be put on areas that LLMs underperform, e.g., Activity 4-4 (addressing impediments) or Activity 5-1 (managing compliance), with the result of high project management performance by the project manager assisted by the LLM.

The responses in the experimental scenarios posed to the LLMs show that the LLMs can respond well to hypothetical scenarios posed as problems in the domain of project management, although these were directed toward human problem-solving skills. This is in line with similar findings from the other exams carried out as well as more complex and unstructured tasks, e.g., in open innovation [28] where high-level management functions of mapping, coordinating, and controlling could be learned and assist humans in context-specific efforts. Others identified several functionalities that could be beneficial to be delegated by the project managers to the AI [53]. The study indicates that project managers recognize the need to address basic project management processes but are also willing to accept technologies that would improve other aspects of the project, e.g., their work as leaders. This is significant and complements this research well, as it comes from the top down, i.e., project managers, and targets AI, while this work comes from the bottom up, i.e., LLMs as a technology to assist and potentially be integrated with the automation of such tasks or decision-making processes.

The empirical results acquired from the test have shown that LLMs can, to a large degree, predict the correct answers. However, only for some of the answers was the reasoning behind it provided. In that sense, it is not always clear (for the multiple-choice questions) if the LLM had the knowledge and provided the right answer or if this was merely a lucky guess. Such limitations are common in LLMs [66] and might depend on the way a question is asked (prompt engineering). This opens Pandora's box with respect to potential adversarial attacks, and even if this is not explicitly attacked, LLMs might not be robust, e.g., they will provide varying answers for the same question, potentially leading to false guidance for project management tasks in specific contexts.

LLMs offer good reasoning competencies [29], [30], and this has been observed in some of the answers given by the models during the test, even if we did not explicitly ask for it. This can be further evaluated by extending the experiments carried out in this work and explicitly asking the LLMs to justify their answers via appropriately crafted prompts. As such, LLMs can even act as a helping hand not only by providing advice for project management tasks, but also the justification for them, or even be used by project participants to potentially explain why some project management decisions taken by the project manager make sense, even if this is not immediately obvious.

While the project management-specific aspects are evaluated, some others, such as social aspects or ethics that may be relevant to decision making, are not sufficiently reflected in the exam questions. In addition, the impact of a wrong decision is merely not getting the respective points, while decisions in the real world have different impacts over time horizons as risk management and decision-making theories consider [67]. As such, there is a need for a more rigorous evaluation that goes beyond exam passing or technical benchmarking [66] and also evaluates the impact implications of the decisions made, including the alignment with human preferences [68]. All of the questions in the test took seconds to answer, even for the complex scenarios and free-text replies. As such, it can be asserted that tasks can be executed faster with LLMs, while a high-quality outcome is achieved in most cases, something that could enable humans to work faster and better.

In light of these considerations, based on the empirical evidence collected from three state-of-the-art LLMs, modern LLMs possess qualities that would enable them to be certified as professional project managers and carry out typical project management actions or at least confidently and correctly advise about them. This is by itself an excellent result, considering that none of the tested LLMs had been trained explicitly in the project management domain. However, as discussed, while correct answers could be provided, there are several factors that may impact their behavior, and therefore, we would suggest that LLMs are further evaluated more thoroughly for project management-related tasks and activities.

VIII. CONCLUSION

The motivation for this work was to investigate the potential of LLMs in relation to project management. Existing general-purpose LLMs, although not trained explicitly in project management, already possess significant competencies that enable them to address well diverse challenges in the domain and can get them pretty far with respect to being certified as professional project managers. The quantifiable experimental results show strong support for the answer to the RQ in that all LLMs would score well on tests that enable a human to be certified by PMI as a project management professional.

Some implications stemming from the results of this research are as follows.

1) The future of LLMs in project management is promising and preliminary results show that they can address well a variety of project management challenges, including complex ones, as demonstrated via the certification preparation questions.
2) Because LLMs have promising reasoning capabilities [29], [30] and exhibit high competencies, their role should be reconsidered and not be seen as a mere tool but potentially part of a dynamic collaborative duo, i.e., that of the project manager and LLM collaborating/interacting toward achieving better results in project management challenges—similar to the collaborative robotics concept.
3) Up to now, project management certification tests assess only knowledge and its utilization in example scenarios,

but because the LLMs have sophisticated competencies via which the humans that interact with it may be able to perform better, it is suggested that the future project management certification exams should also evaluate how well the prospective certified project managers can interact with LLMs (via prompt engineering) to deliver better results (that, e.g., otherwise would not be able to deliver in that quality or time).

Because there is a lack of investigations when it comes to the utilization of LLMs in project management, this work can make some novel contributions. This is the first research work systematically addressing the usage of LLMs in the domain of project management. The results clearly show the superiority today of ChatGPT-4, and for most aspects, these also constitute a significant improvement from its predecessor, ChatGPT-3.5. However, as also witnessed in the tests, this rule cannot be generalized for all tasks, as there were specific cases where this statement did not hold, e.g., ChatGPT-3.5 outperformed ChatGPT-4. This work has advanced the state-of-the-art by providing experimental evidence about the suitability of LLMs in the project management domain and providing a comparison among three popular LLMs based on quantitative data. It has to be kept in mind that the domain of LLMs is a rapidly evolving one and a race is currently underway on advancing the capabilities of the LLMs and making them available for business applications and public interaction. As such, their capabilities need to be continuously monitored and evaluated to track progress.

Several promising future directions could be investigated by considering LLMs in a project management context. There is a need to expand the tested LLMs with additional ones to cover a greater spectrum, and potentially several PMI tests could be assessed. There is also a need to derive metrics and test cases that can be linked to project management competencies and be credibly evaluated in the LLM context. How people interact with LLMs (e.g., via prompt engineering) influences their answers, and as such, a promising research angle might be to see how to fine-tune LLM hyperparameters and ask questions in a manner that leads to consistent and credible answers. For the project management domain, there is a need for certification processes to have new aspects that assess the LLM project manager as a collaborative constellation and measure the enhancements this can bring to project management decisions and actions in practice.

REFERENCES

[1] J. Fiedler and A. Wendler, "Berlin Brandenburg airport," in *Large Infrastructure Projects in Germany: Between Ambition and Realities*, G. Kostka and J. Fiedler, Eds. Berlin, Germany: Springer, 2016, pp. 87–145.

[2] O. Haas and P. Markovič, *Management Information System of the Critical Path of Construction Projects by Way of Example Berlin Brandenburg Airport (BER)*. Cham, Switzerland: Springer Nature, 2023, pp. 465–493.

[3] Y. Zipori, "An analysis of design and digital manufacturing processes in a PLM environment for the aerospace industry," in *Proc. ASME 9th Biennial Conf. Eng. Syst. Des. Anal.*, 2008, pp. 573–574, doi: 10.1115/esda2008-59588.

[4] T. Justinia, "The UK's national programme for it: Why was it dismantled?," *Health Serv. Manage. Res.*, vol. 30, no. 1, pp. 2–9, 2017, doi: 10.1177/0951484816662492.

[5] S. Lauesen, "IT project failures, causes and cures," *IEEE Access*, vol. 8, pp. 72059–72067, 2020, doi: 10.1109/access.2020.2986545.

[6] J. Peppard, "Why do companies' IT projects fail so often?," *Wall Street J.*, 2023. [Online]. Available: https://www.wsj.com/tech/personal-tech/it-failure-corporate-d94c68ff

[7] M. Coccia, "New perspectives in innovation failure analysis: A taxonomy of general errors and strategic management for reducing risks," *Technol. Soc.*, vol. 75, Nov. 2023, Art. no. 102384, doi: 10.1016/j.techsoc.2023.102384.

[8] PMI, *A Guide to the Project Management Body of Knowledge (PMBOK Guide)*, 6th ed. Project Management Institute (PMI), 2017.

[9] S. Bubeck et al., "Sparks of artificial general intelligence: Early experiments with GPT-4," 2023, *arXiv:2303.12712*.

[10] Z. Yang et al., "The dawn of LMMs: Preliminary explorations with GPT-4V(ision)," 2023, *arXiv:2309.17421*.

[11] C. Marnewick and A. L. Marnewick, "Digitalization of project management: Opportunities in research and practice," *Project Leadership Soc.*, vol. 3, Dec. 2022, Art. no. 100061, doi: 10.1016/j.plas.2022.100061.

[12] B. Edwards, "ChatGPT sets record for fastest-growing user base in history, report says," Ars Technica, 2023. [Online]. Available: https://arstechnica.com/information-technology/2023/02/chatgpt-sets-record-for-fastest-growing-user-base-in-history-report-says/

[13] H. Naveed et al., "A comprehensive overview of large language models," Apr. 2024, *arXiv:2307.06435*.

[14] S. Karnouskos, "Symbiosis with artificial intelligence via the prism of law, robots, and society," *Artif. Intell. Law*, vol. 30, no. 1, pp. 93–115, May 2021, doi: 10.1007/s10506-021-09289-1.

[15] J. Yang et al., "Harnessing the power of LLMs in practice: A survey on ChatGPT and beyond," 2023, *arXiv:2304.13712*.

[16] S. Karnouskos, "Large language models in project management," MSc thesis, Fac. Social Sci., Dept. Inform., Umeå Univ., Umeå, Sweden, 2023.

[17] M. R. Ennis, "Competency models: A review of the literature and the role of the employment and training administration (ETA), Office of Policy Development and Research, Employment and Training Administration, U. S. Dept. of Labor., Washington, DC, USA, 2008. [Online]. Available: https://www.careeronestop.org/competencymodel/info_documents/opdrliteraturereview.pdf

[18] F. Draganidis and G. Mentzas, "Competency based management: A review of systems and approaches," *Inf. Manage. Comput. Secur.*, vol. 14, no. 1, pp. 51–64, Jan. 2006, doi: 10.1108/09685220610648373.

[19] J. Sandberg, "Understanding human competence at work: An interpretative approach," *Acad. Manage. J.*, vol. 43, no. 1, pp. 9–25, Feb. 2000, doi: 10.2307/1556383.

[20] M. Bohlouli, N. Mittas, G. Kakarontzas, T. Theodosiou, L. Angelis, and M. Fathi, "Competence assessment as an expert system for human resource management: A mathematical approach," *Expert Syst. Appl.*, vol. 70, pp. 83–102, Mar. 2017, doi: 10.1016/j.eswa.2016.10.046.

[21] European Commission, "The European qualifications framework: Supporting learning, work and cross border mobility: 10th anniversary, Publications Office of the European Union, 2018. [Online]. Available: https://op.europa.eu/s/y6fs

[22] T. Blomquist, A. D. Farashah, and J. Thomas, "Feeling good, being good and looking good: Motivations for, and benefits from, project management certification," *Int. J. Project Manage.*, vol. 36, no. 3, pp. 498–511, Apr. 2018, doi: 10.1016/j.ijproman.2017.11.006.

[23] H. Soroka-Potrzebna, "The importance of certification in project management in the labor market," *Procedia Comput. Sci.*, vol. 192, pp. 1934–1943, 2021, doi: 10.1016/j.procs.2021.08.199.

[24] Z. Uhlir, "The effect of the project manager certification process on the development of project management—A Croatian Perspective," *Procedia—Social Behav. Sci.*, vol. 74, pp. 223–232, Mar. 2013, doi: 10.1016/j.sbspro.2013.03.009.

[25] A. D. Farashah, J. Thomas, and T. Blomquist, "Exploring the value of project management certification in selection and recruiting," *Int. J. Project Manage.*, vol. 37, no. 1, pp. 14–26, Jan. 2019, doi: 10.1016/j.ijproman.2018.09.005.

[26] D. E. Hodgson and S. Paton, "Understanding the professional project manager: Cosmopolitans, locals and identity work," *Int. J. Project Manage.*, vol. 34, no. 2, pp. 352–364, Feb. 2016, doi: 10.1016/j.ijproman.2015.03.003.

[27] L. Crawford, P. Morris, J. Thomas, and M. Winter, "Practitioner development: From trained technicians to reflective practitioners," *Int. J. Project Manage.*, vol. 24, no. 8, pp. 722–733, Nov. 2006, doi: 10.1016/j.ijproman.2006.09.010.
[28] T. Broekhuizen, H. Dekker, P. d. Faria, S. Firk, D. K. Nguyen, and W. Sofka, "AI for managing open innovation: Opportunities, challenges, and a research agenda," *J. Bus. Res.*, vol. 167, Nov. 2023, Art. no. 114196, doi: 10.1016/j.jbusres.2023.114196.
[29] S. J. Han, K. J. Ransom, A. Perfors, and C. Kemp, "Inductive reasoning in humans and large language models," *Cogn. Syst. Res.*, vol. 83, Jan. 2024, Art. no. 101155, doi: 10.1016/j.cogsys.2023.101155.
[30] T. Webb, K. J. Holyoak, and H. Lu, "Emergent analogical reasoning in large language models," *Nature Hum. Behav.*, vol. 7, no. 9, pp. 1526–1541, Jul. 2023, doi: 10.1038/s41562-023-01659-w.
[31] W. X. Zhao et al., "A survey of large language models," 2023, *arXiv: 2303.18223*.
[32] M. T. R. Laskar, M. S. Bari, M. Rahman, M. A. H. Bhuiyan, S. Joty, and J. X. Huang, "A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets," 2023, *arXiv:2305.18486*.
[33] Y. Tay et al., "UL2: Unifying language learning paradigms," 2023, *arXiv:2205.05131*.
[34] OpenAI, "GPT-4 Technical Report," 2024, *arXiv:2303.08774*.
[35] M. Bommarito and D. M. Katz, "GPT takes the bar exam," 2022, *arXiv:2212.14402*.
[36] J. Liu et al., "Benchmarking large language models on CMExam–A comprehensive Chinese medical exam dataset," 2023, *arXiv: 2306.03030*.
[37] J. Kasai, Y. Kasai, K. Sakaguchi, Y. Yamada, and D. Radev, "Evaluating GPT-4 and ChatGPT on Japanese medical licensing examinations," 2023, *arXiv:2303.18027*.
[38] H. Nori, N. King, S. M. McKinney, D. Carignan, and E. Horvitz, "Capabilities of GPT-4 on medical challenge problems," 2023, *arXiv: 2303.13375*.
[39] J. Holmes et al., "Evaluating large language models on a highly-specialized topic," *Radiat. Oncol. Phys.*, 2023, *arXiv:2304.01938*
[40] W. Yeadon and D. P. Halliday, "Exploring Durham university physics exams with large language models," 2023, *arXiv:2306.15609*.
[41] T. Krüger and M. Gref, "Performance of large language models in a computer science degree program," 2023, *arXiv:2308.02432*.
[42] C. Terwiesch, "Would Chat GPT get a Wharton MBA? A prediction based on its performance in the operations management course," Mack Institute for Innovation Management, Wharton School, University of Pennsylvania, Philadelphia, PA, USA, 2023. [Online]. Available: https://mackinstitute.wharton.upenn.edu/wp-content/uploads/2023/01/Christian-Terwiesch-Chat-GTP-1.24.pdf
[43] V. Ratten and P. Jones, "Generative artificial intelligence (ChatGPT): Implications for management educators," *Int. J. Manage. Educ.*, vol. 21, no. 3, Nov. 2023, Art. no. 100857, doi: 10.1016/j.ijme.2023.100857.
[44] T. Wei, H. Wu, and G. Chu, "Is ChatGPT competent? Heterogeneity in the cognitive schemas of financial auditors and robots," *Int. Rev. Econ. Finance*, vol. 88, pp. 1389–1396, Nov. 2023, doi: 10.1016/j.iref.2023.07.108.
[45] S. R. Bowman, "Eight things to know about large language models," 2023, *arXiv:2304.00612*.
[46] L. Ouyang et al., "Training language models to follow instructions with human feedback," 2022, *arXiv:2203.02155*.
[47] OpenAI, "ChatGPT—Release notes," 2023.
[48] Bard, "Bard—Experiment updates," 2023.
[49] L. Chen, M. Zaharia, and J. Zou, "FrugalGPT: How to use large language models while reducing cost and improving performance," 2023, *arXiv:2305.05176*.
[50] R. Mao, G. Chen, X. Zhang, F. Guerin, and E. Cambria, "GPTEval: A survey on assessments of ChatGPT and GPT-4," 2023, *arXiv: 2308.12488*.
[51] T. Eloundou, S. Manning, P. Mishkin, and D. Rock, "GPTs are GPTs: An early look at the labor market impact potential of large language models," 2023, *arXiv:2303.10130*.
[52] Y. K. Dwivedi et al., "Opinion paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy," *Int. J. Inf. Manage.*, vol. 71, Aug. 2023, Art. no. 102642, doi: 10.1016/j.ijinfomgt.2023.102642.
[53] V. Holzmann, D. Zitter, and S. Peshkess, "The expectations of project managers from artificial intelligence: A Delphi study," *Project Manage. J.*, vol. 53, no. 5, pp. 438–455, Jan. 2022, doi: 10.1177/87569728211061779.
[54] A. Barcaui and A. Monat, "Who is better in project planning? Generative artificial intelligence or project managers?," *Project Leadership Soc.*, vol. 4, Dec. 2023, Art. no. 100101, doi: 10.1016/j.plas.2023.100101.
[55] T. Schick et al., "Toolformer: Language models can teach themselves to use tools," 2023, *arXiv:2302.04761*.
[56] H. H. Elmousalami, "Comparison of artificial intelligence techniques for project conceptual cost prediction: A case study and comparative analysis," *IEEE Trans. Eng. Manage.*, vol. 68, no. 1, pp. 183–196, Feb. 2021, doi: 10.1109/tem.2020.2972078.
[57] L. Given, *The SAGE Encyclopedia of Qualitative Research Methods*. 2008.
[58] J. L. Espejel, E. H. Ettifouri, M. S. Y. Alassan, E. M. Chouham, and W. Dahhane, "GPT-3.5, GPT-4, or BARD? Evaluating LLMs reasoning ability in zero-shot setting and performance boosting through prompts," *Natural Lang. Process. J.*, vol. 5, Dec. 2023, Art. no. 100032, doi: 10.1016/j.nlp.2023.100032.
[59] PMI, *PMI Authorized PMP Exam Prep*, 1st ed., Project Management Institute (PMI), 2020. [Online]. Available: https://www.pmhouse.org/pmp-exam-prep-pmi-authorized/
[60] S. Balloccu, P. Schmidtová, M. Lango, and O. Dušek, "Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs," 2024, *arXiv:2402.03927*.
[61] S. Golchin and M. Surdeanu, "Time travel in LLMs: Tracing data contamination in large language models," 2024, *arXiv:2308.08493*.
[62] AlpacaEval, "AlpacaEval leaderboard," 2023. [Online]. Available: https://tatsu-lab.github.io/alpaca_eval/
[63] Y. Dubois et al., "AlpacaFarm: A simulation framework for methods that learn from human feedback," 2024, *arXiv:2305.14387*.
[64] A. d. Wynter, X. Wang, A. Sokolov, Q. Gu, and S.-Q. Chen, "An evaluation on large language model outputs: Discourse and memorization," *Natural Lang. Process. J.*, vol. 4, Sep. 2023, Art. no. 100024, doi: 10.1016/j.nlp.2023.100024.
[65] C. E. Short and J. C. Short, "The artificially intelligent entrepreneur: ChatGPT, prompt engineering, and entrepreneurial rhetoric creation," *J. Bus. Venturing Insights*, vol. 19, Jun. 2023, Art. no. e00388, doi: 10.1016/j.jbvi.2023.e00388.
[66] Y. Chang et al., "A survey on evaluation of large language models," 2023, *arXiv:2307.03109*.
[67] F. Eisenführ, M. Weber, and T. Langer, *Rational Decision Making*. Berlin, Germany: Springer, 2010.
[68] J. Yao, X. Yi, X. Wang, J. Wang, and X. Xie, "From instructions to intrinsic human values—A survey of alignment goals for big models," 2023, *arXiv:2308.12014*.