AI-supported Natural Language Processing in project management – capabilities and research agenda

Helge NUHN
Wilhelm Büchner Hochschule Darmstadt, Hilpertstr. 31, 64295 Darmstadt, Germany
h.nuhn@gpm-ipma.de
Alfred OSWALD
IFST – Institute for Social Technologies GmbH, Germany
a.oswald@ipm-gpma.de
Agnetha FLORE
Center for Digital Innovations Lower Saxony (ZDIN), Germany
a.flore@gpm-ipma.de
Rüdiger LANG
Consileon Business Consultancy, Karlsruhe, Germany
lang_ruediger@hotmail.com

Abstract

AI-based natural language processing (NLP) models show remarkable performance in tasks like question answering or text generation in general. We argue that recent NLP-AI models will play a major role in the transformation of our societies, an endeavor mainly shaped by projects, project managers and project teams. We present the results of an experiment, in which we fed state-of-the-art NLP models like GPT-3 project management-related questions and had an expert team rate the maturity of the answers. Results indicate that the size of the model and the text corpora seem to make substantial difference to the performance. The best model seems to be able to answer most project management-related questions convincingly to the judgement of the expert panel. Attempts to train a model without extensive training data preparation resulted in poorer results. A research agenda is derived and presented.

Keywords: *AI-enhanced Management, Artificial intelligence, Machine Learning, Natural Language processing, Project management*

1. Introduction

'Digital transformation' is an ongoing process of change in all areas of society and business, driven by the emergence of digital technologies. Artificial Intelligence (AI) and Machine Learning (ML) are very effective levers within this transformation.

Sharp definitions of what AI 'is' are hard to devise [23]. For the purpose of this article we align with a very broad AI definition: "AI is computer programming that learns and adapts". Machine learning is the part of the field of artificial intelligence in which large amounts of data are usually used to train AI systems so that they learn to solve complex tasks. Natural language processing means algorithmic work by computers, using any form of algorithms to analyze human-generated texts and extract insights from it. Deep learning alludes to complex and deep networks of computing cells that act in resemblance to neurons and were found to be good universal function approximators [21]. NLP can benefit from deep learning, meaning that intricate patterns in human language are computationally made available for a variety of tasks from analysis to generation.

Apart from a technical perspective of performance, accuracy, precision etc., other dimensions in which the application of AI systems cause long-term outcomes are of interest as well. Peeters et al. [17] suggest next to a purely technology-centric perspective also a human-centric perspective and a collective-intelligence perspective. While the first argues about differences between human and computer cognition, the second is interested in the dynamics of interactions between computer and human intelligence. This perspective gives room for different approaches to analyzing and raising interesting questions about AI use and AI research. These questions become relevant at this moment, because of an increase in the capabilities of AI algorithms on the one hand, and an increase in practical usability of AI models on the other:

Regarding algorithmic capabilities of AI approaches we can state the following. First, research studying autonomous agents that act in environments that require cognitive comprehension, tactical as well as strategic understanding and finally finesse in execution, have strived. Examples include AI systems that drive cars autonomously, win strategy games designed to be played by humans in teams like Dota 3, or board strategy games like Go. Such advances were made possible through the discovery of the capabilities of deep neural networks in reinforcement learning settings for example.

Second, remarkable advances in the field of AI research have been made in three major areas that do not focus on autonomy, strategy and tactics, but human-perceivable output formats. Video and image recognition, segmentation, manipulation, and generation have resulted in stunning IT capabilities like generating pictures based upon very little input information like text prompts [12], other artists' work [10, 1] or mere sketches [9]. Recently, OpenAI release its novel model Dall-E 2, which is capable of retouching parts of images based on text inputs and delivers photorealistic results. Here, both neural networks and other paradigms like transformers play a vital role.

Third, a long-time existing field of scientific research that has benefitted greatly from the development of capable AI algorithmics is text analysis, classification, manipulation, filling-in blanks, question-answering, or ground-up generation of texts. In this field called Natural Language Processing (NLP) we see how the precision e.g. of text summarization algorithms that find the essence of a document has increased tremendously. So much, that it is being employed in business use cases in marketing contexts already, e.g. copy-writing for advertorials. Likewise, techniques for machine generated translations of texts have reached more than acceptable performance. In response, major search engine providers like Google according to recent news articles could consider penalizing content in its ranking algorithm that was created by AI models to ensure quality and relevance of search results[1].

This third field is of specific relevance for business contexts, because of two reasons. For one, much of the daily communication, especially in projects, takes places via written texts. Even if other media are used, texts can meanwhile easily and with high accuracy be extracted from videos or voice recordings, as well as from images and charts. For another, the AI models trained for NLP tasks have recently become very strong, precise and accurate. Above all, the research in the field has led to the publication of pre-trained models, that have very high numbers of parameters and are very usable out-of-the-box. As opposed to other fields of AI research, which only deliver ever better algorithms, pre-trained models are made

[1] https://www.infidigit.com/news/google-claims-that-ai-generated-content-violates-its-guidelines/ (21.04.2022)

available. This considerably lowers the bar to experiment with these models in real-life use case scenarios and to use such models productively within end-user-directed services. Such models hence increase usability of AI greatly.

Despite this high potential that state-of-the-art AI models bear, scientific research into the transition of such technologies into neighboring disciplines is scarce. Therefore, in this article we build an argument how AI system will influence society and why because of this and quite naturally, project management must be within the scope of future AI research urgently. We then problematize the concept of competence in AI-based smart assistants that aim at being of help in the project management. We then explore potential use cases within the project management domain that could be based on NLP-based AI models. To underline our arguments, we present the results of an experiment involving state-of-the-art NLP AI algorithms and their pre-trained models. Lastly, we discuss the findings and derive a research agenda for scientists that helps guide practice-oriented research on how AI NLP models can be of value for the project management discipline.

2. AI and project management from a societal and organizational perspective

Conventional computer programs follow preset rules given by humans. In opposition, AI systems derive rules from data. Especially deep learning algorithms create models that find patterns on different micro- or macroscopic levels within these data. It is known that neural networks act as universal function approximators [21]. Being fed with information that represent the real world, our lives and societal co-existence of humans, human-generated beliefs, values, concepts and prejudices or principles are ingrained within such models [5].

As of their development state today, when such models come into use and interact with human agents, they transport the previously found patterns into the human-computer-interaction that they were designed to participate in. The AI system becomes an agent on itself, with its own capabilities and competencies, and depending on the extent of the implementation with persistent, shorter- or longer-term aims and intents.

Humans can develop meta-competence and therefore are able to work with their knowledge about their own competence and the competence of other agents. Meta-competence is therefore decisive in determining the extent to which AI systems help the digital transformation to become a real social transformation and not just degenerate into more extensive use of technology [15].

In combination with human meta-competences, AI systems can therefore make social patterns of the past and present visible and help support a transformation of organizations and societies.

Also, smaller “societies” like teams will be affected by AI technology, as it becomes more available and useful. Recent language models have shown that they can answer many questions from a variety of different domains – as we also show in this article later. With the number of parameters being the only limiting factor to the substance behind these answers, it may well be that an AI model gains more generalist and wide understanding of numerous topics than any specialized human team member ever could. This way, AI drives the integration of different disciplines such as psychology, social sciences, natural sciences, computer science, mathematics as well as philosophy and others. Consequently, social systems like teams, organizations, or societies will develop substantially differently depending on whether, to what extent, and at what quality levels such AI systems are used. Teams' cooperative capabilities will be enhanced by AI systems and thus teams will become more effective and efficient.

The ability of AI systems to recognize ever more complex patterns very quickly in large data sets of videos and images, text, audio data, numbers make them a technology suited exceptionally for use in settings that are typically cognitively demanding. Both the fact that AI systems have high potential for assisting human cognition and cooperation, as well as the nature of projects being uncharted territory for project managers and project teams, means that AI systems will be put to especially good use in projects.

Projects are forms of temporary organizations for reaching specific goals [13]. They are suitable organizational vessels innovation and organizational change. Therefore, projects challenge those working in them with constantly new situations, whether they be completely new to one of them, a couple of them, or even the entire organization. AI systems will become part of these dynamics, adding another type of agents that interact with humans in newly forming and ever-adapting social networks.

In addition, especially when projects are directed at creating product innovation or introduce new customer services, the projects themselves are exposed to increasing market dynamism and pressure. The other way around however, market pressures have also led to employing projects as forms of organizing within companies in the first place. This reinforcing dynamism is reflected by the notion, that we currently live in a phase of 'projectification' where increasingly more projects are spawned and working within projects becomes the new normal [26, 21].

Projects are an important means of shaping the future; accordingly, the digital transformation will be shaped quite significantly by projects. AI, infusing society and work environments, will lead to three central changes for project management:

- The innovation process will change substantially, as one or more AI systems (AI agents) will significantly expand and change the R&D search space alone or in collaboration with humans [25].
- The (project) management must consider that tasks have to be distributed between humans and AI, alone or together. Decision-making processes will change considerably in this composition [20].
- AI systems will flow into almost all project solutions and AI will thus become a core competence in cooperative project work. At the same time, the impact of AI project solutions on stakeholders, society and nature must be considered [15].

Based on the above-mentioned fundamental demands for the social design of digital transformation and these three key changes in project management, additional AI competencies emerge for project leaders, project members, and other stakeholders.

3. Competences in project managers and AI systems

There is considerable debate among practitioners as well as academics on what makes a good project manager [4]. This debate may partly not be settled because of the high diversity of skills and capabilities that are typically asked for in a project manager. The demanded skills vary with the parameters of the project, whether it be small or large, have many teams involved or just a few people.

Some skills are so basic to successful project management and agreed upon by professional associations like the IPMA that competency catalogues[2] have been developed that help assess what level of proficiency a project manager obtained. These are basis for assessments, which in turn make use of Q&A catalogues. They can as such be considered small, agreed-upon databases of facts and insights that can be used to obtain and test the presence of certain levels of expertise in the field of project management. They are therefore important for project managers and for anybody who wants to deploy resources onto projects, because they would wand to know that they fit the project with adequate expertise.

Following our arguments about the upcoming relevance of more AI models, it is equally important to look at the competences and capabilities of AI algorithms, models, and systems. It has already been established that language models are or act as knowledge bases [16]. They can act as a functional resource to turn to for any project manager or project team member, just like an internet search engine that most likely every reader of these lines uses multiple times a day. The difference is that the results will likely be much more focused. Search engines take the user to numerous virtual places that may or may not contain the answer to the question that the user has. Language models aim at achieving multiple tasks, two of them being question-answering or text completion. As we will show, such modes of operation offer a temptingly easy to use functionality that leaves the research work to the AI algorithm and just accept the AI's answer as the definitive one.

While the adoption of such technology may not be instantaneous, it is expectable that the extent of AI systems being used as project management assistants will increase over time if a certain threshold of usefulness is surpassed and if the assistant functions offer substantial benefit to the user [3].

With increasing extent of knowledge being cast into AI models and increasingly more users relying on such models for their work, the question arises what level of competence to attribute to an AI system. This again leads to follow-up questions, like: what is the level of static knowledge of the "knowledge base" that the AI system? Or: What are the dynamic capabilities of the system when it comes to combining pre-learned knowledge of the business domain with more context specific, novel information?

4. Use-cases for NLP-based AI systems in the project management competency domain

As functionalities and modes of operation of AI algorithms can differ, even if the underlying model remains the same, it is useful to sort potential useful use cases into a framework [14]. Based on workshops with four to five professionals and experts in the field of project management, we come up with a list of potential use cases that would assist projects, project managers and project team members along the three perspectives onto project management competencies according to the IPMA competency baseline ICB 4.0.

The IPMA Individual Competence Baseline [11] distinguishes three major competence areas "perspective", "people" and "practice", which are subdivided into further competence

[2] https://www.ipma.world/individuals/standard/

elements, indicated in the brackets below. In the following, we list some examples of the three competence areas. These examples reflect our current state of knowledge in terms of both number and design. The list will evolve and expand in the future.

Perspective (strategy; governance, structures, and processes; compliance, standards, and regulations; power and interest; culture and values)

- Analyse and compare artefacts (images, videos, speech, text, spreadsheets, etc.) to identify target hierarchies, values, believe systems or other shared mental models of different teams or organizations to display conflicts between these.
- Analyse artefacts and derive related social networks to propose changes to project organizations and related governance structures.
- Analyse conflicts and assess compliance with external or self-set rules and guidelines regarding communication.
- Generate descriptions of vision/goal and mission and value and belief systems of organizations, products or services, create comparisons between them and give advice for improvements.

People (self-reflection and self-management; personal integrity and reliability; personal communication; relationships and engagement; leadership; teamwork; conflict and crises; negotiation; results orientation)

- Analysis of team meetings, emails or chat flows based on the written or spoken word to improve communication and leadership behavior, e.g. point out communication that runs in cycles, or is not focused, or show mental blockades, or is building a team collective mind etc.
- Review effectiveness of self-organization of teams by deriving competing values and goals in meetings or communication artefacts.
- Derivation of risks and transformational hints from verbal and textual statements (meetings, documents, emails, or chats) of stakeholders.
- Support creative processes by giving enhancing impulses into meetings by word clustering or similar word, document analysis or translations from one type of media into others (text to image, image to video, etc.)

Practices (project design; requirements and objectives; scope; time; organization and information; quality; finance; resources, procurement; plan and control; risk and opportunity; stakeholders, change and transformation)

- Create summaries of various lengths of project visions, project charters, progress reports and the like, combining textual descriptions of project endeavours.
- Filter tasks and / or status information out of project email communication
- Generate goal or target hierarchies from analyses of written or verbal communication
- Create work breakdown structures, organizational charts, graphical depictions of planned approaches (one-page summaries) from textual descriptions, sketches, tables in conjunction with available text of vision and goal formulations.
- Create time-plans (tables with dates, GANTT charts, milestone lists, etc.) that rely upon conventional knowledge of typical lengths of phases or work package (i.e.,

workshop series typically last weeks, not hours, workshops last hours not, months), but also on several textual inputs (project charters, work package descriptions) allow for real-time regulation.

- Verification of contract documents and their compliance
- Generation of contract documents based on a predefined contract skeleton
- Generate or process project contract descriptions, requirements, etc.
- Analysis of texts of any form about potential risks
- Translation of texts into graphical representations (e.g., target-hierarchy, planning diagrams, etc.) and the other way around (e.g., formulate texts that describe graphs or diagrams, like organizational charts, visual process models, etc.)
- Translation of textual requirements into effort and cost estimates.
- Summarization of the core results of the temporary organization in various lengths and depths of detail, e.g., for a final project report.
- Create qualitative reviews of vision, goal, target, deliverable attainment, or changes over time

Obviously, text and spoken words play a central role in all the above use cases. If AI systems play an increasingly important role in PM, this requires appropriate artificial intelligence in text processing (Natural Language Processing (NLP)) in both project management and project-specific task domains.

5. Available state-of-the-art NLP models

Some use cases mentioned in the previous section are in part dynamic use cases that are complex and need programmatic logic on top of mere AI models to become effective. However, they stand for a far aim regarding the use of AI systems in the project management domain. To get to such advanced models, some basic properties of such AI systems need to be established. This is what we named static competencies above: knowledge contained in natural language models.

A common language model architecture that currently excels most other AI architectures in a series of benchmarks is GPT. GPT stands for generative pre-trained transformer. It is an AI architecture which, like its name says, makes use of the transformer approach. As opposed to older generations of neural-network-based NLP algorithms and models like RNNs and LSTM approaches, transformers do not only consider the immediate neighborhood of a word when it analyzes it, but uses a targeting mechanism to turn its attention to other words or phrases within a sentence or paragraph. This follows semantics of modern languages much closer and is potentially one reason why this architecture outperforms previous ones.

The real charm of current AI-based NLP models like GPT lies not only in the algorithms that make the models learn and consequently output data, but in the pre-trained models that larger organizations like OpenAI, Google, or AlephAlpha provide after extensive training of parameter-wise huge models with massive datasets. Typically, only larger organizations have the access to the necessary computing power to facilitate the training of the models - such models are computationally expensive to generate, but relatively cheap to use and deploy.

GPT-3, one of the currently most advanced models in the field, is not an entirely free resource. The owning company OpenAI does not distribute the model freely on the internet like its predecessor model GPT-2. According to statements by the company, this is so because they were too intrigued by the performance of the language model and too afraid it might be ill-used that it chose to only give API-based access to it, which can be restricted at any time. The authors of this article were eligible for access after a registration process, pointing out the intended academic use.

One more model was available to the authors. It is based on the same generative pre-trained transformer architecture, but it was trained on a smaller amount of data, aiming at a smaller total model size, and made available freely on a sharing platform for such purposes, namely Hugging Face: GPT-NEO.

6. Experiment aim, method, and measurements

The aim of our experiment was to investigate the ICB competence level of different pre-trained NLP models based on the transformer architecture. The result of this study in turn aims at assessing the capabilities useful for the overall goal of obtaining AI-induced management support for daily work of project teams.

In our experiment, we first develop a maturity model that allows four independent raters to evaluate an AI-generated text as discussed below. We then instruct one AI model to generate texts by feeding a question as an input to the selected model as a prompt. In a third step, raters independently read questions and AI-generated answers and rate the answers according to the maturity model outlined below. Answers are averaged over the four raters and then summarized for all question-answer pairs.

To achieve the study objective, we chose maturity levels as an evaluation method. Maturity models can be understood as a tool that measures the abstract quality "maturity". Maturity is a state of completeness, perfection, or completion [24]. They are commonly used to describe a logical, desired development path for objects of a class, in successive stages. This starts in the initial stage and ends in complete maturity [2]. Maturity models are thus descriptive, determinative, and comparative [19]. Maturity models are a recognized tool and used in strategy development both as a basis of planning, as well as for its evaluation. Thereby, maturity levels (e.g. around competence, capability, level of complexity) are measured within several dimensions with respect to a selected domain [7].

For our study, we used available expert knowledge of maturity models from among the researchers and derived a proprietary model to assess the area of "comprehensibility and correctness of the answer given by the AI language model". Six levels of maturity were defined for the evaluation:

- Level 0: Unrecognizable output
- Level 1: Linguistically acceptable
- Level 2: Linguistically acceptable, content far off-topic
- Level 3: Linguistically acceptable, content not far off-topic
- Level 4: Objective, formally correct in content
- Level 5: Acceptable, convincing answer, coherent in content - according to classic mindset
- Level 6: Acceptable, convincing answer, coherent in content - according to agile mindset.

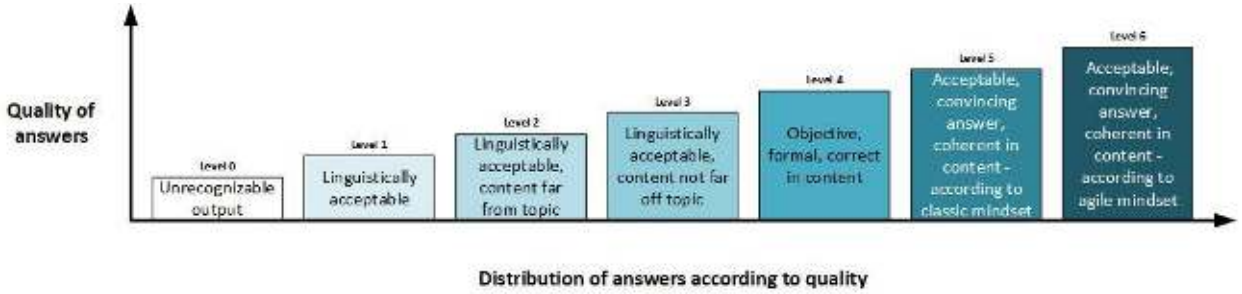
Quality of answers
Level 0
Unrecognizable output
Level 1
Linguistically acceptable
Level 2
Linguistically acceptable, content far from topic
Level 3
Linguistically acceptable, content not far off topic
Level 4
Objective, formal, correct in content
Level 5
Acceptable, convincing answer, coherent in content - according to classic mindset
Level 6
Acceptable, convincing answer, coherent in content - according to agile mindset
Distribution of answers according to quality

Fig. 1: Maturity levels for distribution the quality of responses by AI

To generate the AI answers, we used Google's Colab Pro service[3] as a runtime environment for our experiments, therefore using Python 3.6. For the first experiments we made use of the HappyTransformer[4] package, which is a façade layer for the Hugging Face[5] library and models. As NLP language models we employed GPT-NEO as provided by EleutherAI[6] in a 125 million parameter model for testing and a 1.3 billion parameter model for actual text generation. GPT-NEO was previously trained on a dataset called The Pile, a meta-dataset the contains 22 unique datasets itself [8]. For test purposes, we also used a German language model bert-base-german-cased available through the HappyTransformer package. The expert group discarded the results of the German bert language model as insufficient altogether due to low quality of the results.

Finally, to obtain access to state-of-the-art language models we were able to register with the web service of openAI and used their GPT-3 model to obtain another set of generated answers to our questions.

While the Pile has roughly 800 GB of underlying data, bert-base-german-cased was only trained on 12 GB of data. The Pile consists of books, academic papers, github repositories, websites and chat logs. The German bert model was primarily trained on Wikipedia pages. GPT-3 was trained on a filtered CommonCrawl, several book data sets, as well as the entire Wikipedia, presumably adding up to 2 TB of data, or 499 billion tokens[7].

After the first two rounds of the experiment, namely collecting and rating answers from both GPT-NEO and GPT-3, we decided to take a naive approach onto post-training the GPT-NEO model. For this purpose, we obtained access to a text-only version of the PMBOK Guide (Vol. 4) [18]. Unfortunately, we were not able to obtain a similar resource from the IPMA context and no more current version of either one. Without any further data cleansing steps, we were able to use the text as training material for the GPT-NEO model. Training took place in the same Google Colab Pro environment as discussed above and lasted roughly an hour. After the training, the same 56 questions were fed to the GPT-NEO model as before and results were rated again.

In summary, we employed three NLP models within our experiment

1. GPT-NEO
2. GPT-3
3. GPT-NEO with PMBOK training

[3] https://colab.research.google.com/
[4] https://happytransformer.com/
[5] https://huggingface.co/
[6] https://www.eleuther.ai/
[7] https://lambdalabs.com/blog/demystifying-gpt-3/

We conducted a pre-test with a different test set of questions to check the technical setup as well as to evaluate the maturity model. All questions were translated from German into English for compatibility with the NLP models, using a web service provided by Amazon, Amazon Web Services. The translations were checked and if necessary corrected by a scholar with language proficiency comparable to that of a native speaker. The test question catalogue included 48 questions and was a mix of questions of Wideman[8] and SVR Technologies[9]. The main test question catalogue contained 215 questions and answers and originates from the German Association for Project Management (GPM). It contains questions for different levels of certification in project management as examples for examination questions. Its content is divided into 14 categories:

Table 1: Categories of test question catalogue of GPM

Strategy	Results orientation
Governance, structures and processes	Project design
Compliance, standards and regulations	Requirements and goals
Power and interests	Scope of services and deliverables
Culture and values	Process and deadlines
Self-reflection and self-management	Organization, information and documentation
Personal integrity and reliability	Quality
Personal communication	Costs and financing
Relationship and commitment	Resources
Leadership	Procurement
Teamwork	Planning and control
Conflicts and crises	Opportunities and risks
Versatility	Stakeholders
Negotiations	Change and transformation

As a result of workshops within the expert group, two questions within each category were selected according to their relative importance and the expected insights that AI-generated answers to these questions would allow to be gained. The number of question-answer-pairs was intentionally reduced. This way, expert raters could undergo their task without the danger of fatigue-induced or other errors. Therefore, 56 questions remained within the catalog that were used to feed the AI algorithms with.

For each experiment, the minimum, maximum, mean, median and standard deviation were determined for the average expert rating per question. Furthermore, Fleiss' Kappa was calculated for the inter-rater reliability measurement [6].

7. Results and discussion

We present results as shown in table 2. The best results were achieved using GPT-3 as language model. The poorest were produced by GPT-NEO trained with the PMBOK content, while original GPT-NEO model ranked in the middle, but close to the mean values that its trained counterpart generated.

[8] http://www.maxwideman.com/papers/questions/questions.pdf
[9] https://svrtechnologies.com/63-best-project-management-questions-and-answers-pdf/

GPT-3 shows a remarkable average of averages of 4.29, representing a quality level corresponding to a qualitative assessment being “objective, formally correct”. Even the minimum value ø min for GPT-3 with 3.68 was higher than any maximum value ø max of all other tests. The poor result of the test GPT-NEO with PMBOK Training is demonstrated by an average value ø average of 1.34 and a maximum value ø max of 1.54.

Inter-rater agreement was classified as fair or poor. However, Fleiss’ Kappa is generating more conservative scores with increasing number of rating categories. The standard deviation indicates that the raters were in most cases only one or two maturity levels apart from one another, indicating rather acceptable convergent validity of the maturity model.

Table 2: Experiment results, best results in boldface

	(Pre-Test)	GPT-NEO	GPT-3	GPT-NEO with PMBOK Training
ø min	(1.63)	1.29	**3.68**	0.77
ø max	(3.13)	2.20	**4.66**	1.95
ø average value	(2.31)	1.76	**4.29**	1.34
ø median	(2.24)	1.79	**4.25**	1.32
ø standard deviation	(1.50)	**0.91**	0.98	1.18
Fleiss Kappa (Inter Rater Reliability)	(0.24)	**0.34**	0.29	0.19
	(Fair)	Fair	Fair	Poor

Figure 2 graphically demonstrates the locality, spread and skewness groups of the numerical data through their quartiles for our four tests. The boxplot is also based on the data of the average values for each question as rated by the four experts. It clearly shows the superiority of the GPT-3 language model, whereas the results for the GPT-NEO model were substantially lower and even decreased by training.

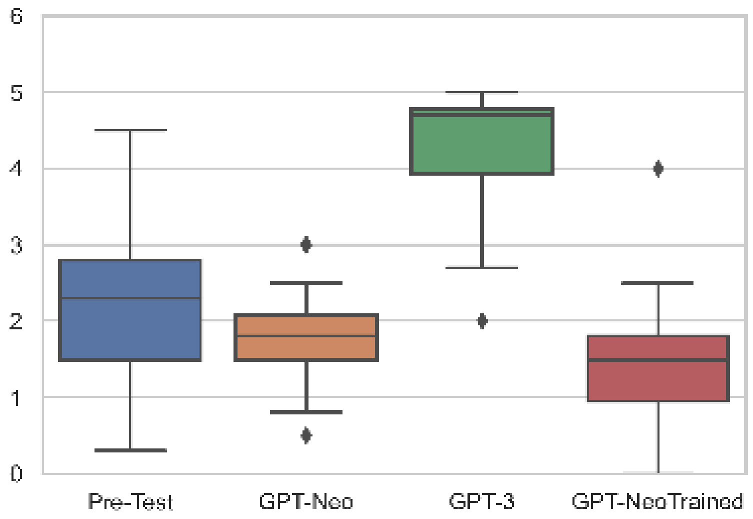

Fig. 2: Boxplot of rating results

Generally, the models that were applied "as-is", i.e., without any further attempts to train them performed reasonably well in many cases. Asked for the relevance of the Six Sigma concept for project management it stated:

Question (of the pre-test sample questions): *What is Six Sigma and how is it important in project management?*

GPT-NEO: *Six Sigma is an approach to project planning and management. Its main principle is to reduce the time and uncertainty of the project, by eliminating the weakest link in the chain. The concept of Six Sigma has been applied in different fields.*

This answer received a rating of 4.5 with a span of 2.

However, the most observed "error" that was consistently produced by the models GPT-NEO model were strings contained special characters that could not be explained by ascii to utf-8 conversions or similar encoding incompatibilities. This problem could be traced back to formatting problems that already happened during the AWS-based translation of the inputs and was recalled by the NLP model verbatim but also occurred deliberately within originally produced answers without any mirroring of input phrases or words. While this is a minor problem to fix for technically skilled users of the models, it did influence the experts' ratings. It led to systematically lower ratings as was discussed in ex-post between the raters. Notice the string "*â€™*" in the following answer

Question: *How do you deal with underperforming project team members?*

GPT-NEO: *One of the best ways to handle an under performing project member is to make sure that they have a healthy level of accountability. Youâ€™ll want to be sure to give them a chance to re-evaluate their role.*

Also, the GPT-NEO model had to be set to only output phrases of a fixed maximum length. Therefore, it is hard to tell whether the really interesting part of an answer was simply yet to come, as shown in the following example that is an answer to the question what to consider when delegating tasks as a project manager.

GPT-NEO: *I am looking at the following scenario:*
You are a contractor for a large company that has a single building. This building is being renovated, and you are being paid in part by the company to manage the renovation. You have a very

Another noticeable pattern was to find passages of text that were seemingly taken from publications as-is, or from FAQ repositories. Notice the "A:" followed by a paragraph.

GPT-NEO: *This article describes the types of work that must be undertaken by the project personnel.*

GPT-NEO: A:
You may have found this question in the context of Bitcoin and Bitcoin mining, but there are many other scenarios as well, where confidence is necessary.
Trust is important

When people use a service that others have built, they

On slightly fewer occasions, the NEO model also produced counter questions, but not answers, to the posed questions:

GPT-NEO: *What are the different types? Can I get some tips for starting a project in this kind of context?*

In other instances, quite obviously seemed to copy content directly from websites, blogs, chats etc. The copied content was seemingly copied without adaptation to the specific content of the question. The following link links to a valid web domain, but not to a currently active web page.

GPT-NEO: *What can we do to combat it? We just have to figure out what is causing it, and what can be done about it.*
(Source: http://www.npr.org/2011/06/11/1350

GPT-NEO: *â€¢ **The schedule and cost plans should be reviewed throughout the planning and scheduling processes.***
5. 4.2.4.1 Reviewing the scope of the schedule
To review the purpose and scope for the

On many occasions, colloquial language was produced which can very likely be linked back to the pre-training process and the corresponding data sets used within it.

GPT-NEO: *There are many questions you should avoid like the plague when youâ€™re seeking a creative answer. These include:*
Does this solve the problem? Does it offer the right solution? Is it the best possible answer? Do you feel satisfied

While we did not encounter offensive language or intensely toxic statements, we did find instances of clichéd responses:

GPT-NEO: *Males are: (i) able to reproduce their own species (ii) highly competitive, (iii) easily satisfied, and (iv) are generally very docile The females are (i), (v), and are highly [line breaks removed for readability]*

While such statements could possibly be valid in their original contexts – think of a description of a desired online game character that was discussed in an online forum – it is a rather noteworthy answer to the original question, which asked to draw (sic!) Maslow' Need Pyramid.

Other types of errors that occurred frequently within the answers given by GPT-NEO include references to people or (mostly IT) projects.

In general, GPT-3 produced much less errors of this kind. Answers given were usually consistent and coherent within the answer. Practically never did counter questions, special characters of abruptly ending sentences occur. Also, obvious copies of texts from websites, blogs, FAQs or chats/forums were not noted.

Question: *What is project management?*

GPT-3: *There is no single definition of "project management," but at its core, project management is the process of planning, organizing, and executing a series of tasks in order to achieve a specific goal. Typically, project management is used in the context of larger projects that require coordination between multiple people and teams.*

Limitations

To our knowledge no research on project management domain knowledge contained within language models has been published before, making this a unique and novel contribution to what we deem a new field of research. As there are consequently no agreed standards for experiments like this, much of the research reported here is subject to limitations.

For one, we unfortunately could not rely on previously developed and validated maturity models or comparable instruments of measurement of the quality of the AI-generated answers. During the experiment, we noticed several shortcomings of the maturity model itself. For example, maturity level six was seldomly used and the differentiation between levels 5 and 6 are profoundly subject to subjectivity. With this regard, the maturity model should be adapted, and a proper code book should be generated for further research into this direction. This would also help convergent and discriminant validity further.

Also, while the selection of language models followed a stringent argument, the selection of post-tests material was coincidental. With increased efforts, more current literature from different schools of thought regarding the project management discipline could be used to enhance the training steps. In addition, no pre-processing of the additional learning texts took place, while some errors, likely due to scanning and digitizing, were easily identifiable upon quick visual inspection. This pre-processing, we assume, would have the potential to increasing ex-post training of the language models drastically. Many of the occurrences of references to chapters like '5.5.6.4' can be traced back to the pattern of chapter references within the PMBOK guide. We consciously chose a way of minimal pre-processing of additional learning material to simulate likely future use, that can in many cases not rely on machine learning experts to train and fine-tune available language models, however.

8. Research agenda

Given the rapid development of the field of AI models with elaborate levels of proficiency in the generation of text that is relevant to both society and project management in specific, we suggest more researchers turn to experimentation with AI models and potential use cases in the field. As we progressed through our experiments, we noticed several open questions that arise from the quite novel field in conjunction with the field of project management. The following questions should therefore be addressed not strictly, but preferably in the order that they are presented in. While there may be other approaches that would render parts of the proposed questions irrelevant, we argue that deep learning-based models are probably here to stay due to their demonstrated power and relative ease of use.

Assessment of pre-trained knowledge

- How can general knowledge, domain-specific knowledge and especially project management-specific knowledge be assessed and benchmarked for pre-trained models? Can this process be automated?

- What types of knowledge are contained within existing AI models? E.g., sub-domain areas of the project management body of knowledge?
- What would a standard benchmark look like to assess knowledge competences within pre-trained models (structurally / procedurally)?
- Can pre-training text corpora, independent of specific algorithms and models be rated with regard to the domain-specific competency that they teach these models?

Training / Learning

- Objective standards: How should machine learning domain-specific knowledge be done from a processual and structural perspective?
- What can be expected from training AI models? I.e. how many training runs are helpful, should models be trained with multiple different data sets and what is the consequence? Can existing knowledge be overwritten and models be "convinced" of alternate truths?
- Can language models be enabled to critically reason between two or more positions on a certain question that they learned?
- How reliable is training? E.g. can models be expected to learn something on the first attempt? What if conflicting factual knowledge is contained in the model already?
- Are there tipping points in obtained knowledge that abruptly change an overall perspective of a language model onto a topic?
- How stable are the results that can be expected? I.e. does continual learning potentially let learned knowledge subside?
- Does mislearning take place, under what circumstances and how to avoid it?

Evaluation and interpretation (after post-training)

- How can learning success be assessed and benchmarked?
- What is the influence of randomization, temperature and seeds, within text-generation models on given statements regarding factual knowledge?
- Can AI models develop several "characters" that can be discerned like humans could be? E.g. are there models that argue more or less risk-averse?

With further steps into the corresponding research direction, surely the list would grow rapidly. We expect it to be necessary or at least very helpful to team up with computer scientists, data analysts, but also with linguists and cognitive psychologists to tackle further research in this area.

9. Conclusion

In this article we made the point that AI NLP models have made remarkable technological progress within the recent years. A point has been reached, where some of the most advanced language models can tackle tasks that were previously exclusively solvable by humans. Such pre-trained models were published openly, they are publicly available or at least available with negligible results for the use cases that are discussed here. In this state already, they show potential for generating value for project management practitioners and theorists alike. We have shown that based on our knowledge of current AI model capabilities, many practical use cases are thinkable that could enhance project management endeavors. To underline our cases, we presented the results of a computer experiment, involving state-of-the-art natural language model architectures. We have shown that GPT-3 as one of the most advanced models to date reaches very high levels of expert rated competency attributions. So much so, that having a respective language model could potentially pass written tests on the subject. Other models struggle heavily both in form and content to convince experts of their domain-specific knowledge. Taken a naive approach to training actually worsened the results instead of improving them. It is notable that none of the models had been trained by their creators and AI professionals with project management as a discipline in mind. This underlines how general the knowledge of our discipline is taken altogether, or put differently, how much project management is contained such general text corpora that represent a breadth of our society. The sheer potential that AI-based assistants have for more effective and efficient project management is the root of our call for further research, which we will follow and hope others to join in.

References

1. Abdal, Rameen, Peihao Zhu, Niloy Mitra, and Peter Wonka. „StyleFlow: Attribute-conditioned Exploration of StyleGAN-Generated Images using Conditional Continuous Normalizing Flows.". 2020.
2. Becker, J., Knackstedt, R., & Poeppelbuss, J. (2008). Dokumentationsqualität von Reifegradmodellentwicklungen. Arbeitsbericht Des Instituts Für Wirtschaftsinformatik, (121), 1–50. Retrieved from http://www.ifib-consult.de/publikationsdateien/2009.pdf
3. Bentley, F., Luvogt, C., Silverman, M., Wirasinghe, R., White, B., & Lottridge, D. M. (2018). Understanding the Long-Term Use of Smart Speaker Assistants. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *2*, 1–24.
4. Chipulu, M., Neoh, J.G., Ojiako, U. and Williams, T. "A Multidimensional Analysis of Project Manager Competences," IEEE Transactions on Engineering Management, vol. 60, Art. no. 3, Aug. 2013, doi: 10.1109/tem.2012.2215330.
5. Christian, B. (2020) The Alignment Problem: Machine Learning and Human Values: W. W. Norton & Company, kindle edition
6. Fleiss, J. L. (1971) "Measuring nominal scale agreement among many raters." Psychological Bulletin, Vol. 76, No. 5 pp. 378–382
7. Flore, A. (2020). Reifegradmodell für Smart Grids: Bewertung der Migrationspfade anhand von zwei Fallstudien; https://www.shaker.de/de/content/catalogue/index.asp?lang=de&ID=8&ISBN=978-3-8440-7749-0&search=yes
8. Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S., & Leahy, C. (2020). The Pile: An 800GB Dataset of Diverse Text for Language Modeling. arXiv Preprint arXiv:2101.00027.
9. Ghosh, Arnab, et al. „Interactive Sketch & Fill: Multiclass Sketch-to-Image Translation." Interactive Sketch & Fill: Multiclass Sketch-to-Image Translation. 2019.
10. Härkönen, Erik, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. „GANSpace: Discovering Interpretable GAN Controls." arXiv, 2020.
11. IPMA (2015) Individual Competence Baseline for Project Management, Version 4.0: IPMA International Project Management Association
12. Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2021). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. arXiv Preprint arXiv:2107.13586.
13. Lundin, Rolf A., and Anders Söderholm. „A theory of the temporary organization." 11 (1995): 437-455.

14. Nuhn, H. (2021) Organizing for temporality and supporting AI systems – a framework for applied AI and organization research, Lecture Notes in Informatics, GI e.V
15. Oswald, A. (2022) The Whole – More than the Sum of Its Parts! Self-Organization – The Universal Principle! in Ding R, Wagner R, Bodea CN (editors) Research on Project, Programme and Portfolio Management – Projects as an Arena for Self-Organizing: Lecture Notes in Management and Industrial Engineering, Springer Nature
16. Petroni, F., Rocktäschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A. H., & Riedel, S. (2019). Language Models as Knowledge Bases?
17. Peeters, M. M. M., van Diggelen, J., van den Bosch, K., Bronhorst, A., Neerinex, M. A., Schraagen, J. M., Raaijmakers, S. (2021) Hybrid Collective Intelligence in a Human-AI Society, in AI & Society Journal, March 2021
18. PMI Project Management Institute (2008), A GUIDE TO THE PROJECT MANAGEMENT BODY OF KNOWLEDGE. Project Management Institute, Inc. Newton Square, PA, USA.
19. Poeppelbuss, J., & Röglinger, M. (2011). What makes a useful maturity model? A framework of general design principles for maturity models and its demonstration in business process management. Ecis, 1–12. Retrieved from http://aisel.aisnet.org/ecis2011/28/
20. Puranam, P. (2021) Human–AI collaborative decision-making as an organization design Problem, Journal of Organization Design (2021) 10:75–80
21. Scarselli, F., & Tsoi, A. C. (1998). Universal approximation using feedforward neural networks: A survey of some existing methods, and some new results. *Neural Networks*, *11*, 15–37.
22. Schoper, Y., Wald, A., Ingason, H., & Fridgeirsson, T. (2017). Projectification in Western economies: A comparative study of Germany, Norway and Iceland. International Journal of Project Management, 36. doi: 10.1016/j.ijproman.2017.07.008
23. Schuett, J., & others. (2019). A legal definition of AI. arXiv Preprint arXiv:1909.01095.
24. Simpson, J., & Weiner, E. (1989). The Oxford English Dictionary.
25. Vasilescu, D.-C., Filzmoser, M. (2021) Machine invention systems: a (r)evolution of the invention process?, Journal AI & Society, January 2021
26. Wald, Andreas Erich; Spanuth, Thomas; Schneider, Christoph; Schoper, Yvonne (2015). Towards a Measurement of "Projectification": A Study on the Share of Project Work in the German Economy. Advanced Project management (Vol 4) Flexibility and Innovative Capacity. ISBN: 978-3-924841-72-0. *GPM Deutsche Gesellschaft für Projektmanagement e.V.*. Chapter 1.2. s 19 - 36.