Neural Text Generators in Enterprise Modeling: Can ChatGPT be used as Proxy Domain Expert?

Kurt Sandkuhl
The University of Rostock
Rostock, Germany *kurt.sandkuhl@uni-rostock.de*

Balbir S. Barn
Middlesex University
London, UK *b.barn@mdx.ac.uk*

Souvik Barat
Tata Consultancy Services Research
Pune, India *souvik.barat@tcs.com*

Abstract

Enterprise modeling is concerned with the systematic development of a comprehensive and holistic representation of an enterprise (an enterprise model) to support organisational initiatives. Domain experts have an essential role in enterprise modeling projects (EM), as they provide the required domain knowledge or specifics of the organisation under consideration. The paper investigates if neural text generators (large language models) can reduce the dependency on domain experts for certain tasks in enterprise modeling. The main contributions of this paper are (1) a systematic literature analysis on neural text generator use in EM, (2) the identification of potential for applying large language models in EM, and (3) findings from quasi-experiments comparing output of ChatGPT and domain experts for the same EM task.

Keywords: enterprise modeling, large language model, ChatGPT, conceptual modeling, proxy domain expert.

1. Introduction

Enterprise modeling (EM), in general, is concerned with the systematic development of a comprehensive and holistic representation of an organisation (an enterprise model) to support organisational initiatives, such as identifying improvement potential, supporting operational processes, changing business models or adopting technological innovations (cf. Section 2.1). Recent advances in the field of artificial intelligence (AI) have resulted in opportunities to use AI for various tasks in EM, such as machine learning integrated into decision modeling [3], into recommender systems in business process modeling [14], or the use of graph neural networks for assisting modellers [18]. However, using AI techniques as a source of domain knowledge or as a substitute for subject matter experts has not attracted much research (see Section 4).

In enterprise modeling projects, domain experts are an important resource, as they contribute knowledge about the application domain under consideration in general or the organization under investigation in particular (cf. section 5). In most enterprises, highly-experienced domain experts who can provide this domain knowledge are very busy and not easily available, which can delay modeling projects. Neural text generators such as Large language models (LLM) based on the GPT-3 architecture (cf. section 2.2), can be a tool to reduce the workload of domain experts and, therfore support EM as such. Our hypothesis is that ChatGPT and similar technologies can assist in EM by providing general domain knowledge or gathering basic facts. The main objective of this work is to further investigate this topic by exploring the potential and the limits of using ChatGPT as proxy (substitute) for domain experts. The focus is on what tasks in enterprise modeling could be supported by ChatGPT, what prompts to use to collect

the required domain knowledge and to establish the accuracy of the information provided by ChatGPT.

The main contributions of this paper are (1) a systematic literature analysis on the use of neural text generators in EM, (2) the identification of application potential for LLMs in EM, and (3) findings from quasi-experiments comparing output of ChatGPT and domain experts for the same EM task. The paper is structured as follows: Section 2 briefly introduces the background for our work. Both enterprise modeling and neural text generation such as that produced from LLMs is covered. Section 3 introduces the research approach applied in our work. Section 4 presents the results of the systematic literature analysis undertaken. Section 5 analyses the general application potential of LLMs in EM. Section 6.1 introduces the experiment design and section 6.2 the experiment results. Section 7 discusses the results followed by the final section on concluding remarks.

2. Background

2.1. Multi-perspective Enterprise Modeling

EM is addressing the "systematic analysis and modeling of processes, organisation structures, products structures, IT-systems or any other perspective relevant for the modeling purpose" [22]. The variety and dynamics of methods, languages and tools supporting EM is visible in work on research roadmaps and future directions, originating both from the information systems community (see, e.g., [16]) and from scholars in industrial organisation (e.g., [21]).

Enterprise Modeling (EM) is meant to support organisations in coping with a broad range of challenges, including managing organisational change in dynamic business environments, aligning of business goals and information systems to support these goals, as well as explicating and consolidating business knowledge from diverse stakeholder groups thus facilitating organisational learning. The role of Enterprise Modeling usually is to provide methods, tools, and practices for capturing and visualising the current ("as-is") situation and to develop the future ("to-be") situation. In particular, a model of the current situation forms one of the fundamentals for supporting future development of organisations.

Given the complexity of enterprises, in the course of modeling an enterprise, there is the need to understand, analyse, capture and represent what is relevant for different stakeholders and/or modeling purposes. In this context, there seems to be an agreement in the academic literature related to enterprise modeling that a key feature of an enterprise model is that it includes various perspectives. Among the most prominent ones is [5] and [15] to use EM as a problem-solving tool. Here, EM is only used for supporting the discussion among a group of stakeholders trying to analyse a specific problem at hand.

2.2. Neural Text Generation - the rise of Large Language Models

The development in large language models and their evolution has been widely documented and the reader is directed to key texts such as [4]. The pre-training of LLMs is task-agnostic [9].

LLMs present new opportunities for experimentation and prototyping with Artificial Intelligence (AI) as pre-training ensures that enough information is encoded such that customisation is possible, in-context and at run-time to enable handling of new tasks through prompts expresssed in natural language [23]. GPT-3 with its Chatbot frontend - ChatGPT[1] can solve a variety of tasks that have so far included summarization, translation, grammar correction, email composition and others [7]. The so-far free availability of ChatGPT and the very simple and powerful prompt based front-end to GPT-3 has led to many domains of application. In higher (tertiary) education, there is a fulsome debate about the potential of academic misconduct as well as the

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

opportunities such as that described in [2]. In medicine, the use of ChatGPT performance has been evaluated in AI assisted medical education for the United States Medical Licensing Exam (USMLE) [12]. The evaluation here is particularly interesting as a means of adjudicating on the performance of the LLM using a scoring system for accuracy, concordance and insight.

Prompt engineering [13] is now a critical component of study for LLMs as it encompasses the techniques by which end-users can use LLMs to perform prediction tasks where the original input x is modified using some form of template whose unfilled slots are populated using the probabilistic models encoded in the LLM such that a final prediction output y is obtained. A comprehensive review of prompting methods is available in [13]. Prompts are often described as zero-shot or few-shot. A zero-shot prompt describes the intention of the task requirement in natural language. E.g. a prompt asking ChatGPT to ask if a Volkswagen Beetle is a car forms a simple classification task. Few-shot prompts are those that demonstrate to the LLM the required pattern (desirable inputs and outputs) to follow in order to fine tune the LLM to produce the desired prediction. An example typically has a context and a desired completion (for example an English sentence and the French translation).

The fluid response of LLMs to prompts given to the system means that prompt based prototyping allows non-Machine Learning (ML) experts to prototype ML functionality at lower cost and without the need to train models up front. Effectively, augmenting input with answered prompts becomes in-context learning.

3. Research Method

Work presented in this paper is part of a research program aiming at developing methodical-technical support for enterprise modeling based on artificial intelligence techniques. In this context, this work explores the use of neural text generators in enterprise modeling with a specific focus on exploring the potential and the limits of using ChatGPT as a proxy (substitute) for domain experts. The main research question is: **In enterprise modeling, how can neural text generation be used as a substitute for domain experts?** This question can be refined into the sub-questions RQ 1.1: In what areas of enterprise modeling could neural text generation potentially be used? And RQ 1.2 For the identified areas of EM, how consistent and complete is the output of ChatGPT with the information provided by domain experts? The research method used to answer the research questions is a combination of literature review, conceptual-deductive work and quasi-experiments.

The literature search aimed at identifying related work and results from other scholars to be taken into account when investigating the potential of neural text generators. For this step, we used Kitchenham's approach for systematic literature reviews (SLR). Kitchenham [10] suggests six steps, which we briefly introduce in the following and document in detail in section 4. The first step is to develop research questions (RQ) to be answered by the SLR. The process of paper identification starts with defining the overall search space (step 2), which basically consists of determining the literature sources to take into account in the light of the research questions. Paper identification continues with the population phase (step 3). In this step, the search string is developed and applied by searching the literature sources. Afterwards, the step "paper selection" follows by defining inclusion and exclusion criteria and a manual selection of relevant papers found in the population phase (step 4). The data collection phase (step 5) has its focus on extracting the information relevant for answering the research question from the set of identified relevant papers. The last step is the analysis of data and interpretation, i.e., to answer the research question defined in step 1 by using collected data of relevant papers.

Critically, as the SLR returned no previous work on identifying EM areas suitable for neural text analysis, we structured the field of EM along the tasks to perform during a modeling project and the sub-models to produce. Based on this structure, we identified potential areas for LLM use (see section 5). This is the argumentative-deductive part of our work.

In the final step, two of the identified potential areas were selected for further investiga-

Table 1. Results of the literature analysis

Query	No. of Hits Scopus	AISeL	IEEE Xplore	Relevant
(("neural text" OR "ChatGPT") AND ("enterprise modelling" OR "conceptual modelling" OR "process modelling))	0	0	0	0
(("neural text") AND ("modelling" OR "modeling"))	79	4	17	0
(("text generator") AND ("modelling" OR "modeling"))	31	0	5	1
Total				1

tion by conducting quasi-experiments. A controlled experiment in software engineering and information systems development is "a randomised or quasi-experiment in which individuals or teams (the study units) conduct one or more [. . .] tasks for the sake of comparing different populations, processes, methods, techniques, languages or tools (the treatments)" [19]. In our work, we perform a quasi-experiment; the study units are ChatGPT and domain experts, and the treatments are different modeling tasks. A quasi-experiment is "an experiment in which units are not assigned to conditions randomly" [6]. The experiment does not aim at testing a specific hypothesis but is exploratory research to answer the research questions defined. The experiment design is described in detail in section 6.

4. Literature Analysis

This section describes the results of a systematic literature review (SLR) that follows the procedure proposed by Kitchenham. Starting point for the SLR is the research question (RQ) What previous scientific work is visible in publications about using neural text generators in enterprise modeling? Based on the research question, the literature search started with an initial search in Google Scholar using “neural text” and “enterprise modelling” as the main keywords. Based on the initial results, which showed no exact hits but a broad bandwidth of potentially relevant areas, synonyms and associated terms for these two main keywords were identified. For neural text we used as synonyms neural text generator and ChatGPT; for enterprise modeling we used conceptual modeling, process modeling and modeling. Importantly, we accounted for different ways to spell modeling (c.f. modelling).

The synonyms for enterprise modeling were chosen from previous experience in the field. The synonyms for neural text were derived from the search results in Google Scholar. The literature databases selected for the analysis were Scopus, AISeL and IEEE Xplore to ensure a good coverage of the fields computer science and business information systems. In Scopus, we searched title, abstract and keywords, in AISeL all fields and in IEEE Xplore, all meta-data.

As visible in Table 1, the queries specifically addressing neural text or ChatGPT in enterprise, conceptual and process modeling did not return any hits. Neural text in modeling resulted in 91 unique hits aggregated from all three databases. The majority of these papers are from the fields of speech synthesis, text-to-speech, development of neural language models, neural text generators and neural text classifiers, as well as the use of neural text generators in document processing. Some papers also originate from dialogue modeling and text rewriting, However, none of the 91 papers contributes to our research question and is considered relevant.

Furthermore, the query addressing text generators in modeling returned 33 unique hits. These hits address a variety of topics from very diverse areas, such as metamodels for writing textual transformations, SysML and Simulink integration, text generation for requirements

modeling and social media, or data augmentation. The only relevant paper [1] addresses the translation between different business process modeling languages and uses natural-language text generation for improving the understandability and user acceptance of this translation. Here, the focus is on model to text generation, which is not addressing our research field. Thus, there is a large gap in research studies providing additional significance to this research.

5. Application potential of Neural Text Generators in EM

The investigation of the potential of ChatGPT as a proxy for domain experts has to start from the role of domain experts in enterprise modeling and their expected contribution. Stirna and Persson [20] define the contributions of domain experts in general as “supplying domain knowledge, knowledge about organisation units involved [...]; examining and evaluating the results of enterprise modelling, and integration of modelling results of different teams into a consistent whole.” These contributions are required for different aspects of an enterprise that are also called viewpoints or perspectives. According to Frank, most enterprise modeling approaches include several perspectives that address concerns of different stakeholder groups and potentially require different domain experts [8]. An analysis by Vernadet showed that frequently used perspectives are goals, organisation structure, process, products and IT and resources [21].

In addition to different contributions expected from domain experts and various perspectives, the different modeling phases require different ways of participation from the domain experts. [11] concludes that the most relevant modeling phases to be distinguished are scoping of the modeling project ("scoping"), preparation of the modeling project ("preparation"), modeling of the current situation (“as is"), analysis of the “as is” and modeling of alternatives for addressing identified change needs (“change alternatives”), and modeling of the future situation for the selected alternative (“to be”). In scoping and preparation, the domain expert commonly has the task to provide relevant knowledge on the application domain and the organization in general. In the “as is” modeling, additional knowledge about the enterprise under consideration is the most important contribution of the domain experts. The results of the modeling process have to be examined for accuracy and completeness. In the process of analysis and finding alternatives, creativity in designing feasible and acceptable changes is most important. In modeling the “to be” situation, the domain experts have to make sure that the different perspectives add to a consistent whole.

In total, this results in a variety of different stages that potentially could be examined for the potential of ChatGPT support. The focus of this work is on supporting the domain experts’ role in (a) the preparation of the modeling project and (b) the identification of alternatives for change. The underlying conjecture for this decision is that analytical tasks in the early modeling phases are more suitable than the more creative later phases, i.e. the idea is to contrast the analytical preparation work with the more creative work of defining alternatives for change.

In addition to the domain experts’ knowledge contribution, we also have to observe the actual modeling task. Development of a model consists of at least four elementary tasks: identifying the model elements in every perspective, i.e., identifying concepts and relations between them; identifying the relations between elements of the different perspectives, and refinement of model elements if required. Table 2 shows the phases of an EM project as rows and the different contributions of domain experts as columns.

6. Experiments on ChatGPT use in EM

6.1. Experiment Design

To investigate the potential of ChatGPT, we designed two quasi-experiments. The first experiment (E1) focuses on the preparation phase for modelling the current situation in an enterprise. In the preparation phase, the aim is to prepare the modelling team for the upcoming modelling

Table 2. Potential application areas of neural text generators in EM

Phases of EM project	*supply of domain knowledge*	**Task of Domain Experts** *integrate modeling results*	*evaluate results*
Scoping	overview about organization in general and problem areas to investigate		
Preparation	application domain and the organization in general		
As-is modeling	perspectives relevant for the scope (e.g., goals, organisation structure, process, products, IT, resources)	models developed for the perspectives	individual models and inter-model integration
Alternatives for change	potential changes; how realistic and accepted are they?		
To-be modeling	all perspectives relevant for the change	models developed for specifying the change	individual models and inter-model integration

project by identifying and collecting relevant information from the organization and the application domain under consideration. Basic domain knowledge in the modelling team is necessary to prepare stakeholder interviews and other elicitation activities. Furthermore, this knowledge helps during scoping of the project. A common practice in modelling projects is either to permanently integrate an expert familiar with the application field into the modelling team or to associate such an expert during the preparation phase. In this context, accuracy of the information and completeness of the areas to investigate during the modelling project are essential.

The second experiment (E2) focuses on preparing changes in an organization by identifying different alternatives for the future situation. For this purpose, modelling projects typically use interviews with selected stakeholders or modelling workshops with domain experts from the organization under consideration. Here, relevance and feasibility for the organization under consideration are essential. In both experiments, correct use of the selected modelling language is another issue.

As an application domain for both experiments, we selected the management of a higher education institution (HEI). E1 focused on common business operations in a university; E2 was directed to the task of improving the rating of the HEI and business rules suitable to implement these goals. For modelling language, we selected 4EM [17], a multi-perspective EM language used in many universities for teaching EM. 4EM distinguishes between the goal/problem, business process, actors and resources, business rule, products and services, concepts, and technical components perspectives. The 4EM meta-model defines the concepts and relations for all perspectives.

The general setup for both experiments was as follows:

1. Both, ChatGPT and the domain expert, were asked to provide information for the same task. The prompt for ChatGPT included information about the notation of the 4EM modelling language, whereas the domain expert had some experience with 4EM and only received the information about the task.

2. The results of ChatGPT and the domain expert were analysed by researchers conducting the experiment in five steps:

   (a) Differences in terminology: in case ChatGPT and the domain expert expressed the same information with different words, the terminology was harmonized and the

resulting changes documented

(b) information in the domain expert's result not contained in the ChatGPT output was identified

(c) information in the ChatGPT output not contained in the domain expert's result was identified

(d) information contained in both results also was identified

(e) information obtained from domain experts are provided to ChatGPT as examples for better outcome (i.e., few-shots learning).

3. The domain expert was asked to evaluate the ChatGPT output:

   (a) For the changes made to harmonize terminology (step 2.a), the domain expert was asked to confirm the correctness of the changes. In case the changes were not correct, the changes were reverted, and the processes restarted from step 2.b

   (b) For the information from ChatGPT that is not contained in the domain expert's result, the domain expert was asked to decide if this information was not accurate (h – hallucination), accurate and missing in the domain expert's result (a/a – accurate and additional) or accurate but out of scope (a/o – accurate + out of scope)

   (c) For the information from the domain expert's result but not in the ChatGPT, the domain expert was asked to decide if the missing information has to be considered as mandatory (m) or optional (o) information.

   (d) For the information contained in both, the domain expert's result and ChatGPT, the domain expert was asked to determine if the meaning and intention could be seen as identical (i), similar to a large extent (s) or significantly different (d).

4. The researchers participating in the experiment used the ChatGPT output and the domain expert's result and prepared a separate 4EM model for both. Here the result could be that there are no model mistakes, missing relations, missing concepts.

Step 3.b in the above process basically judges accuracy of ChatGPT output by identifying hallucinations, missing and additional information. Step 3.c judges completeness. Step 4 evaluates the suitability for modelling.

For E1, the task to be performed by the domain expert and ChatGPT is to describe the core operational processes of a HEI including the information required and produced by each process. For E2, the task is to define goals and organisational rules to implement them for improving the ranking of the HEI under consideration.

6.2. Experiment Results

This section presents the results of both experiments in two different subsections by describing the steps defined in the experiment design (see 6.1).

Experiment 1: University business processes

In step 1 of the experiment, the prompts for ChatGPT were developed:

- Prompt E1-1: Now I want to focus on Business as usual operations, BPM model, of XYX university. It essentially describes What are the business processes? How do they handle information? and material? Essentially, A business process is assumed to consume input in terms of information and/or material and produce output of information and/or material. BPM components are process, external process, information set, and material

set. Processes and external processes produce and consume information or material sets. Processes can be decomposed into sub-processes. It can have relationships with model elements from other models such as goal model and concept model, e.g., relationship Information and Material Sets of Course is "referring to" Student Concept. Produce BPM for XYZ university.

- Prompt E1-2: This is not a process model - this is a concept model. Focus on processes of a university, their sub-processes and steps involved those processes. Regenerate BPM of XYZ university

Prompt E1-2 was necessary as E1-1 resulted in a list of concepts (e.g., process, external process, role), attributes of processes and associations between concepts. An excerpt of the output of ChatGPT after prompt E1-2 is depicted in figure 1.

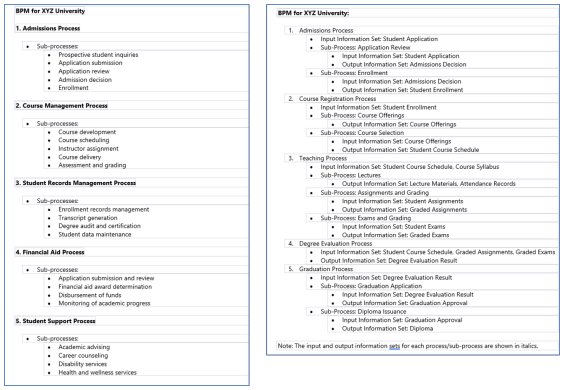

Fig. 1. Excerpt from ChatGPT output for prompt E1-2: left: university business processes; right: elements of 4EM business process model

In step 2, no differences in terminology had to be corrected. As a result of step 2, we produced a table comparing the output of ChatGPT with the result of the domain expert. The domain expert used this table and the output of ChatGPT to perform step 3, i.e., to compare the results and evaluate ChatGPT output according to step 3. The result is shown in table 3. The table shows that the domain expert identified more university business processes than ChatGPT (e.g., programme management or human research management); most of them were seen as mandatory for university operations. The processes identified by both, ChatGPT and domain expert, were all seen as similar or even identical.

Table 3. E1: Comparison of high-level university business processes of ChatGPT and domain expert

Domain Expert	ChatGPT	Domain Expert evaluation
Application		in ChatGPT, this is part of admission)
Admission	Admission	s- similar
Student Management	Student Records Management	s - similar
Financial Management	Financial Aid Management	s- similar
Programme Management		m - mandatory
Course Management	Course Management	i - identical
Student Career Service	Student Support	s- similar
Quality Management		o - optional as separate process
Grants Management	Research Management	s - similar
Human Resource Management		m - mandatory
Facility and Resource Management		o - optional as separate process

Table 4. E1: Comparison of the goal for improving the ranking of ChatGPT and domain expert

Domain Expert	**ChatGPT**	**Domain Expert evaluation**
To improve research impact and visibility within the next 3 years (sub-goals: To increase the number of high-ranked publications by 10% in 3 years; To increase the number of citations by 15% in 3 years; To increase the average h-index of the university's researchers)	Increase research output and quality; Enhance the reputation of faculty and staff	d - significantly different
To increase research funding from competitive resources by 10% within 3 years (sub-goals: To increase basic research funding from ERC; To increase direct funding from industry)	Increase research output and quality	d - significantly different
To improve student rating (sub-goals: To improve internationalization of course of study programmes; To improve average rating in student surveys and CHE rating)	Improve student satisfaction and retention	d - significantly different
	Strengthen partnerships with industry and other universities	a/a - accurate and additional

Experiment 2: Improve university ranking

Similar to experiment 1, we first developed the prompts for ChatGPT:

- Prompt E2-1: I am trying to capture vision and strategy of an organization using a goal model (GM) as follows: it has concepts of "Goal", "Opportunity", "Problem" (i.e., "Threat" and "Weakness"), "Cause", and "Constraint". Goals are refined by subgoals and they are connected with its parent goal using "AND" and "OR" relationships. All concepts may have binary relationships with other concepts of the type "supports" and "hinders". Produce vision and strategy of university XYZ that aims to improve its ranking using above goal model - please be less verbose and detailed outcome. I want all concepts should be labelled and those labels should be used while describing all relationships.
- Prompt E2-2: Elaborate goals and their relationships.
- Prompt E2-3: Yes. Now I want to produce policy and rules for XYZ University using Business Rules Model (i.e., BRM). Similar to goal model GM, BRM has concept of "Rule", "IS Component" or "Technical Component", and "Process". Rules may be related to each other with binary relationships and with symbolic relationships of types "AND", "OR", and "AND/OR". Rules may have inter-model relationships of types "rule hinders goal" where goal is from Goal model, "rule directs use of an IS Component", and "rule triggers process" (from process model). Produce BRM for XYZ university using less verbose term.

Similar to E1, there were no differences in terminology to be corrected (step 2) and the domain expert again used a table with ChatGPT output vs. result of the domain expert to compare and evaluate (step 3). The result is shown in table 4.

The domain expert interpreted the task much wider than ChatGPT and defined goals including sub-goals. Although some goals proposed by ChatGPT are not relevant for the university the domain expert had in mind when the goals were developed, these additional goals were judged by the domain expert as "potentially relevant for many other universities". The table also shows that some ChatGPT goals don't exactly match the expert's goals but can be related and would contribute to them.

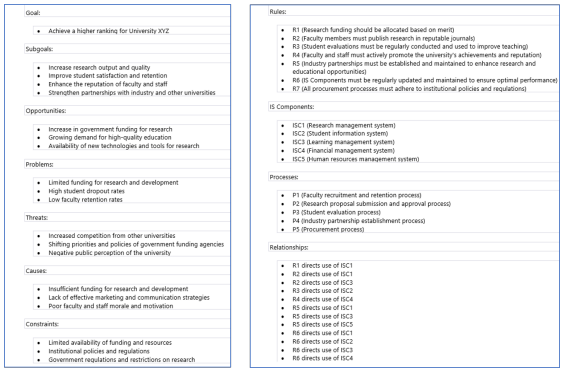

Fig. 2. Excerpt from ChatGPT output for prompt E2-1 and E2-3: left: university business processes; right: elements of 4EM business process model

Figure 2 shows an excerpt of the ChatGPT output. On the left is visible that ChatGPT also produced threats, weaknesses and opportunities, which are part of the 4EM meta-model. The expert did not include these aspects, but evaluated the ChatGPT output again as "potentially relevant for many universities". Furthermore, the business rules proposed by ChatGPT (on the right in Figure X) were considered by the expert as consistent to the goals and "inspiring". We also used the ChatGPT results of both experiments and developed enterprise models in 4EM language according to the output. As shown in the excerpt in figure Y, the output also included the 4EM concepts and relations. The model development showed that the concepts and relations were complete, i.e., the output described a valid 4EM model.

7. Discussion

Evaluation of the domain expert confirms for both experiments that the output of ChatGPT, in the large, is accurate and relevant. In E1 it confirms that it can be used to prepare modelling projects, in E2 that it contributes useful inspiration to developing change alternatives.

Although our work resulted in a number of findings, it also has many limitations that concern various aspects of the experiment and the process of enterprise modelling:

Task: we strongly believe that the utility of ChatCPT and the pertinence of the output provided depends on the modelling task. The conjecture is that modelling of general processes or general features of an application domain can expect more support from ChatGPT than modelling specific or even unique processes or structures of a certain enterprise - basically because there obviously is more information in the training corpus for ChatGPT for the general task. As a consequence, the tasks we defined for the experience affect the quality of the results and changing the task might change the results. This aspect needs further investigation in future work, for example, by investigating various tasks covering the continuum between "very general" and "very specific".

Domain expert: the domain expert has an important role in our experiment as both, the source of domain knowledge used as the "reference" to compare the ChatGPT output against and instance to judge accuracy of the ChatGPT output. Although different domain experts will have a joint view on the application domain there still might be differences when it comes to details. Thus, changing the domain expert might actually affect the results. We tried to address this issue in our experiment by involving a second expert to confirm the first expert's view.

Neural text generator: in our experiment, we used ChatGPT. Another neural text generator might have produced different output. Similarly, advances in ChatGPT may also lead to different output. Although we do not expect substantial differences between the generators, this still should be investigated in future work.

Prompt: the prompts used in our experiment were developed in an explorative rather than systematic way. It cannot be ruled out that there is a possibility to improve the prompts to achieve more relevant and complete output. Model-driven prompt design, together with accompanying toolsets that are designed for domain specific use in the enterprise context is likely be a fruitful future research area.

Perspective of enterprise model: our experiment included process modelling, goal modelling, concept modelling and business rule modelling. Generating output for and modelling of other perspectives (e.g., products or organization structures) has to be part of future work.

Result of modelling task: so far, we aimed at generating textual output that included the required information for developing the actual model. In future work, also generating the model in an appropriate visual modelling langauge should be investigated.

8. Summary

The work presented in this paper addressed the questions what tasks in enterprise modelling could be supported by ChatGPT, what prompts to use to collect the required domain knowledge and how accurate the information provided by ChatGPT is. To answer this question, we identified phases and tasks in enterprise modelling that require substantial contributions from domain experts. Two areas were selected for further investigation: the preparation phase of EM projects and the identification of change alternatives, including business rules to apply. For the application domain and tasks investigated in two quasi-experiments, the results show that ChatGPT can coexist with domain experts to improve productivity, completeness and precision. ChatGPT can help in the preparation phase collection of general information about the application domain and even common business processes and their information flow. But the results should neither be considered as complete nor covering the specifics of an enterprise. For the latter, domain experts are still needed. For the identification of change alternatives, ChatGPT proved in our task of improving university ranking as a source of inspiration for the domain expert, both for goals and for business rules. However, similar to the preparation phase, this did not include the actual situation in an enterprise. Furthermore, the 4EM output included all information required for a valid 4EM goal, business rule and business process model.

The discussion section already identified a number of areas for future work that basically result from current limitations. In summary, a broader investigation of the utility of ChatGPT for more and different application cases seems relevant and required to understand the potential and limits better. We consider the contribution of this paper as confirmation that research in this field is relevant and promising.

References

1. L. Ackermann. Language-centric approaches for improving business process model acceptance. volume 2196, pages 51–55, 2018. cited By 1.
2. Balbir Barn. Chatgpt could be your ally – really!, Jan 2023.
3. Dominik Bork, Syed Juned Ali, and Georgi Milenov Dinev. Ai-enhanced hybrid decision management. *Business & Information Systems Engineering*, pages 1–21, 2023.
4. Tom et al. Brown. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.
5. Janis Bubenko Jr, Anne Persson, and Janis Stirna. An intentional perspective on enterprise modeling. *Intentional perspectives on information systems engineering*, pages 215–237, 2010.
6. Thomas D Cook, Donald Thomas Campbell, and William Shadish. *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin Boston,

MA, 2002.

7. Luciano Floridi and Massimo Chiriatti. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694, 2020.
8. Ulrich Frank. Multi-perspective enterprise modeling: foundational concepts, prospects and future research challenges. *Software & Systems Modeling*, 13(3):941–962, 2014.
9. Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International Conference on Machine Learning*, pages 9118–9147. PMLR, 2022.
10. Barbara Kitchenham. Procedures for performing systematic reviews. *Keele, UK, Keele University*, 33(2004):1–26, 2004.
11. John Krogstie. *Quality of business process models*. Springer, 2016.
12. Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, et al. Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models. *PLoS digital health*, 2(2):e0000198, 2023.
13. Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023.
14. Arvind Nair, Xia Ning, and James H Hill. Using recommender systems to improve proactive modeling. *Software and Systems Modeling*, pages 1–23, 2021.
15. Anne Persson and Janis Stirna. An explorative study into the influence of business goals on the practical use of enterprise modelling methods and tools. *New Perspectives on Information Systems Development: Theory, Methods, and Practice*, pages 275–287, 2002.
16. Kurt Sandkuhl, Hans-Georg Fill, Stijn Hoppenbrouwers, John Krogstie, Florian Matthes, Andreas Opdahl, Gerhard Schwabe, Ömer Uludag, and Robert Winter. From expert discipline to common practice: a vision and research agenda for extending the reach of enterprise modeling. *Business & Information Systems Engineering*, 60:69–80, 2018.
17. Kurt Sandkuhl, Janis Stirna, Anne Persson, and Matthias Wißotzki. *Enterprise modeling*. Springer, 2014.
18. Nikolay Shilov, Walaa Othman, Michael Fellmann, and Kurt Sandkuhl. Machine learning for enterprise modeling assistance: an investigation of the potential and proof of concept. *Software and Systems Modeling*, pages 1–28, 2023.
19. Dag IK Sjøberg, Jo Erskine Hannay, Ove Hansen, Vigdis By Kampenes, Amela Karahasanovic, N-K Liborg, and Anette C Rekdal. A survey of controlled experiments in software engineering. *IEEE transactions on software engineering*, 31(9):733–753, 2005.
20. Janis Stirna and Anne Persson. Enterprise modeling. *Cham: Springer*, 2018.
21. Francois Vernadat. Enterprise modelling: Research review and outlook. *Computers in Industry*, 122:103265, 2020.
22. Francois B Vernadat. Enterprise modelling and integration: From fact modelling to enterprise interoperability. *Enterprise inter-and intra-organizational integration: Building international consensus*, pages 25–33, 2003.
23. Tongshuang Wu, Ellen Jiang, Aaron Donsbach, Jeff Gray, Alejandra Molina, Michael Terry, and Carrie J Cai. Promptchainer: Chaining large language model prompts through visual programming. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–10, 2022.