Conversational Process Modelling: State of the Art, Applications, and Implications in Practice

Nataliia Klievtsova[1], Janik-Vasily Benzin[2], Timotheus Kampik[3], Juergen Mangler[2], and Stefanie Rinderle-Ma[2(✉)]

[1] Austrian Center for Digital Production, Vienna, Austria
nataliia.klievtsova@acdp.at
[2] TUM School of Computation, Information and Technology, Technical University of Munich, Garching, Germany
{janik.benzin,juergen.mangler,stefanie.rinderle-ma}@tum.de
[3] SAP Signavio, Berlin, Germany
timotheus.kampik@sap.com

Abstract. Chatbots such as ChatGPT have caused tremendous hype lately. For BPM applications, it is often not clear how to apply chatbots to generate business value. Hence, this work aims at the systematic analysis of existing chatbots for their support of conversational process modelling as a process-oriented capability. Application scenarios are identified along the process life cycle. Then a systematic literature review on conversational process modelling is performed. The resulting taxonomy serves as input for the identification of application scenarios for conversational process modelling, including paraphrasing and improvement of process descriptions. The application scenarios are evaluated for existing chatbots based on a real-world test set from the higher education domain. It contains process descriptions as well as corresponding process models, together with an assessment of the model quality. Based on the literature and application scenario analyses, recommendations for the usage (practical implications) and further development (research directions) of conversational process modelling are derived.

Keywords: Conversational process modelling · Chatbots · Process Descriptions · Process Models

1 Introduction

AI-powered chatbots *"have a considerable impact in many domains directly related to the design, operation, and application of information systems"* and at the same time need to be handled with care [70], as models provide you with information without considering their own technology's limitations. Business process management as an information systems discipline seems a viable candidate to benefit from chatbots and hence from the recent advances in large

C. Di Francescomarino et al. (Eds.): BPM 2023 Forum, LNBIP 490, pp. 319–336, 2023.
https://doi.org/10.1007/978-3-031-41623-1_19

language models, in particular, when supporting users in creating and improving process-related content, most prominently process models and process descriptions. Process models enable participants to understand the processes in which they are involved [17] and to improve business performance [6]. However, errors in the process models may have adverse business consequences [24], and may lead to problems during process execution and quality issues [15].

Currently the creation of process models is often based on the interaction between domain experts having the knowledge of the process and process modellers/analysts capable of process modelling and analysis techniques. Hence, the acquisition of as-is models can consume up to 60% of the time spent on process management projects [29]. The overarching question of this work is thus how and to which degree chatbots can replace the process modeller/analyst when creating process models through **conversational modelling (CM)** with the domain expert.

CM means conversation flow modelling where the chatbot can receive and interpret inputs from the user (i.e., follow-up questions, unexpected inputs, or changes of topic) and provide appropriate responses that keep the conversation coherent [49].

This question can be broken down into the following research questions:
RQ1 How can CM methods/tools be employed for process modelling?
RQ2 Which CM methods/tools exist for process modelling?
RQ3 How can we evaluate CM methods/tools with respect to process modelling?
RQ4 Which implications do Chatbots have for BPM modelling practice/research?

RQ1 – RQ4 are tackled as follows: Based on the concept of conversational process modelling, initial application scenarios are posed based on the process life cycle (cf. Sect. 2). These initial application scenarios provide the keywords for the subsequent literature review (cf. Sect. 3) which aims at refining the scenarios along a taxonomy of existing approaches. For evaluating existing chatbots, a test set of process descriptions, process models, and quality assessment is collected and prepared (cf. Sect. 4.1). The systematic analysis of the chatbots (cf. Sect. 4.2) along with the refined application scenarios are conducted based on key performance indicators and provide the basis for deriving practical implications and research directions in conversational process modelling (cf. Sect. 5).

2 Conversational Process Modelling

Only few papers address conversational modelling, mostly by focusing on the design of virtual human agents (aka chatbots), e.g., [49,61]. However, there is no common understanding of conversational **process** modelling yet and we hence provide informal Concept 1 which takes up characteristics of conversational modelling regarding the participants in the conversation, i.e., the domain expert and the chatbot, and the iterative nature of the conversation.

Concept 1. (Conversational process modelling) *describes the process of creating and improving process models and process descriptions based on the iterative exchange of questions/answers between domain experts and chatbots.*

Concept 1 reflects the overarching goal of conversational process modelling, i.e., to enable process modelling and improvement based on interaction between the domain expert and the chatbot, instead of interaction between the domain expert and the process analyst/modeller. This goal constitutes the first pillar to analyze the BPM life cycle w.r.t the process modelling scenarios where conversational process modelling can be applied. The second pillar reflects the assumption that conversational process modelling is exclusively based on domain expert/chatbot interaction and does not employ any other tool. In the conclusion, we will sketch how conversational process modelling can be extended if the chatbot usage is augmented by other tools such as process simulation tools.

In the following, Concept 1 is fleshed out for application scenarios along the BPM life cycle as provided in [27]. The BPM life cycle is chosen as it provides a systematic structuring of the different process-oriented tasks and capabilities towards creating business value.

Process discovery subsumes a range of methods to create process models (and is not to be confused with process discovery as the process mining task is necessarily based on event logs). The typical input in a process discovery project consists of textual process descriptions gathered based on interviews or workshops. Based on the process descriptions, process models are created by process modellers/analysts. We identified the following steps as suitable for being supported by chatbots: (1) gathering the process descriptions for creating the process model. This also includes the preparation of the process descriptions, i.e., to increase the quality of the process description in terms of, for example, being precise, e.g. through automatic paraphrasing. (2) taking a process description as an input and producing a process model (accompanied by the process description). Here, the chatbot can be employed for analyzing the text and extracting process model relevant information such as activities and their relations as well as actors [12]. Finally (3) assessing a process model (with the accompanying process description), regarding model quality based on quality metrics such as cohesion [72] and guidelines such as number of elements or label style [8].

The **process analysis** phase builds the bridge between the as-is process model created in the process discovery phase and the to-be model created in the process redesign phase. It is concerned with the qualitative and quantitative assessment of a process models. A qualitative analysis comprises, for example, an assessment whether or not certain activities can be automated; this can then be analogously reflected by an action recommendation, e.g., if the automation potential is not fully exploited, yet. The chatbot can support this assessment based on the extracted activities in the process discovery phase. The results of the qualitative assessment can then be used in the process redesign phase for corresponding redesign actions. Quantitative process analysis comprises, for example, detecting bottlenecks based on process simulations. As mentioned before, for this work, we assume that the chatbot is used without invoking further tools

and systems such as a process simulator. Hence, quantitative process analysis does not include tasks for conversational process modelling at this stage, but for future work as discussed in Sect. 4.3.

Process redesign comprises the definition of the redesign goal which again is considered a managerial task. The chatbot can support the domain expert by proposing existing redesign methods such as Lean Six Sigma, as well as in querying models (cf. [56]) or applying the redesign instructions. Especially important is refactoring of process descriptions, based on existing guidelines on process model refactoring or catalogues of process smells such as [73].

The phases of **process implementation** and **process monitoring** are considered as a part of future work of conversational process modelling as they will require the invocation of additional tools and systems such as a process engine or process-aware information system.

Table 1 summarizes the initial application scenarios for conversational process modelling along the process life cycle phases and steps which constitute the input for the subsequent literature and test set based analyses.

Table 1. Application Scenarios and Chatbot Tasks along Process Life Cycle

# application	input	output	chatbot task
1. gather information	process description	process description	paraphrase
2. process modelling	process description	process model, process description	extract
3. assure model quality	process model, process description, process modelling guidelines and metrics	quality issues, refined process model, refined process description	compare and assess
4. select redesign method	collection of process models and process descriptions	redesign method, selection of process models and process descriptions	select method, query models
5. apply redesign method	collection of process models and process descriptions, redesign method	collection of process models and process descriptions	query and refactor models

The BPMN model depicted in Fig. 1 assembles and refines the application scenarios, together with their input, outputs, and related chatbot tasks as summarized in Table 1 into a generic process model for conversational process modelling, reflecting its interactive and iterative characteristics: at first, the domain expert provides a process description which is refined (→ paraphrase) and the results are displayed (→ extract). Then an assessment of the result quality is conducted (→ compare and assess). If the quality is insufficient, the process models/descriptions are refined (→ query, refactor), possibly based on a specific method (→ select method), until the quality reaches a sufficient level.

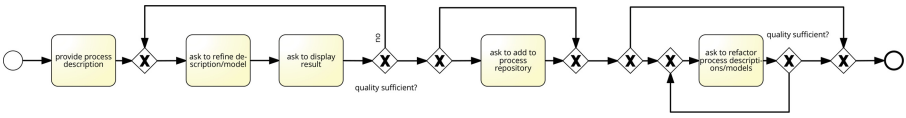


Fig. 1. The Process of Conversational Process Modeling (modeled in BPMN using SAP Signavio)

3 State of the Art

The literature analysis consists of two steps, i.e., i) a pre-review based on the initial application scenarios and life cycle phases summarized in Table 1 and based on the outcome of the pre-review, ii) a more generalized review including, for example, NLP-based methods for the extraction of model information from process descriptions. i) and ii) follow the guiding principles of [37].

i) Pre-review: The pre-review is conducted based on the keywords resulting from building the cross product of the application scenarios and keyword "chatbot" summarized in Table 1, e.g., `‘‘process modelling’’ chatbot`. These keywords are then used in the title search (allintitle) on google.scholar.com[1]. Next, we use the keywords resulting from the cross product of application scenario and chatbot task, e.g., `‘‘process modelling’’ paraphrase` and the keywords resulting from the cross product of keyword "conversational" and the application scenarios (allintitle), e.g., `conversational ‘‘process modelling’’`. In order to broaden the pre-review, we repeated the search for application and chatbot, but without keyword "process". Most of these searches result in 0 or a couple of hits, which were rejected due to quality issues or domain irrelevance.

The pre-review did not yield deeper insights into techniques, opportunities, and limitations of conversational process modelling. The results rather point towards generalizing the keywords used for the search, particularly covering NLP-based methods. Hence, for the **ii) second search**, we used https://scholar.google.com to produce Table 2. It shows the list of 52 papers relevant for a wide variety of relevant topics. Selection of the papers for the list was done based on the existence of the enumerated keywords (Selection Criteria) in the abstract or the title (for the first 20 hits).

In the following, we will discuss the literature collected in Table 2 regarding five fundamental questions that partly correspond to the research questions and partly to the pointers derived from the pre-review.

How do chatbots work, and what are important areas of application? A chatbot is a type of human-computer interaction, used to simulate conversations to solve particular user problems [3]. Chatbots work by processing language input from humans (furthermore referred to as natural language processing (NLP) [21, 50]), and reacting to it. The interpretation of human input is achieved through a set of rules [20, 26, 40], or by utilizing large language models (LLMs) [42], which

[1] last accessed 2023-03-23 and 2023-03-26 respectively.

are trained to understand the meaning/intent/context [18,44] and generate new content based on different statistical and probabilistic techniques. According to [51] the main areas of chatbot application are human resources, e-commerce, learning management systems, customer service, and sales.

How are responses generated? After receiving user input, the chatbot processes it into a machine-readable form and based on that input generates a natural language output utilizing different types of response generation methods [77]. Chatbot systems can be divided into six categories, based on the type of response generator [44]. (1) template-based: response is selected from the list of predefined pairs of query patterns; (2) corpus-based: converts user query to a structured query language (SQL) query and passes it to utilized techniques of professional knowledge management (i.e., database, ontology); (3) intent-based: task-oriented system, which based on user query tries to recognise user intent with the help of advanced NLU techniques; (4) RNN-based: RNN-based (Recurrent Neural Network) chatbot generates response query directly from the user query with the help of the model, trained on dialogue data set; (5) RL-based: RL-based (Reinforcement Learning) chatbots use rewarding and punishing functions to achieve the desired behaviour; (6) hybrid-based: a combination of approaches listed above to achieve better performance or to overcome limitations, faced by using one approach only.

How can response generation be implemented? All of the above types utilize some type of knowledge graph to formalize the configuration [7,76] and the intended output format of the conversation [4,55]. The knowledge graph is either accessed by simple querying languages such as AIML or SPARQL, or it is encoded as part of a neural network through training. So responses are either queried explicitly or generated implicitly as part of a neural network. Both approaches have different strengths and weaknesses. For conversation-related applications such as entertainment, neural networks work well, but for other applications with special output, other approaches are still valid solutions. Low-code solutions to control explicit responses [25] as well as BPMN-based solutions to encode potential progressions of a conversation [60] have been proposed. One example of such a system is PACA [41]. Automatically learning from user interactions can be achieved not only for neural networks (e.g., reinforcement learning) but also by encoding interactions automatically into rules, such as in [5,36].

Can chatbots deal with business processes? According to the survey of chatbot integration [9], 2 out of 347 chatbot systems support the business process interface pattern, i.e., [34,43] that convert BPMN process models into dialog models/chatbots. Currently, there are no chatbots that are able to generate BPMN models themselves. However, interest in the generation of models from various types of document sources has recently increased [29,31,64]. Referring to [32] as an input for business process model generation use case diagrams, business rules, standard operating procedures, and plain unstructured text are considered. Based on the approaches mentioned above, the following 3 steps for creating BPMN can be summarized [12,66]: (1) Sentence Level Analysis:

Table 2. Literature Queries, Hits, and Selections

Query (allintitle:)	Hits	Selection Criteria	#	List
chatbot technology overview	1		1	[3]
Natural language processing	10400	automated NLP	2	[21,50]
nlp Chatbot Development	7	deep learning	1	[59]
chatbots business processes	2	capability to learn	1	[36]
Chatbot integration	32	chatbot integration	1	[9]
quark chatbot	1		1	[34]
((Chatbots) OR (chatbot)) Process Models	2	process model	1	[43]
reasoning processes descriptions	3		1	[67]
"process model generation"	15	text	1	[29]
generating BPMN diagram	2	text	1	[64]
business process (model) OR (models) generating	34	Natural Language,document sources	2	[31,32]
extracting business process language models	2	NLP, language model	2	[12,66]
AI based language models	2	NLP, LMs	1	[42]
large language models	628	NLP, BPMN	3	[52,75] , [38]
BOMN generation	22	NLP, LMs	1	[48]
"process extraction" from text	6	text, textual information	1	[10,11,13]
"knowledge graphs" chatbots	5	NLP, LMs	1	[4,7,54,55,76]
chatbots BPMN modelling	0	—	—	—
chatbots graph generation	0	—	—	—
((model based) OR (model-based))	12	NLP, BPMN, UML	1	[28]
generate graphs chatbots	0	—	—	—
generate graphs plain text	0	—	—	—
BPMN modelling chatbots	0	—	—	—
low-code chatbot development	1		1	[25]
generating texts models	2	process model	1	[39]
declarative process model generation	0	—	—	—
process models chatbot	1	—	1	[5]
process conversational agents	7	BPMN	2	[41,60]
rule based chatbots	5	natural language, AIML	3	[20,26,40]
chatbot designs	4	natural language	2	[18,44]
Process Models Chatbots	1	—	1	[43]
mining models from text	11	process model	1	[45]
automatic generation bpmn	5	from BPMN, process model	3	[14,23,62]
text information extraction	539	unstructured text, semi-structured text	7	[22,33,53,57,58,68,69]
text data augmentation methods	8	methodology	1	[79]
data augmentation approaches nlp	1		1	[2]
easy data augmentation techniques	4	data augmentation	3	[30,63,74]
automatic machine translation paraphrasing	3	paraphrasing	2	[71,78]
paraphrasing automatic evaluation	7	bleu, english	2	[16,35,78]

extraction of basic BPMN artefacts such as tasks, events, and actors; (2) Text Level Analysis: exploration of relationships between basic items, e.g., gateways. (3) Process Model Generation: create a syntactically correct model, that captures the semantics of the input. [67] proposes a machine-readable intermediate format generated out of natural language (either through automatic or manual annotation). The result is then easy to interpret by computers.

How can we evaluate chatbots with respect to BPM modelling? Currently there are no gold standard data sets that can be used to evaluate and compare the efficiency of process extraction from unstructured text [10]. In [29] a set of 47 text-model pairs from industry and textbooks are introduced, which could be converted with an accuracy of 77% (up to 96% of similarity for some cases) from text to model. In [39], 53 model-text pairs were used to evaluate performance of a novel model-to-text transformation method. To avoid the necessity of constant creation of new datasets by hand, data augmentation techniques (increase of the training set size with the help of the modified copies of already existing data set items) can be used [2,79]. Another important tool is paraphrasing [35], which is about generating similar texts from a source. Such texts are generally recognized as lexically and syntactically different while remaining semantically equal.

4 Performance of Current Generation LLMs for Conversational Process Modelling

In order to assess the performance of conversational process modelling tools and answer RQ3, it is necessary to come up with a data set, an evaluation method, and a set of KPIs. Extending the three steps, which are required to create a BPMN model (see Sect. 3), a fully integrated conversational process modelling toolchain would contain: (a) extraction of tasks from textual descriptions, (b) extraction of logic such as decisions or parallel branchings from textual descriptions, (c) creation and the layout of a BPMN model, and (d) the application of modifications for refinement of BPMN models. As a fully integrated conversational process modelling tool does not exist yet, in this paper we concentrate on how well current LLMs, namely GPT models text-davinci-001 (GPT1), text-davinci-002 (GPT2), text-davinci-003 (GPT3) from openai.org playground[2], as well gpt 3.5 turbo (GPT3.5) from writesonic.com[3], perform for extracting tasks for textual description (see (a) above). Task extraction is a starting point of the conversational process modelling toolchain, as the task is an atomic element of the process flow, which represents a unit of work that should be performed [65].

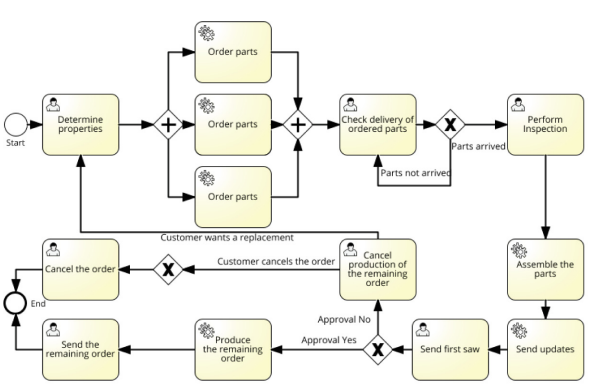
4.1 Test Set Generation

The test set [46] utilized in this paper contains 21 textual process descriptions from 6 topics or domains. For each process description between 8 and 11 BPMN

[2] last access: 2023-03-29.
[3] last access: 2023-03-29.

process models have been created by modelling novices. These models represent different possible ways of interpreting the textual process description. Each model has at least one start and end event, 3 exclusive gateways, one parallel gateway, and an average of 14 tasks. Some models also contain sub-processes, pools, and lanes. Each model was evaluated by a modelling expert using a quality value from 0 to 5, to reflect, on how well the textual description has been transformed into a BPMN model, i.e., all tasks and decisions from the textual description are in the BPMN, tasks which can run in parallel have been correctly identified, and the BPMN model is well-formed.

An example of a textual description and an associated interpretation, i.e., the BPMN model, can be seen in Fig. 2.



Category: Manufacturing
Title: Chainsaw

You produce custom chainsaws on demand. Your chainsaws have at least 5 properties such as length of the "guide bar", chain width, electric or motor chainsaw. After your customer told you the properties, you can start ordering the parts from various online sources (in parallel). After the parts arrive you have to do a manual inspection of all parts, and then assemble the parts. During production, you regularly send updates to your customer. After producing the first saw you send it to your customer. If he likes it, the rest of his order are produced.

Fig. 2. Textual Description And BPMN Model From the Evaluation Dataset

4.2 Evaluation

In this section, we will use the following KPIs and discuss their impact on conversational process modelling approaches: **KPI1** - Text Similarity; **KPI2** - Set Similarity; **KPI3** - Set Overlap; **KPI4** - Restricted Text Similarity; **KPI5** - Restricted Set Similarity; **KPI6** - Restricted Set Overlap; **KPI7** - Average Augmented Task Extraction Prevalence and Similarity (GPT3 only). All results, including non-averaged data, are also available in [47].

Prompt engineering and KPIs: KPIs 1–3 are used to assess task extraction from original process descriptions. This is realized by passing the following prompt "Considering following $< process\_description >$ return the list of main tasks in it" to the LLMs. For assessment based on KPIs 4–6, the original prompt is changed to the "Considering following $< process\_description >$ return the list of main tasks (each 3–5 words) in it" to improve the granularity of extracted tasks and to refine the quality of obtained tasks' labeling. KPI7 is used to evaluate how stable task extraction is, performed by LLMs by extending the set of original process descriptions by utilizing different paraphrasing algorithms.

Task extraction from associated models is realised by parsing .XML documents and extracting relevant BPMN activities, keeping their sequence in the process flow.

As the basis for each similarity measurement we utilize contextual (BERT) and non-contextual (TD-IDF) vectorisers with a cosine similarity metric [19]. The contextual and non-contextual approaches will be denoted as C and NC.

For **KPI1**, each LLM (GPT1, GPT2, GPT3, GPT3.5) is instructed to extract the tasks from the original process descriptions. The answer is then compared to the original text to assess the completeness of the extraction. The results are depicted in Table 3. For this KPI, GPT3.5 is the most successful LLM.

Table 3. Text Similarity (**KPI1**): Comparison of tasks extracted by LLM and original text using contextual (BERT) and non-contextual (TD-IDF) vectorisers

Method	gpt1	gpt2	gpt3	gpt3.5
non-contextual	0.46	0.65	0.60	0.63
contextual	0.76	0.80	0.78	0.84

Table 4 shows the results for **KPI2**. The four LLMs are instructed to extract tasks from each textual description. This set of tasks is then compared to the set of tasks, extracted from each BPMN model mentioned above (see Sect. 4.1). As for every textual description multiple BPMN models exist, the results are averaged per textual description. The averages are then again averaged for all textual descriptions. GPT3 is successful for this KPI with 74% extraction rate.

Table 4. Set Similarity (**KPI2**): Comparison of tasks extracted by LLM with tasks extracted from BPMN Models. For each text a set of n tasks is extracted. Each text has 8–11 associated models from which again a set m of tasks can be extracted. Each set n is compared with all sets m, yielding a set of similarities which is averaged for similarity methods contextual (C) and non-contextual (NC)

LLM	C	NC	avg. # of tasks extracted from texts	avg. # of tasks extracted from models
gpt 1	0.72	0.32	7.6	12
gpt 2	0.71	0.32	6.7	12
gpt 3	0.74	0.35	7.7	12
gpt 3.5	0.73	0.36	8.5	12

For **KPI3**, the goal is to quantify the overlap between extracted tasks from the original text and associated to its models: (1) how similar are individual tasks, and (2) how many tasks exist only in one of the two extractions. The results are shown in Table 5 and show that between 6 and 7 tasks extracted from the models are also found in the text, while about 6 tasks could not be found in the extracted text. When looking at it from the point of view of the

Table 5. Set Overlap (**KPI3**): Each task extracted from the text is compared (for each associated model) with task extracted from the model. If the similarity is bigger than a threshold, a task is deemed common, else it is deemed to only occur in either the model or the text.

LLM	similarity	common model	common chat	only in model	only in chat
gpt 1	C	6.6	4.5	5.2	3.2
gpt 1	NC	5.9	4	5.9	3.6
gpt 2	C	6.2	4.1	5.6	2.6
gpt 2	NC	5.6	3.6	6.2	3
gpt 3	C	6.7	4.6	5.1	3
gpt 3	NC	6.7	4.6	5.1	3
gpt 3.5	C	7	4.7	4.9	3.8
gpt 3.5	NC	6.5	4.4	5.4	4.1

tasks extracted from the text, the ratio becomes 4:3. So almost 50% of the tasks are not similar between the model and text (see discussion for details).

KPI4 focuses on restricting the number of words per extracted task, to coax the bot into extracting more tasks, as generally, the number of extracted tasks from the text is lower than the number of tasks contained in the models (see discussion for more details). Table 6 shows that this decreases the similarity when comparing text (due to stronger paraphrasing), but **KPI5** (cf. Table 7) and **KPI6** (cf. Table 8) show an increase in the number of tasks by one while not decreasing similarity when compared to the tasks from the model.

Table 6. Restricted Text Similarity (**KPI4**): Task names are allowed to only have 3–5 words, cmp. Table 3.

method	gpt1	gpt2	gpt3	gpt3.5
non-contextual	0.24	0.47	0.38	0.27
contextual	0.70	0.77	0.73	0.73

Table 7. Restricted Set Similarity (**KPI5**): Task names are allowed to only have 3–5 words, cmp. Table 4.

LLM	C	NC	avg. # of tasks extracted from texts	avg. # of tasks extracted from models
gpt 1	0.73	0.32	7.6	12
gpt 2	0.74	0.33	7.7	12
gpt 3	0.73	0.32	8.3	12
gpt 3.5	0.75	0.30	8.5	12

Finally, for **KPI7**, we assessed the effects of paraphrasing on prevalence and similarity, i.e., how stable LLMs are for task extraction with similar input. We use nine different algorithms for paraphrasing text [2] (rewriting sentences using synonyms), which is, for example, useful to clean up textual descriptions from humans. The results are displayed in Table 9, and show that especially the contextual similarity does not decrease significantly, while the number of extracted tasks even improves in comparison to the original text.

Table 8. Restricted Set Overlap (**KPI6**): Task names are allowed to only have 3-5 words, cmp. Table 5.

LLM	similarity	common model	common chat	only in model	only in chat
gpt 1	NC	6	4	5.7	3.5
gpt 2	NC	6.4	4.2	5.4	3.5
gpt 3	NC	7	4.7	4.8	3.5
gpt 3.5	NC	6.8	4.6	5	3.8

Table 9. Average Augmented Task Extraction Prevalence and Similarity GPT3 (**KPI7**): for nine different paraphrasing methods, the average number of tasks, and similarity measures are calculated. The second row holds the value of the original text from Table 6

	Original	SR	DL	SW	IN	NLPaug	TDE	TRU	TES	EMB
avg. # tasks	8.25	8.10	8.43	7.48	8.19	8.10	7.57	7.86	8.62	8.29
C similarity	0.73	0.69	0.69	0.68	0.70	0.70	0.70	0.67	0.70	0.70
NC similarity	0.38	0.20	0.22	0.25	0.21	0.21	0.21	0.19	0.21	0.22

4.3 Discussion

Tables 3, 4, 5, 6, 7, 8 and 9 clearly show that GTP3 currently supports task extraction the best, beating GPT3.5. The potential reason for GPT3 success could be the size of the model (175 billion parameters over 1,3 billion for GPT3.5). GPT3.5 model is optimized for a chat and may not be as effective for more complex language tasks [1].

Another important insight is that manually designed and refined models contain additional tasks that cannot be directly extracted from the original text but exist due to a humans ability to "read between the lines" or reason about task granularity. GPT extracts tasks exactly as written in text but does not have the capability to reason when it makes more sense to have multiple small tasks instead of a big one. We tried to coax GPT3 into extracting more tasks by restricting the number of words describing a task (i.e., its label), which increased the average number of extracted tasks slightly by 1, as can be seen in Table 7.

On average, GPT extracted a third less tasks than existed in the model. When strictly looking at the capability of extracting tasks from the original text, GPT3, on average, achieves a text similarity of 80%. The interpretation of this value is difficult. It could mean that the LLM missed about 20% of the tasks or, alternatively, that 20% of the text are just the filler words that have been ignored by the LLM. Together with the observation that the LLM does not like to split up tasks, the 30% less tasks extracted from the text in comparison to the model, hint at a possible explanation.

5 Conclusion: Practical Implications and Research Directions

From the state-of-the-art discussion in Sect. 3 and the results of the evaluation presented in Sect. 4.3, the following two main managerial implications can be derived:

1. For the chatbot application scenarios "gather information" and "process modelling" (cf. Table 1), chatbots are in principle ready to be applied in practice as-is, yet the results have to be taken with a grain of salt, i.e., the domain expert should always check the results. However, the lack of an appropriate, human-readable output format, e.g., a BPMN process model, limits the space of early adopters in a company significantly to experts at the intersection of their domain and computer science. This limitation is particularly unfortunate, as it counteracts the goal of conversational process modelling to minimize the necessary technical skills of the domain expert.
2. For the chatbot application scenarios "compare and assess", "select method, query models", and "query and refactor models", off-the-shelf chatbots are not yet ready to be applied due to their inability to output process models and to understand process model semantics.

As business process modelling has become an important tool for managing organizational change and for capturing requirements of software, the first managerial implication is that conversational process modelling can already have a significant business impact. Considering that the central problem in this area – the acquisition of as-is models – consumes up to 60% of the time spent on process management projects [29], chatbot-based partial automation can be sufficiently impactful, even if substantial human refinement is required.

The second managerial implication is that future research should focus on integrating the strong language capabilities of chatbots with the specialized capabilities of existing knowledge-based tools. The integrative research direction is more promising than chatbot training with specialized process modeling training sets featuring native process models, e.g., process models in BPMN format and a number of semantic targets, such as information on the existence of deadlocks in a process model. First, training of the chatbot with respect to business process models ignores the vast existing modeling knowledge encoded into existing tools. Second, semantics are clearly defined and encoded in existing tools such that training chatbots with the aim of understanding formal semantics is futile unless it serves as an intermediate step that unlocks further value.

To conclude, while advanced tasks such as model querying, refinement, and analysis presumably require domain-specific solutions, the interplay of traditional, knowledge based approaches to business process modeling can relatively straight-forwardly be augmented by machine learning-based chatbots to facilitate tedious tasks such as information gathering and basic model creation.

References

1. Openai documentation: models overview. https://platform.openai.com/docs/models/
2. Data augmentation approaches in natural language processing: a survey. AI Open **3**, 71–90 (2022). https://doi.org/10.1016/j.aiopen.2022.03.001
3. Adamopoulou, E., Moussiades, L.: An overview of chatbot technology. In: Artificial Intelligence Applications and Innovations, pp. 373–383 (2020)
4. Ait-Mlouk, A., Jiang, L.: KBot: a knowledge graph based chatbot for natural language understanding over linked data. IEEE Access **8**, 149220–149230 (2020). https://doi.org/10.1109/ACCESS.2020.3016142
5. Alman, A., Balder, K.J., Maggi, F.M., van der Aa, H.: Declo: a chatbot for user-friendly specification of declarative process models. In: Best Dissertation Award, Doctoral Consortium, and Demonstration & Resources Track at BPM 2020, vol. 2673, pp. 122–126. CEUR-WS.org (2020). https://ceur-ws.org/Vol-2673/paperDR12.pdf
6. Alotaibi, Y.: Business process modelling challenges and solutions: a literature review. J. Intell. Manuf. **27**(4), 701–723 (August 2016). https://doi.org/10.1007/s10845-014-0917-4,https://ideas.repec.org/a/spr/joinma/v27y2016i4d10.1007_s10845-014-0917-4.html
7. Avila, C.V.S., Franco, W., Maia, J.G.R., Vidal, V.M.P.: CONQUEST: a framework for building template-based IQA chatbots for enterprise knowledge graphs. In: Métais, E., Meziane, F., Horacek, H., Cimiano, P. (eds.) NLDB 2020. LNCS, vol. 12089, pp. 60–72. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-51310-8_6
8. Avila, D.T., dos Santos, R.I., Mendling, J., Thom, L.H.: A systematic literature review of process modeling guidelines and their empirical support. Bus. Process. Manag. J. **27**(1), 1–23 (2021). https://doi.org/10.1108/BPMJ-10-2019-0407
9. Baez, M., Daniel, F., Casati, F., Benatallah, B.: Chatbot integration in few patterns. IEEE Internet Comput. **25**(03), 52–59 (2021). https://doi.org/10.1109/MIC.2020.3024605
10. Bellan, P., Dragoni, M., Ghidini, C.: A qualitative analysis of the state of the art in process extraction from text. In: DP@AI*IA (2020)
11. Bellan, P., Dragoni, M., Ghidini, C.: Process extraction from text: state of the art and challenges for the future. arXiv preprint arXiv:2110.03754 (2021)
12. Bellan, P., Dragoni, M., Ghidini, C.: Extracting business process entities and relations from text using pre-trained language models and in-context learning, pp. 182–199 (09 2022). https://doi.org/10.1007/978-3-031-17604-3_11
13. Bellan, P., Ghidini, C., Dragoni, M., Ponzetto, S.P., van der Aa, H.: Process extraction from natural language text: the pet dataset and annotation guidelines. In: Workshop on Natural Language for Artificial Intelligence (2022)
14. Belo, O., Gomes, C., Oliveira, B., Marques, R., Santos, V.: Automatic generation of ETL physical systems from BPMN conceptual models. In: Model and Data Engineering, pp. 239–247 (2015)
15. de Brito Dias, C.L., Dani, V.S., Mendling, J., Thom, L.H.: Anti-patterns for process modeling problems: an analysis of BPMN 2.0-based tools behavior. In: Business Process Management Workshops (2019)
16. Callison-Burch, C., Cohn, T., Lapata, M.: Parametric: an automatic evaluation metric for paraphrasing. In: Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pp. 97–104 (2008)

17. de Camargo, J.V., et al.: A complementary analysis of the behavior of BPMN tools regarding process modeling problems. In: Augusto, A., Gill, A., Bork, D., Nurcan, S., Reinhartz-Berger, I., Schmidt, R. (eds.) Enterprise, Business-Process and Information Systems Modeling, pp. 43–59. Springer International Publishing, Cham (2022). https://doi.org/10.1007/978-3-031-07475-2_4
18. Cañizares, P.C., Pérez-Soler, S., Guerra, E., de Lara, J.: Automating the measurement of heterogeneous chatbot designs. Appl. Comput., 1491–1498 (2022)
19. Chandrasekaran, D., Mago, V.: Evolution of semantic similarity-a survey. ACM Comput. Surv. **54**(2), 41:1–41:37 (2021). https://doi.org/10.1145/3440755
20. Choa, N., Limb, Y., Limc, J.: Research design to compare the impacts of two different types of chatbots on mobile shopping behavior: rule-based and natural language processing-based. Editorial Board, p. 43
21. Chowdhary, K.: Fundamentals of Artificial Intelligence. Springer India (2020). https://doi.org/10.1007/978-81-322-3972-7, https://books.google.at/books?id=8SfbDwAAQBAJ
22. Ciravegna, D., et al.: Adaptive information extraction from text by rule induction and generalisation (2001)
23. Cossentino, M., Lopes, S., Sabatucci, L.: A tool for the automatic generation of MOISE organisations from BPMN. In: WOA, vol. 1613, p. 69 (2020)
24. Dani, V.S., Freitas, C.M.D.S., Thom, L.H.: Recommendations for visual feedback about problems within BPMN process models. Softw. Syst. Model. **21**(5), 2039–2065 (2022). https://doi.org/10.1007/s10270-021-00972-0
25. Daniel, G., Cabot, J., Deruelle, L., Derras, M.: Xatkit: a multimodal low-code chatbot development framework. IEEE Access **8**, 15332–15346 (2020). https://doi.org/10.1109/ACCESS.2020.2966919
26. Dihyat, M.M.H., Hough, J.: Can rule-based chatbots outperform neural models without pre-training in small data situations: a preliminary comparison of AIML and Seq2Seq. In: Workshop Semantics Pragmatics Dialogue, pp. 1–3 (2021)
27. Dumas, M., La Rosa, M., Mendling, J., Reijers, H.A.: Fundamentals of Business Process Management. Springer, Berlin, Heidelberg (2013). https://doi.org/10.1007/978-3-642-33143-5
28. Ed-douibi, H., Cánovas Izquierdo, J.L., Daniel, G., Cabot, J.: A model-based chatbot generation approach to converse with open data sources. In: Web Engineering, pp. 440–455 (2021)
29. Friedrich, F., Mendling, J., Puhlmann, F.: Process model generation from natural language text. In: Advanced Information Systems Engineering. pp. 482–496 (2011)
30. Fu, K., Lin, J., Ke, D., Xie, Y., Zhang, J., Lin, B.: A full text-dependent end to end mispronunciation detection and diagnosis with easy data augmentation techniques. arXiv preprint arXiv:2104.08428 (2021)
31. Honkisz, K., Kluza, K., Wiśniewski, P.: A concept for generating business process models from natural language description. In: Liu, W., Giunchiglia, F., Yang, B. (eds.) KSEM 2018. LNCS (LNAI), vol. 11061, pp. 91–103. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-99365-2_8
32. Indahyanti, U., Djunaidy, A., Siahaan, D.: Auto-generating business process model from heterogeneous documents: a comprehensive literature survey. In: Electrical Engineering, Computer Science and Informatics, pp. 239–243 (2022). https://doi.org/10.23919/EECSI56542.2022.9946460
33. Jiang, J.: Information extraction from text. Mining Text Data, pp. 11–41 (2012)
34. Kalia, A.K., Telang, P.R., Xiao, J., Vukovic, M.: Quark: a methodology to transform people-driven processes to chatbot services. In: Service-Oriented Computing, pp. 53–61 (2017)

35. Kauchak, D., Barzilay, R.: Paraphrasing for automatic evaluation. In: Human Language Technology Conference of the NAACL, pp. 455–462 (2006)
36. Kecht, C., Egger, A., Kratsch, W., Röglinger, M.: Quantifying chatbots' ability to learn business processes. Inf. Syst. **113**, 102176 (2023). https://doi.org/10.1016/j.is.2023.102176
37. Kitchenham, B.: Procedures for Performing Systematic Reviews. Keele University Technical Report TR/SE-0401, Keele University (2004)
38. Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y.: Large language models are zero-shot reasoners. arXiv preprint arXiv:2205.11916 (2022)
39. Leopold, H., Mendling, J., Polyvyanyy, A.: Generating natural language texts from business process models. In: Advanced Information Systems Engineering, pp. 64–79 (2012)
40. Lim, Y., Lim, J., Cho, N.: An experimental comparison of the usability of rule-based and natural language processing-based chatbots. Asia Pacific J. Inf. Syst. **30**(4), 832–846 (2020)
41. Lins, L.F., Melo, G., Oliveira, T., Alencar, P., Cowan, D.: PACAs: process-aware conversational agents. In: Business Process Management Workshops, pp. 312–318 (2022)
42. Liu, Z., Roberts, R.A., Lal-Nag, M., Chen, X., Huang, R., Tong, W.: AI-based language models powering drug discovery and development. Drug Discovery Today **26**(11), 2593–2607 (2021)
43. López, A., Sànchez-Ferreres, J., Carmona, J., Padró, L.: From process models to chatbots. In: Giorgini, P., Weber, B. (eds.) CAiSE 2019. LNCS, vol. 11483, pp. 383–398. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-21290-2_24
44. Luo, B., Lau, R.Y.K., Li, C., Si, Y.W.: A critical review of state-of-the-art chatbot designs and applications. WIREs Data Min. Knowl. Discov. **12**(1), e1434 (2022). https://doi.org/10.1002/widm.1434
45. M. Riefer, S. Ternis, T.T.: Mining process models from natural language text: a state-of-the-art analysis. In: Wirtschaftsinformatik, pp. 9–11 (2016)
46. Mangler, J., Klievtsova, N.: Dataset: textual process descriptions and corresponding BPMN models (2023). https://doi.org/10.5281/zenodo.7783492
47. Mangler, J., Klievtsova, N.: Evaluation details: GPT capabilites for extracting tasks from textual process descriptions (2023). https://doi.org/10.5281/zenodo.8063211
48. Maqbool, B., et al.: A comprehensive investigation of BPMN models generation from textual requirements-techniques, tools and trends. In: Information Science and Applications, pp. 543–557 (2019)
49. McTear, M.F.: Conversational modelling for chatbots: current approaches and future directions (2018)
50. Meurers, D.: Natural language processing and language learning. Encyclopedia Appl. Linguist, 4193–4205 (2012)
51. Miklosik, A., Evans, N., Qureshi, A.M.A.: The use of chatbots in digital business transformation: A systematic literature review. IEEE Access **9**, 106530–106539 (2021). https://doi.org/10.1109/ACCESS.2021.3100885
52. Min, B., et al.: Recent advances in natural language processing via large pre-trained language models: a survey. arXiv preprint arXiv:2111.01243 (2021)
53. Mooney, R.J., Bunescu, R.: Mining knowledge from text using information extraction. ACM SIGKDD Explor. Newsl. **7**(1), 3–10 (2005)
54. Omar, R., Mangukiya, O., Kalnis, P., Mansour, E.: ChatGPT versus traditional question answering for knowledge graphs: current status and future directions towards knowledge graph chatbots. arXiv preprint arXiv:2302.06466 (2023)

55. Patsoulis, G., Promikyridis, R., Tambouris, E.: Integration of chatbots with Knowledge Graphs in eGovernment: the case of getting a passport. In: 25th Pan-Hellenic Conference on Informatics, pp. 425–429 (2021)
56. Polyvyanyy, A. (ed.): Process Querying Methods. Springer (2022). https://doi.org/10.1007/978-3-030-92875-9
57. Rahman, S., Kandogan, E.: Characterizing practices, limitations, and opportunities related to text information extraction workflows: a human-in-the-loop perspective. In: Human Factors in Computing Systems, pp. 1–15 (2022)
58. Rau, L.F., Jacobs, P.S., Zernik, U.: Information extraction and text summarization using linguistic knowledge acquisition. Inf. Process. Manage. **25**(4), 419–428 (1989)
59. Rawat, B., Bist, A.S., Rahardja, U., Aini, Q., Ayu Sanjaya, Y.P.: Recent deep learning based NLP techniques for chatbot development: an exhaustive survey. In: Cyber and IT Service Management, pp. 1–4 (2022). https://doi.org/10.1109/CITSM56380.2022.9935858
60. Rooein, D., Bianchini, D., Leotta, F., Mecella, M., Paolini, P., Pernici, B.: aCHAT-WF: generating conversational agents for teaching business process models. Softw. Syst. Model. **21**(3), 891–914 (2022). https://doi.org/10.1007/s10270-021-00925-7
61. Rossen, B., Lind, S., Lok, B.: Human-centered distributed conversational modeling: efficient modeling of robust virtual human conversations. In: Ruttkay, Z., Kipp, M., Nijholt, A., Vilhjálmsson, H.H. (eds.) IVA 2009. LNCS (LNAI), vol. 5773, pp. 474–481. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-04380-2_52
62. Ryniak, C., Burgert, O.: Automatic generation of checklists from business process model and notation (BPMN) models for surgical assist systems. Curr. Dir. Biomed. Eng. **6**(1), 20200005 (2020). https://doi.org/10.1515/cdbme-2020-0005
63. Santing, L.: Easy data augmentation techniques for traditional machine learning models on text classification tasks. B.S. thesis (2021)
64. Sholiq, S., Sarno, R., Astuti, E.S.: Generating BPMN diagram from textual requirements. J. King Saud University - Comput. Inf. Sci. **34**(10), 10079–10093 (2022). https://doi.org/10.1016/j.jksuci.2022.10.007
65. Silver, B.: BPMN Method and Style: With BPMN Implementer's Guide. Cody-Cassidy Press (2011). https://books.google.at/books?id=mLDYygAACAAJ
66. Sintoris, K., Vergidis, K.: Extracting business process models using natural language processing (NLP) techniques. In: Business Informatics, vol. 01, pp. 135–139 (2017). https://doi.org/10.1109/CBI.2017.41
67. Sànchez-Ferreres, J., Burattin, A., Carmona, J., Montali, M., Padró, L.: Formal reasoning on natural language descriptions of processes. In: Hildebrandt, T., van Dongen, B.F., Röglinger, M., Mendling, J. (eds.) BPM 2019. LNCS, vol. 11675, pp. 86–101. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-26619-6_8
68. Soderland, S.: Learning information extraction rules for semi-structured and free text. Mach. Learn. **34**, 233–272 (1999)
69. Tang, B., Wu, Y., Jiang, M., Chen, Y., Denny, J.C., Xu, H.: A hybrid system for temporal information extraction from clinical text. J. Am. Med. Inform. Assoc. **20**(5), 828–835 (2013)
70. Teubner, T., Flath, C., Weinhardt, C.: Welcome to the era of chatGPT. Bus. Inf. Syst. Eng. (2023). https://doi.org/10.1007/s12599-023-00795-x
71. Thompson, B., Post, M.: Automatic machine translation evaluation in many languages via zero-shot paraphrasing. arXiv preprint arXiv:2004.14564 (2020)
72. Vanderfeesten, I.T.P., Reijers, H.A., van der Aalst, W.M.P.: Evaluating workflow process designs using cohesion and coupling metrics. Comput. Ind. **59**(5), 420–437 (2008). https://doi.org/10.1016/j.compind.2007.12.007

73. Weber, B., Reichert, M., Mendling, J., Reijers, H.A.: Refactoring large process model repositories. Comput. Ind. **62**(5), 467–486 (2011). https://doi.org/10.1016/j.compind.2010.12.012
74. Wei, J., Zou, K.: EDA: easy data augmentation techniques for boosting performance on text classification tasks. arXiv preprint arXiv:1901.11196 (2019)
75. Witteveen, S., Andrews, M.: Paraphrasing with large language models. arXiv preprint arXiv:1911.09661 (2019)
76. Wittig, A., Perevalov, A., Both, A.: Towards bridging the gap between knowledge graphs and chatbots. In: Web Engineering, pp. 315–322 (2022)
77. Meyer von Wolff, R., Nörtemann, J., Hobert, S., Schumann, M.: Chatbots for the information acquisition at universities – a student's view on the application area. In: Følstad, A., et al. (eds.) CONVERSATIONS 2019. LNCS, vol. 11970, pp. 231–244. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-39540-7_16
78. Yoshimura, R., Shimanaka, H., Matsumura, Y., Yamagishi, H., Komachi, M.: Filtering pseudo-references by paraphrasing for automatic evaluation of machine translation. In: Machine Translation, pp. 521–525 (2019)
79. Yu, J., Choi, J., Lee, Y.: Mixing approach for text data augmentation based on an ensemble of explainable artificial intelligence methods. Neural Process. Lett. 1–17 (2022). https://doi.org/10.1007/s11063-022-10961-z