Generating BPMN Healthcare Process Models Using LLMs

Adam Grík[1], Silvia Putekova[2], Anna Petrulova[3], Marek Rabincak[4], Karol Nemoga[5], and Fedor Lehocki[1](✉)

[1] Faculty of Informatics and Information Technologies, Slovak University of Technology in Bratislava, Bratislava, Slovakia
{xgrik,fedor.lehocki}@stuba.sk
[2] Faculty of Health Care and Social Work, Trnava University, Trnava, Slovakia
[3] National Institute of Cardiovascular Diseases, Bratislava, Slovakia
[4] Psychiatric Hospital Michalovce, Michalovce, Slovakia
[5] Mathematical Institute SAS, Bratislava, Slovakia

Abstract. Business Process Model and Notation (BPMN) is widely used in healthcare to formalize clinical workflows, clinical pathways, and organizational procedures. In practice, however, these processes are primarily documented in extensive, unstructured textual artifacts such as clinical guidelines and standardized healthcare procedures. Transforming such documents into BPMN models typically requires substantial manual effort in the early modeling phases, which can reduce time efficiency. Recent advances in LLM-based chatbots offer promising capabilities for analyzing and synthesizing complex textual documents. Nevertheless, existing BPMN generation approaches typically depend on intermediate representations and deterministic parsers, introducing additional architectural complexity and reducing practical usability. This paper proposes a multi-step LLM-based workflow for directly generating BPMN process models encoded in XML from unstructured healthcare procedure documents using a conversational chatbot. The workflow combines multiple prompting techniques and strategies to improve generation robustness and technical correctness. The approach is evaluated on five official healthcare procedure documents issued by the Ministry of Health of the Slovak Republic from the domains of psychology and psychiatry. The generated BPMN models are assessed using a dual evaluation strategy that combines qualitative expert evaluation with objective XML-based metrics. Experimental results demonstrate that the proposed workflow consistently produces formally correct, complete, and clear BPMN models, outperforming a zero-shot prompting baseline according to the evaluated quality criteria. The findings indicate that the proposed approach is well suited for supporting early-stage process modeling from complex healthcare documents.

Keywords: BPMN · healthcare · large language models · chatbots · process modeling

M. Särestöniemi et al. (Eds.): NCDHWS 2026, CCIS 3010, pp. 508–524, 2026.
https://doi.org/10.1007/978-3-032-28819-6_33

1 Introduction

1.1 Motivation and Problem

Business Process Model and Notation (BPMN) is accepted as a notation in healthcare for representing clinical workflows, clinical pathways, and organizational procedures in a structured and interpretable form. In practice, however, healthcare processes are primarily described in extensive textual documents such as standardized procedures and clinical guidelines. Manually analyzing these documents and transforming them into formal process models can be time-consuming and typically requires close collaboration with domain experts.

Recent advances in large language model (LLM)–based chatbots, such as ChatGPT, demonstrate strong capabilities in analyzing and synthesizing complex textual documents. Prior research highlights the potential of conversational AI systems to support healthcare services by assisting in the interpretation and processing of medical documentation [1]. These capabilities make LLM-based chatbots a promising tool for supporting the early stages of process modeling from textual healthcare documents as well as support educational efficiency by targeting healthcare professional's knowledge gaps related to application of clinical pathways in practice [2].

Empirical studies indicate that LLMs can perform document analysis tasks substantially faster than human analysts while achieving comparable qualitative coverage [3]. In this context, LLM-based chatbots are not intended to replace domain experts, but rather to accelerate initial phases of process modeling, such as document analysis and the generation of preliminary BPMN drafts, which can subsequently be refined and validated by experts.

Despite this potential, the direct application of LLM-based chatbots to structured process modeling raises several open challenges.

One major challenge is that existing approaches for applying LLMs to process modeling predominantly rely on intermediate representations, such as parser-based abstractions, tabular or schema-driven representations, or domain-specific process modeling languages. These additional transformation layers increase the overall complexity of the solution and introduce dependencies on auxiliary tools and environments, which can limit the practical accessibility of such approaches in applied settings. Approaches that attempt direct generation of BPMN process models encoded in XML from natural language documents remain limited. Consequently, there is a lack of systematic methodologies for directly generating BPMN process models from unstructured healthcare procedure documents.

Another important challenge arises from inherent limitations that complicate their use for structured process modeling tasks. Their outputs are non-deterministic and sensitive to prompt formulation, with measurable variability observed even under deterministic decoding settings [4]. This variability poses challenges for generating consistent, complete, and technically valid BPMN models, particularly in the context of complex healthcare processes, and therefore necessitates explicit mechanisms for controlling and stabilizing the generation process.

Together, these challenges highlight the need for structured approaches that enable BPMN model generation directly from textual healthcare documents using LLM-based systems, while incorporating controlled prompting strategies to guide the generation process.

1.2 Related Work

Recent research has explored the use of LLMs to support process-oriented reasoning in the healthcare domain, particularly for extracting, interpreting, and navigating clinical workflows from textual sources.

In [5] authors investigate the automated extraction of clinical pathway information from guideline documents using LLMs, proposing a conceptual intermediate model that bridges unstructured text and formal process representations. Their findings demonstrate the suitability of LLMs for supporting early-stage pathway extraction from clinical text.

Similarly, [6] demonstrates the use of LLMs for generating diagnostic pathways through sequential decision-making prompts, effectively modeling clinical reasoning processes for differential diagnosis. Their results show that the use of structured prompting strategies substantially improves the quality and coherence of the generated diagnostic pathways.

The inverse problem is addressed in [7] by focusing on the interpretation and querying of existing clinical process models using LLMs, demonstrating that LLMs can reliably understand and navigate complex clinical pathway graphs.

Complementary findings are reported in [8], demonstrating that LLMs can effectively comprehend and reason over BPMN models. Similarly, [9] explores the use of LLMs for generating process descriptions from BPMN models, further supporting the notion that LLMs possess inherent understanding of BPMN structure and process semantics.

Beyond healthcare, multiple approaches have investigated the generation of BPMN models using LLMs in general process modeling contexts.

Most LLM-based BPMN generation approaches do not produce BPMN process models encoded in XML directly, but instead rely on intermediate representations (e.g., structured text, JSON, tabular formats) that are later transformed into BPMN using deterministic parsers, introducing additional architectural complexity [10–15].

Direct generation of BPMN process models encoded in XML using LLMs without intermediate representations has been explored only to a limited extent. In [16], a zero-shot prompting approach for direct BPMN XML generation is proposed; however, it does not address the challenges posed by complex, unstructured documents nor the need for generation stability and consistency.

Finally, the evaluation of BPMN models generated by LLMs remains an underexplored area. A systematic approach to this problem is introduced in [17] through the 3C framework, which evaluates such models beyond syntactic validity. This framework represents one of the first efforts to assess the quality of LLM-generated BPMN models in a more comprehensive manner.

In summary, existing research demonstrates that LLMs are capable of understanding, extracting, and reasoning about process-related information, and can support process modeling tasks in both healthcare and general domains. However, current approaches to BPMN model generation largely employ multi-stage pipelines based on intermediate representations, while direct generation of BPMN process models encoded in XML from unstructured documents remains insufficiently explored, particularly in real-world healthcare settings. These gaps motivate the workflow-based approach and evaluation strategy proposed in this work.

1.3 Contributions

This work presents a multi-step LLM-based workflow for directly generating executable BPMN XML from complex, unstructured healthcare documents, combining structured information extraction, schema-guided prompting, and repair strategies. The applicability of the workflow is demonstrated on official healthcare procedure documents from the domain of psychology and psychiatry. A dual evaluation approach is introduced, combining qualitative 3C-based assessment with objective XML-based metrics to assess both visual and formal properties of LLM-generated BPMN models. Finally, the proposed workflow is empirically compared with a zero-shot prompting baseline.

2 Materials and Methods/Study Scenario

2.1 LLM-Based BPMN Generation Workflow

The proposed workflow represents a sequence of multiple steps used to generate BPMN process models in XML format from structured textual documents. Instead of relying on a single prompt, the workflow decomposes the generation task into several steps, each corresponding to a specific prompting strategy applied during the generation process.

The workflow consists of a sequence of clearly defined steps, as illustrated in Fig. 1.

Role-based Prompting serves as the initial step of the workflow. In this step, the chatbot is explicitly instructed to assume the role of an expert in BPMN process modeling. By defining the model's role upfront, this step establishes a consistent modeling perspective and constrains the model's responses to the domain of process modeling. The primary purpose of this step is to promote the use of domain-specific terminology and structural reasoning aligned with BPMN principles.

RAG[1] Initialization is performed to load and provide the relevant healthcare documents used as input for the BPMN generation process. In this step, the structured textual documents describing the target process are made available to the large language model as contextual information. This ensures that

[1] RAG stands for *Retrieval-Augmented Generation*.

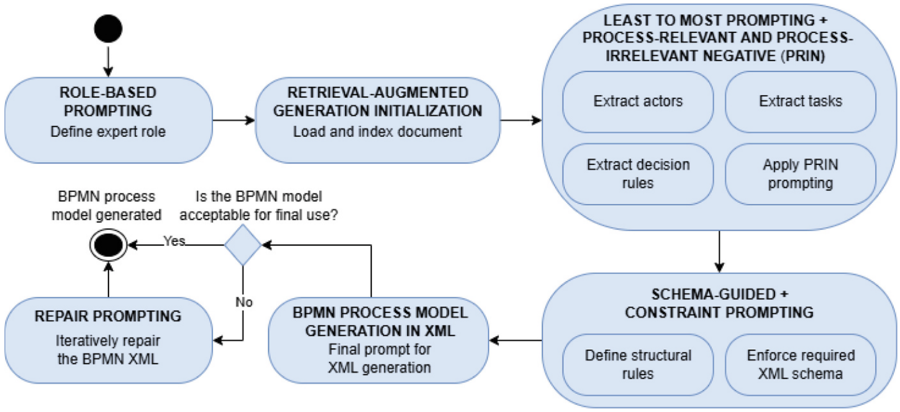

Fig. 1. Multi-step workflow

subsequent steps are grounded in the source documents from which the process model is generated.

Least-to-Most Prompting with PRIN[2] applies an incremental prompting strategy to decompose the process modeling task into a sequence of smaller and more manageable sub-tasks. Instead of generating the complete process model at once, the workflow progressively extracts individual process elements step-by-step. Specifically, the model is prompted to extract process actors, tasks, and decision points in separate steps. During each extraction step, a PRIN-inspired prompting strategy is applied by explicitly requesting both elements that belong to the target process and elements that are present in the source documents but should be excluded from the final process model [18]. This strategy helps clarify decision boundaries and reduce ambiguity in the generated outputs.

Schema-Guided and Constraint-Based Prompting introduces explicit structural guidance to ensure that the generated BPMN representation conforms to the expected XML format. In this step, an example of a valid BPMN XML structure is provided to the model, which helps align the model's output with the expected schema and establishes the required syntactic layout and mandatory elements. In addition, explicit constraints are applied to enforce the inclusion of required diagram and layout information, ensuring that the final process model is not only syntactically valid but also structurally complete. Together, schema guidance and constraint-based prompting ensure alignment of the generated BPMN XML with the expected structural and syntactic requirements.

BPMN XML Generation generates the final BPMN process model in XML format using all information, constraints, and intermediate results obtained from the preceding workflow steps.

[2] PRIN stands for *Process-Relevant and Process-Irrelevant Negative prompting*.

Repair Prompting is applied after a manual review and execution check of the generated BPMN process model, when syntactic, structural, or layout-related issues are identified. The workflow does not prescribe explicit repair rules, a fixed number of repair iterations, or strict acceptance thresholds for determining when a model is considered sufficiently complete or correct. Instead, the application of repair prompting is guided by expert judgment based on the observed issues.

Furthermore, the decision to apply repair prompting is context-dependent and considers practical aspects such as time efficiency and the complexity of the required modifications. In cases where identified issues are minor, for example limited to layout adjustments or simple structural refinements, manual correction of the BPMN model may be more efficient than invoking additional LLM-based repair iterations.

2.2 Evaluation Methodology

The 3C-Based Evaluation Framework was originally proposed in [17]. In this work, the framework is adapted and applied to assess the visual quality of BPMN process models generated by the proposed LLM-based workflow. The evaluation is performed manually by a human expert and is based solely on visual inspection of the rendered BPMN diagrams, without considering the underlying XML structure.

The framework evaluates LLM-generated process models across three dimensions: Correctness, Completeness and Clarity. *Correctness* reflects the semantic and logical validity of the generated BPMN process model, including the correctness of control flow and decision logic. *Completeness* assesses the extent to which the generated model captures all relevant process elements and control-flow paths described in the source document. *Clarity* evaluates the readability and interpretability of the BPMN model from a human perspective, focusing on the transparency of the process structure and flow.

Within each dimension, the framework defines multiple evaluation criteria with assigned importance levels (High, Moderate, Low) and corresponding weights. Each criterion is assessed using a binary true/false evaluation, indicating whether the generated model satisfies the criterion or not. Each generated BPMN process model is evaluated independently for each of the three dimensions. The final score for each dimension is computed as a weighted average of the individual criteria, resulting in a value in the range $\langle 0,1 \rangle$.

For the purposes of this evaluation, only a subset of the original 3C criteria was applied, as some criteria are not applicable in the context of the proposed workflow. Examples include criteria that assume explicitly defined control-flow elements in the prompt or criteria that require comparison against a reference (gold-standard) process model, neither of which are available in our setting.

Together, these dimensions enable a qualitative assessment of BPMN models that captures semantic correctness and human interpretability, which cannot be fully evaluated using purely syntactic or structural metrics.

While the 3C-based framework provides a structured approach to model quality assessment, it also exhibits certain limitations. In particular, the evaluation relies on expert judgment, which introduces an inherent degree of subjectivity.

Objective XML-Based Metrics are introduced to complement the qualitative 3C-based evaluation by providing a quantitative and reproducible assessment of generated BPMN process models. In contrast to visually driven evaluations in the original 3C framework, the proposed metrics operate directly on the BPMN XML representation and are independent of visual layout and human interpretation. This allows the assessment to focus on structural correctness, formal completeness, and minimal technical prerequisites for model clarity.

To maintain conceptual consistency with the qualitative evaluation, the objective metrics are organized according to the same three dimensions—Correctness, Completeness, and Clarity—and evaluated using a set of ratio-based measures derived from the BPMN XML representation, as summarized in Table 1.

All metrics are computed as ratio-based measures, yielding values in the range $\langle 0,1 \rangle$. For each evaluation dimension, the final dimension score is computed as the arithmetic mean of all metrics belonging to that dimension, with equal importance assigned to each metric.

Table 1. Objective XML-based metrics

Dimension	Metric	Formula
Correctness	Element connectivity	$\frac{\text{connected elements}}{\text{total elements}}$
	Balanced gateways	$\frac{\text{balanced gateways}}{\text{total gateways}}$
	Valid sequence flows[a]	$\frac{\text{valid sequence flows}}{\text{total sequence flows}}$
Completeness	Task coverage	$\frac{\text{tasks in XML}}{\text{tasks extracted}}$
	Decision coverage	$\frac{\text{decisions in XML}}{\text{decisions extracted}}$
Clarity	Labeled elements (name+id)	$\frac{\text{labeled elements}}{\text{total elements}}$
	Labeled outgoing flows	$\frac{\text{labeled outgoing flows}}{\text{total outgoing flows}}$
	BPMN-DI[b] coverage	$\frac{\text{elements with BPMN-DI}}{\text{total elements}}$

[a] A *sequenceFlow* is a BPMN element defining the execution order between connected flow nodes [19].
[b] BPMN Diagram Interchange [19]

For correctness-oriented metrics, connected elements denote BPMN tasks, gateways, and events that are referenced as either source or target in at least one sequenceFlow element. Balanced exclusive gateways refer to *exclusive (XOR) gateways* that are structurally well-formed with at least one incoming and at least two outgoing sequence flows, or as merge points with at least two incoming and one outgoing sequence flow. Valid sequence flows are defined as sequenceFlow elements whose source and target references correspond to existing BPMN elements.

Completeness-oriented metrics measure coverage with respect to elements identified during the extraction steps, without penalizing additional elements present in the generated model. In this study, completeness is assessed using task coverage and decision coverage, where decisions correspond to exclusive (XOR) gateways in the BPMN process model.

Clarity-oriented metrics focus on minimal technical prerequisites for readability, including the presence of element labels and corresponding BPMN-DI representations.

While the objective XML-based metrics provide a reproducible and expert-independent assessment of structural and formal properties, they do not capture semantic correctness or higher-level interpretability of BPMN models. For this reason, they are used as a complementary evaluation alongside the qualitative 3C-based framework.

2.3 Study Scenario

The proposed workflow was evaluated using a set of official healthcare procedure documents issued by the Ministry of Health of the Slovak Republic [20–24]. The selected documents originate from the domain of psychology and psychiatric care and include standardized procedures and comprehensive psychological management guidelines, describing process steps such as diagnosis, intervention, treatment, and prevention. The source documents are used in Tables 2 and 3 and are referenced throughout the study using anonymized identifiers DOC1–DOC5[3] [20–24]. The complete original versions of all evaluated documents are publicly available in a dedicated GitHub repository[4].

Each document was processed independently using the proposed LLM-based BPMN generation workflow. The same workflow configuration and prompting strategies were applied across all provided documents to ensure consistent experimental conditions. All experiments were conducted using ChatGPT (version 5.2) as the underlying LLM. The generated BPMN XML models were imported into Camunda Modeler for rendering and inspection, ensuring that all evaluated models were executable and visually interpretable in a standard BPMN environment.

For comparison, the same document set was also processed using a single-step zero-shot prompting baseline inspired by the approach reported in [16].

All generated BPMN process models were evaluated using both the qualitative 3C-based framework and the objective XML-based metrics described in the previous section.

---

[3] DOC1 – Comprehensive Psychological Management of an Adult Patient with Chronic Pain; DOC2 – Comprehensive Psychological Management of a Patient with a Somatic Disease; DOC3 – Agoraphobia; DOC4 – Post-Traumatic Stress Disorder; DOC5 – Comprehensive Psychological Management of an Adult Patient with Depressive Disorder and Recurrent Depressive Disorder.

[4] https://github.com/xgrik/llm-bpmn-generation-artifacts.

2.4 Prompt Building

This section outlines representative examples and skeleton formulations of the prompting strategies that constitute the proposed workflow. The original and exact prompt formulations, as used in the experiments, are publicly available in a dedicated GitHub repository.[5]

Role-Based Prompting strategy is used to initialize the interaction context by explicitly defining the modeling perspective and scope of the LLM before any process extraction or generation is performed. A representative example of the role-based prompt is shown in Listing 1.1.

Listing 1.1. Representative role-based prompt

```
You are an expert in BPMN 2.0 process modeling with experience in
    healthcare workflows and system-oriented process automation.

Your task is to support the transformation of structured healthcare
    documents into formal BPMN process models that capture both human
    and system-supported activities.

Constraints:
- strictly follow BPMN modeling principles,
- use precise process-oriented and healthcare terminology,
- ground all modeled elements in the provided source documents,
- do not introduce assumptions or unsupported process steps,
- focus on structured process modeling rather than narrative
    descriptions.
```

RAG Initialization prompting strategy presents the source healthcare document to the LLM and instructs it to assess whether the document is suitable for BPMN-based process modeling. A representative example of the RAG initialization prompt is provided in Listing 1.2.

Listing 1.2. Representative RAG initialization prompt

```
The following document describes a healthcare-related procedure titled:
"[DOCUMENT_TITLE]".

Your task is to carefully read and analyze the document and decide
    whether it contains sufficient procedural information to be
    represented as a BPMN process model.

In your assessment, focus on whether the document:
- describes a sequence of activities or actions,
- includes decision points, roles, or responsibilities,
- implies control-flow relations between steps.
```

[5] https://github.com/xgrik/llm-bpmn-generation-artifacts.

```
Do not generate any BPMN elements at this stage. Only analyze the
    document and state whether it is suitable for BPMN-based process
    modeling, with a brief justification.
```

Least-to-Most Prompting with PRIN strategy operationalizes the incremental extraction of process elements by applying the same PRIN-based prompt structure across multiple extraction steps (actors, tasks, and decision points), while varying only the targeted element type. A representative example of the PRIN-based extraction prompt is shown in Listing 1.3.

Listing 1.3. Representative PRIN-based extraction prompt

```
Based on the provided healthcare process document, identify all
    [PROCESS_ELEMENT_TYPE] that should be included in the BPMN process
    model.

Include only elements that:
- actively influence the execution or control flow of the process,
- represent concrete responsibilities, actions, or decisions,
- are explicitly supported by the source document.

For each identified element:
- provide a brief description,
- reference the part of the document it is derived from.

Additionally, identify elements mentioned in the document that should
    NOT be included as [PROCESS_ELEMENT_TYPE] in the BPMN model.

These may include:
- contextual or informational mentions,
- high-level goals or recommendations,
- assessments or statements that do not affect process execution.

For each excluded element, briefly justify why it is not relevant for
    BPMN modeling.
```

Schema-Guided and Constraint-Based Prompting strategy constrains the XML generation step by providing an explicit BPMN schema template and a set of structural constraints that the model must satisfy when producing the final BPMN process representation. A representative example of the schema-guided and constraint-based prompt is shown in Listing 1.4.

Listing 1.4. Representative schema-guided and constraint-based prompt

```
Use the provided BPMN 2.0 XML schema example as a mandatory structural
    reference
for generating the BPMN process model.
```

```
The schema example includes:
- [INSERT XML SCHEMA HEADER AND NAMESPACES, e.g., Camunda-compatible
    BPMN definitions],
- [INSERT MINIMAL EXECUTABLE PROCESS STRUCTURE],
- [INSERT BASIC BPMN-DI ELEMENTS REQUIRED FOR DIAGRAM RENDERING].

The generated BPMN XML must comply with the structure and namespaces
    defined in the provided schema example.

The model must additionally satisfy a set of explicit structural and
    modeling constraints, such as:
- [INSERT REQUIRED STRUCTURAL, CONSISTENCY OR DIAGRAMMATIC
    CONSTRAINTS, e.g., executable process definition, valid sequence
    flow references, BPMN-DI coverage]

Do not introduce any process elements that were not identified in
    previous extraction steps. Use the schema and constraints to guide
    the generation, ensuring that the resulting BPMN XML is
    structurally complete and executable. But do not generate any BPMN
    elements at this stage.
```

BPMN XML Generation step synthesizes all previously extracted process elements and constraints into a single, executable BPMN process model, instructing the LLM to generate only the final XML representation without additional explanations. A representative example of the BPMN XML generation prompt is shown in Listing 1.5.

Listing 1.5. Representative BPMN XML generation prompt

```
Generate the final BPMN 2.0 process model in XML format using
    exclusively the process elements and control-flow logic identified
    in previous workflow steps.

Synthesize all extracted actors, tasks, decision points, and
    constraints into a single, coherent, and executable BPMN process
    model.

Requirements:
- use only previously identified process elements,
- do not introduce any new actors, tasks, events, or gateways,
- reflect both human-performed and system-supported activities,
- ensure the output is executable and suitable for deployment.

Return only the final BPMN XML without explanations or additional text.
```

Repair Prompting strategy serves as a corrective fallback mechanism that can be invoked when validation tools or manual inspection reveal issues in the generated BPMN XML, such as schema violations or diagram rendering problems. A representative example of a repair prompting formulation is provided in Listing 1.6.

Listing 1.6. Representative repair prompt

```
The following BPMN XML contains errors identified during validation or
    manual inspection.

[INSERT VALIDATION ERRORS, EXECUTION ERRORS, OR LAYOUT ISSUES HERE]

Your task is to repair the existing BPMN XML while preserving the
    original process structure and semantics.

Requirements:
- fix only the reported issues,
- do not introduce new process elements,
- do not remove valid existing elements,
- preserve all previously identified actors, tasks, and decision
    points,
- ensure the repaired model remains executable and structurally
    consistent.

Return only the corrected BPMN XML without explanations or additional
    text.
```

3 Results and Discussion

The primary focus is on evaluating the effectiveness and robustness of the proposed multi-step LLM-based BPMN generation workflow using both the qualitative 3C-based framework and the objective XML-based metrics. A baseline zero-shot prompting approach is included as a reference point to contextualize the observed improvements and highlight the contribution of the proposed workflow. Both approaches were evaluated on the same set of five healthcare procedure documents.

For evaluation purposes, only BPMN models generated prior to the application of repair prompting were considered, in order to assess the baseline performance of the generation workflow without iterative post-hoc corrections.

An example of a final BPMN process model generated from DOC1, including the complete process representation and a brief descriptive explanation, is provided in a dedicated GitHub repository[6].

3.1 3C-Based Results

The 3C-based evaluation provides a qualitative, expert-driven assessment of the visual clarity, perceived correctness, and completeness of BPMN process models generated by an LLM, as summarized in Table 2.

[6] https://github.com/xgrik/llm-bpmn-generation-artifacts.

Table 2. 3C-based results

	Weight	WORKFLOW					ZERO-SHOT				
Clarity		DOC 1	DOC 2	DOC 3	DOC 4	DOC 5	DOC 1	DOC 2	DOC 3	DOC 4	DOC 5
No unlabeled elements	3	☑	☑	☑	☑	☑	X	☐	☐	☑	☐
No diagram layout issues (eg. spacing, zigzag)	2	☑	☐	☐	☐	☑	X	☑	☑	☐	☐
No overlapping flows	2	☐	☐	☐	☐	☐	X	☐	☐	☐	☐
All outgoing arcs of (X)OR splits have labels	1	☑	☑	☑	☑	☑	X	☐	☐	☑	☐
		0.75	**0.5**	**0.5**	**0.5**	**0.75**	**X**	**0.25**	**0.25**	**0.5**	**0.38**
Correctenss											
No syntactical and behavioral correcteness violations (eg. deadlocks)	3	☑	☑	☑	☑	☑	X	☐	☐	☑	☐
No structural correctness violations (eg. disconnected elements or paths)	3	☑	☑	☑	☑	☑	X	☐	☐	☑	☐
No reduntant flows/elements (no hallucinations introduced)	2	☑	☑	☑	☑	☐	X	☐	☐	☑	☐
		1.0	**1.0**	**1.0**	**1.0**	**0.75**	**X**	**0.0**	**0.0**	**1.0**	**0.0**
Completness											
No missing start/end events	3	☑	☑	☑	☑	☑	X	☐	☑	☑	☑
No missing decision outcomes (eg. Both 'Yes' and 'No' paths are modeled)	2	☑	☑	☑	☑	☑	X	☐	☐	☑	☑
		1.0	**1.0**	**1.0**	**1.0**	**1.0**	**X**	**0.0**	**0.6**	**1.0**	**1.0**

Across all evaluated documents, the workflow achieves consistently high scores in the Correctness and Completeness dimensions, indicating that the generated BPMN models are structurally sound, free of disconnected elements and deadlocks, and capture all essential process components, including decision outcomes and start/end events.

The Clarity dimension exposes the primary limitation of the generated models. While all models consistently contain labeled elements and properly labeled outgoing decision flows, recurrent issues related to diagram layout and overlapping sequence flows are observed. This indicates that the main challenge for the LLM does not lie in conceptual or structural process modeling, but rather in the visual arrangement and routing of BPMN elements. These issues are purely visual in nature and do not affect the logical correctness or completeness of the underlying process models.

The zero-shot prompting baseline exhibits substantially lower and less stable performance across the evaluated dimensions. In particular, deficiencies in diagram layout generation and incomplete or missing BPMN-DI information frequently hinder meaningful qualitative assessment. In one evaluated case (DOC 1), the generated BPMN XML did not satisfy the minimal requirements for visual interpretability, as the model lacked BPMN-DI information required to render the process diagram. As a result, the process model could not be visually inspected, which directly affected both Clarity and Correctness evaluations.

This qualitative evaluation, which follows the 3C framework, was conducted using a subset of its original criteria, excluding those that require a reference process model, explicit control-flow specifications in the prompt, or lane-based modeling.

3.2 XML-Based Results

Table 3 presents the results of the objective XML-based evaluation, which provides a quantitative assessment of the formal structure, completeness, and minimal technical prerequisites of the generated BPMN models, independent of visual inspection.

Table 3. XML-based results

	WORKFLOW					ZERO-SHOT				
Clarity	DOC 1	DOC 2	DOC 3	DOC 4	DOC 5	DOC 1	DOC 2	DOC 3	DOC 4	DOC 5
Labeled elements	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Labeled outgoing flows	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	1.0	0.0
BPMN-DI Coverage	1.0	1.0	1.0	1.0	1.0	0.0	0.17	0.56	1.0	0.85
	1.0	**1.0**	**1.0**	**1.0**	**1.0**	**0.66**	**0.72**	**0.52**	**1.0**	**0.62**
Correctenss										
Element connectivity	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Balanced getaways	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Valid sequence flow	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	1.0	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**
Completness										
Task coverage	0.94	0.92	1.0	1.0	1.0	-	-	-	-	-
Decision coverage	0.83	0.71	0.75	0.91	1.0	-	-	-	-	-
	0.89	**0.82**	**0.88**	**0.96**	**1.0**	**-**	**-**	**-**	**-**	**-**

Across all evaluated documents, the proposed workflow achieves consistently maximal scores in both the Correctness and Clarity dimensions. These results indicate the formal soundness and technical completeness of the workflow-generated BPMN XML.

In terms of Completeness, the workflow exhibits minor variability. This indicates that while the workflow successfully transfers the majority of previously extracted process elements into the final BPMN model, some information loss may occur during the multi-step generation process. Nevertheless, completeness scores remain consistently high across all documents.

In contrast, the zero-shot prompting approach shows pronounced deficiencies in the Clarity dimension. In particular, incomplete or missing BPMN-DI information and inconsistently labeled outgoing decision flows are observed across multiple documents. These findings align with the qualitative 3C-based evaluation and demonstrate that the limitations of zero-shot prompting are not purely perceptual but are reflected in the underlying XML structure.

Correctness metrics reach perfect scores for the zero-shot models, indicating syntactically valid BPMN XML. However, the absence of reliable BPMN-DI information significantly limits the practical usability and visual interpretability of the generated models.

Completeness metrics were not evaluated for the zero-shot approach, as no explicit extraction of process elements was performed.

3.3 Discussion

Although both approaches can generate syntactically valid BPMN XML, notable differences emerge in their stability, completeness, and suitability for practical BPMN modeling.

The workflow demonstrates a high degree of robustness across all evaluated documents, yielding consistently structured, fully visualizable BPMN models with complete BPMN-DI information. In contrast, the zero-shot approach exhibits pronounced variability: some generated models achieve acceptable structural quality, while others suffer from incomplete diagram information or missing visual representations altogether. This variability significantly limits the reliability of zero-shot prompting for systematic BPMN model generation.

A key distinction lies in the consistent generation of BPMN-DI data. The proposed workflow reliably produces models that can be directly visualized and manipulated in standard BPMN modeling tools, which is essential for practical adoption by domain experts and process modelers. Although zero-shot prompting can generate formally valid XML, the absence or incompleteness of BPMN-DI information frequently renders the resulting models difficult or impossible to interpret visually, reducing their applicability in real-world modeling scenarios. Our presented approach is subject to several limitations that should be considered when interpreting the results. Use of specific large language model for process model generation - while it can provide useful insights at the given point in time the results can change as the technology continues to evolve. The selection was motivated by community adoption and its' extensive coverage in recent studies. Another limitation is focus on the processes in single clinical domain that can create bias toward specific process types. Furthermore, the study is based on a single LLM (ChatGPT), which may limit the generalizability of the results across different model architectures [25]; future work will therefore explore comparative evaluations involving multiple LLMs. Additionally, LLM-based generation is inherently sensitive to prompt formulation, where even minor changes in wording can lead to variations in output [4]; to mitigate this, the exact prompts used in the experiments are provided to ensure transparency and reproducibility.

Overall, the results indicate that the primary strength of the proposed workflow lies not only in formal correctness, but in its ability to consistently generate complete, executable, and tool-compatible BPMN models. At the same time, the observed layout-related issues suggest that visual optimization remains a challenging aspect of LLM-based process modeling and represents a direction for further improvement.

4 Conclusion

This work investigated the use of LLM-based chatbots to support process modeling from complex healthcare documents and introduced a multi-step workflow for directly generating BPMN process models encoded in XML. The results demonstrate that the proposed workflow produces structurally correct, complete, and consistently visualizable BPMN models, outperforming a zero-shot prompting baseline in terms of stability, clarity, and practical usability.

Despite these improvements, limitations remain. In particular, the visual layout quality of generated diagrams is not always optimal, and semantic correctness with respect to clinical intent is not formally verified and still requires expert review. Future work will therefore focus on validating generated process models against the original clinical procedures in collaboration with domain experts, enabling a systematic assessment of clinical fidelity and differences between workflow-based and zero-shot generation. Additional directions include improving layout-aware generation strategies, incorporating expert feedback into the workflow, and extending evaluation with reference models to support deeper semantic validation of generated BPMN models.

Acknowledgements. This work was funded by the European Union through the NextGenerationEU initiative under the Recovery and Resilience Plan for Slovakia, project No. 09I05-03-V02-00084, *Digital solutions in support of mental health in patients with CHF (DigiMent)*. The research was also supported by EU's DEP Sustainable Healthcare with Digital Health Data Competence and by the NATO Science for Peace and Security Program under project SPS G8765 - "PerRsonalised Micro Environmental change Alert - PRIME Alert".

References

1. Javaid, M., Haleem, A., Singh, R.P.: ChatGPT for healthcare services: an emerging stage for an innovative perspective. BenchCouncil Trans. Benchmarks Stand. Eval. (2023)
2. Udvardi, O., Sikula, M., Batalik, P., Lang, J.: Personalized learning support. In: 2025 IEEE Digital Education and MOOCS Conference (DEMOcon), pp. 1–6 (2025)
3. Prescott, M.R., et al.: Comparing the efficacy and efficiency of human and generative AI: qualitative thematic analyses. JMIR AI (2024)
4. Atil, B., Chittams, A., Fu, L., Ture, F., Xu, L., Baldwin, B.: LLM stability: a detailed analysis with some surprises (2024)
5. Grathwol, D., van der Aa, H., López, H.A.: Automating pathway extraction from clinical guidelines: a conceptual model, datasets and initial experiments. In: International Conference on Cooperative Information Systems. Springer (2024)
6. Castagnari, E., Muyama, L., Coulet, A.: Prompting large language models for supporting the differential diagnosis of anemia. In: 2024 2nd International Conference on Foundation and Large Language Models (FLLM). IEEE (2024)
7. Yu, Y., et al.: Using large language models to retrieve critical data from clinical processes and business rules. Bioengineering (2024)
8. Kourani, H., et al.: Leveraging large language models for enhanced process model comprehension. Decis. Support Syst. (2025)
9. Gomes, F.C.: Generating process descriptions from BPMN models: towards an approach with large language models (2024). https://lume.ufrgs.br/handle/10183/287897
10. Kourani, H., Berti, A., Schuster, D., van der Aalst, W.M.: Process modeling with large language models. In: International Conference on Business Process Modeling, Development and Support. Springer (2024)
11. Kourani, H., Berti, A., Schuster, D., van der Aalst, W.M.: Evaluating large language models on business process modeling: framework, benchmark, and self-improvement analysis. Softw. Syst. Model. (2025)
12. Köpke, J., Safan, A.: Efficient LLM-based conversational process modeling. In: International Conference on Business Process Management. Springer (2024)

13. Grohs, M., Abb, L., Elsayed, N., Rehse, J.-R.: Large language models can accomplish business process management tasks. In: International Conference on Business Process Management. Springer (2023)
14. Nour Eldin, A., Assy, N., Anesini, O., Dalmas, B., Gaaloul, W.: Nala2bpmn: automating BPMN model generation with large language models. In: International Conference on Cooperative Information Systems. Springer (2024)
15. Daclin, N., Mallek-Daclin, S., Zacharewicz, G.: Generative AI for business model generation (GAI4BM): from textual description to business process model. In: The 10th International Food Operations and Processing Simulation Workshop (2024)
16. Toxtli, C., Li, W.: Automating automation: using LLMs to generate BPMN workflows for robotic process automation. In: World Congress in Computer Science, Computer Engineering & Applied Computing. Springer (2024)
17. Drakopoulos, P., Malousoudis, P., Nousias, N., Tsakalidis, G., Vergidis, K.: Do LLMs speak BPMN? An evaluation of their process modeling capabilities based on quality measures. Computation **14**(1), 10 (2026)
18. Ahn, J.J., Yin, W.: Prompt-Reverse Inconsistency: LLM Self-Inconsistency Beyond Generative Randomness and Prompt Paraphrasing. arXiv preprint arXiv:2504.01282 (2025)
19. Object Management Group (OMG). Business Process Model and Notation (BPMN) Version 2.0.2 (2014)
20. Ministry of Health of the Slovak Republic. Agorafóbia (2021)
21. Ministry of Health of the Slovak Republic. Komplexný psychologický manažment dospelého pacienta s chronickou bolesťou (2021)
22. Ministry of Health of the Slovak Republic. Komplexný psychologický manažment dospelého pacienta s depresívnou poruchou a recidivujúcou depresívnou poruchou (2021)
23. Ministry of Health of the Slovak Republic. Komplexný psychologický manažment pacienta so somatickým ochorením (2021)
24. Ministry of Health of the Slovak Republic. Posttraumatická stresová porucha (2021)
25. Liang, P., et al.: Holistic evaluation of language models. arXiv preprint arXiv:2211.09110 (2022)