Dear Editors and Reviewers,

Please find enclosed the second revised version of our manuscript:

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

We sincerely thank you for your continued engagement and for the constructive feedback provided during the review process. In this second revision, we have carefully addressed the remaining concerns and further refined the manuscript accordingly.

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

|  |  |
| --- | --- |
| **REVIEWER CONCERNS** | **AUTHOR EXPLANATIONS ON REVISIONS** |
| **Reviewer 1** | |
| *What is still not clear to me is why the scores for e.g. semantic faithfulness need to be automatically computed instead of assessing them manually (this could be done by some independent expert). Given the rather small number of answers to compare, this should be easily feasible and I would find it much more meaningful...* | We fully understand and agree that, given the relatively small number of answers under comparison, a human assessment could feasibly provide interpretation. However, we argue that:   * The human assessment is already represented in the discussion on each result (around Tables 1-8), with RAGAs being an additional “sensory” instrument to achieve measurements that may inform future analysis and can be adapted to larger scale experiments of the kind reported here. Most importantly, RAGAs discerns different facets of answer deviations – e.g. between inference inability and user question misinterpretation. The human interpretation around Tables 1-8 was refined by reducing the (sometimes redundant) numeric reporting, making more explicit the human scrutiny; * It is also a secondary objective of our research to refine an evaluation protocol that is both diversified in distinct indicators and can be reused for similar but larger scale assessments - this is now made more explicit when motivating the problem and objectives (in Introduction) and in the evaluation rationale explanations (highlighted paragraphs in the first part of Section 4). |
| *It would also allow the discussion of examples to be shortened to their essence, avoiding explanation of scores that does not contribute much in my opinion. This should be corrected before final publication.* | The text around the result tables (1-8) was refined (and slightly reduced to their essence) to eliminate reproducing the numbers already shown in the tables, and instead to highlight better the human interpretation of results. |
| **Reviewer 2** | |
| *The presentation remains highly technical.*  *The methodology and experiment sections contain dense details that may be difficult for non-expert readers to follow. Some figures, tables, and long technical explanations could be moved to an appendix or made more concise.* | We introduced brief explanations for key technical terms (RAG, the meaning of metrics) and reduced some “dry” numeric reporting (around Tables 1-8) to place more emphasis on human interpretation. As a consequence, the highlighted areas in Subsection 3.3 and Section 4 are slightly more accessible to non-technical readers - this is however limited, considering the journal profile and targeted audience.  We kept in the body of the paper those experiment results on which the main discussion is grounded, since the metric-based analysis and interpretation are at the core of this investigation. The reading experience would be affected by distancing all figures/tables from their direct referential comments. We keep in appendix the corner cases that are not be impacted in this manner.  We will also consider further copy-editing recommendations from the journal, in case of acceptance. |
| *Limited discussion of limitations.*  *The paper would benefit from a more explicit reflection on the study’s constraints,* *such as the reliance on GPT-4 only and the potential variability in responses due to the stochastic nature of LLMs.* | Within Section 6, we have expanded the conclusions with additional recognized limitations (as highlighted) |
| *Practical implications could be expanded.*  *The discussion on how these findings impact BPM practitioners and tool developers could be more detailed.* | We expanded the discussions in Section 5 with additional statements on pragmatic implications (as highlighted) |