Explanatory Capabilities of Large Language Models in Prescriptive Process Monitoring (Extended Abstract)*

Kateryna Kubrak[1], **Lana Botchorishvili**[2], **Fredrik Milani**[1], **Alexander Nolte**[3,4], **Marlon Dumas**[1,2]
[1] University of Tartu, Estonia
[2] Apromore, Australia
[3] Einhodven University of Technology, The Netherlands
[4] Carnegie Mellon University, USA
{kateryna.kubrak,fredrik.milani, marlon.dumas}@ut.ee, a.u.nolte@tue.nl

Abstract

Prescriptive Process Monitoring (PrPM) systems recommend interventions in ongoing business process cases to improve performance. However, performance gains only materialize if users follow the recommendations. Prior research has shown that users are more likely to follow recommendations when they understand them. In this paper, we explore the use of Large Language Models (LLMs) to generate explanations for PrPM recommendations. We developed a prompting method based on typical user questions and integrated it into an existing PrPM system. Our evaluation indicates that LLMs can help users of PrPM systems to better understand the recommendations. However, the explanations fall short in addressing the underlying "why" and do not always support users in assessing the trustworthiness of the recommendations.

1 Introduction

Prescriptive Process Monitoring (PrPM) is a family of techniques that recommend interventions in ongoing cases of a business process to optimize performance [Kubrak *et al.*, 2022]. These techniques commonly rely on Artificial Intelligence (AI) to predict undesirable case outcomes and propose interventions [Bozorgi *et al.*, 2023]. While predictions have become more accurate and useful, research suggests users often rely on their own judgment and ignore recommendations [Dees *et al.*, 2019], which prevents potential performance benefits from materializing.

One reason for why users might not follow recommendations is that they do not understand the rationale behind them [Dees *et al.*, 2019]. There are existing approaches to explain recommendations, but these often rely on plots and numbers (e.g., statistical measures or measures of feature importance) [Galanti *et al.*, 2023], which users struggle with understanding [Rizzi *et al.*, 2024] or does not match their information needs [Rizzi *et al.*, 2024]. A more promising approach is dialogue-based systems where users can ask questions to, iteratively build understanding [Laato *et al.*, 2022; Cambria *et al.*, 2023]. Large Language Models (LLMs) appear to be particularly suitable in this context because they can enable a dialogue between a system and a user [Feldhus *et al.*, 2022], elaborate on plots and numbers, and answer follow-up questions [Feldhus *et al.*, 2022].

Our research objective (RO) is *to design and evaluate an approach for LLM-based explanations of recommendations generated by prescriptive process monitoring techniques*. To pursue this objective, we designed a prompting method that enables an LLM to elaborate on and explain PrPM recommendations. Prompts are natural language specifications of instructions for an LLM [Bellan *et al.*, 2022]. First, we identified the needs of end users to understand explanations in the context of PrPM by analyzing the Explainable AI Question Bank (XAIQB) [Liao *et al.*, 2020]. XAIQB is a set of prototypical questions for designing user-centered explanations. By contextualizing these questions to the PrPM setting (e.g., "What are the recommendations prescribed by PrPM techniques?"), we derived a set of explanation requirements and mapped them to appropriate explanation strategies. Based on the requirements, we designed a prompting method composed of context, data description, conversational rules, task, and examples. The examples were grounded in the mapped question categories from the previous phase. To evaluate the method, we implemented an LLM-based chatbot on top of an existing PrPM system.

Thus, the contribution of this paper is two-fold. The first is a prompting method to present explanations of recommendations in PrPM. The second is insights into potential benefits and challenges of designing LLM-based systems for enhancing explainability in PrPM systems.[1]

*This paper summarizes Kubrak, K., Botchorishvili, L., Milani, F., Nolte, A., Dumas, M. (2024). Explanatory Capabilities of Large Language Models in Prescriptive Process Monitoring. In: BPM 2024. LNCS, vol 14940. Springer, Cham. https://doi.org/10.1007/978-3-031-70396-6_23

[1]All supplementary material, including the full mapping of explainability questions (Section 2.1), prompt engineering report (Section 2.2), and questions and answers dataset (Section 3) are available at: https://doi.org/10.6084/m9.figshare.25415290.v1

2 Prompting Method

2.1 Elicited Requirements

To elicit requirements, we used XAIQB [Liao *et al.*, 2020], that categorizes explainability questions into 10 groups. We tailored these questions to the specific context of PrPM and mapped them to potential explanations, serving as examples to be included in the prompt.

Category	Questions	Ways to explain	Prototypical output
Data	What is the size of the event log?	Number of cases in the event log	The event log consists of [number] of cases.
Performance	Why should I believe that the predictions are correct?	Provide performance metrics for the models (accuracy, precision, recall)	The accuracy of recommendations is on average [number].
Output	What do the different recommendation types mean?	Describe the differences between the techniques	Next activity: A next activity is a type of a recommendation that is prescribed by an algorithm that predicts [...]

Table 1: [Excerpt] Mapping of explainability questions and ways to explain. For full mapping, see supplementary material.

Based on the mapping, we elicited the following functional (FR) and non-functional (NR) requirements:

- FR1: The chat's answers should contain correct data from the PrPM outputs.
- FR2: The chat's answers should contain relevant content based on recommendation type.
- NR1: The chat should always provide a response.
- NR2: The chat should respond to the user's question within near-real time.

FR1 refers to the need for the LLM to query correct data from the database to include in the prototypical output (e.g., the number of cases). FR2 relates to giving correct information about the techniques that prescribe the different recommendation types. In the PrPM system where the prompting method was integrated to, PrPM techniques in the background produce three different types of recommendations (guiding, correlation-, or causality-based (see PrPM system description in [Kubrak *et al.*, 2023a]). Thus, the LLM would have to correctly match the information.

2.2 Prompting Method

The initial prompt, drawing from the literature, consisted of context, data description, general conversational rules, task, and examples. Context included specifying the domain (process mining) and details about the PrPM system, such as which techniques are used, the workflow, and the input parameters. Data description related to the structure of the database connected to the PrPM system (described in the next subsection). The task for to the LLM was to answer questions about the PrPm system's recommendations and query the database to obtain the required data for the answers.

To answer the user's questions, the LLM needs to query the database where case data and recommendations are stored (FR1). We conducted three tests to identify how to best represent the query examples in the prompt. For the tests, we used three variations: #1 no examples, #2 example question and steps for making the query, #3 example question, steps for making the query, and the query itself. Prompts #1 and #2 produced incorrect queries by returning the entire case data and an empty response, respectively. For both these prompts, the LLM sometimes queried the collection of files instead of the collection of cases in the database structure. Variation #3 produced correct responses. Therefore, we designed the prompting method to include a question, steps, and a query since it proved to work correctly.

Table 2 details the overall structure of the prompt. FR2 did not require querying the database because explanations for different recommendation types are already provided as text within the component "Examples" in the table.

Component	Text (excerpt)
Context	[PrPM system] uses three algorithms to generate prescriptions for business processes [...] The [PrPM system] workflow involves: Uploading an event log. Defining column types. Setting parameters [...] The key parameters are: Case Completion: An activity that marks the end of a case, e.g., 'Application completed' [...]
Data description	- Description of MongoDB files collection - Description of cases collection
General conversational rules	When answering, use simple language for the explanations. Do not mention the database or show raw data in your responses. [...]
Examples	QUESTION: What is the size of the event log? ANSWER: The event log consists of `<nr_of_cases>` of cases. QUERY: `collection:'cases', aggregate:['$match': 'event_log_id':<EVENT_LOG_ID>, '$count': 'number_of_cases']` STEPS: Run the query with function `query_db` to find the number of cases in this event log.
Task	Your role is to answer questions about [PrPM system] recommendations and query the database for specific case or event log info.

Table 2: [Excerpt] Components of the prompt with text excerpts of each component. For full prompt and full prompt engineering report, see supplementary material.

2.3 PrPM System Integration

To evaluate the prompting method, we developed an LLM-based chatbot on top of a PrPM system (see Fig. 1). The system prescribes recommendations which are stored in the database. For each case, there may be up to three recommendation types prescribed (guiding, correlation-based, causality-based [Kubrak *et al.*, 2022]). Users upload an event log and receive recommendations for ongoing cases. The generated recommendations are displayed in the UI.

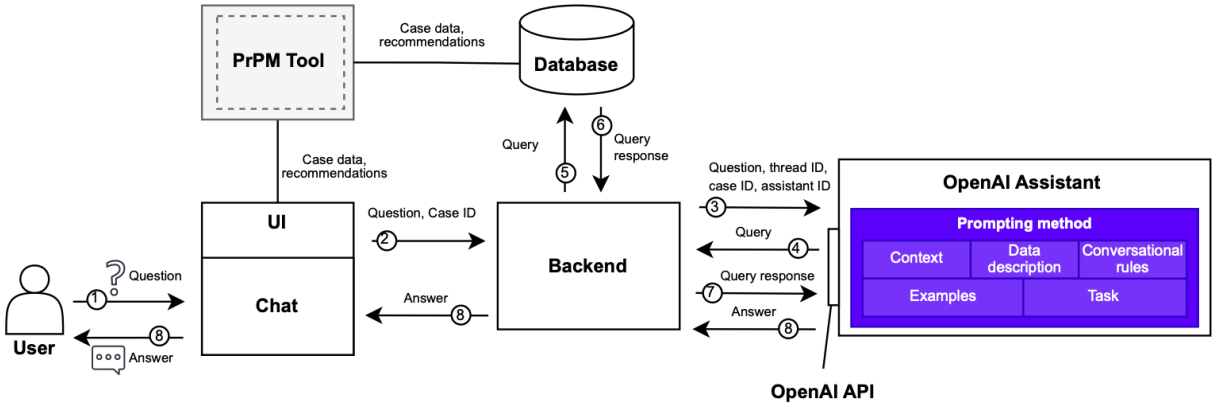

Figure 1: High-level overview of the interaction between the user with the chatbot within a PrPM system. The chatbot displays the answers generated by the LLM based on the prompt.

When the user asks a question (1), the case ID and the question are sent to the backend (2). It then uses the OpenAI API thread endpoint whenever the user creates a new thread (chat). For each question, the backend creates a new run using OpenAI API endpoint (3). The run is configured to include the thread ID (specific to a case) and assistant ID. The backend also configures the run to overwrite the assistant instructions (the prompt) by appending the event log, case structures, and their respective IDs to the run instructions. If a question requires querying the database, OpenAI provides the backend with the function arguments (4) and the backend queries the database (5). The backend then sends the function output to OpenAI (7), which takes the question and function output and produces the answer (8).

3 Evaluation

The goal of the evaluation was to (1) assess users' perception of generated explanations based on the prompting method, and (2) assess users' interaction with the chat. To pursue these goals, we used a mixed-methods approach that consisted of contextual interviews and a survey. The participants were process analysts, working with process analysis internally at a company or as a consultant (12 in total). We used a synthetic event log of a claim management process. The participants were able to ask their questions in the chat. For each interview, we analyzed (1) the spoken interaction between the participant and the interviewee, and (2) the conversation between the participant and the chat. The participants' questions and the chat's responses were coded. The coding scheme for the participants' questions was based on the question categories from XAIQB. To evaluate the chatbot's explanations, we applied a combined deductive and inductive coding approach. Following common practice in explanation research [Hoffman *et al.*, 2023], we reviewed literature on evaluating textual explanations [Zemla *et al.*, 2017; Nauta *et al.*, 2023] and identified relevant characteristics. We then refined these through open coding of the data. The final categories were: *Coherency, Relevance to the question, Completeness, Correctness*, and *Compactness.* Two authors of the paper conducted the coding independently. They each coded a portion of the dataset, and through multiple rounds arrived at an acceptable agreement score. Cohen's Kappa for questions coding was 0.65 (substantial) and for explanations coding between 0.47-0.5 (moderate).

The results of the evaluation are organized into; participant's questions, chat's responses, and patterns in the participant-chat interactions. The complete question and answer dataset is available in the supplementary material.

Participants' Questions. Most questions (55%) focused on clarifying PrPM *"Outputs"*, including terms like "CATE score" and "intervention". Participants also asked about case outcomes and the predicted end of the case. Questions on the *"Why"* accounted for 18% and focused on understanding why a recommendation was made, especially when multiple options were presented. A smaller number of questions (12%) were in category *"Others"* that related to contextual or statistical information about the case. The participants asked comparatively few questions related to the categories of *"How"* (7%), *"What if"* (4%), *"Data"* (2%), and *"How to be that"* (1%). We did not record any questions in categories *"Performance"* (this category refers to the performance of the techniques, e.g., their accuracy), and *"How to still be this"*.

Chat's Responses. The chat provided timely responses (NR2), but failed to provide a response once (NR1). However, this was due to an error in the PrPM system back-end. Coding of explanations revealed high levels of *Completeness* (94%), *Coherency* (98%) and *Relevance to the question* (94%), but *Correctness* (75%) and *Compactness* (85%) showed room for improvement. Some errors stemmed from either skipping database queries or misinterpreting retrieved data, which affected the trustworthiness of responses.

Interaction Patterns. The majority (8/12) first studied the case and recommendations, and then formulated a question to clarify something they did not understand. One participant opened the chat right away and asked a general question about the case performance before taking a closer look at the case. Two participants started the conversation by asking about issues in the case they should address. Last, one par-

#	The explanations…
Q1	…help me understand how [prototype] works.
Q2	…of how [prototype] works are satisfying.
Q3	…have sufficient detail.
Q4	…seem complete.
Q5	…tell me how to use [prototype].
Q6	…are useful to my goals.
Q7	…show me how accurate [prototype] is.
Q8	…let me judge when I should trust and not trust [prototype].

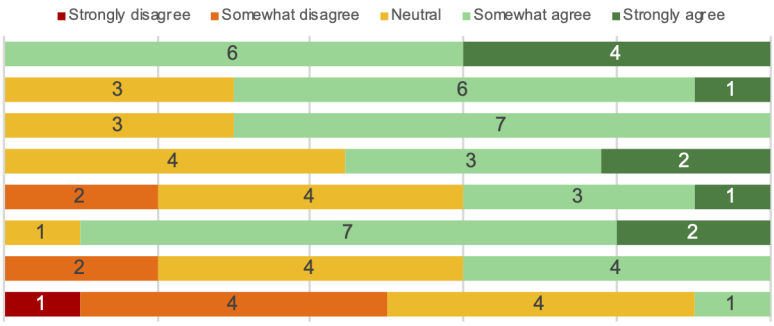

Figure 2: Survey results. Prototype refers to the used PrPM system.

ticipant reversed the interaction with the chat by asking how it could help them, basing the next question on its response. Related to the last approach, two other participants later suggested adding a feature with pre-defined clickable questions to the chat. While the chat handled user disagreements by correcting itself, it occasionally reinforced incorrect assumptions made by the users.

Perception of Explanations. The survey was filled out by 10 out of 12 participants, indicating a response rate of 83%. The participants' perception of the explanations was generally positive, with a few being rather neutral (see Fig. 2). A comparatively low indicator is the participants' trust in the recommendations provided by the PrPM system. This could be related to the correctness of the explanations.

4 Discussion

Our main findings indicate that users seek more detailed and contextualized information when interpreting the PrPM recommendations. Participants frequently asked questions related to why a specific recommendation was given. This suggests the need for explanations that go beyond simple activity labels and their parameters. This observation is aligned with previous research suggesting that users do not follow recommendations because they do not understand the rationale behind them [Dees *et al.*, 2019].

We also identified actionable areas for improvement during the Kairos evaluation. More specifically, we noted that participants expressed interest in guidance through pre-defined or suggested questions when interacting with Kairos. This is in agreement with with prior work on user support in conversational systems [Lee *et al.*, 2023]. Therefore, refining the prompts to highlight high-priority information, such as frequently asked questions from key categories, and providing a glossary of PrPM-specific terminology can further enhance user understanding [Jessen *et al.*, 2023].

Our research highlights several implications for both academia and industry: the need to integrate case performance data, to analyze recommendations across multiple cases rather than just single cases, to consider additional user groups, to strengthen the rationale behind recommendations, and to ensure the correctness of PrPM outputs. First of all, we noted that several participants asked specific questions about case performance (e.g., cycle time and case performance in comparison to others). Such data help users gain contextual information around the recommendations. However, this requires either ensuring that the LLM would be able to calculate, for example, cycle time, or ensure access to case performance data. Second, our evaluation focused on recommendations for individual cases. A future direction is to expand the PrPM system to include aggregated process-level insights, which could be more valuable for analysts[Milani *et al.*, 2022]. Third, our evaluation was conducted with one of the three end-user groups for PrPM, i.e., process analysts (see [Kubrak *et al.*, 2023b]). However, we observed that many questions were about *Output* and *Why*. This information is also valuable for operational workers who make case-specific decisions [Dees *et al.*, 2019]. Therefore, another avenue for future research is to conduct an evaluation of the LLM explanations for specific-case recommendations with operational workers. Fourth, participants asked why a specific recommendation was given. Addressing this could involve integrating explainability techniques with LLMs, e.g., using causal graphs [Fahland *et al.*, 2025], SHAP values [Galanti *et al.*, 2023], or counterfactuals [Hsieh *et al.*, 2021]. Finally, correctness emerged as a key area for improvement. While the chat corrected itself when prompted, hallucinations still occurred. Future work could explore verification layers [Ji *et al.*, 2023] or evaluate different LLMs to enhance reliability.

5 Conclusion

We presented and evaluated an approach for generating LLM-based explanations of recommendations in PrPM. To evaluate the prompting method, we implemented an LLM-based chat on top of a PrPM system. The implications for research point towards the need for further development of causal recommendations in PrPM, as well as causal explanations. Future research of explanations in PrPM may use the guidance of questions asked in our study to cater the explanations to the end-user needs. Practical implications include adding template questions to the chat, improving the prompt specifically for most-asked questions, and enabling questions about case performance-related information.

In future work, we aim to focus on refining the approach to elicit clearer justifications for recommendations (i.e., bringing out the "why"). We also plan to compare LLM-generated explanations with established explainable AI methods to explore their combined potential. Further, we aim to expand the evaluation to operational workers.

Acknowledgments

This research is supported by the Estonian Research Council (PRG1226) and the European Research Council (PIX Project).

References

[Bellan *et al.*, 2022] Patrizio Bellan, Mauro Dragoni, and Chiara Ghidini. Extracting business process entities and relations from text using pre-trained language models and in-context learning. In *EDOC*, volume 13585 of *LNCS*, pages 182–199. Springer, 2022.

[Bozorgi *et al.*, 2023] Zahra Dasht Bozorgi, Irene Teinemaa, Marlon Dumas, Marcello La Rosa, and Artem Polyvyanyy. Prescriptive process monitoring based on causal effect estimation. *Inf. Syst.*, 116:102198, 2023.

[Cambria *et al.*, 2023] Erik Cambria, Lorenzo Malandri, Fabio Mercorio, Mario Mezzanzanica, and Navid Nobani. A survey on XAI and natural language explanations. *Inf. Process. Manag.*, 60(1):103111, 2023.

[Dees *et al.*, 2019] Marcus Dees, Massimiliano de Leoni, Wil M. P. van der Aalst, and Hajo A. Reijers. What if process predictions are not followed by good recommendations? In *BPM (Industry Forum)*, volume 2428 of *CEUR Workshop Proceedings*, pages 61–72. CEUR-WS.org, 2019.

[Fahland *et al.*, 2025] Dirk Fahland, Fabiana Fournier, Lior Limonad, Inna Skarbovsky, and Ava J. E. Swevels. How well can a large language model explain business processes as perceived by users? *Data Knowl. Eng.*, 157:102416, 2025.

[Feldhus *et al.*, 2022] Nils Feldhus, Ajay Madhavan Ravichandran, and Sebastian Möller. Mediators: Conversational agents explaining NLP model behavior. *CoRR*, abs/2206.06029, 2022.

[Galanti *et al.*, 2023] Riccardo Galanti, Massimiliano de Leoni, Merylin Monaro, Nicolò Navarin, Alan Marazzi, Brigida Di Stasi, and Stéphanie Maldera. An explainable decision support system for predictive process analytics. *Eng. Appl. Artif. Intell.*, 120:105904, 2023.

[Hoffman *et al.*, 2023] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. Measures for explainable AI: explanation goodness, user satisfaction, mental models, curiosity, trust, and human-ai performance. *Frontiers Comput. Sci.*, 5, 2023.

[Hsieh *et al.*, 2021] Chihcheng Hsieh, Catarina Moreira, and Chun Ouyang. Dice4el: Interpreting process predictions using a milestone-aware counterfactual approach. In *ICPM*, pages 88–95. IEEE, 2021.

[Jessen *et al.*, 2023] Urszula Jessen, Michal Sroka, and Dirk Fahland. Chit-chat or deep talk: Prompt engineering for process mining. *CoRR*, abs/2307.09909, 2023.

[Ji *et al.*, 2023] Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. Towards mitigating LLM hallucination via self reflection. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Assoc. for Comp. Ling.: EMNLP 2023*, pages 1827–1843, Singapore, 2023. ACL.

[Kubrak *et al.*, 2022] Kateryna Kubrak, Fredrik Milani, Alexander Nolte, and Marlon Dumas. Prescriptive process monitoring: Quo vadis? *PeerJ Comput. Sci.*, 8:e1097, 2022.

[Kubrak *et al.*, 2023a] Kateryna Kubrak, Lana Botchorishvili, Fredrik Milani, Marlon Dumas, Alexander Nolte, Mahmoud Shoush, and Zhaosi Qu. Kairos: A tool for prescriptive monitoring of business processes. In *ICPM Doctoral Consortium / Demo*, volume 3648 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2023.

[Kubrak *et al.*, 2023b] Kateryna Kubrak, Fredrik Milani, Alexander Nolte, and Marlon Dumas. Design and evaluation of a user interface concept for prescriptive process monitoring. In *CAiSE*, volume 13901 of *LNCS*, pages 347–363. Springer, 2023.

[Laato *et al.*, 2022] Samuli Laato, Miika Tiainen, A. K. M. Najmul Islam, and Matti Mäntymäki. How to explain AI systems to end users: a systematic literature review and research agenda. *Internet Res.*, 32(7):1–31, 2022.

[Lee *et al.*, 2023] Yoonjoo Lee, Tae Soo Kim, Sungdong Kim, Yohan Yun, and Juho Kim. DAPIE: interactive step-by-step explanatory dialogues to answer children's why and how questions. In *CHI*, pages 450:1–450:22. ACM, 2023.

[Liao *et al.*, 2020] Q. Vera Liao, Daniel M. Gruen, and Sarah Miller. Questioning the AI: informing design practices for explainable AI user experiences. In *CHI*, pages 1–15. ACM, 2020.

[Milani *et al.*, 2022] Fredrik Milani, Katsiaryna Lashkevich, Fabrizio Maria Maggi, and Chiara Di Francescomarino. Process mining: A guide for practitioners. In *RCIS*, volume 446 of *LNBIP*, pages 265–282. Springer, 2022.

[Nauta *et al.*, 2023] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. *ACM Comp. Surv.*, 55(13s):295:1–295:42, 2023.

[Rizzi *et al.*, 2024] Williams Rizzi, Marco Comuzzi, Chiara Di Francescomarino, Chiara Ghidini, Suhwan Lee, Fabrizio Maria Maggi, and Alexander Nolte. Explainable predictive process monitoring: a user evaluation. *Process Science*, 1(1):3, 2024.

[Zemla *et al.*, 2017] Jeffrey C Zemla, Steven Sloman, Christos Bechlivanidis, and David A Lagnado. Evaluating everyday explanations. *Psychonomic bulletin & review*, 24:1488–1500, 2017.