






Process Modeler vs. Chatbot: Is Generative AI Taking Over Process Modeling?

Nataliia Kliedtsova¹, Janik-Vasily Benzin¹, Juergen Mangler¹,
Timotheus Kampik², and Stefanie Rinderle-Ma¹

¹ Technical University of Munich, TUM School of Computation, Information and Technology, Garching, Germany

{firstname.lastname}@tum.de

² SAP Signavio, Berlin, Germany

{firstname.lastname}@sap.com

Abstract. Large language models (LLMs) have become a promising tool for automating complex tasks such as process model generation from text. In order to evaluate the capabilities of LLMs in generating process models, it is crucial to provide means to assess the output quality. A few studies have already provided key performance indicators for assessing aspects such as completeness of the models in a quantitative way. In this paper, we focus on the qualitative assessment of generated process models generated by LLMs based on a user survey. By analyzing user preferences, we aim to determine whether LLM-generated process models meet the needs and expectations of experts. Our analysis reveals that 60% of users, regardless of their modeling experience, prefer LLM-generated models over human-created ground truth models.

Keywords: Business Process Engineering and Management · Process Modeling · Generative AI · Large Language Models · User study

1 Introduction

Process descriptions are textual descriptions of organizational routines that can serve as, e.g., manuals and learning materials for participants, as well as foundations for process optimization, redesign, automation, and execution. Process descriptions have to be comprehensible to multiple stakeholders from diverse backgrounds and knowledge cultures [10]. However, due to their flexible nature they often leave space for interpretation, resulting in arguably subjective representations and a lack of objectivity [3] leading to ambiguity.

Thus, process descriptions are transformed into *process models* [26] in order to improve the clarity of the described processes and to enable their analysis, facilitate decision-making about the processes, and aid the development of process software and documentation, as well as workflow management [19]. However, transforming process descriptions into models can be challenging, as domain experts and process modelers have to continuously communicate using different languages (i.e., domain-specific natural language vs. modeling languages) to exchange their knowledge and vision [25].

The inherent ambiguity in process descriptions complicates this transformation, as the modeler must decide which aspects to include, the level of abstraction, and the perspectives to consider [3]. These decisions can negatively impact the comprehension and application of the models [18]. For instance, if the modeler omits important details or includes unnecessary information, the resulting model can either be too simplistic or overly complex. A model that is too abstract may lack sufficient detail for practical application.

Therefore, it is crucial to ensure that domain knowledge is accessible during modeling [18]. One proposed solution is to equip domain experts with a modeling tool that allows them to create process models using natural language instructions through a conversational user interface, such as a chatbot [8,11]. LLMs such as ChatGPT have emerged as potential candidates for this task due to their capability to understand unstructured natural language text.

In order to evaluate whether domain experts can independently design high-quality process models using generative AI, it is paramount to assess the quality of these process models compared to those produced by humans. In previous work [20,21], we have suggested quantitative metrics such as model completeness and structural similarity. In this work, we conduct a survey with process modelers of different skills, focusing on qualitative aspects of the model and user satisfaction. One finding is that 60% of the participants—regardless of their modeling experience—prefer LLM-generated models over human-created ground truth models.

The paper is structured as follows: Section 2 describes the survey design and Sect. 3 the survey results. The latter are discussed in Sect. 4, followed by related work in Sect. 5, and a conclusion in Sect. 6.

2 Survey Design

Following [12], this section presents details of the survey design in Sect. 2.1 and the questionnaire design in Sect. 2.2.

2.1 Goal, Participants, and Data Collection

The main purpose of the survey is to perform a qualitative evaluation of process models generated by LLMs. The study participants are a group of students and professionals with different programming and modeling background. Their task is to select the one process model out of a list of given process models, which corresponds best to the provided process description (see Figure 1). All data were collected electronically. The survey was performed anonymously and online by means of a questionnaire designed in Microsoft Forms³.

The collected data are then analyzed to identify preferences among respondents regarding selected process models. Statistical methods (i.e., tests of independence and goodness-of-fit tests) are applied to determine which models are most frequently selected as the best match to the provided descriptions.

³ <https://forms.office.com/e/Y55jyNuPi2>

2.2 Questionnaire Design

The survey questionnaire is divided into three parts. First, we collect demographic information about participants regarding their modeling and programming experience. Second, we provide a brief description of the standardized BPMN (Business Process Model and Notation) notation and a simple process model to rate the general level of understanding of this description. We also ask which additional information is potentially required by the modeler to create a good process model. Finally, we provide several process descriptions and multiple process models associated with them. Figure 1 illustrates an example from the survey, where one model is always generated by a human modeler (see Data Set), while the others are generated by an LLM given the process description and a prompt (see Prompts and generated models).

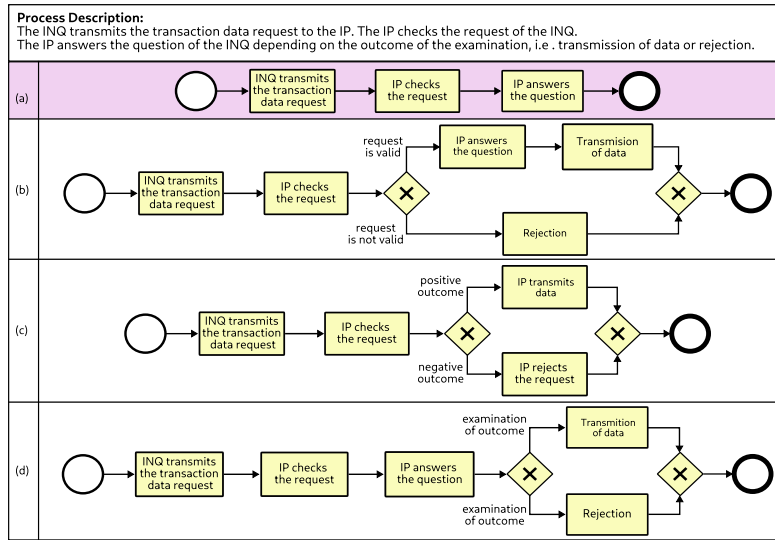


Fig. 1. Process Description and Associated Process Models, where (a) is a ground truth model; (b-d) are LLM-generated models

Participants are then asked to select the model from the set of proposed models that best matches the process description. In the following, we describe the data set with process descriptions and human-modeled process models as well as the approach to LLM-generated process models.

Data Set: We use the PET data set⁴ [4] for process model generation. The PET data set only contains textual process descriptions. As ground truth models, we thus utilize BPMN process models created manually based on the existing human annotation of activities, gateways, and control flow provided in the PET dataset.

⁴ <https://huggingface.co/datasets/patriziobellan/PET>

Seven examples from the PET dataset are taken over from [5], as they are of different lengths and complexity.

Model Representation: The context window of an LLM (i.e., the total amount of text, including both the user’s prompt and the model’s generated output), typically ranging from 1,000 to 8,000 tokens, limits processing due to fixed sequence lengths used during training. As LLMs scale up, they can generate more extensive and longer responses. However it can impact the output quality, potentially introducing issues such as hallucinations and incurring significant costs. Therefore, it is necessary to use a simplified representation of traditional XML-based BPMN 2.0 models to facilitate efficient process model generation and visualization. In this survey, as introduced in [21], we utilize two intermediate process model representations for model generation: Mermaid.js⁵ (MER) and Graphviz⁶ (GV).

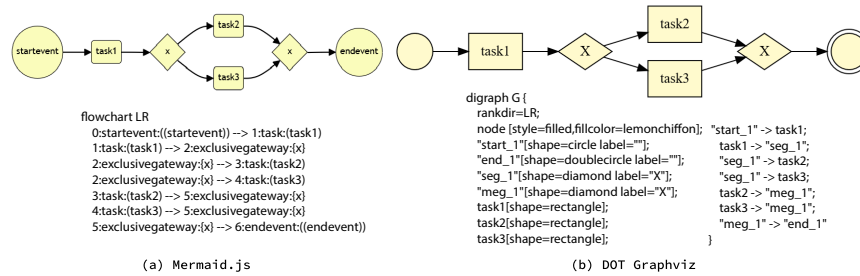


Fig. 2. Selected Model Representations of Text Fragment given in TF1

Selected representations contain the orientation of the graph and custom structure for all nodes and edges in it. For every node, the specific features (e.g., type, color) are assigned to represent a particular BPMN element. For example, the simple process description

“After task1, either task2 or task3 are conducted.” (Text Fragment TF1) can be converted by an LLM into the process model as shown in Fig. 2.

Prompts and Generated Models: All models generated in the scope of this survey are based on the zero-shot principle, utilizing GPT-4 and three types of prompts developed in [21].

Figure 3 shows which prompts were used to generate models for the survey and their structure. Each prompt consists of three parts: [1] a process description, [2] some additional information, and [3] the actual task that should be solved by the LLM, i.e. “generate a graph” (cmp.⁷). The processes description [1] and the actual task to be performed [3] by the LLM are included in each prompt type.

⁵ <https://mermaid.js.org/>

⁶ <https://graphviz.org/doc/info/lang.html>

⁷ https://github.com/com-pot-93/convermod/tree/main/prompt_engineering

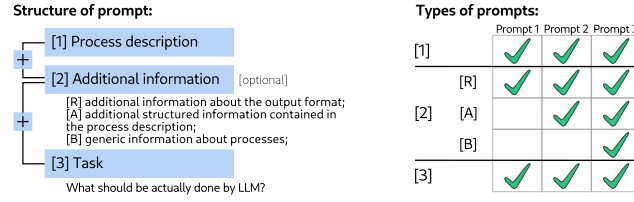


Fig. 3. Prompt Structure and Utilized Prompt Types

Multiple, optional pieces of information (see Fig. 3 [R,A,B]) can also be passed to the LLMs to influence the quality of the generated model.

All generated models contain two categories of BPMN elements: (a) flow objects (start and end events, tasks, exclusive and parallel gateways) and (b) connecting objects (sequence flows). Only models that were evaluated as correct and complete in [21] are taken into account⁸. Based on seven selected process descriptions from the PET data set, the two model representations (MER and GV), and the three prompts (cf. Fig. 3), 42 models were generated. Models identical to the ground truth and those that were incorrect or incomplete (only 26% of all generated models) were excluded from the survey, as they would provide no basis for distinction or could lead to biased results. Identical models do not provide any basis for distinction. Incorrect or incomplete models do not provide accurate representations of relevant process descriptions, which could lead to confusion among participants or biased results. Hence, including these models would not yield meaningful insights into preferences, as the choices would be arbitrary or redundant. In the end, 5 examples and 19 models were used for the survey.

3 Survey Results

Participant background: A total of 40 respondents took part in the survey. Around 60% of respondents are familiar with various graphical modeling languages (e.g., UML, ER, or BPMN). 80% of them have more than 3 years of modeling experience or were applying modeling languages in class and industry projects and could be considered as confident modelers. Other participants either have no modeling experience or have few modeling skills. 29 out of 40 participants have more than 3 years of programming experience. 17.5% of all participants are not familiar with markup languages. Out of those familiar with markup languages, 15% have used MER and 42% have used GV.

Prompt engineering: The assessments of completeness and correctness performed in [21] showed that GPT4 yields the best results using Prompt 1 (P1). The second best results were achieved by Prompt 2 (P2). In order to gauge these results with human intuition, participants are asked to select one of the proposed

⁸ <https://github.com/com-pot-93/convermod/tree/main/survey/models>

Table 1. Percentage Distribution of Frequencies For Additional Information

Type	-	R	A	B	R+A	R+B	A+B	R+A+B	Other
%	12.5	7.5	22.5	2.5	22.5	7.5	5	7.5	12.5

types of additional information or their combination in the prompt template described in Figure 3. The options were: [R] a set of rules of how to represent a particular text as a graph; [A] an explicit list of process activities; [B] a summary of the BPMN standard in addition to a textual process description.

Most of the respondents selected [A] or [R]+[A] (22.5% each). Notably, [R]+[A] corresponds to the achieved results, as Prompt 2 uses [R]+[A] and provides the second-best results. In contrast, only 7.5% of all respondents consider using [R] independently, despite the fact best results are achieved by using P1, which utilizes [R] (see Tab. 1).

Furthermore, 12.5% of all respondents suggest using other combinations of proposed information types like [A]+[B] or [R]+[B]. About 12.5% of all participants suggest including additional information or methods as process model examples or a workshop with a domain expert (again, see Tab. 1).

Model representation: Respondents are also asked to rate their level of understanding when utilizing [R] (i.e., MER and GV) and [B] with respect to the 5-point scale in the survey. Only 2 participants rated [B] as poor. Out of 26 respondents, that were rating [R], 23% (MER) and 38% (GV) consider them as unclear (i.e., rated as poor or very poor). Nevertheless, 70% of respondents rate MER [R] as good or very good. At the same time, only 30% of participants acknowledge GV [R] as good or very good.

Interestingly, a majority of participants (17) find MER representation to be easier to read and understand and express a preference for using it over GV. In contrast, only 20% (5 participants) prefer GV representation, and 15% (4 participants) did not have any preference. However, it is essential to highlight that the representation preference distribution might be linked to the prior experience and comprehension of the provided information (i.e., [R]), observed during its evaluation (see Sect. 5.4).

Let us note that in a highly integrated prototype for using LLMs to create graphical models out of text, neither GV nor MER [R] would be visible to the user. Even so, another thing to highlight is that LLMs demonstrate a reduced occurrence of syntax errors using GV’s representation, but generate semantically better models using MER representation.

Hypothesis testing: To assess whether LLM-generated models could be deemed to be of similar quality as models designed by a human modeler, we ask participants to read a process description and to select the one of several proposed models that, in their opinion, corresponds best to it. For every process description, participants could select one model out of 2–5 associated models, where one model is always the ground truth model and the others are generated by LLMs (cf. Fig. 1).

To determine whether the actual data conforms to an expected pattern or distribution, the Pearson’s chi-squared goodness-of-fit test is used. This test is chosen because we are examining just one categorical variable and comparing its observed frequencies to expected frequencies. Mostly, sources suggest to apply the chi-squared test for larger sample sizes, but there is no agreement on a “large” and “small” sample size definition or its boundaries [7]. According to [29] no expected frequency should fall below 5.

For every case, all collected responses ($N = 40$) fall into three groups: ground truth model (**ground**), LLM-generated model (**llm**), or other (**other**, i.e., a model suggested by a user and not included in the proposed options). The “I don’t know” option was removed from the observations since, participants tend to select it for all the cases (or its majority) due to limited experience, lack of motivation, or time constraints. Only a small number of all responses (10%) correspond to the “I don’t know” option.

Our null hypothesis is based on the assumption that the majority of people prefer the ground truth models over the LLM-generated ones and states the probabilities for each group as follows:

$$H_0 : P = (P_1, P_2, P_3) = (0.8, 0.15, 0.05) \quad (1)$$

where P_1 is the probability that the ground truth model is selected, P_2 stands for the LLM-generated model, and P_3 for “other”.

Our alternative hypothesis suggest that the distribution of responses differs from our null hypothesis. In other words, we suspect that people do not predominantly select the ground truth models, and there is a change in the distribution among the three groups:

$$H_1 : P \neq (0.8, 0.15, 0.05) \quad (2)$$

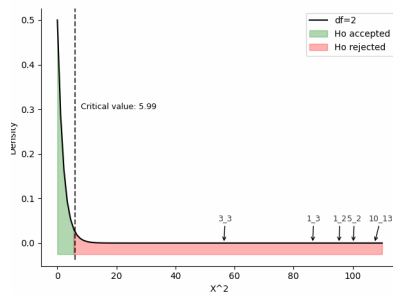


Fig. 4. Pearson’s Chi-Squared Test Results: Model Preference

To determine the validity of the null hypothesis and whether it should be rejected, we compute the critical value for the goodness-of-fit test and compare it to the value obtained for each case.

For every model, the goodness-of-fit is significantly larger than its critical value, and the p-value is well below the selected level of significance ($p=0.0001$). Hence, we reject the null hypothesis in all cases. It appears that only 19% of the responses recognize the ground truth models as the most suitable for the provided process descriptions. In contrast, 69% of the responses identify the models generated by LLMs as the most suitable.

Modeling experience: Since only half of the participants possess modeling experience and are familiar with BPMN, we aim to investigate whether an

association exists between the choice of model type (**ground**, **llm**, or **other**) and the level of modeling experience (i.e., (**ne**) no experience, (**crb**) learned in class or from book, (**cp**) used on a class project, (**pi**) used on one project in industry, (**mp**) used on multiple projects in industry). Our inference is that individuals with more modeling experience are more likely to choose **other** or **ground** choices, while those with less experience tend to prefer **llm** models.

To investigate whether there is a true relation between modeling experience and selected model type, we employ two tests of independence: the Pearson’s chi-squared test and the Fisher’s exact test. In both cases, our null hypothesis states that these variables are independent from each other:

$$H_0 : A \perp B, \quad (3)$$

where A is type of the chosen model and B is the level of modeling experience. Conversely, the alternative hypothesis claims that there is a connection between A and B.

Considering the individual results for both Fisher and Pearson tests (see Tab. 2), there is not enough evidence to reject the null hypothesis in 9 out of 10 cases. This indicates that we do not have sufficient confirmation to claim a significant connection between the chosen model type and a level of modeling skills. In simpler terms, these two variables are not considered to be dependent.

Table 2. Relationship Between Modeling Experience and Selected Model Type: p-values (LS = level of significance)

Case	LS	1.2	1.3	3.3	5.2	10.13	Total
Fisher	0.05	0.387	0.129	0.03	0.336	0.06	0.044
Pearson	0.05	0.306	0.179	0.053	0.292	0.111	0.054

However, when we consider all cases collectively as a single sample, the overall picture undergoes a substantial shift. According to the Fisher test, the null hypothesis is expected to be rejected, suggesting a potential dependency between modeling experience and choice of a model type. Based on the Pearson test, the null hypothesis remains valid, but the p-value is only slightly higher than the initially set level of significance. However, these results require careful interpretation. Combining all process models into a single sample for the independence test may compromise the assumption of independent observations, as multiple process models were evaluated by the same individuals.

Intriguingly, the current relationship between model type and modeling experience differs from our initial expectations. Participants with no modeling experience (**ne**) and those with a more academic background (**crb** and **cp**) tend to choose **ground** models more frequently compared to individuals with real-life experience (**pi** and **mp**). Conversely, more experienced respondents show a preference for the **other** option over the **ground** models. However, all groups consistently vote for **llm** models. The distribution of **llm** model selections remains consistent ($\sim 60\%$) across all groups (see Fig. 5).

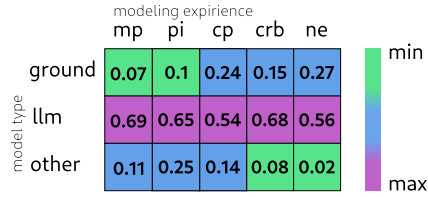


Fig. 5. Percentage Distribution of Frequencies For Modeling Experience and Selected Model Type

Concerns: It is essential to mention that these results should be considered with caution, due to relatively small sample size. With half of the respondents being limited familiar with the subject matter, it is important to acknowledge the potential influence of various response biases.

Manually filtering out incorrect or incomplete LLM-generated models can bias the results in favor of LLM-generated models by showcasing only the most accurate and complete outputs, which does not reflect the full range of the LLM’s performance.

In addition, the variability caused by the probabilistic nature of the LLM and the lack of control over its evolution (i.e., release of new versions that impact significant changes in its output) can lead to reproducibility issues [28]. Furthermore, survey respondents’ engagement and data quality can be compromised by survey length and complexity, potentially resulting in rushed or incomplete responses. Misinterpretation of questions and the influence of social desirability bias may introduce inaccuracies.

4 Discussion

Around 60% of the participants prefer models generated by LLMs over the models created based on human annotations or suggesting their solutions (see Sect. 3). This preference remains consistent regardless of modeling experience. These results highlight the capabilities of LLMs in generating models with a quality comparable to human-generated models. Still, our results should be interpreted with caution: the examples from the PET dataset may be overly simplistic compared to real-world use cases, and LLMs may perform exceptionally well on these simplified examples such that the same performance might not be achieved in more complex cases. Additionally, while LLMs demonstrate high performance in model generation for simple process descriptions, their efficiency is questionable when higher levels of abstraction or multiple perspectives are involved.

However, while the overall preference leans towards LLM-generated models, it is crucial to examine the cases in which human-generated models were chosen. Understanding these scenarios can provide insights into the limitations of LLMs and areas where human intuition and expertise still play a crucial role. Models created by humans might be preferred in scenarios involving high complexity

where human intuition and judgment play a significant role and scenarios requiring deep domain-specific knowledge or industry expertise.

Furthermore, the number of BPMN elements utilized during model generation was restricted to the basic elements (see Sect. 2.2). Introducing a greater variety of model elements may compromise the quality of the results. During the survey, respondents were aware that proposed models were generated by humans or LLMs. This could significantly influence the survey results, potentially introducing bias into the evaluation process. Participants' perceptions and expectations can affect their judgments. For instance, they might assume that LLM-generated models are more simple or sophisticated or that human-generated models are more detailed and better structured, influencing their choice respectively.

5 Related Work

According to [16] business process models can be generated from different sources, such as business rules, standard operating procedures, spreadsheets and unstructured text. In the scope of this paper we focus on unstructured text (i.e., T2M transformation), as not everyone understands specific formats and notations, but essentially everyone understands at least one natural language [9].

The idea of transforming unstructured text into a structured, diagram-based representation such as UML (Unified Modeling Language), ER diagrams (Entity-relationship), BPMN, and DECLARE (Declarative Process Modeling), is not novel. Most existing methods rely on text pattern search, rule-based approaches, or semantic analysis (e.g., [14,30,13]). [27,1] propose techniques for automatic annotation of textual process descriptions utilising classical supervised machine learning-based approaches. Nonetheless, due to the absence of suitable publicly available datasets containing real-life complex data, applying these machine learning techniques becomes challenging [6].

Hence, the rise and evolution of natural language processing (NLP) holds additional promise for research, particularly when applied in the context of utilizing LLMs. In [6,20], language models are utilized to extract entities and relationships from textual process descriptions and in [15,21] a method for generating a process model in a pre-specified intermediary notation as an output format of LLMs (particularly GPT4) is proposed and show-cased. Additionally, several online tools for conversational process modeling were introduced in [23,22].

However, not only the generation of models is a challenge, but also their evaluation presents significant difficulties and demands considerable effort. Mostly, the evaluation tends to be more quantitative, focusing on measurable aspects like model complexity, execution time, and compliance with syntactical rules, rather than qualitative semantic evaluation [24,17,2]. In [22], the effectiveness of the tool and its technological acceptance are also taken into consideration. Yet, during evaluation it must be ensured that the models accurately reflect the intended processes and are useful to stakeholders (i.e., model correctness regarding provided process description and stakeholder's satisfaction with it).

6 Conclusions and Future Research Directions

The results of the evaluation indicate that chatbots for conversational process modeling are ready to be applied in practice. The LLMs demonstrate a strong capability to handle model generation tasks, confirmed by the high accuracy of the generated models—74% were both complete and correct in representing the provided process descriptions. The survey results show that distinguishing a ground truth model from a set of LLM-generated models is not straightforward, and interestingly, the ability to select the correct model does not depend on modeling experience.

The probabilistic nature of LLMs necessitates that domain experts check the results. However, since acquiring as-is models consumes up to 60% of the time spent on process management projects [13], chatbot-based partial automation can be sufficiently impactful, even if substantial human refinement is required.

Future research will focus on integrating the strong language capabilities of chatbots into the iterative process of model generation, where users create process models with the help of chatbots and then refine them. Additionally, the set of utilized business process elements should be extended to include at least pools and lanes. There is also a necessity to enhance the ability of LLM-generated models to handle more complex use cases and integrate diverse viewpoints based on stakeholder perspectives and modeling purposes.

References

1. Ackermann, L., Neuberger, J., Jablonski, S.: Data-driven annotation of textual process descriptions based on formal meaning representations. In: *Advanced Inf. Syst. Engineering*. pp. 75–90 (2021)
2. 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>
3. Beerepoot, I., Ciccio, C.D., Reijers, H.A., et al., S.R.: The biggest business process management problems to solve before we die. *Comput. Ind.* **146**, 103837 (2023)
4. Bellan, P., van der Aa, H., Dragoni, M., Ghidini, C., Ponzetto, S.P.: PET: an annotated dataset for process extraction from natural language text tasks. In: *Business Process Management Workshops*. pp. 315–321 (2022)
5. Bellan, P., Dragoni, M., Ghidini, C.: A qualitative analysis of the state of the art in process extraction from text. In: *Italian Association for Artificial Intelligence*. pp. 19–30. *CEUR Workshop Proceedings* (2020)
6. Bellan, P., Dragoni, M., Ghidini, C.: Extracting business process entities and relations from text using pre-trained language models and in-context learning. In: *Enterprise Design, Operations, and Computing*. pp. 182–199 (09 2022)
7. Campbell, I.: Chi-squared and Fisher–Irwin tests of two-by-two tables with small sample recommendations. *Statistics in Medicine* **26**(19), 3661–3675 (2007)
8. Casciani, A., Bernardi, M.L., Cimitile, M., Marrella, A.: Conversational systems for ai-augmented business process management. In: *Research Challenges in Information Science*. pp. 183–200 (2024)
9. Dalianis, H.: A method for validating a conceptual model by natural language discourse generation. In: *Advanced Inf. Syst. Engineering*. pp. 425–444 (1992)

10. Dehnert, J., van der Aalst, W.M.P.: Bridging the gap between business models and workflow specifications. *Int. J. Cooperative Inf. Syst.* **13**(3), 289–332 (2004)
11. Dumas, M., et al.: AI-augmented business process management systems: A research manifesto. *ACM Transactions on Management Inf. Systems* **14**, 1 – 19 (2022)
12. Fowler Jr, F.J.: *Survey research methods*. Sage publications (2013)
13. Friedrich, F.: *Automated Generation of Business Process Models from Natural Language Input*. Master’s thesis, Humboldt-University zu Berlin (2010)
14. Ghose, A., Koliadis, G., Chueng, A.: Process discovery from model and text artefacts. In: *Services Computing Workshops*. pp. 167–174 (2007)
15. Grohs, M., Abb, L., Elsayed, N., Rehse, J.: Large language models can accomplish business process management tasks. *CoRR* **abs/2307.09923** (2023)
16. Honkisz, K., Kluza, K., Wisniewski, P.: A concept for generating business process models from natural language description. In: *Knowledge Science, Engineering and Management*. pp. 91–103 (2018)
17. Kahloun, F., Ghannouchi, S.A.: Improvement of quality for business process modeling driven by guidelines. *Procedia Computer Science* **126**, 39–48 (2018)
18. Kannengiesser, U., Oppl, S.: *Business processes to touch: Engaging domain experts in process modelling*. vol. 1418 (09 2015)
19. Kesari, M., Chang, S., Seddon, P.B.: A content-analytic study of the advantages and disadvantages of process modelling. In: *Australasian Conference on Information Systems, ACIS 2003, Perth, Australia, November 26-28, 2003* (2003)
20. Klievtsova, N., Benzin, J., Kampik, T., Mangler, J., Rinderle-Ma, S.: Conversational process modelling: State of the art, applications, and implications in practice. In: *Business Process Management Forum*. pp. 319–336 (2023)
21. Klievtsova, N., Mangler, J., Kampik, T., Benzin, J.V., Rinderle-Ma, S.: How can generative ai empower domain experts in creating process models? In: *Wirtschaftsinformatik* (2024)
22. Koepke, J., Safan, A.: Efficient llm-based conversational process modeling. In: *Business Process Management Workshops* (2024)
23. Kourani, H., Berti, A., Schuster, D., van der Aalst, W.M.P.: Promoai: Process modeling with generative AI. *CoRR* **abs/2403.04327** (2024)
24. de Oca, I.M.M., Snoeck, M., Reijers, H.A., Rodríguez-Morffi, A.: A systematic literature review of studies on business process modeling quality. *Information and software technology* **58**, 187–205 (2015)
25. Odeh, Y.: *Bpmn in engineering software requirements: An introductory brief guide*. In: *International Conference on Information Management and Engineering* (2017)
26. Polyvyanyy, A., Smirnov, S., Weske, M.: Business process model abstraction. In: *Handbook on Business Process Management, Introduction, Methods, and Information Systems*, pp. 147–165. *Int. Handbooks on Inf. Syst.*, Springer (2015)
27. Qian, C., Wen, L., Kumar, A., Lin, L., Lin, L., Zong, Z., Li, S., Wang, J.: An approach for process model extraction by multi-grained text classification. In: *Advanced Information Systems Engineering*. pp. 268–282 (2020)
28. Sallou, J., Durieux, T., Panichella, A.: Breaking the silence: the threats of using llms in software engineering. In: *Proceedings of the 2024 ACM/IEEE 44th International Conference on Software Engineering, NIER@ICSE 2024, Lisbon, Portugal, April 14-20, 2024*. pp. 102–106. ACM (2024). <https://doi.org/10.1145/3639476.3639764>
29. VanVoorhis, C.W., Morgan, B.L., et al.: Understanding power and rules of thumb for determining sample sizes. *Tutorials in quantitative methods for psychology* **3**(2), 43–50 (2007)
30. Yue, T., Briand, L.C., Labiche, Y.: An automated approach to transform use cases into activity diagrams. In: *Modelling Foundations and Appl.* pp. 337–353 (2010)