This document is part of the supplementary material for the paper titled 'LLM-Enabled Business Process Coherence Checking Based on Multi-Level Process Documentation' by Schulte, M.; Franzoi*, S.; Köhne, F.; vom Brocke, J. submitted for publication to the journal Process Science (https://link.springer.com/journal/44311).

Section A:

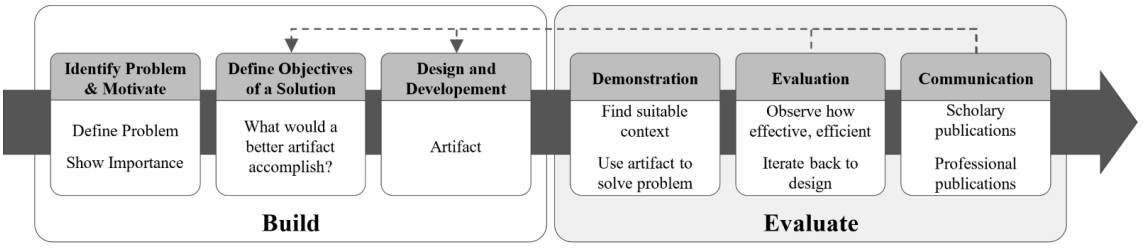

Fig. 1 Design Science Research, adapted from Peffers et al. (2007) and Sonnenberg and vom Brocke (2012)

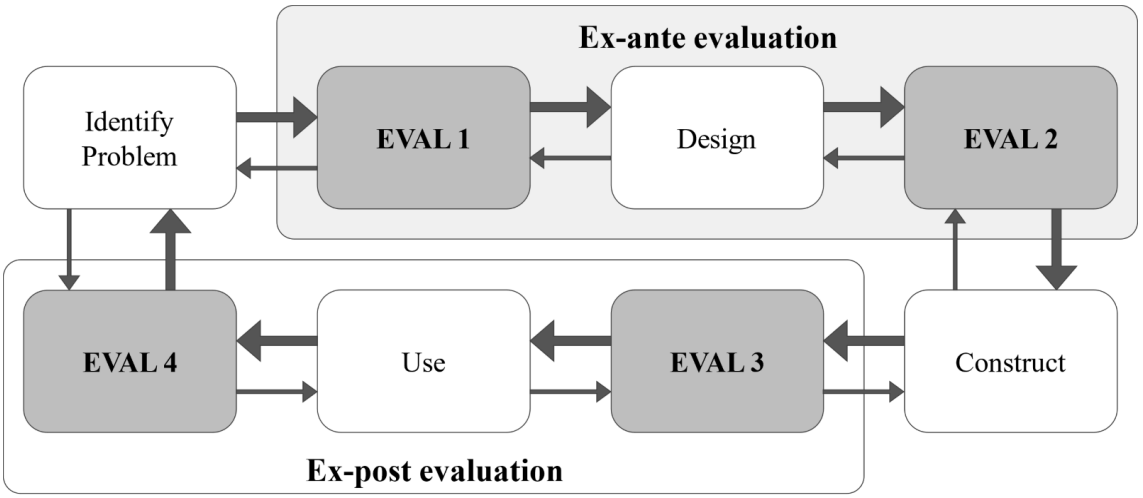

Fig. 2 Evaluation Activities within the DSR Process adapted from Sonnenberg and vom Brocke (2012)

Interview Method

The methodological design of the semi-structured interviews, a key component of the ex-ante evaluation phase, is designed with flexibility in mind to allow for the collection of rich, in-depth insights. The semi-structured format allowed for a natural flow of dialogue, structured by an interview guide, yet open enough to allow for adaptation and exploration based on participants' responses (Kallio et al., 2016). This approach is instrumental in assessing both the theoretical soundness and practical applicability of the artifact. In addition, the interviews, characterised as both artificial due to their controlled environment for demonstrating the PoC and formative in their intention to refine the artifact, aimed to elicit qualitative insights, complemented by quantitative assessments via a Likert scale on pre-defined criteria (Venable et al., 2016).

Participants in these interviews were selected from German-speaking BPM researchers and process owners, reflecting the language considerations necessary for accurate communication and understanding, so all interviews were conducted in German. This setting provided an authentic context for feedback and ensured comfort and clarity in discussing complex BPM concepts and artifact functionality.

The procedure for these interviews followed a structured yet adaptive format, following the research on semi-structured interviews conducted by Kalio et al. (2016). Beginning with obtaining consent for the recording and use of the data, the sessions included an initial explanation of the paper concept and an overview of the

intended functionality and design of the artifact. This was followed by a presentation of the design objectives and a demonstration of the proof of concept based on data from literature. The demonstrated proof of concept is artificial in nature and was conducted in a controlled environment (Venable et al., 2016).

Furthermore, the sessions were distinctly formative, as the feedback received was aimed at improving the design specifications of the artifact (Venable et al., 2016). Specifically, the evaluation criteria of understandability, feasibility, applicability and operationality were all derived from the Eval 2 phase, as outlined in the framework by Sonnenberg and Vom Brocke (2012). Participants engaged in a critical evaluation of these specifications, discussing each criterion and subsequently rating them on a Likert scale from 1 to 7. This addition of a quantitative element to the qualitative data not only ensured a robust evaluation of the artifact from both a practical and theoretical perspective, but also facilitated a dynamic interaction that could prompt immediate reconsideration or confirmation of specific design aspects.

By incorporating this methodology, the semi-structured interviews served as a central tool in bridging theoretical research and practical application, allowing for a rich, multi-faceted evaluation of the artifact's potential impact and effectiveness in real-world BPM settings. This approach underscored the importance of engaging both researchers and practitioners in discussions that challenge and refine the artifact, ultimately enhancing its relevance and utility in improving BPM practices.

Experiment Method

The methodological design of the controlled experiment, essential for the ex-post evaluation phase, is designed to rigorously evaluate the performance of the artifact. Guided by the principles of Engström et al. (2020), the experiment followed an iterative process focused on optimisation without direct stakeholder involvement, ensuring unbiased empirical validation. Quantitative in nature, the experiment involved multiple iterations to measure and improve the performance of the artifact at different stages of optimisation. Anchored in Sonnenberg and vom Brocke (2012) ex-post evaluation framework, the controlled environment facilitated a thorough and systematic evaluation.

The data for the experiment was obtained from a selection of the dataset created by Sànchez-Ferreres et al. (2018), as described in Section E ‘Dataset Creation’. This dataset was enriched by a BPM expert, who made specific changes to the BPMN documents and documented them in a structured manner according to the Business Process Change Classification Framework introduced in Section C. This enrichment created a robust ground truth for the experiment, providing a solid basis for evaluating the performance of the artifact across different iterations of the software.

The experimental design followed the principles of artificial and formative evaluation as defined by Venable et al. (2016), ensuring precise control. Following Engström et al. (2020), the experiment was conducted under controlled conditions, allowing for rigorous empirical validation. Quantitative metrics were collected during each iteration, with each iteration focusing on a specific optimisation technique, such as BPMN-specific optimisations or prompt engineering techniques. This structured approach enabled a comprehensive assessment of the artifact's readiness for real-world application.

Throughout this iterative process, the evaluation criteria of efficiency, effectiveness and robustness, which are derived from Sonnenberg and vom Brocke (2012), were consistently applied. These criteria are further explained and defined in Section E. The analysis of quantitative data from the performance metrics informed each round of optimisation, facilitating continuous improvement and ensuring that the artifact evolved incrementally based on empirical evidence. This iterative approach maintained alignment between practical outcomes and theoretical constructs, thereby enhancing the reliability and applicability of the artifact.

Adhering to this structured and iterative experimental methodology, the artifact was continuously refined, underlining its practical utility and theoretical robustness. This rigorous evaluation process confirmed the artifact's readiness for wider use in real-world BPM environments. Through methodical iterations, the artifact was fine-tuned to deliver improved performance and reliability, making it well prepared for practical deployment.

This document is part of the supplementary material for the paper titled 'LLM-Enabled Business Process Coherence Checking Based on Multi-Level Process Documentation' by Schulte, M.; Franzoi*, S.; Köhne, F.; vom Brocke, J. submitted for publication to the journal Process Science (https://link.springer.com/journal/44311).

Focus Group Method

The focus group method used in this paper serves as a qualitative, ex-post evaluation to assess the performance and utility of the artifact in real-world settings as part of the Eval 4 phase as introduced by Sonnenberg and vom Brocke (2012). In line with the DSR approach, the focus group discussions aim to confirm findings and gather in-depth insights that complement the quantitative results obtained from the controlled experiments. This method follows the principles outlined by Tremblay et al. (2010) for confirmatory focus groups, which are particularly useful for validating the practical applicability and significance of the artifact. In addition, this approach follows the best practices outlined by Schulze et al. (2023) and complements the qualitative semi-structured interviews conducted earlier in the research.

Two focus groups were conducted, each comprising 4-5 participants, with each session lasting approximately one hour. According to Tremblay et al. (2010), this sample size and duration is sufficient to provide a robust basis for evaluating the utility of the artifact. The purpose of conducting two separate focus groups is to ensure the inclusion of a variety of perspectives, thereby increasing the comprehensiveness of the evaluation. Participants were carefully selected from experienced BPM and AI consultants with academic backgrounds ranging from bachelor's degrees to professorships. This composition of participants ensured that the discussions captured a wide range of viewpoints and expertise, which is critical for validating the generalisability and practical utility of the artifact.

To overcome spatial and temporal constraints, the focus groups were conducted online. This approach has several advantages, as highlighted by Schulze et al. (2023): it increases participation by being more convenient for participants, reduces costs and time, and provides better data recording opportunities while promoting inclusivity. The flexibility of online settings allows for a more robust and comprehensive evaluation, allowing participants to fully engage without the logistical challenges associated with face-to-face meetings. A potential disadvantage of online focus groups is the reduction in social interaction, which can hinder the depth and complexity of discussion and the natural ease of communication that is more easily achieved in face-to-face meetings.

In these focus groups, the functionality of the artifact was demonstrated using naturalistic data from diverse industries. Four different use cases were demonstrated, each involving a variety of naturalistic datasets. This demonstration highlighted not only the effectiveness of the artifact, but also its generalisability across different documentation formats and industrial contexts. By providing concrete examples, the demonstrations facilitated a clearer understanding of the potential applications and benefits of the artifact in real-world business process management scenarios.

The evaluation criteria for the focus groups are derived from the Eval 4 phase described by Sonnenberg and vom Brocke (2012) and include fidelity with real world phenomenon, impact on artifact environment and user, and applicability. These criteria ensure a comprehensive assessment of the artifact's relevance and effectiveness in practice. Within the focus groups, each participant independently and anonymously ranked these evaluation criteria on a Likert scale from 1 to 7 using a simple online survey. A more detailed explanation of these criteria can be found in Section E.

The focus group discussions were designed to evaluate the artifact in the context of the three realities proposed by Sun and Kantor (2006): real tasks, real systems, and real users. This approach ensures that the evaluation is grounded in real-world conditions and provides insights that are both applicable and valuable to the ongoing refinement of the artifact.

By integrating the focus group method into the evaluation strategy, this paper aims to achieve a balanced and thorough evaluation, incorporating both quantitative data from controlled experiments and qualitative insights from real-world interactions. This comprehensive approach ensures that the artifact is not only theoretically robust, but also practically viable and impactful in diverse business process management environments.