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 E:

Dataset Creation. For the empirical validation in this thesis, specifically during the experiment phase as defined in the DSR evaluation process, a comprehensive dataset from multiple process repositories was used. This novel Process Coherence Checking Dataset is central to assessing the effectiveness and robustness of LLMs in identifying and verifying the coherence of multi-level process documentation.

The initial dataset, derived from the research of Sànchez-Ferreres et al. (2018)[7], includes a wide range of process models originally compiled by the BPM Academic Initiative and further detailed by Eid-Sabbagh et al. (2012). These models, enriched with textual descriptions compiled by expert process modelers, form the basis of a dataset for checking the coherence of documented business processes. In addition, model-text pairs of the Friedrich et al. (2011) dataset, noted for its use in recent studies by Grohs et al. (2023), Klievtsova et al. (2023), and Sonbol et al. (2023), are included to increase the diversity of source materials. This dataset is valued for its heterogeneous composition - including sources from the academic, industrial, educational, and public sectors.

The careful selection of these datasets allows the experiment to thoroughly test the performance of the artifact in assessing consistency and coherence, as specified in the thesis research questions. This choice of diverse datasets is essential to validate the artifact's effectiveness and adaptability in real-world BPM settings, which aligns well with the thesis' goal of improving BPM practices through innovative technology.

In its original composition, the dataset consisted of 25 model pairs from Eid-Sabbagh et al. (2012) and 49 model pairs from Friedrich et al. (2011). To maintain balance in the new dataset, an equal number of models was selected from both sources to ensure balanced representation. However, only models that could be compiled without errors using the bpmn.io VS Code plugin, as introduced in Section C, were included. In addition, the models had to be syntactically correct according to BPMN 2.0 standards[8]. To ensure that all relevant Business Process Change Dimensions could be accurately tested, only models where all dimensions were applicable, as introduced in Section C, were included. This requirement necessitated the presence of swim lanes to capture organizational change. In addition, the dataset was curated to be heterogeneous in terms of modelling style, scope and topic. Models containing artifacts in multiple languages were excluded to maintain consistency. These restrictive criteria filtered the dataset down to 12 model-text pairs, with an equal split of 6 pairs from each data source.

Once selected, the model-text pairs were organized into individual subfolders to facilitate structured and efficient analysis. Subsequently, the models were customized to simulate realistic business scenarios by a BPM expert with no prior knowledge of the change detection mechanism used, thereby ensuring that the changes were made independently and without bias.

The changes introduced were categorised by the BPM expert according to the Change Relevance Categories outlined in Section C and include all relevant change dimensions, namely task, data, control flow and organization. To reflect the inherent complexity of realistic changes, different types of changes were combined rather than isolated. In addition, one third of the dataset contained no relevant changes, but only unrelated or negligible changes, or no changes at all. This control group was essential to test the system's ability to avoid flagging non-relevant changes in accordance with the final design specification.

All changes were meticulously documented in a tabular format, which was then converted to a JSON schema for automated use in the experiment. This format ensures ease of integration and consistency of data processing.

Process Name	**Short Process Description**	**Original Data Source**	**Relevant Changes Made**
Client Acquisition	Management and Marketing Process for Client Acquisition	Eid-Sabbagh et al. (2012)	Task (x2), Control Flow (x2)
Credit	Credit Acceptance Process	Eid-Sabbagh et al. (2012)	-
Dispatching	Computer Hardware Store Goods Dispatching	Friedrich et al. (2011)	Organization (x2), Task, Control Flow
Hospital	Patient Treatment Process	Friedrich et al. (2011)	Data (x2), Organization, Task
Hotel	Hotel Room Service Process	Friedrich et al. (2011)	Data (x2), Organization, Task (x2), Control Flow
Invoice	CRM Invoicing Process	Eid-Sabbagh et al. (2012)	-
Part Production	Simple Parts Production Process in an Organization	Friedrich et al. (2011)	Organization (x2), Control Flow (x2)
Purchasing	Purchasing Process of an Organization	Eid-Sabbagh et al. (2012)	Data, Task (x4), Organization
Recourse	Insurance Claim Recourse Checking	Friedrich et al. (2011)	Control Flow (x2)
Replacement Parts	Replacement Part Procurement Process	Eid-Sabbagh et al. (2012)	Data
Web Design	Website Designing Process	Eid-Sabbagh et al. (2012)	-
Zoo	Zoo Ticketing Process	Friedrich et al. (2011)	-

Tab. 1 Experiment Data Summary.

The original dataset was released under the GNU General Public Licence, which guarantees the freedom to use, study, modify and share the data. Our customised dataset will be released under the same licence, encouraging open collaboration and ensuring that any derivative works adhere to the same copyleft terms. This licensing strategy is in line with the thesis' commitment to promoting open, innovative research in BPM.

Experiment Evaluation Criteria. In the context of the experimental validation of the artifact presented in this thesis, it is crucial to apply rigorous evaluation criteria in order to comprehensively assess its performance. The criteria adopted are derived from the Eval 3 phase proposed by Sonnenberg and vom Brocke (2012), which focuses on validation in an artificial setting. The chosen evaluation criteria are Effectiveness, Robustness and Efficiency. *Effectiveness* refers to the *accuracy* of the artifact in identifying changes compared to the ground truth established by a BPM expert. Accuracy is quantified as a percentage, with 100% indicating perfect identification of all relevant changes. The metric penalises for hallucinated changes, defined as additional changes not present in the ground truth. Similarly, failure to identify relevant changes results in a significant penalty. Correctly identifying relevant changes results in minor penalties if the dimension is different from that identified by the expert. Misclassifying negligible or unrelated changes as relevant, especially with incorrect dimensions, further reduces the accuracy score. Conversely, correctly ignoring changes that the BPM expert considers negligible or unrelated increases the accuracy percentage. This nuanced approach to scoring, developed in collaboration with the BPM expert involved in the dataset creation used in the experiment, ensures that the performance of the artifact is rigorously evaluated against a well-defined benchmark.

Robustness is assessed by examining the *consistency* of the accuracy of the artifact over multiple runs of the experiment. This criterion is measured by calculating 1 minus the standard deviation of the accuracy values obtained for the processes during the experiment. Each of the twelve processes in the dataset was subjected to five separate runs for each iteration of the experiment, resulting in a total of 60 data points per iteration. A higher percentage value, defined as 100% being the optimal value, indicates enhanced consistency and reliability in the performance of the artifact. This suggests that the results are consistent and replicable under different conditions. Assessing robustness in this way ensures that the artifact is not only effective in isolated instances, but also reliable across repeated trials, providing confidence in its applicability to real-world BPM environments.

Efficiency is a critical measure in the evaluation of the artifact, specifically evaluated by the API *cost per process* for each experimental run. This cost is directly related to the number of tokens used, assuming a single model is used throughout the evaluation. The API cost provides a proxy for resource usage, including factors such as time and computational load, as well as the environmental impact of these computations (Shekhar et al., 2024). Unlike direct measures of time, which can be affected by varying demand and load conditions external to the system, API cost provides a stable and consistent measure of efficiency. In addition, the number of tokens used is also reported to allow comparisons of efficiency in the light of fluctuating prices or newer model developments. However, price remains an important metric as it is directly related to the computational resources required, which is also strongly correlated with environmental impact.

In summary, the evaluation criteria of efficiency, effectiveness and robustness provide a comprehensive framework for validating the performance of the artifact. By rigorously applying these criteria, we can ensure that the artifact is not only theoretically sound, but also practically robust, reliable and ready for use in real-world BPM environments.

Experiment Result Analysis. The data demonstrates that the mean accuracy and consistency have increased with each iterative refinement, indicating highly effective performance. Moreover, the costs per process run were reduced by over 50% from the first to the second iteration, indicating a significant optimisation. Although there was a slight increase in cost in the third and fourth iterations due to the incorporation of additional sophisticated prompting techniques, the overall cost-effectiveness remained significantly better than the baseline established in the first iteration.

A Repeated Measures Analysis of Variance (ANOVA) was performed to statistically validate the observed improvements in accuracy (Park et al., 2009). Accuracy was chosen as the measure of statistical significance because consistency is directly derived from it and because, according to the expert interviews, it is more meaningful for artifact performance than efficiency, which directly correlates with shorter prompting techniques. This test is particularly suited to our experimental setup, which involves multiple measurements of the same dataset under different conditions. Specifically, each of the twelve process data points was tested five times over four iterations. This within-subject design requires an analysis method capable of handling the dependency and correlation within the data from each source, thus ensuring more sensitive detection of change and improved statistical power. Repeated Measures ANOVA effectively decomposes the total variance observed in the data into components due to different experimental conditions and the inherent variability of repeated measurements.

The Repeated Measures ANOVA test produces an F statistic that compares the variance between iterations with the variance within each iteration. In this study, the analysis yielded an F value of 2.9359 with 3 degrees of freedom for the numerator and 33 degrees of freedom for the denominator. A p-value of 0.0477 was obtained, indicating the probability that the observed differences occurred by chance. Given the conventional alpha level of 0.05, which is a widely accepted threshold for statistical significance, the null hypothesis - that there is no difference in accuracy across the iterations - can be rejected. This result confirms that the improvements in accuracy across iterations are statistically significant.

To further validate these findings, a Linear Mixed Effects Model was applied. This model is particularly advantageous as it takes into account both fixed effects (the iterations) and random effects (the variability within each process instance). The inclusion of random effects allows the model to deal with the hierarchical structure of the data, where measurements are nested within process instances. The linear mixed effects model gave a scale of 0.0056 and a log likelihood of 34.7307.

The scale parameter of 0.0056 indicates the within-group variance, showing a relatively low level of inherent variability in the repeated measures, implying consistent results within each iteration. The log-likelihood value of 34.7307 represents the goodness of fit of the model, with higher values indicating a better fit. The model results confirm that iteration has a significant fixed effect on accuracy, confirming that the observed accuracy improvements are not due to random variation, but rather to the systematic changes introduced in each iteration. By capturing both the fixed effects of iterations and the random effects of repeated measurements, the Linear Mixed-Effects Model provides a robust validation of the performance improvements of the artifact across iterations. The consistency between the results of the Repeated Measures ANOVA and the linear mixed effects model reinforces the robustness of the observed accuracy improvements over successive iterations. The detailed calculations for both the Repeated Measures ANOVA test and the Linear Mixed-Effects Model can be found in the codebase of the artifact.

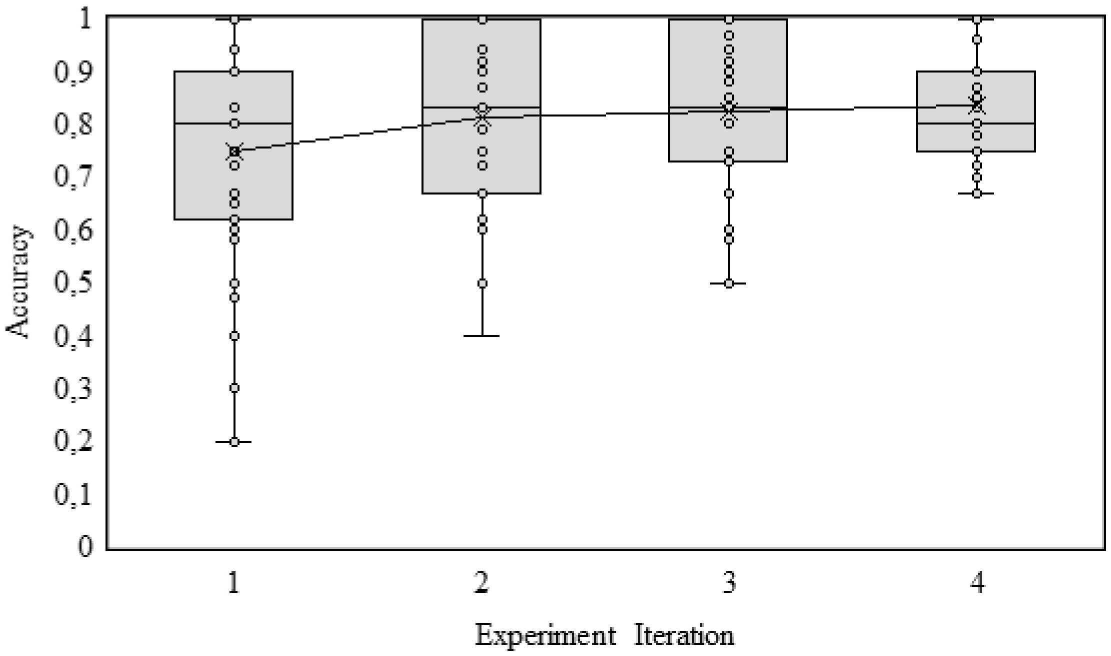

Fig. 3 Experiment Accuracy Visualization.

The accuracy of aProCheCk improved progressively over the iterations, as shown in the boxplot diagram depicted in Figure 9. The average accuracy increased consistently over the iterations, with the spread of accuracy scores decreasing, indicating more stable performance. While the median accuracy improved slightly from the first to the second and third iterations, it showed a slight decrease in the fourth iteration. This trend highlights the ability of the artifact to provide reliable results, although perfect accuracy remains unattainable due to the complex and subjective nature of coherence checking. In this context, 100% accuracy would imply an overfitting of the dataset, as coherence decisions inherently involve a degree of subjectivity.

Consistency, measured by 1 minus the standard deviation of accuracy values over multiple runs, also showed significant improvements. The average consistency increased with each iteration, starting from an already high baseline. This reduction in variability further highlights the ability of the artifact to produce reliable results under varying conditions. However, due to the non-deterministic nature of LLMs, achieving 100% consistency is not feasible. Nevertheless, the high levels of consistency observed are indicative of the robustness of aProCheCk, making it suitable for use in real-world BPM environments where consistent performance is critical.

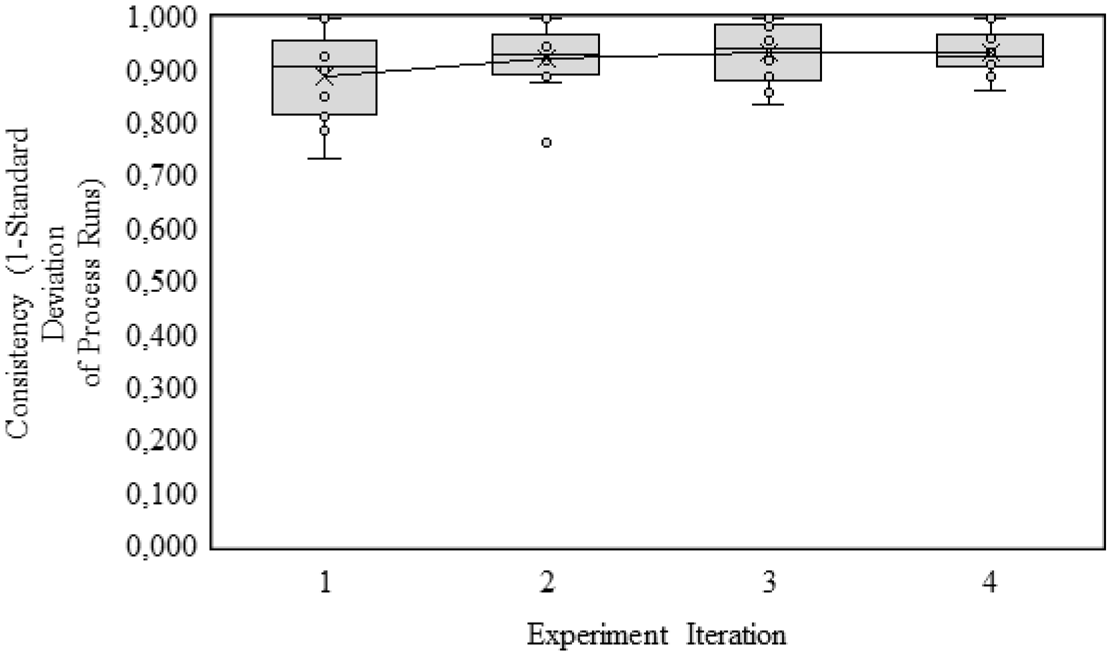

Fig. 4 Experiment Consistency Visualization.

API costs showed significant reductions, particularly between the first and second iterations. The introduction of BPMN-specific optimisations, including removal of irrelevant visual data and pre-checking for identical files after visual data removal, resulted in fewer API calls and reduced token usage. This first optimisation step significantly reduced the cost per process run, effectively halving the API cost. These changes resulted in fewer API calls overall and smaller calls with fewer tokens, improving cost efficiency.

As long as only one model is used, as in this experiment, the API costs have a strong correlation with the number of tokens used. By reducing the number of tokens per API call, the optimisation has a direct impact on the overall cost. Although the third and fourth iterations saw a slight increase in cost due to the additional computational resources required for advanced prompt engineering techniques, the overall API cost remained significantly lower than the initial baseline. For more detailed comparisons, the tokens used per iteration are shown in Fig. 24 in the Appendix. These results highlight the effectiveness of the

BPMN-specific measures in optimising token usage and reducing costs, demonstrating the efficiency of the artifact in managing computational resources for practical use.

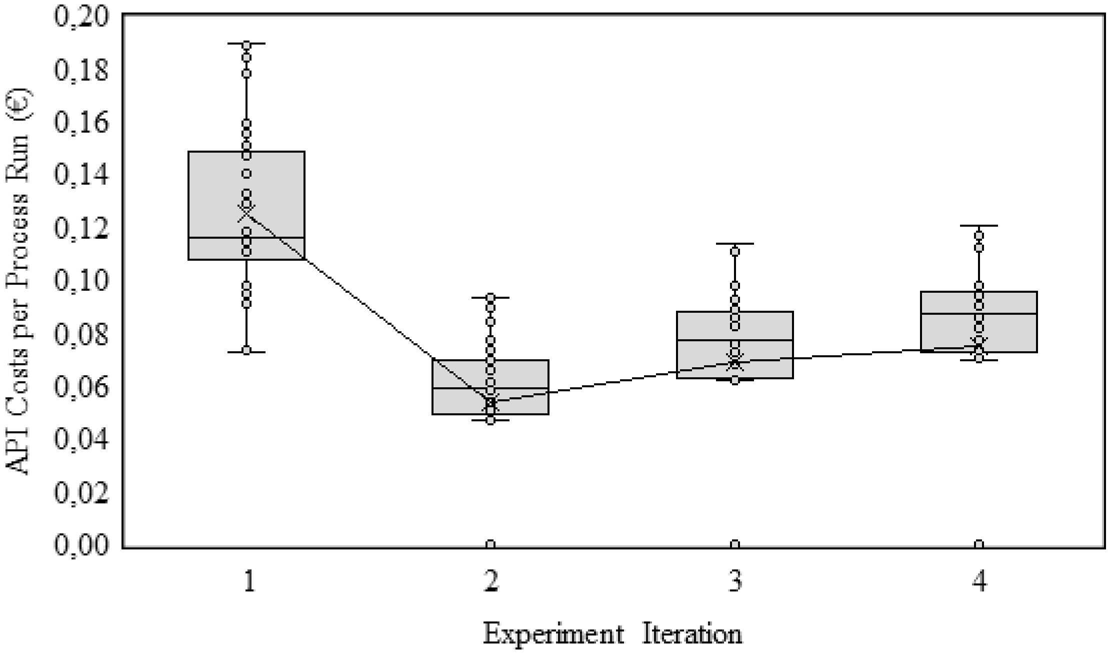

Fig. 5 Experiment API Costs Visualization.

The overall performance of aProCheCk was greatly improved by the optimizations introduced during the iterative refinement process. The use of structured prompting was instrumental in breaking down complex coherence checking tasks into more streamlined sub-problems. This approach is thought to have improved the accuracy of results and reduced the likelihood of hallucinations. Advanced prompting techniques, such as few-shot prompting and chain of thought, further enhanced the artifact's decision-making capabilities. Few-shot prompting provided richer context and exemplar output, while chain-of-thought prompting improved logical reasoning, both of which contributed to higher accuracy and consistency of the artifact.

Despite the impressive performance, certain limitations were observed during the experiment runs. The handling of special cases involving uncommon BPMN elements or non-standard modelling practices proved problematic. For example, the addition of unrelated new objects in the credit process led to hallucinated connections, while less common elements such as conditional connectors in the invoice and parts production processes also caused difficulties. These limitations indicate the need for further problem-specific prompt engineering to improve accuracy and consistency.