Dear Editors and Reviewers,

Please find enclosed our revised paper:

*Revisiting Business Process Analysis through the lens of Large Language Models: Prompting experiments with BPMN process serializations*

We are grateful for your insightful comments, as well as for your improvement suggestions, which we found very informative and helpful. We have incorporated your recommendations in our revised paper. Consequently, you will find both content and structural changes in the new manuscript.

We have highlighted in yellow the new content, as well as the new references. Explanations are indicated in the table below on how each reviewer concern was addressed.

|  |  |
| --- | --- |
| **REVIEWER CONCERNS** | **AUTHOR EXPLANATIONS ON REVISIONS** |
| **Reviewer 1** | |
| *The precise motivation for the research is only partially clear to me… I really miss a clear statement of the objective of this research…. Vaguely formulated research question…* | Explicit research questions are also formulated now in the Introduction. We also made it more explicit (in Section 2, highlighted paragraphs) that  - our investigation is motivated by a call to action found in the literature [10], to revisit the BPM lifecycle through the lens of the capabilities that LLMs can bring to its various phases  - we now focus on the Process Analysis phase, where the challenge lies in reconciling the deterministic nature of querying structured BPMN process serializations with the generative abilities of LLMs |
| *In my opinion, many of the queries that were tried out in the evaluation can best be answered by looking at a graphical model. (Reviewer 1)* | Process querying is a long-standing field of research with many tools and languages developed for it, some of them catalogued in reference [16]. For all those approaches it is true that answers can typically be identified visually by looking at a diagram. However:  - all querying approaches aim for the general case of information retrieval across a repository of content, not just minimal examples (also true for data queries, e.g. SQL)  - on the other hand, our paper does not look at scalability and real case repositories, we need to start form small and first understand how minimalist examples can be handled by LLMs; future work will of course look at more realist repositories |
| *I wonder how the test queries were selected and what was the general strategy behind this selection. This should be more clearly described in the paper*  *It [the paper] mainly shows how RDF representations of processes lead to more useful answers / summaries. This conclusion has been reached via a number of test questions whose selection is not very clearly motivated. (Reviewer 1)* | The research questions (now formulated in the Introduction) give some initial hints on this, in the sense of separating the two major tasks: prompts that navigate chains of relationships, labels and construct types in a complex realistic example vs. prompts that tackle minimalist un-labelled examples reduced to distinct workflow patterns.  Afterwards, just before Table 1/Section 4.1 the highlighted section goes into more detail regarding the retrieval patterns targeted by the prompting strategy. |
| *I am not very convinced about the use of the RAGA scores.*  *For instance, Case I in Table 1 gets a a faithfulness score of 1.0 which your discussion initially supports. Later, you say that the answer introduces additional details not supported by the context (e.g. the activity "activate bot"), which should lead to a lower faithfulness score. Another example is the discussion of answer similarity for Case I in Table 2 where you state that the high score is based on the comprehensiveness of the answer - but the ground truth is not at all comprehensive! In conclusion, I think you should discuss the scores more critically... (Reviewer 1)* | We have taken into account the reviewer’s observations and have re-examined our interpretation of the RAGAs scores. As for the reviewer’s doubts regarding the results:   * faithfulness, which takes into account the generated answer and the context, scores 1 in Table 1 is because the RDF context includes “Activate bot” activity, so the generated answer – which mirrors that context – naturally achieves perfect consistency. However, to calculate answer relevancy (with a score at around 0.78), the reverse-engineered question variants do not fully match the prompt’s intent, penalizing the inclusion of unnecessary details (such as “Activate bot”); * the high semantic similarity in Table 2 shows that the overall meaning between the generated answer and the ground truth aligns well when compared via cosine similarity of their vector embeddings – even though the ground truth itself is not comprehensive, as you stated. |
| *Please do not report them [the RAGAS scores] with 6 decimal places (Reviewer 1)* | We have rounded the scores to two decimals. |
| *Since you are discussing all answers in detail anyway, why not rely on a human assessment of these (or other) metrics instead of automatically computed ones?*  *For instance, it is somehow counterintuitive that answers that are identical to the ground truth receive very low values for some scores (e.g. faithfulness=0 for Case I in Tables 6 and 7, but faithfulness=1 in Table 8, in all cases answer is identical to ground truth) (Reviewer 1)* | While automatically computed metrics provide more objective insights, we also rely on our own human interpretation when defining the ground truths and when we interpret the RAGAs scores. Therefore the human and the automatic assessments are somewhat balanced. In future works we may devise different evaluation paths, but for now we got a sense that RAGAs metrics are advocated as state of the art measuring approaches in the literature.  Normally, “Faithfulness” only measures how factually consistent an answer is with its context – it does not take the ground truth into account, which is instead used to calculate the score for other metrics like ”factual correctness” (previously labelled as “answer correctness”). In the revised version, we have explained, when necessary, that the discrepancy between the low faithfulness score and a supposed correct answer – which is, indeed, counterintuitive – is due to the differences in formatting between the context (structured data represented in RDF/XML) and the answer (unstructured text). We tried to address this by modifying the metric to create “semantic faithfulness”, where both the answer and the retrieved context are converted into vector embeddings and their cosine similarity is computed, since this approach accounts for the broader meaning of both structured and unstructured data, allowing for a more accurate evaluation of how well they align. |
| *In terms of presentation, I find that the paper is far too long, given its rather minor contribution.*  *In particular, it seems to me that Tables 7-19 rather consistently show how the RDF-based variant (Case I) delivers more accurate answers than Case II. Even if there are some subtleties / additional findings, I feel that you could summarise all main findings of these Tables in a rather short section and put the tables themselves into an appendix.*  *The paper could be substantially shortened to avoid repetition of findings across test cases. When shortened, it might be suitable for a submission at a conference. (Reviewer 1)* | We have created the Appendix section with all the tables containing the test queries that demonstrate that Case I outperforms Case II. In the Appendix, we present scenarios in which GPT‑4 responded with “I do not know” or “The provided context does not contain enough information” when processing minimalist BPMN XML models (Tables 9 - 16). Tables 17 and 18 in the appendix refer to another category of scenarios in which the responses were partially correct for the BPMN XML case. Meanwhile, for the RDF case, the responses generated by the GPT model are consistently correct in all scenarios/align closely with the ground truth.  We have also included explanations – detached from the RAGAs documentation – regarding how the metrics were calculated.  Overall this being an on-line journal with no hard page count limits, we will consult in case of acceptance with the editor on how much to keep in the paper and how much to reference in the Github link containing ancillary details (footnote 5). |
| **Reviewer 2** | |
| *The methodology lacks clarity about how LLM inferencing was conducted.*  *The abstract mentions the ChatGPT interface, but the manuscript inconsistently refers to “LLM services” and “GPT services,” leaving the reader unsure whether standard ChatGPT functionality or custom GPTs were used (Reviewer 2)* | We reduced such ambiguities by consistently referring to the use of GPT-4 via the API for the reported experimentation, therefore the analysis in this draft strictly refers to this set-up (and the experimentation was redone through the API avoiding the ChatGPT front-end and customizations). LLMs services in general are mentioned where the discourse makes general statements. |
| *The specific models utilized are not disclosed.*  *Clear documentation of the experimental setup, including the version of the LLM and its configuration, is essential for reproducibility and transparency (Reviewer 2)* | Besides documenting our GPT-4 configuration (temperature, embedding model, token limit and other presets) in Section 3.3, we also inserted a GitHub repository link (footnote 5) with our Python scripts and other files used for our experiments. |
| *The section titled “Large Language Models and the BPM lifecycle” appears to function as a background chapter, but is far too brief for a journal article. It lacks critical engagement with existing literature and does not establish a clear research gap (Reviewer 2)* | We expanded Section 2 by incorporating more recent literature and defining the research gap. Specifically, in spite of the emerging capabilities of LLMs across all phases of the BPM lifecycle, previous studies have mostly focused on LLM performance on processing unstructured text or visual content, rather than structured serializations (specifically, BPMN processes serialized using RDF and XML). |
| *The paper fails to articulate its research gap clearly, hypotheses, or research questions.*  *Without this foundation, the study's objectives and outcomes are ambiguous, making it difficult to evaluate its scientific merit (Reviewer 2)* | The Introduction now formulates explicit research questions, contextualized according to the call for action on LLM-BPM lifecycle in ref. [10] and by gaps identified in the expanded literature coverage form Section 2. |
| *The tables are not adequately referenced in the text, and their presentation is suboptimal. Consider relocating tables to the appendix if they are not directly essential to the narrative.*  *Figures 1–3 and potentially Figure 4 could also be moved to an appendix, as their content adds limited value to the main discussion. Conversely, Figures 5 and 6 could be expanded upon to better engage readers less familiar with BPM concepts (Reviewer 2)* | We relocated most of the tables to the Appendix section, while keeping in the main paper body those directly relevant to the discussion within the text.  Figs. 5-6 contain internal captions distinguished for each minimalist example, briefly explaining the nature/pattern of each example for readers lacking BPMN expertise. We believe this diversity of patterns is relevant to remain in the main body of the paper, as some tables also refer those internal captions. |
| *The paper lacks a dedicated discussion section. Introduce a discussion section to interpret the findings and connect them to broader implications (Reviewer 2)* | We introduced a dedicated “Discussion” section (5) that synthesizes the results and connects them to broader implications. Moreover, within “Experimental Outcomes and Evaluation”, each scenario is followed by an in-depth analysis on how the metrics were calculated (in accordance to the RAGAs documentation which is also available publicly for more details). |
| *The conclusion does not sufficiently explore future research opportunities.*  *For example, the potential of continuous pre-training on domain-specific knowledge graphs, meta-models, or XML/RDF schemas is an intriguing direction that is not mentioned but warrants inclusion (Reviewer 2)* | We included the recommended future research opportunities at the very end |
| *Issues such as incomplete sentences (e.g., the last sentence in the abstract) and unnecessarily long sentences hinder readability. Rigorous proofreading is needed (Reviewer 2)* | General discourse quality was improved by adjusting or shortening phrases in the revised version. |
| *Caution is advised when describing LLMs as “understanding” or “reasoning.” The paper should reflect recent literature, which highlights the limitations of LLMs in these areas (Reviewer 2)* | We have updated the wording to minimize the use of “understanding” and to reflect the current literature concerns about how much LLMs actually understand (ref. [39] was added on this) and a statement on this reluctance was added in the conclusions. |
| *SPARQL snippets should not span across pages, as this disrupts the flow and readability of technical content (Reviewer 2)* | We improved the layout for better reading flow, however final layout will be aligned in case of acceptance with the publication template (current layout should not be seen as final). |
| ***Reviewer 3*** | |
| *It is not made explicit which XML schema is used.*  *I assume RDF is compared with the ".bpmn" representation of the models. This should be then made explicit writing that RDF export is compared to BPMN schema in XML (Reviewer 3)* | We have now explicitly stated that we compare BPMN models serialized in RDF to those serialized in the XML schema prescribed by the standard BPMN 2.0 XML, as available in Signavio. |
| *The presentation of the experiment results should be written clearer.*  *It took me some time to identify that the prompts are in the first line above the tables. At least the label "Prompt 1:" etc should be written in bold, but I suggest to not only write them in the table (Reviewer 3)* | We added a brief description for each table, bolded the prompt and at times, referenced the prompt in our analysis, below the tables. |
| *The table formatting is not clear.*  *There should be a label what the rows mean. And for the key metrics it is not necessary to mention them twice per row. One possibility would be to have three columns: the first column containing the what the rows mean and the second and third the values for each case (Reviewer 3)* | We have updated the table formatting by restructuring it as suggested. |
| *It is not clear, how the values for the key metrics are calculated (Reviewer 3)* | We have included explanations on how the values for these metrics are calculated, especially for the cases where the RAGAs values may seem unclear. |
| *It is also suspicious that they are calculated to six digits after the period (Reviewer 3)* | All scores are now rounded to two digits. |
| *The experiment consists of 19 prompts, but it is not clear, why exactly these prompts are chosen (Reviewer 3)* | Just before Table 1/Section 4.1 the highlighted section provides more detail regarding the retrieval patterns that were targeted by the prompting strategy. |
| *For each of the 19 prompts an analysis is described, but the conclusion is very short.*  *There should be a learning,* *synthesizing the results for the prompts by making clear for which kinds of prompts which key metrics are higher for RDF and XML and for RDF and BPMN XML (Reviewer 3)* | We synthesized the results for our prompt in a Discussion section (5). |
| *At the end of Section 3.2 a reference to BPM Analyse is missing (Reviewer 3)* | In this revised version, we have removed BPMN Analyst (https://chatgpt.com/g/g-vEgsodrkw-bpmn-analyst) and streamlined our approach to the GPT-4 model via the API to avoid bias possibly introduced by the analyst configuration. |