Human-Centered Evaluation of an LLM-Based Process Modeling Copilot: A Mixed-Methods Study with Domain Experts

Chantale Lauer
chantale.lauer@uni-ssarland.de
Saarland University
Saarbrücken, Saarland, Germany
German Research Institute for Artificial Intelligence
Saarbrücken, Saarland, Germany

Peter Pfeiffer
peter.pfeiffer@5plus.ai
5Plus GmbH
Würzburg, Bavaria, Germany

Nijat Mehdiyev
nijat.mehdiyev@dfki.de
German Research Institute for Artificial Intelligence
Saarbrücken, Saarland, Germany
Saarland University
Saarbrücken, Saarland, Germany

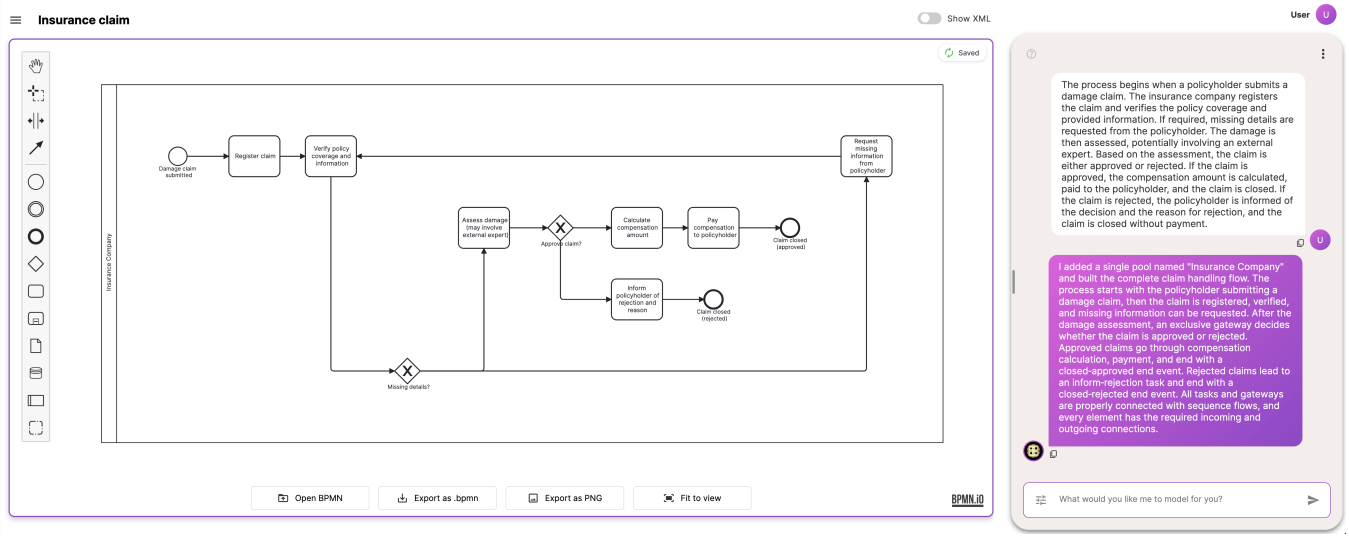

Figure 1: User interface of KICoPro with an example process.

Abstract

Integrating Large Language Models (LLMs) into business process management tools promises to democratize Business Process Model and Notation (BPMN) modeling for non-experts. While automated frameworks assess syntactic and semantic quality, they miss human factors like trust, usability, and professional alignment. We conducted a mixed-methods evaluation of our proposed solution, an LLM-powered BPMN copilot, with five process modeling experts using focus groups and standardized questionnaires. Our findings reveal a critical tension between acceptable perceived usability (mean CUQ score: 67.2/100) and notably lower trust (mean score: 48.8%), with reliability rated as the most critical concern (M=1.8/5). Furthermore, we identified output-quality issues, prompting difficulties, and a need for the LLM to ask more in-depth clarifying questions about the process. We envision five use cases ranging from domain-expert support to enterprise quality assurance. We demonstrate the necessity of human-centered evaluation complementing automated benchmarking for LLM modeling agents.

CCS Concepts

• **Human-centered computing** → **Human computer interaction (HCI)-Empirical studies in HCI**; **Human computer interaction (HCI)**.

Keywords

Business Process Management, Conceptual Modeling, Large Language Models, Human-Centered Evaluation

CHI26, Barcelona, ES

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

ACM Reference Format:
Chantale Lauer, Peter Pfeiffer, and Nijat Mehdiyev. 2026. Human-Centered Evaluation of an LLM-Based Process Modeling Copilot: A Mixed-Methods Study with Domain Experts. In *Proceedings of conference on Human Factors in Computing Systems (CHI26).* ACM, New York, NY, USA, 7 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Large Language Models (LLMs) are increasingly used in business process management (BPM), especially for generating Business Process Model and Notation (BPMN) models from natural language descriptions [10, 13, 27, 28]. Recent systems such as ProMoAI [13], BPMN Chatbot [11], and Camunda BPMN Copilot [6] illustrate the potential of conversational process modeling. KICoPro[1] exemplifies such a conversational system that transforms textual process descriptions into BPMN models. This capability could broaden access to process modeling beyond dedicated experts [28].

However, existing evaluations have focused mainly on automated benchmarks of syntactic, semantic, and pragmatic quality [6, 12]. While such metrics are important, they do not capture whether these systems are actually usable and trustworthy in practice. In particular, conversational BPMN generation may create a gulf of execution [23], where users must infer how to formulate prompts to achieve their modeling goals, and may increase cognitive load when they have to structure descriptions, decompose complex processes, and verify generated process models. These issues are especially relevant because BPMN models are human readable communication artifacts that must support collaborative understanding and review [24]. As Liao and Varshney argue, human-centered evaluation is therefore needed to understand integration into work practices, trust calibration, and interaction breakdowns [19].

We address this gap through a focus group study with process modeling experts, complemented by standardized questionnaires measuring trust, usability, and task-specific performance. Our research questions are:

- **RQ1:** How do process modeling experts perceive the usability and design of an LLM-based modeling copilot?
- **RQ2:** What are the strengths and weaknesses of LLM generated BPMN models from experts' perspectives?
- **RQ3:** What usage scenarios do experts envision for LLM based process modeling tools?

2 Related Work

Prior research shows that LLMs can support the generation and refinement of process models from textual descriptions. ProMoAI [13] demonstrated strong performance of GPT-4 in process model generation, error resolution, and feedback integration, while BPMN Chatbot [11] showed that high model correctness can be achieved with substantially lower token usage. Overall, these studies highlight the potential of LLM based process modeling. But the evaluation is mainly based on quality metrics rather than user experience or trust. Commercial tools such as the Camunda BPMN Copilot [6] and BPMN Assistant [20] further reflect growing interest in LLM-assisted process modeling. Overall, the current work emphasizes technical performance over human factors in real-world use cases.

2.1 Automated Evaluation Frameworks

Several frameworks have been proposed for systematically evaluating LLM generated BPMN models. Kourani et al. [12] developed a comprehensive benchmark consisting of 20 diverse business processes, evaluating 16 state-of-the-art LLMs and revealing significant performance variations across LLMs. Recent work has also introduced structured evaluation frameworks assessing external quality across established BPM quality dimensions [6, 15, 16].

Drawing from ISO 9126 and systematic reviews of process model quality [25], these frameworks typically evaluate clarity, correctness, and completeness. However, as Sànchez Gonzàlez et al. [25] argue, process model quality is inherently multidimensional, including pragmatic, human-centered aspects. However, those are not fully captured by automated metrics alone.

2.2 Human Centered Evaluation of AI Systems

The Human Centered Interaction (HCI) community has increasingly emphasized the need for human-centered evaluation of AI systems [1, 19, 29]. Trust in AI systems has emerged as a critical factor in effective human AI collaboration [5, 17]. Amershi et al. [1] proposed guidelines for human AI interaction that emphasize supporting user understanding, enabling appropriate trust, and facilitating effective correction, aspects that require human evaluation to assess. Scharowski et al. [26] validated the Trust Scale for the AI Context (TAI), demonstrating its psychometric quality for measuring trust in AI applications, including chatbots.

For conversational AI, Holmes et al. [9] developed the Chatbot Usability Questionnaire (CUQ), recognizing that traditional usability measures may not adequately capture the characteristics of conversational systems. The CUQ has been validated across multiple contexts [8] and adopted in studies of professional applications [22].

3 KICoPro

KICoPro is a web-based conversational BPMN modeling tool that enables users to generate and iteratively refine BPMN models via natural language interaction. The system was developed as a research prototype to explore the potential of LLM-based assistance for process modeling tasks, with particular attention to supporting iterative refinements common in professional practice.

The system architecture combines a chat-based frontend, shown in Figure 1, with an LLM backend and BPMN rendering components. Users interact through a conversational interface where they can describe business processes in natural language (right-hand side in Figure 1), request modifications to generated models, ask questions about the current process model state, and request explanations of process modeling decisions. The system maintains a conversation history, enabling multi-turn interactions in which users can reference and refine previous outputs. The process models are displayed in the BPMN modeler (left-hand side in Figure 1), which also allows the users to modify the BPMN models.

KICoPro supports natural-language processing of domain-agnostic process descriptions, generates BPMN models that are visualizable within conversations, supports iterative refinement through targeted modifications, maintains persistent chat histories across sessions, enables BPMN file imports for extension, and offers compatibility with BPMN standard export.

At the time of evaluation, certain BPMN elements were not fully supported by the system's layout engine, including lanes (for representing organizational roles), message flows (for inter process communication), and some annotation types. These limitations were communicated to participants as part of the evaluation setup.

[1] https://www.kicopro.com/

4 Mixed-Method Evaluation

We conducted a mixed-methods evaluation combining a focus group workshop with standardized questionnaires to capture both qualitative insights into expert users' experiences and quantitative measures of usability, trust, and task-specific performance. This methodological triangulation follows established practices in HCI research for evaluating complex interactive systems [4], where neither purely qualitative nor purely quantitative approaches alone can capture the full picture of user experience. The qualitative component enables deep exploration of user perceptions, challenges, and envisioned use cases, while the quantitative component provides standardized measures that enable comparison with benchmarks and other systems.

4.1 Participants

Five process modeling experts participated in the evaluation ($n = 5$), an overview is given in Table 1. All participants worked extensively with BPMN models in their professional roles within the same organization, thereby representing a homogeneous sample of domain experts. Self-reported engagement with process models was mostly several days per week, indicating that process modeling constituted a core professional activity for all participants. Participant ages ranged across four decades (20s through 50s), representing diverse career stages. All participants reported prior experience with chatbot systems, with an average chatbot usage frequency of 2 to 3 days per week, indicating familiarity with conversational AI interfaces.

Table 1: Participant characteristics ($n = 5$).

ID	Age Group	BPMN Work (days/week)	Chatbot Exp.	Chatbot Use (days/week)
P1	20-30	<1	Yes	2-3
P2	20-30	2-3	Yes	2-3
P3	20-30	<1	Yes	2-3
P4	31-40	4-5	Yes	3-5
P5	51-60	4-5	Yes	<1

Focusing on expert rather than novice users was a deliberate methodological choice aligned with our research questions. Experts identify subtle quality issues missed by novices, provide essential oversight for organizational adoption, and articulate workflow integration needs.

4.2 Procedure

The evaluation was conducted in several phases, with a structured focus-group workshop lasting approximately three hours as the main component. It followed established guidelines for focus groups with expert participants [14, 21].

Phase 1 - Kickoff (remote, 1 hour) Participants received a brief kickoff session introducing the tool and its functionalities. They were then asked to share their expectations for a modeling chatbot without prior exposure to the prototype. This phase served dual purposes: orienting participants to the evaluation context and capturing baseline expectations prior to hands-on interaction.

Phase 2 - Hands-on self exploration (2 weeks): Following the kickoff, participants engaged in semi-structured, extensive hands-on exploration of our proposed solution. The participants were provided with two representative process descriptions, of varying complexity, to model, but were also encouraged to explore freely using their own process descriptions and modification requests.

Phase 3 - Focus Group (In-Person, 3 1/2 hours): We conducted a semi-structured focus group where participants reflected on their experiences with our proposed solution, addressing interface usability, BPMN quality, limitations, improvements, and professional use cases. Discussions were documented through detailed observer notes and participant contributions on a shared whiteboard.

Phase 4 - Questionnaires (30 minutes): The participants completed standardized online questionnaires after the focus group, assessing usability, trust, and task-specific performance based on their overall hands-on experience.

4.3 Instruments

4.3.1 Chatbot Usability Questionnaire (CUQ). We employed the 16-item Chatbot Usability Questionnaire (CUQ) [9] to evaluate perceived usability of the conversational BPMN copilot. As CUQ is specifically designed for chat-based interfaces, it overcomes limitations of traditional scales like System Usability Scale (SUS) [3] by addressing conversational characteristics across five dimensions: personality and engagement, onboarding and welcome, navigation and ease of use, understanding and communication, and error handling and overall assessment. Participants responded on a 5-point Likert scale (1 = strongly disagree to 5 = strongly agree). Eight negative items were reverse-scored before being aggregated into a 0–100 score, with higher values indicating superior usability. Through validation studies [8], a benchmark mean of 68/100 was established.

4.3.2 Trust Assessment. Trust was measured using an 8-item scale adapted from the Trust Scale for the AI Context (TAI) [7, 26], which has been validated for measuring trust in AI-based chatbot systems with strong psychometric properties. The scale captures multiple dimensions of trust in AI systems: confidence in functionality, predictability, reliability, security, efficiency, suspicion, comparative competence, and decision support. Each item was rated on a 5-point Likert scale, with the total score converted to a percentage scale (range: 0-100%) for interpretability.

4.3.3 Tool-Specific Quality Assessment. We developed an eight-item questionnaire to assess BPMN model quality and task-specific capabilities not captured by generic measures. The items covered understanding of textual descriptions, representation of key activities, correct sequencing, structural clarity, implementation of modifications, process explanations, professional suitability, and the level of detail. The responses, given on a five-point Likert scale, were converted to percentages. Additionally, two open-ended questions captured positive and negative use cases: "I like using the tool for..." and "The tool did not achieve good results in the following use cases:...".

4.4 Data Analysis

All quantitative questionnaire data were analyzed descriptively (means, ranges, distributions, and standard deviation). Given the small sample size ($n = 5$), we report descriptive statistics following study recommendations [18]. The small sample precludes claims about statistical significance but enables identification of patterns.

Qualitative data from the focus group, workshop documentation, and open-ended questionnaire responses were analyzed using thematic analysis following Braun and Clarke's six-phase approach [2]. The initial coding captured discrete observations, which were iteratively clustered into candidate themes based on conceptual similarity. Those themes were refined against the full dataset for coherence and organized around the study's three research questions, highlighting both common patterns and distinctive perspectives.

5 Results

5.1 Qualitative Results

Thematic analysis of focus group discussions, workshop documentation, and open-ended questionnaire responses yielded seven major themes organized around the three research questions, revealing seven major themes, as well as five potential use cases.

RQ1: Usability and Design Perceptions

Theme 1: Intuitive Interface, Opaque Prompting. The interface was praised as simple and intuitive, with familiar chat-based operations (text input, process model viewing, and new conversations) posing no difficulties. However, a prompting paradox revealed a gulf of execution: users understood the overall goal of generating BPMN models from text, but were unsure how to formulate prompts that would reliably produce useful results. They struggled with input formulation, e.g., balancing detail vs. conciseness, structuring multi part processes, and including domain context, which hindered the adoption. Error handling exacerbated this, as participants lacked remediation guidance for layout errors or suboptimal outputs.

Theme 2: Response Latency as Workflow Barrier. The process modeling time was perceived as overly long, disrupting the iterative workflow essential to professional process modeling. Participants reported that input output latency impeded rapid iteration for refinement and exploration, disrupting the process modeler's cognitive rhythm.

RQ2: Strengths and Weaknesses of LLM-Generated BPMNs

Theme 3: Varying Output Quality Participants report that the quality of generated BPMN models is sometimes insufficient. In particular, longer process descriptions tend to yield lower-quality BPMN models that capture only a subset of the described process.

Theme 4: Chunking as Emergent Coping Strategy. Participants discovered chunking as a coping strategy to improve output quality for long descriptions, with iterative, piecewise prompting outperforming comprehensive inputs. However, this strategy increased cognitive load, as users had to segment their mental model of the process into manageable fragments, maintain coherence across iterations, and mentally reassemble the resulting model.

Theme 5: Imprecise Modifications. Modification requests proved unreliable due to incorrect implementations. Key issues included unsupported BPMN elements (lanes, annotations) without notification, with participants stressing that complete palette coverage is needed. Additionally, modification requests often trigger issues such as insertions with unexpected connections or unintended modifications to previously unchanged parts.

Theme 6: Absent Clarification Dialogue. It was noted that the LLM failed to ask clarifying questions for ambiguous inputs. Professional process modeling requires an iterative dialogue to resolve descriptive gaps, a capability absent in the system. This lack of clarifying interaction leads the LLM to generate complete outputs from incomplete inputs via implicit rather than explicit assumptions.

Theme 7: Convention Violations. Participants observed that process modeling conventions were not followed, including both BPMN 2.0 standards (e.g., a task has one incoming and one outgoing sequence flow) as well as organization-specific guidelines (e.g., labels need to follow a specific pattern) intended to ensure consistency across an enterprise process portfolio. They expected the tool to be configurable with enterprise conventions and capable of performing quality assurance against them.

RQ3: Envisioned Use Cases

Analysis of open-ended responses and focus group discussions revealed five distinct use cases that the participants envisioned for LLM-based process modeling tools.

Use Case 1: Support for Domain Experts (Non-Modelers). The system aids BPMN-inexperienced domain experts by generating diagrams from natural language, enabling specialist refinement or direct documentation, also addressing the "blank page problem" by providing an initial structure. However, it was stressed that non-experts need higher output quality, as they cannot readily detect errors, unlike skilled process modelers.

Use Case 2: Quality Assurance Bot. The copilot could validate process models against BPMN standards and organizational conventions, while automatically correcting identified issues.

Use Case 3: Process Creation from Visual Inputs. A key workshop requirement was image-to-BPMN conversion, covering photographs of hand-drawn sketches. This addresses the common need to formalize informal process model sketches.

Use Case 4: Enterprise-Integrated Local Deployment. A local LLM variant was envisioned, which is trained on the organization's process portfolio, enabling the recognition of existing processes, pattern-based suggestions, enterprise-wide consistency checks, and subprocess reuse.

Use Case 5: Process Optimization and Support. The system should support process improvement by identifying optimization potential, suggesting enhancements, and asking probing questions. This extends beyond modeling to active process analysis, leveraging LLM reasoning to detect inefficiencies human modelers might miss.

Negative Use Cases. Participants also clearly identified contexts in which the current system performed poorly, thereby highlighting its present limitations. Long and complex process descriptions often led to unsatisfactory results (P1, P3, P4), and reconstructing already known processes from memory also proved challenging (P5). Moreover, for experts working on complex processes, the tool was perceived as slowing work down rather than accelerating it (P2). These negative cases help bound expectations for current capabilities and identify priority areas for improvement.

5.2 Quantitative Results

This section summarizes the results from the questionnaires concerning usability, trust, and task-specific quality. Table 2 summarizes the scores across all assessment dimensions.

Table 2: Summary of quantitative assessment scores.

Measure	Mean	SD	Min	Max	Ref.
Usability (CUQ)	67.2	7.1	56.25	73.44	68.0†
Trust	48.8	10.5	31.25	59.38	60.0‡
Task Quality	54.4	17.8	25.0	65.63	–

†CUQ benchmark. ‡Suggested acceptable threshold.

5.2.1 Chatbot Usability (CUQ). The mean CUQ score across all participants was 67.2 ($SD = 7.1$), falling marginally below the established benchmark of 68 for chatbot usability. Individual scores ranged from 56.25 to 73.44, indicating moderate variability in usability perceptions across participants. The median score (70.31) slightly exceeded the mean, suggesting one lower outlier.

Table 3: Chatbot Usability Questionnaire (CUQ) item-Level results.

Item	Statement	Mean	SD	Int.	Critical
Q1 (+)	Realistic personality	3.0	1.00	○	
Q2 (–)	Robotic behavior	2.4	0.89	+	
Q3 (+)	Welcoming onboarding	**4.4**	0.55	+	
Q4 (–)	Unfriendly behavior	**1.2**	0.45	++	
Q5 (+)	Clear purpose explanation	2.2	0.45	-	↯
Q6 (–)	No purpose explanation	3.2	1.30	○	
Q7 (+)	Easy to use	**4.0**	0.71	+	
Q8 (–)	Confusing to use	2.4	1.34	+	
Q9 (+)	User input understanding	3.2	1.30	○	
Q10 (–)	Input recognition failure	2.6	0.89	○	
Q11 (+)	Useful responses	**3.8**	0.45	+	
Q12 (–)	Irrelevant responses	**1.4**	0.55	++	
Q13 (+)	Good error handling	3.4	1.14	○	
Q14 (–)	Unable to handle errors	1.8	1.30	-	↯
Q15 (+)	Overall ease of use	**4.0**	1.22	+	
Q16 (–)	Complex usage	2.0	0.71	+	

Scale: 1 = strongly disagree, 5 = strongly agree. Item: (+) = positive; (-) = negative.
Int.: ++ = highly positive; + = positive; ○= neutral; - = negative; – = highly negative;

Analysis of individual CUQ items, shown in Table 3, revealed important patterns in usability perceptions. The highest-rated items related to the initial user experience and basic interaction mechanics. Participants found the chatbot welcoming at first setup (Q3: $M = 4.4$, $SD = 0.55$), easy to use (Q7: $M = 4.0$, $SD = 0.71$; Q15: $M = 4.0$, $SD = 1.22$), and not unfriendly (Q4: $M = 1.2$, $SD = 0.45$) or excessively complex (Q16: $M = 2.0$, $SD = 0.71$). Responses were generally perceived as useful, appropriate, and informative (Q11: $M = 3.8$, $SD = 0.45$), and the system was not seen as producing irrelevant responses (Q12: $M = 1.4$, $SD = 0.55$).

However, participants showed uncertainty about the system's transparency: the item "clear purpose explanation" scored below neutral (Q5: $M = 2.2$, $SD = 0.45$), while "no purpose explanation" exhibited high variability (Q6: $M = 3.2$, $SD = 1.30$), indicating disagreement. This suggests that despite ease of use, users lacked clarity on system capabilities, which aligns with qualitative findings. The understanding of user input showed mixed perceptions (Q9: $M = 3.2$, $SD = 1.30$), with uncertainty about input recognition failures (Q10: $M = 2.6$, $SD = 0.89$). This high variability suggests that comprehension depends on input characteristics or strategies. Also, error handling received moderate ratings (Q13: $M = 3.4$, $SD = 1.14$; Q14: $M = 1.8$, $SD = 1.30$), again with notable variability.

5.2.2 Trust Assessment. The mean trust score, computed as the average of participants' aggregated item scores, was 48.8% ($SD = 10.5\%$), indicating moderate trust in the system. However, this score falls notably below a suggested threshold of 60% for acceptable trust and substantially below the usability score, revealing a usability-trust gap. Individual scores ranged from 31.25% to 59.38%, with participant P5 showing notably lower trust than others.

Table 4: Trust scale item-Level results.

Item	Statement	Mean	SD	Int	Critical
Q1 (+)	Functions as intended	3.6	0.55	+	
Q2 (+)	Predictable results	2.4	0.55	○	
Q3 (+)	**Correctness reliability**	**1.8**	**0.45**	-	↯
Q4 (+)	Safe to rely on	2.4	0.55	○	
Q5 (+)	Operational efficiency	3.2	0.84	○	
Q6 (-)	System suspicion (rev.)	3.8	1.10	+	
Q7 (+)	Outperforms novice human	3.8	0.84	+	
Q8 (+)	Suitable for decision-making	2.6	0.89	○	

Scale: 1 = strongly disagree, 5 = strongly agree. Item: (+) = positive; (-) = negative.
Int.: ++ = highly positive; + = positive; ○= neutral; - = negative; – = highly negative;

Item-level analysis in Table 4 revealed critical patterns in trust formation. The participants expressed moderate confidence that the system functions well overall (Q1: $M = 3.6$, $SD = 0.55$) and believed the system could perform the task better than an inexperienced human user (Q7: $M = 3.8$, $SD = 0.84$).

However, critical trust dimensions received notably lower ratings. Reliability received the lowest rating (Q3: $M = 1.8$, $SD = 0.45$), indicating that the participants could not consistently rely on correct outputs. As this showed the lowest variability, it suggests consensus. Predictability was also rated low (Q2: $M = 2.4$, $SD = 0.55$), as was the sense of security when relying on the system (Q4: $M = 2.4$, $SD = 0.55$). The system's suitability for decision-making received moderate ratings (Q8: $M = 2.6$, $SD = 0.89$).

Notably, participants showed low suspicion of the system (Q6: $M = 3.8$, $SD = 1.10$) and moderate efficiency ratings (Q5: $M = 3.2$, $SD = 0.84$). This indicates no distrust of malicious intent but rather concerns about output reliability and correctness, revealing a nuanced trust profile where the system is seen as well-intentioned but inconsistent.

5.2.3 Tool-Specific Quality Assessment. The mean score for the tool-specific quality assessment was 54.4% ($SD = 17.8\%$), indicating moderate perceived quality of generated BPMN models. Individual scores ranged widely from 25.0% to 65.63%, with participant P5 rating quality substantially lower than others. As it was the same

participant who showed lowest trust and usability scores, it suggests consistency in that individual's negative experience.

Table 5: Tool-specific quality assessment results.

Item	Statement	Mean	SD	Int.	Critical
Q1 (+)	Text understanding	2.8	1.30	-	↯
Q2 (+)	Process representation	3.0	0.71	○	
Q3 (+)	Correct flow order	3.4	1.34	-	↯
Q4 (+)	Structural clarity	3.2	0.84	○	
Q5 (+)	Modification handling	3.2	1.30	-	↯
Q6 (+)	Description correctness	3.4	0.55	○	
Q7 (+)	Application suitability	3.4	1.34	-	↯
Q8 (+)	**Level of detail**	**3.8**	**0.45**	+	

Scale: 1 = strongly disagree, 5 = strongly agree. Item: (+) = positive; (-) = negative.
Int.: ++ = highly positive; + = positive; ○= neutral; - = negative; – = highly negative;

The item-level analysis (Table 5) revealed that the level of detail received the highest rating (Q8: $M = 3.8$, $SD = 0.45$), suggesting that generated process models captured appropriate granularity. The system's ability to provide correct descriptions of modeled processes was also rated moderately well (Q6: $M = 3.4$, $SD = 0.55$), as were activity flow and ordering (Q3: $M = 3.4$, $SD = 1.34$) and overall application appropriate quality (Q7: $M = 3.4$, $SD = 1.34$).

Text understanding received low ratings with high variability (Q1: $M = 2.8$, $SD = 1.30$), indicating inconsistent input comprehension. This suggests comprehension depends on input characteristics or participant discovered interaction strategies. Also, modification handling showed similar patterns (Q5: $M = 3.2$, $SD = 1.30$), indicating variable success for change requests.

6 Discussion

Our results show that human-centered evaluation reveals system quality aspects beyond automated metrics [6, 12]. Specifically, our study uncovered a usability trust gap: a pleasant interaction does not guarantee confidence in output reliability for professional use. This observation aligns with trust calibration research distinguishing affective responses from cognitive assessments of system capability [17]. These findings, together with the other results reported in this study, motivate design implications that can only be identified through human-centered evaluation. Accordingly, future systems should be assessed using both automated benchmarks for scalability and human evaluation to surface interaction breakdowns, trust needs, and organizational fit that inform design.

6.1 Design Implications

Prompting Guidance. The prompting paradox indicates a need for explicit input formulation support by making system expectations transparent, thus reducing the gulf of execution [23]. This could include example conversations, templates for common process types and documentation of what input features correlate with good output quality.

Progressive Disclosure and Chunking Support. The success of chunking suggests designing systems that can incrementally construct a process model from a large process description, rather than requiring users to chunk the input themselves, thereby reducing users' cognitive load and improving outcomes. This could include staged information prompts, piece-by-piece building with intermediate validation, and state visualization.

Proactive Clarification. Future systems should detect input ambiguities and ask clarifying questions before generating output.

Convention Configuration. Organizational conventions necessitate customization capabilities for professional deployment. Systems should encode conventions, validate outputs against them, and learn from existing process portfolios.

Confidence Communication. Given reliability concerns, systems should communicate output confidence, highlight uncertainties or likely errors, and direct review to problematic elements. This transparency supports appropriate trust calibration [17].

Latency Management. Response time concerns suggest the need for progress indication during generation, streaming output where feasible, and interaction design that accommodates processing delays without disrupting cognitive flow.

6.2 Limitations

Several limitations constrain our findings' generalizability. The small ($n = 5$), single-organization expert sample limits applicability to novices and other contexts. Moreover, our expert focus limits conclusions about democratization potential. The evaluation captured a snapshot of an evolving system, in which technical limitations (e.g., missing lanes) have since been addressed. Prior chatbot experience may have shaped participant expectations differently than for first-time users. Also, the workshop format may not reflect extended use patterns. Despite these, our qualitative insights complement automated evaluations and inform larger, more diverse studies.

7 Conclusion

We conducted a mixed methods evaluation of our LLM-based conversational BPMN modeling system, with five domain experts using a focus group workshop and questionnaires. Our quantitative analysis revealed a usability trust gap: although usability scores approached established benchmarks, trust in output reliability lagged, with reliability emerging as the primary concern. The qualitative analysis identified seven key themes in experts' experiences, including the prompting paradox (knowing the tool's function but not its effective use), output quality issues, particularly for long textual descriptions, and an absent clarifying dialogue. Participants envisioned use cases, including initial drafts for domain experts, quality assurance of existing models, and image-to-BPMN support.

Our work shows that human-centered evaluation reveals critical LLM toolaspects of LLM tool quality, such as trust calibration, interaction challenges, coping strategies, and alignment with professional practice automated benchmarks. Therefore, we argue, that a comprehensive evaluation must consist of an automated technical assessment combined with a human-centered practical investigation. Understanding experts' experiences, trust, and adaptation is essential for practical adoption, as these tools are intended for use in professional domains.

Acknowledgments

Parts of this work were conducted within the projects KICoPro (FKZ: 01IS24053C) and EINHORN (FKZ: 01IS24048C), funded by the Federal Ministry of Research, Technology and Space (BMFTR).

References

[1] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. 1–13. doi:10.1145/3290605.3300233

[2] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3 (2006), 77–101. Issue 2. doi:10.1191/1478088706QP063OA

[3] John Brooke. 1996. SUS: A 'Quick and Dirty' Usability Scale. *Usability Evaluation in Industry* (1996), 189–194.

[4] John W. Creswell and Vicki L. Plano Clark. 2017. *Designing and Conducting Mixed Methods Research* (3rd ed.). SAGE.

[5] Ewart J. de Visser, Marieke M. M. Peeters, Malte F. Jung, Spencer Kohn, Tyler H. Shaw, Richard Pak, and Mark A. Neerincx. 2020. Towards a Theory of Longitudinal Trust Calibration in Human–Robot Teams. *International Journal of Social Robotics* 12 (2020), 459–478. doi:10.1007/s12369-019-00596-x

[6] Panagiotis Drakopoulos, Panagiotis Malousoudis, Nikolaos Nousias, George Tsakalidis, and Kostas Vergidis. 2026. Do LLMs Speak BPMN? An Evaluation of Their Process Modeling Capabilities Based on Quality Measures. *Computation* 14, 1 (2026), 10. doi:10.3390/computation14010010

[7] Robert R. Hoffman et al. 2023. *Metrics for Explainable AI.* CRC Press.

[8] Samuel Holmes, Raymond Bond, Anne Moorhead, Jane Zheng, Vivien Coates, and Michael McTear. 2023. Towards Validating a Chatbot Usability Scale. In *Design, User Experience, and Usability. HCII 2023 (Lecture Notes in Computer Science, Vol. 14033)*. Springer, 321–339. doi:10.1007/978-3-031-35708-4_24

[9] Samuel Holmes, Anne Moorhead, Raymond Bond, Huiru Zheng, Vivien Coates, and Michael McTear. 2019. Usability Testing of a Healthcare Chatbot: Can We Use Conventional Methods to Assess Conversational User Interfaces?. In *Proceedings of the 31st European Conference on Cognitive Ergonomics (ECCE '19)*. 207–214. doi:10.1145/3335082.3335094

[10] Nataliia Klievtsova, Janik Bezin, Timotheus Kampik, Jürgen Mangler, and Stefanie Rinderle-Ma. 2023. Conversational Process Modelling: State of the Art, Applications, and Implications in Practice. In *Business Process Management Forum*. Springer, 319–336. doi:10.1007/978-3-031-41623-1_19

[11] Julius Köpke and Aya Safan. 2024. Introducing the BPMN-Chatbot for Efficient LLM-Based Process Modeling. In *Proceedings of the BPM 2024 Demos & Resources Forum (CEUR Workshop Proceedings, Vol. 3758)*. 86–90.

[12] Humam Kourani, Alessandro Berti, Daniel Schuster, and Wil M.P. Van der Aalst. 2025. Evaluating Large Language Models on Business Process Modeling: Framework, Benchmark, and Self-Improvement Analysis. *Software and Systems Modeling* (2025). doi:10.1007/s10270-025-01318-w

[13] Humam Kourani, Alessandro Berti, and Wil M. P. van der Aalst. 2024. Process Modeling With Large Language Models. *arXiv preprint arXiv:2403.07541* (2024).

[14] Richard A. Krueger and Mary Anne Casey. 2014. *Focus Groups: A Practical Guide for Applied Research* (5th ed.). SAGE.

[15] Chantale Lauer, Peter Pfeiffer, Alexander Rombach, and Nijat Mehdiyev. 2025. Conversational Business Process Modeling using LLMs: Initial Results and Challenges. In *EMISA 2025 Proceedings*. GI.

[16] Chantale Lauer, Peter Pfeiffer, Alexander Rombach, and Nijat Mehdiyev. 2026. Assessing the Business Process Modeling Competences of Large Language Models. arXiv:2601.21787 [cs.SE] https://arxiv.org/abs/2601.21787

[17] John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors* 46, 1 (2004), 50–80. doi:10.1518/hfes.46.1.50_30392

[18] James R. Lewis and Jeff Sauro. 2016. *Quantifying the User Experience (2nd ed.).*

[19] Q. Vera Liao and Kush R. Varshney. 2021. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. *arXiv preprint arXiv:2110.10790* (2021).

[20] Josip Tomo Licardo, Nikola Tankovic, and Darko Etinger. 2025. BPMN Assistant: An LLM-Based Approach to Business Process Modeling. *arXiv preprint arXiv:2509.24592* (2025).

[21] David L. Morgan. 1996. Focus Groups. *Annual Review of Sociology* 22, 1 (1996), 129–152. doi:10.1146/annurev.soc.22.1.129

[22] Nguyen, Michelle Hoang and Sedoc, João and Taylor, Casey Overby. 2024. Usability, Engagement, and Report Usefulness of Chatbot-Based Family Health History Data Collection: Mixed Methods Analysis. *Journal of Medical Internet Research* 26 (2024), e55164. doi:10.2196/55164

[23] Don Norman. 2013. *The Design of Everyday Things: Revised and Expanded Edition.* Basic Books.

[24] Object Management Group. 2014. *Business Process Model and Notation (BPMN), Version 2.0.2.* Technical Report. OMG. https://www.omg.org/spec/BPMN/2.0.2

[25] Laura Sánchez-González, Félix García Rubio, Francisco Ruiz González, and Mario Piattini Velthuis. 2010. Measurement in business processes: a systematic review. *Business Process Management Journal* (2010), 114–134. doi:10.1108/14637151011017976

[26] Nicolas Scharowski, Sebastian A.C. Perrig, Nick von Felten, Lena Fanya Aeschbach, Klaus Opwis, Philipp Wintersberger, and Florian Brühlmann. 2025. To Trust or Distrust AI: A Questionnaire Validation Study. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. 361–374.

[27] Wil M. P. van der Aalst. 2013. Business Process Management: A Comprehensive Survey. *ISRN Software Engineering* 2013 (2013), 1–37. doi:10.1155/2013/507984

[28] Maxim Vidgof, Stefan Bachhofner, and Jan Mendling. 2023. Large language models for business process management: Opportunities and challenges. In *International conference on business process management*. Springer, 107–123.

[29] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-Examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. 1–13. doi:10.1145/3313831.3376301