A Work Project, presented as part of the requirements for the Award of a Master's degree in Business Analytics from the Nova School of Business and Economics.

AI-DRIVEN BUSINESS PROCESS MANAGEMENT SYSTEMS: INTEGRATING LARGE LANGUAGE MODELS WITH PROCESS ANALYTICS AND SIMULATION

Alexander Efremov

Work project carried out under the supervision of:

Rodrigo Belo

08/01/2025

As organizations face rising complexity in their processes, analytics and Artificial Intelligence (AI) become key enablers for Business Process Management (BPM). Recent progress in Generative AI, especially Large Language Models (LLMs), opens more opportunities to address BPM use cases. Yet, the implementation and development of holistic AI-driven systems is scarcely explored. This work develops a proof-of-concept system, integrating LLMs with prescriptive analytics for the BPM domain. Specifically, the system incorporates an LLM-based architecture and a custom-developed process simulation engine. It demonstrates the capability of utilizing LLMs to enable conversational process simulation and serves as a framework for future AI-driven solutions.

AI-driven BPM, Large Language Models (LLMs), Process Analytics, Process Simulation, Prescriptive Analytics, What-If-Analysis, Business Process Management (BPM)

I would like to extend my heartfelt gratitude to Rodrigo Belo, whose guidance and advice were invaluable in refining this abstract and far-reaching topic into a focused and concise study.

I also wish to thank [Company name redacted], a leading strategy and innovation consultancy, for providing the technical infrastructure, resources, and expertise for this study. Special thanks to my advisors from the business side, for introducing me to this exciting vision for AI in BPM and encouraging me to pursue this ambitious challenge.

This work used infrastructure and resources funded by Fundação para a Ciência e a Tecnologia (UID/ECO/00124/2013, UID/ECO/00124/2019 and Social Sciences DataLab, Project 22209), POR Lisboa (LISBOA-01-0145-FEDER-007722 and Social Sciences DataLab, Project 22209) and POR Norte (Social Sciences DataLab, Project 22209).

1 Introduction

Every collection of organizational activities, transforming inputs to outputs can be described as a process (Davenport and Short 1990). Business Process Management (BPM), the organizational practice that concerns the identification, modeling, measurement, and optimization of these business processes, is therefore an essential backbone of successful business operations (Melão and Pidd 2000). Over time, BPM has moved from optimization through process reengineering to a continuous, data and technology-driven approach, increasingly utilizing data analytics and Artificial Intelligence (AI) (Armistead, Pritchard, and Machin 1999; Chen, Chiang, and Storey 2012; Dumas et al. 2023).

Data centric technologies have primarily enabled the analysis of structured data such as data tables or performance logs (Chen, Chiang, and Storey 2012). Specific technologies like process mining have also made use of the unique characteristics of a process, allowing analytics practitioners to gain transparency and insights into their processes like never before (Reinkemeyer 2020). Furthermore, the simulation of processes allows to prescriptively analyze the process to answer "what-if" questions (Gröger, Schwarz, and Mitschang 2014). Further evolutions have been leveraging AI for prediction tasks, and the recent developments in generative AI (GenAI) have further unlocked the potential to also analyze unstructured data (Vaswani et al. 2023). GenAI technologies, such as OpenAIs Large Language models (LLMs) models have shown great performance in understanding and even solving natural language tasks (Brown et al. 2020). These models further enable users to interact with them conversationally, removing many technical barriers that other analytics technologies impose (Sallam 2023).

GenAI has made significant surges in industry adoption rates in the past two years. According to a recent industry study by McKinsey, out of the 1,365 respondents 72% have indicated the adoption of AI technologies and 67% indicated the adoption of GenAI technologies in their organizations (McKinsey 2024).

However, the same report suggests that adoption in highly process-driven and BPM related functions such as “Service operations” or “Supply chain and inventory management” are rather low with 16% and 6% respectively (McKinsey 2024). Among the reasons for this unmatched potential are the risks that GenAI systems pose, such as their tendency to hallucinate incorrect statements and follow illogical reasoning (Kim et al. 2023; OpenAI et al. 2024). Errors that most deterministic process analytics techniques such as process mining and simulation do not exhibit. Furthermore, researchers and practitioners have developed and adopted isolated use cases utilizing AI for the BPM domain, such as creating business models from natural language or processing unstructured invoices, but other proposed holistic systems are still lacking in development and adoption (Dumas et al. 2023; Kampik et al. 2023).

For instance, analysts managing a Purchase-to-Pay (P2P) process, covering all steps from posing a purchase request to receiving and paying for the item, are conducting many manual and time-consuming BPM tasks. To make decisions they rely on manually analyzing the data or creating spreadsheet models based on gut assumptions. These limitations further highlight the need for a conversational AI system, allowing to quickly describe simulation scenarios, execute them, and utilize the result analysis in decision-making.

This work addresses these challenges by developing a proof-of-concept (PoC) system, which integrates LLMs with prescriptive process analytics techniques i.e., process simulation. The system aims to achieve the objectives of providing a modular and extensible LLM-based framework, integrated process analytics techniques including simulation, and finally providing a user-friendly and interactive user interface. To achieve this, the following research questions (RQ) are posed:

- RQ 1: How can LLMs be leveraged effectively for a system architecture capable of analyzing natural language queries and triggering deterministic BPM tools?
- RQ 2: How can LLMs further help facilitate human centric process analytics, such as simulation, to improve decision-making?

This work follows a simplified Design Science Process outlined by Peffers et al. (2007): The background research presented in section 2 covers the problem identification, section 3 defines the motivation and the objectives of the PoC system, section 4 presents the design and development of the PoC system. Section 5 demonstrates and evaluates the system's capabilities, with section 6 finally discussing the results of the work. This written work presents the demonstration aspect of the design science process.

2 Background

2.1 Business Process Management

A process can be described as a collection of activities that transform a given input into a desired output. Using that definition, most activities in a business organization can be described as processes, therefore it is unsurprising that an effort is allocated to optimize processes in a business context. Early efforts focused on achieving that goal by radically overhauling business processes using methodologies commonly labeled as "Business Process Reengineering". Later, companies and scientists moved toward more sustainable measures and practices collectively known as Business Process Management (BPM) (Melão and Pidd 2000).

BPM focuses on modeling, measuring, and optimizing business processes: After initially identifying all relevant processes in the organization, modeling such as the creation of graphical flowcharts is a standard practice in organizations and helps understanding and communicating the process (Bandara, Gable, and Rosemann 2005). Further, BPM concerns itself with the measurement and evaluation of processes as they are executed, ideally against operational and strategic targets linked to the business strategy (Smart, Maddern, and Maull 2009). Lastly, BPM practices, stemming from their purpose to bring value to the organization, aim to optimize and notably automate business processes using technology like Robotic Process Automation (RPA), or recently, Artificial Intelligence (AI) (van der Aalst, Bichler, and Heinzl 2018). BPM is supported by many organizational entities such as process strategies, employee roles organized in process teams, as well as performance plans and KPIs (Armistead, Pritchard, and Machin 1999). As processes and BPM have grown with complexity many different roles emerged inside and outside the process teams (Margherita 2014). This makes BPM a highly interconnected and interdisciplinary field inside an organization. Furthermore, as business processes are considered a generic factor as well as a strategic asset in most companies, the improvement of BPM, just like BPM itself, is a never-ending optimization journey (Armistead, Pritchard, and Machin 1999; McCormack and Johnson 2001). Companies are still facing significant challenges managing this complexity as well as the need to balance the proper amount of change and innovation focusing on their key processes to stay relevant and successful in their respective field (Trkman 2010).

2.2 Analytics and AI in BPM

The utilization of Information Systems (IS) in BPM has been proposed and well-studied as soon as the field emerged (Fiedler, Grover, and Teng 1995). IS technology is a key component to BPM, ranging from execution supporting systems such as workflow engines or automation software to more performance related analytics technology (Margherita 2014). Thanks to database-driven based systems like Enterprise Resource Planning tools (ERP) or Customer Relationship Management systems (CRM) such as SAP or Salesforce, a huge amount of structured data is available to analyze in order to support BPM goals (Chen, Chiang, and Storey 2012; Smart, Maddern, and Maull 2009; Reinkemeyer 2020). Classical analytics, like monitoring dashboards or predictive forecasting, enable data-driven optimization and even process innovation (Chen, Chiang, and Storey 2012; Mikalef and Krogstie 2020). Another technique called process mining makes use of process data specific characteristics, such as their sequential nature, and allows users to model process flows from different systems into so-called event logs, visually analyze those event logs to discover process variants and deviations, as well as analyze performance metrics and calculate KPIs (Chen, Chiang, and Storey 2012; Reinkemeyer 2020). Process mining thereby allows to gain new insights into business processes to generate value by increasing process efficiency and productivity, and boosting customer satisfaction (Badakhshan et al. 2022). While process mining performs great for understanding processes, additional technologies and methodologies are needed to leverage that information into useful insights or actions beyond obvious optimization potentials (Gröger, Schwarz, and Mitschang 2014). An example of prescriptive techniques, which focus on answering "what-to-do" questions, is business process simulation, which allows practitioners to run a digital process model through simulated time (Greasley 2019). It also allows quick adjustments of model parameters to improve the model or to test out different scenarios such as e.g. the improvement of processing time in a process step or allocation of additional resources to a different step (Greasley 2019). This allows for the predictive and prescriptive answering of questions regarding not only what happened in a process but also what will happen in it, and what the BPM practitioner should do to lead to it or prevent it from happening (Greasley 2019; Gröger, Schwarz, and Mitschang 2014).

When combining the amount of structured data available in processes with recent advances in the availability, quality, and costs of Machine Learning (ML) techniques, Artificial Intelligence (AI) has also gained a foothold in the BPM community (Dumas et al. 2023). Classic ML models can predict the likelihood of process delays, forecast process outcomes based on historical data, as well as identify anomalies (Kaftantzis et al. 2024). These cases work well with the structured data available in the context of process IS, or also with the data leveraged by applying process mining (Khan et al. 2023; Kaftantzis et al. 2024). However, another type of AI is required to process unstructured data such as images or natural language. Recent advances in Generative AI (GenAI) have enabled AI-systems to also make use of this type of data, such as e-mail communication, customer complaints or invoices (Dwivedi et al. 2023). One very prominent example has been the development of Large Language Models (LLMs), large-scale ML-models trained on a vast amount of language data capable of processing and generating natural language (Vaswani et al. 2023). LLM-based systems like ChatGPT, provide users with conversational interfaces that empower them to formulate natural language queries for natural language answers (OpenAI et al. 2024). They have further demonstrated the ability to approach novel and complex problems they have not been specifically trained for, providing answers and approaches to new and atypical scenarios (Brown et al. 2020). These capabilities have enabled LLMs to solve a variety of tasks, but they do come with significant risks as well: Due to their non-deterministic nature, LLMs have the tendency to "hallucinate" information, such as confidently providing false information or making errors in logical reasoning and should therefore be used cautiously and with safeguards (Kim et al. 2023; OpenAI et al. 2024).

The use of ML and LLMs in the BPM domain is well-studied. Generally, there are two types of papers to be found. One side focuses on specific use cases aided or enabled by AI models such as the creation of process diagrams from natural language, assessing automation potentials of process tasks using LLMs (Grohs et al. 2023) or translating process mining artifacts into natural language to enable question answering (Berti and Qafari 2023). In addition, more general approaches on how to best utilize AI holistically in BPM practices. One approach by BPM practitioners puts LLMs at the core of an "AI-augmented BPM System", assessing the potential of AI to enable self-adaptive automations, how Explainable AI can further increase the trust into such systems, and how recent advances in LLMs enable proper interactivity with human agents (Dumas et al. 2023). Another recent paper follows that call for AI-centered BPM systems and proposes the concept of "Large Process Models" – AI-driven systems that leverage LLMs in combination with BPM proven IS, process specific knowledge, and human agents for feedback and control (Kampik et al. 2023). Further, both approaches emphasize the importance of the chat-based nature often encountered in LLM solutions, as natural language conversations reduce technical knowledge barriers such as the requirement to understand business process modeling languages (Dumas et al. 2023; Kampik et al. 2023).

Summarized, BPM is a critical organizational practice, required for all organizations and showing the ability to generate value. Properly leveraging IS as well as AI can have a transformational effect on BPM. Recent research has both proposed holistic AI-driven BPM and demonstrated the ability of AI to solve isolated BPM use cases. Most of these recent advances are thanks to GenAI, specifically LLMs, which come with some challenges regarding reasoning and hallucinations. Nonetheless, LLMs are able to process unstructured data, come up with novel solutions, and converse with humans in natural language.

3 Motivation and objectives of the solution

As shown in section 2, lots of practical research has been done on specific use cases in BPM addressable by utilizing LLMs. Furthermore, some research has been proposing and designing more holistic systems incorporating LLMs as well as other IS and Analytics (see section 2.2). However, while prescriptive analytics techniques and AI have been successfully integrated for other fields such as the military domain (van Oijen and de Marez Oyens 2023) or Operations Research (Sun et al. 2024), to the best of the author's knowledge, no existing work has specifically combined these approaches in a holistic system for BPM. This gap highlights a need for a holistic, AI-driven BPM system, which properly integrates the capabilities of LLMs with the deterministic nature of prescriptive process analytics, such as simulation.

Addressing this need could further bridge the gap between isolated successful AI use cases with the proposed comprehensive systems, as well as bring further notice to the potential of prescriptive techniques. Therefore, this work aims to develop a **proof-of-concept system** that addresses the following three system objectives (SO):

1. **Modular LLM-based Framework:** Design a modular architecture that leverages LLMs for prompt dissection, action creation, and execution BPM functionalities, ensuring adaptability and scalability. (SO1)
2. **Integrated Process Analytics and Simulation:** Integrate BPM functionalities to enable prescriptive analytics, including what-if scenario simulation and analysis, seamlessly usable by the modular framework. (SO2)
3. **User-Friendly Interactivity:** Develop a natural language-driven, chat-based user interface that allows non-technical users to interact with the system, validate their ideas quickly, and gather insights generated by the system. (SO3)

4 Design and Development

To address the research gap outlined in section 2, as well as to achieve the objectives set out in section 3, this work now presents a PoC for an AI-driven BPM system, which integrates LLMs with process analytics and simulation.

4.1 System architecture

The system is designed with a modular architecture centered around three core components: a **frontend,** an **LLM-based AI agent**, and **a simulation engine**. The three decoupled components ensure the modularity and scalability of the system by allowing easy adjustments or extensions of each component.

Frontend: The user is able to interact with the system through a Streamlit-based chat interface, similar to ChatGPT. Users can input queries into the system in natural language, which are then captured by the frontend. In turn, the frontend presents the user with natural language responses, as well as visualizations such as Gantt charts. The interface is designed to make the complex process simulation as accessible and comprehensible as possible to non-technical users.

LLM-based AI agent (backend): After a query has been captured by the frontend, it is passed to the central orchestrating component: the AI agent. The agent is built using the LangChain framework and leverages GPT-4o, a state-of-the-art LLM. They enable the AI agent to interpret user inputs and identify appropriate backend actions. Based on these actions, the AI agent is then capable of triggering the required tools from the backend. Lastly, the results from the backend tools are interpreted and transformed into natural language explanations and passed to the frontend. Supporting the agent is a custom prompt template and a message history providing better context in interpreting queries.

Simulation engine (backend): The backend includes a custom-built simulation engine, implemented by using the SimPy library. The engine consists first of the developed simulation class, which allows the modeling and execution of sequential process simulations. Input parameters allow the adjustment of simulation parameters like resource capacities. Second, built upon this class is the custom process simulation tool, which encapsulates the functionalities of the simulation engine and makes them accessible for the AI agent.

Data integration and supporting functions: The system uses a lightweight folder-based structure for operational data, and outputs. For operational data, the system saves temporary results in XML, JSON, or natural language texts. For outputs, such as images or charts, the system creates session-specific folder structures, allowing users to recall past sessions.

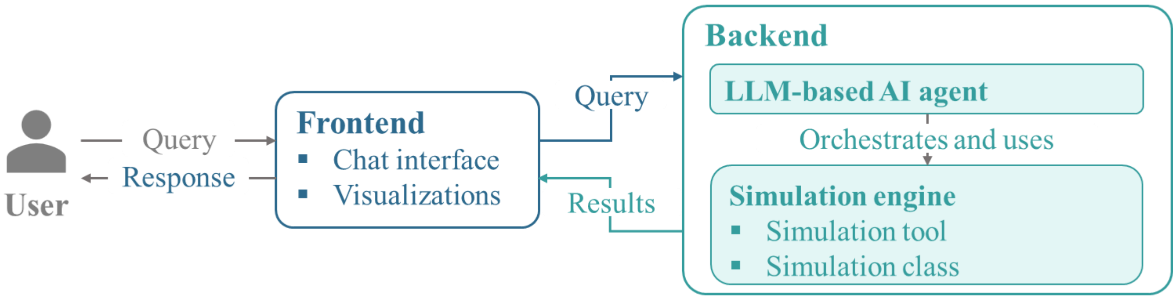


Figure 1: High level system architecture, showing the flow from query through frontend, backend, and back to frontend.

4.2 LLM-based AI agent

An LLM-based framework, represented by the AI agent, is the key component of the system. It orchestrates between the user and all other functions of the system such as the simulation engine and supporting functionalities such as logging or data storage.

Technological foundation:

The framework is primarily built using the LangChain library. It provides the core infrastructure to define and organize tools for the AI agent. Custom tools, software functions to be usable by the agent, are encapsulated in the library's interface, which provides standardized input/output schemas and allows the definition of custom tool descriptions. Further, it handles the creation of a "chain of actions". If an action matches a tool's description, the library provides the functionality to execute this tool.

Enabling LangChain is an LLM connection, for this system it is OpenAI's GPT-4o. The LLM allows the parsing of the user query into a chain of actions and can also be used as a standalone action to e.g. summarize and interpret tool results. Smaller LLMs, such as a small text-embedding model are used in the prompt creation described further on.

Development contributions:

Building on the foundation, the system utilizes a custom prompting template. The template offers the AI agent a contextual prefix, explaining its role in the system and the focus on the BPM domain. It further provides examples of how user queries should be interpreted. For instance, in the P2P example, where the user inputs the query *"What would happen to our P2P process if we reduce our warehouse staff by two people?",* the template ensures the agent understands the intent behind this query. An example picker class, built around the small text embedding LLM, allows the agent to pick the example most relevant to the query as an additional prompt for further interpretation and tool selection.

Utilizing LangChain's tool interface, custom descriptions are created for the functions of the simulation engine (see section 4.3). These descriptions emphasize the BPM domain of the tool, its expected inputs, and provided outputs. This allows the LLM to properly interpret, if the tool is appropriate for an action. For example, the agent selects the *ProcessSimulationTool* and passes a string parameter instructing the tool to *"Simulate the P2P process with two fewer people in the warehouse"*. Furthermore, a session-specific memory structure is implemented, allowing the passing of the session history to the agent's prompts, and the saving of interactions and images into a distinct folder.

4.3 Simulation engine

The simulation engine is the custom-developed component enabling prescriptive process analytics in the system. It enables the modeling, execution, and analysis of process simulations. The engine is built around a ProcessSimulation class and a tool usable by the AI agent.

Technological foundation:

The engine is developed using the SimPy library, a process-based, discrete-event simulation framework. SimPy provides the tools for modeling sequential workflows, resources, and simulation environments. SimPy allows the modeling of workflows by defining process steps as tasks and process sequences as flows. Tasks can get a mean and standard deviation duration, which is used in the simulation environment to determine their execution length. Lastly, SimPy offers resource modeling to simulate limiting constraints like employees or equipment. The shared resources are allocated to tasks and need to be requested by simulation instances before they can start a task.

Additionally, the engine uses the PM4PY process mining library to extract process diagrams and task durations from event logs. For parsing natural language descriptions of resource adjustments for the ProcessSimulation tool a GPT-4o interface is used.

Development contributions:

Utilizing this foundation, the system implements the custom ProcessSimulation class on top of the SimPy capabilities. Custom functions allow the use of the process mining outputs to generate the simulation model, including the task and process flows, as well as the task durations extracted from the XML process diagram. Other functions enable easy adjusting of task resources, durations, and simulation parameters such as how many instances are simulated over which spread of time.

Once a simulation model is created, the class provides functions to execute and analyze it. This includes functions to plot a Gantt chart visualizing the simulation instances, showing wait and processing times for each step, and summarizing results in both structured (e.g. dictionaries and JSONs) and unstructured natural language. Additionally, the class offers tools for analyzing and visualizing resource utilizations.

Encapsulating this class is the ProcessSimulation tool, designed for interaction with the AI agent. When invoked by the agent, it first connects to an event log and a resources excel (demo data descriptions can be found in Appendix B) and creates a simulation model using the class methods. Afterwards it interprets the string instruction passed by the AI agent, in the P2P example *"Simulate the P2P process with two fewer people in the warehouse"* and calls a custom parsing function to interpret the requested adjustments. The parsing function extracts the current resources from the model and passes them together with a custom prompt, and the resource change instruction to GPT-4o. Based on that, GPT-4o returns the adjustment in a structured JSON format, which in turn is imported into the simulation model. The tool finally runs the model and returns analysis summaries to the AI agent and visualizations to the frontend. The simplified flow from instruction by the AI agent to the response from the ProcessSimulation tool to the AI agent is illustrated in Figure 2.

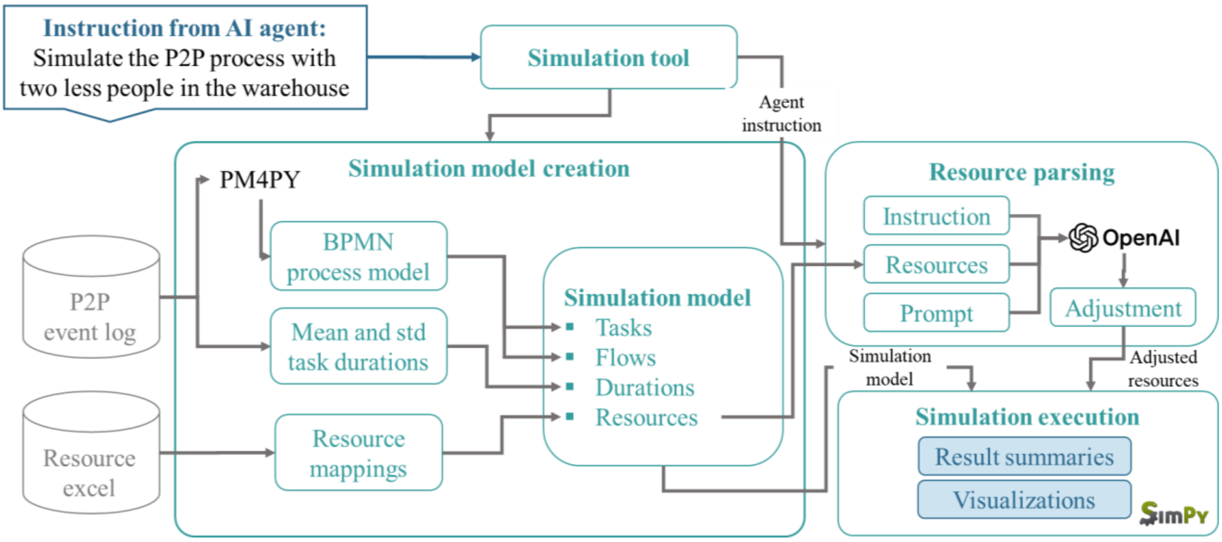


Figure 2: Simplified processing of the AI agent instruction in the simulation engine. Components of the simulation engine are encased in green, data is shown in grey, and final results passed to AI agent are highlighted in blue.

4.4 Frontend and data integration

The last enabling components of the system are its user-centric frontend and lightweight data integration layer to connect it to both human users and relevant process data.

Technological foundation:

The frontend is developed using Streamlit, a Python-based framework for building data-centric web applications. The chat-based interface is designed to be similar to industry standards such as the web interface of ChatGPT. The interface supports multi-turn conversation, i.e., the user can input follow-up queries, as well as the provision of visualizations from inside other tools. GPT-4o is used to generate appropriate session titles.

Development contributions:

The interface is customized to adjust to corporate design standards such as logos and naming. In addition, the session management of Streamlit is extended to generate folder structures for each interaction. Using this, prior interactions in the same window can be utilized as context for the AI agent.

Data integration is handled in a very lightweight manner, with operational data saved in an internal temporary data folder, and session specific data in the generated folders. For demonstration purposes, an event log (Appendix B.2) is available for the simulation engine. The final interface is displayed in Figure 3, with an example including responses in Appendix A.1.

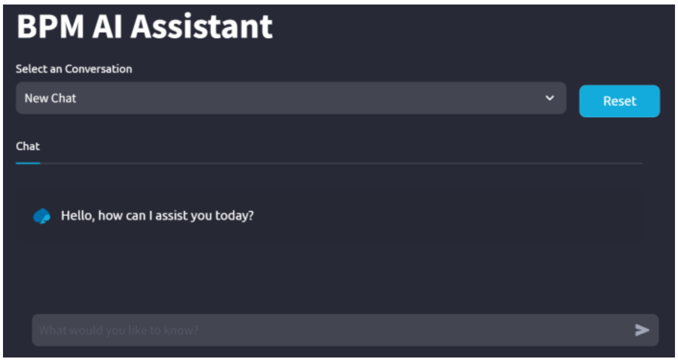


Figure 3: Screenshot showing the chat-based user interface with a drop-down for prior interactions and a query input field

4.5 Exemplary Query flow

To show the interoperability of the system, the following is an abstracted flow of a user query:

1. The user enters a natural language query such as *"What would happen to our Purchase-to-Pay process if we reduce our workers in the warehouse to 2 people?"* in the chat window.
2. The AI agent uses GPT-4o and the prompt template and generates a chain of actions, including the use of the process simulation tool with the instruction *"Simulate the P2P process with 2 fewer people in the warehouse"*.

3. The simulation tool uses the integrated event log and resources excel to create a simulation model. It then transforms the instruction into simulation parameters.
4. The simulation tool executes the simulation, displays a Gantt chart of the results to the users, and passes a natural language summary of the results to the AI agent.
5. The AI agent again uses GPT-4o to analyze the results given by the simulation tool, with the generated analysis finally displayed to the user in the chat window.

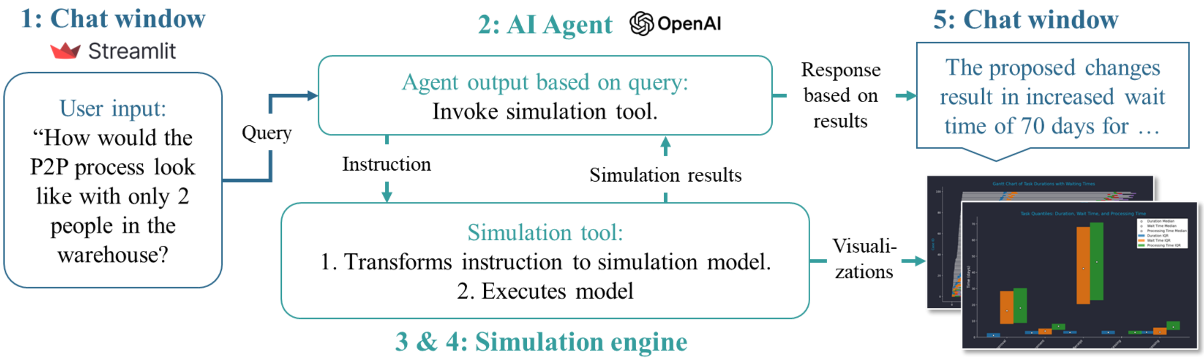


Figure 4: Simplified query flow from entering a query in the Streamlit frontend to the response and the visualizations.

The query flow, as shown in Figure 4, demonstrates the seamless integration of frontend, AI agent, and simulation engine. This integration allows the system to capture human queries in the frontend, generate a reasonable chain of actions via the AI agent, execute simulations using the engine, and deliver insights to the user.

5 Results

This section presents the results of the development of the PoC system outlined in section 4. It further presents testing results of the system against a methodology designed to assess the achievement of the objectives outlined in section 3. The evaluation methodology focused on testing the modularity of the system's framework, its integration of prescriptive analytics and simulation functionalities, and its user-friendly interactivity. It also evaluates the overall effectiveness and efficiency of the system's core functionality of enabling prescriptive analytics in a conversational format. The methodology and its objectives are summarized in Table 1:

Table 1: Developed and applied evaluation framework

Objective	Evaluation Focus	Evaluation methodology
Modular LLM-based Framework (SO1)	How well does the system process user queries and dissect them into a reasonable chain of actions?	Test 50 queries, including ambiguous and atypical cases which might be able to be handled by the system and map expected actions such as tool invocations. Run the prompts and calculate the accuracy of the system (number of correctly processed queries/ total number of queries).
Integrated Process Analytics and Simulation (SO2)	How well is the Process Simulation Engine integrated into the system?	Simulate 30 test cases by mapping generated instructions from the LLM-based Framework to expected parameters to be used in the simulations. Run the test cases and evaluate the accuracy in parsing the queries into simulation parameters.
User-Friendly interactivity (SO3)	How easy is it to use the Chat-based system?	Conduct five user studies with management consultants. The participants are given three tasks that should be solvable using the system and fill out a standardized survey to evaluate the "System Usability Scale".
All (Overall performance)	How effective is the system at providing prescriptive process analytics in a conversational format?	Test 50 queries ranging from basic to complex, evaluating three dimensions for each query correct tool use, accurate parameter parsing, and helpfulness of the response. Run the test with no additional tools, five, and 10 tools defined. Conduct those tests with and without the prompt template to evaluate its usefulness.

5.1 Evaluation of the Modular LLM-based Framework

To evaluate the modularity and the decision-making capabilities of the LLM-based framework, a series of test queries were tested. The queries were passed to an instance of the LLM-based AI agent, and the resulting internal chain of actions was evaluated against expected results, primarily expected tool usage. The dataset compromised 50 distinct queries, ranging from straightforward requests like *"Simulate the Purchase-to-Pay process with only three people in the warehouse"* to complex like *"Simulate reducing all teams to half staff except warehouse"*. To evaluate tool use robustness, four additional dummy tools with appropriate descriptions were created and added to the AI agent tools. Their descriptions can be found in Appendix C.1.2.

To evaluate the queries against the expected results, the system uses a custom callback handler functionality, allowing the extraction of the internal chain of actions generated by the agent. An example of the evaluation can be found in Appendix C.1.1. Running all queries resulted in the following performance: The system called the correct tool for 46 out of the 50 total queries, resulting in an overall accuracy of 92%. All queries resulting in wrong tool usage were related to a created dummy process mining tool. In each of them, the system did not use a tool but opted to answer the question by utilizing GPT-4o. Full results are found in Appendix C.1.3.

5.2 Evaluation of the integrated Process Analytics and Simulation

The integration of the designed prescriptive analytics tools including the simulation engine was assessed by examining how the prompts generatable by the AI agent are assessed and processed into simulation relevant parameters for the simulation functions.

To test this, 30 test cases were designed, which consisted of a short instructive query like those defined in the prompting template, as well as expected parameters for simulation resources. For example, for the instruction *"Reduce warehouse staff by 2"* the expected adjustment reduces the staff from four to two. Complex cases also evaluated multiple resource changes at once, as well as instructions with no specific numbers such as "Double all resources".

Out of the 30 test cases evaluated, 20 ran successfully, properly parsing the resource adjustments. Five resulted in expected and designed errors, not allowing the engine to reduce resources below zero. The last five resulted in unexpected parameter adjustments or unexpected errors. Most of these happened with scaling instructions e.g. asking for the tool to halve all resources. In one case, the parsing was not able to adjust multiple resources, only adjusting two out of the three expected. Extensive results can be found in Appendix C.2.

5.3 Evaluation of the User-friendly Interactivity

To evaluate the usability and interactivity outlined in objective 3, a short user study was conducted. Thanks to the developed frontend participants could test out queries related to the expected functionality of the system. Then they filled out a standardized "System Usability Scale" (SUS) questionnaire, a common benchmark in Human-Computer-Interaction and IS research (Bangor, Kortum, and Miller 2008; Lewis 2018).

The study was conducted with five management consultants with BPM and process analytics experience. Participants were given a study overview (Appendix C.3.1) which described the context of the system, the process data integrated in the system, and three tasks which they should solve using it. Two small hints were given to optimize query behavior. After the tasks, they filled out the standard SUS questionnaire. Appendix C.3.2 - C.3.4 provides more information on participants, their answers, and the score calculation.

The scores ranged between 90 and 95 points, with an average score of 93. Junior level consultants and the manager had a slightly lower score than the two senior level consultants. For all questions except one, four participants answered the question with the most favorable option with one picking the option just below that. The question *"I need to learn a lot of things before I could get going with the AI assistant"* was the divisive exception, with two participants picking the best option *"Strongly disagree"* and three people selecting *"Partially agree"*.

All participants were able to complete the tasks. They struggled with formulating the first query, not knowing which queries would trigger the simulation tool in the backend. After success with the first task, all of them achieved the rest of the tasks with more speed and ease.

5.4 Evaluation of overall performance

Finally, to evaluate the system's overall performance, 50 queries were designed in line with the advertised capabilities of the system:

1. **Basic queries (20),** e.g., *"Simulate the P2P process with one fewer finance clerk."*
2. **Advanced queries (10),** complex scenarios requiring complex resource adjustments or multiple simulations. Example: *"Compare Scenario A: Warehouse staff reduced by 1 and Scenario B: Approvers increased by 2. Analyze which scenario reduces bottlenecks."*
3. **Difficult Analytical queries (20),** requiring more nuanced analyses and multiple simulations such as e.g. *"What is the percentage change in wait time for 'Invoice Processing' when payment processors are increased by 1?"*

To further test the systems adaptability and scalability, all tests were executed under three conditions: With no additional dummy tools, with five tools defined, and with 10 tools defined. Descriptions of the designed dummy tools can be found in Appendix C4.1. Lastly, to demonstrate the benefit of the prompt template, all tests were run with and without the prompt.

The resulting 300 responses were evaluated as correct, if the system executed the ProcessSimulation tool, including multiple executions when required, used the correct parameters for each instance, and provided a response that was deemed helpful and relevant. The evaluation results are illustrated in Figure 5:

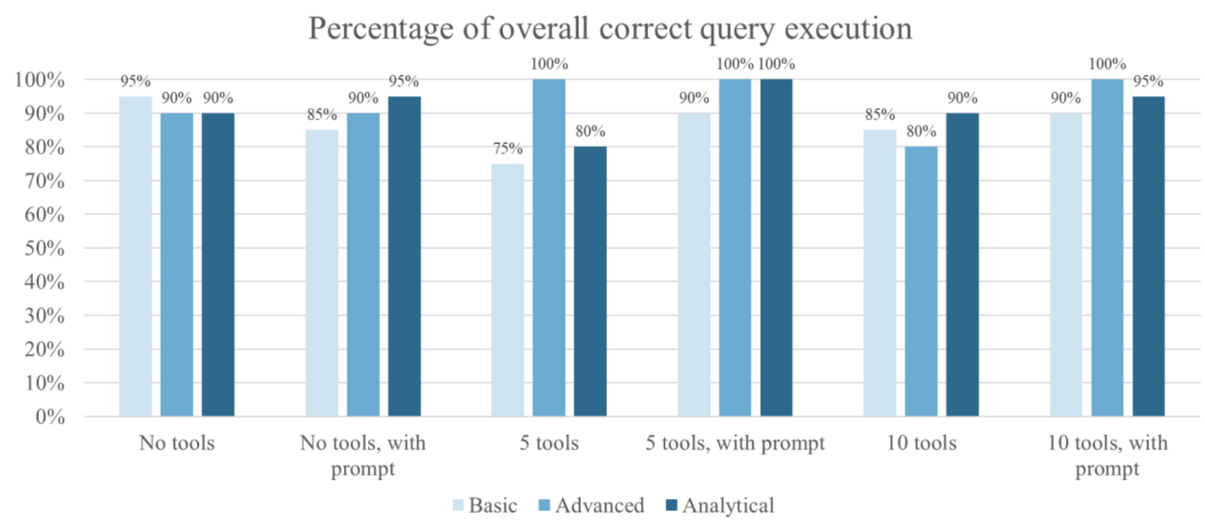


Figure 5: Grouped bar chart of test performance dissected by query difficulty (bar color) and testing scenario (group label)

Overall, the system with prompt template shows a performance of at least 90% query correctness, except for basic queries in the scenario with no tools, where it is 85%. It also performs equally or better than the system with no prompt template for all queries except the same scenario. The prompt template further allowed the system to handle almost all the analytical and advanced queries, even when more tools were introduced. Notably it enabled the system to infer the need for different scenarios that were not explicitly stated, such as simulating a benchmark for percentage calculations or simulating five scenarios for when a range of possible resource adjustments was given. Results for all queries can be found in Appendix C.4.2.

6 Discussion

This work aimed to address the gap in research and industry of integrating LLMs with prescriptive analytics techniques, specifically simulation, in the BPM domain. Guided by two research questions concerning the integration of LLMs into an effective system architecture as a central orchestrating unit (RQ1) as well as the help LLMs can provide enabling human-centric process analytics and simulation (RQ3), this work developed a PoC system and evaluated it against three objectives: the achievement of a modular LLM-based framework (SO1), the integration of process analytics and simulation (SO2), and the implementation of a user-friendly, interactive interface (SO3).

6.1 Achievement of objectives

Modular LLM-based framework (SO1):

The first objective focused on the creation of a modular system architecture capable of utilizing LLMs for human query dissection and the generation of actions with the potential to trigger other tools. The framework-specific test with 50 queries (section 5.1) resulted in a 92% accuracy and therefore indicated a very strong performance of the LangChain and GPT-4o powered framework's capability to select and trigger appropriate functionalities based on human queries. The overall system tests (section 5.4) further support this, where scaling the tools from 5 to 10 did not impact the great performance of the system ranging from one outlier of 85% accuracy to otherwise 90-100% accuracy even for complex queries. This accuracy aligns with LangChain's benchmarks for single tool use, where state-of-the-art LLMs like Claude-2.1, GPT-3.5-turbo and GPT4 performed with 85%, 95%, and 90% percent respectively ("Benchmarking Agent Tool Use" 2023). Strong performance with the tool selection, therefore, enables the system to reliably decide on and perform actions outside an LLMs capability and customizable to BPM practices. Further, it supports its modularity by providing an easy framework for integrating tools into the system. Nonetheless, some queries were not dissected as expected, indicating that more work needs to be done e.g. by refining the prompting template.

The designed framework developed as the evaluated system architecture, able to call and execute BPM related tools based on prompting and tool description, is a proof of concept for systems proposed and outlined by research such as Dumas et al. and Kampik et al. (Dumas et al. 2023; Kampik et al. 2023). It aligns with the gap identified in section 3, being a scalable and modular system able to call deterministic functions. Therefore, this work provides an answer to RQ1: "How can LLMs be leveraged effectively for a system architecture capable of analyzing natural language queries and triggering appropriate deterministic tools".

Integrated Process Analytics and Simulation (SO2):

The second objective focused on the integration of process analytics and simulation into the system, notably to allow prescriptive analytics of what-if scenarios. The objective-specific tests (section 5.2) revealed mixed results across the 30 test cases. 20 cases were properly parsed into correct simulation parameters, including more complex resource adjustments. Of the other cases five failed due to design constraints such as preventing negative resource adjustments, showcasing a robustness to real world accuracy but also an inability of the system to handle those cases more elegantly such as asking the user to enter correct parameters. The five remaining failed due to unexpected issues in parsing the parameters, specifically the inability of the simulation agent to handle mathematical scaling such as multiplying or dividing all resources. The additional overall tests (section 5.4) did not struggle as much with the resource parsing, but also mishandled the cases with mathematical scaling. These cases could be fixable with better prompting or more guard-railed prompt reception from the simulation tool.

While the success of the integration was not as clear as that of the overall system, it nonetheless shows the potential of integrating the custom simulation engine using LLMs. By passing natural language queries to the simulation engine and in turn using GPT-4o to parse it into an engine readable JSON format, the integration shows how ambiguous, non-predefined input parameters can be handled by such systems, leading to expected results in 83% of cases. Clearly, more robust handling is needed to fully achieve the objective, but by providing a way to match natural language with appropriate prescriptive analytics, the RQ2, “How can LLMs further help facilitate human centric process analytics such as simulation?” can partly be addressed.

User-Friendly Interactivity (SO3):

Lastly, the third objective is to provide a natural language-driven frontend interface, facilitating an easy and interactive way to use prescriptive process analytics functions. The usability study resulted in an exceptional mean SUS score of 93. Benchmarking this score against a variety of systems it is safely in the fourth quartile of over 200 studied systems (Bangor, Kortum, and Miller 2008). All participants successfully achieved their tasks, showing that chat-based interfaces are great at facilitating effective analytics through a user-friendly interface.

One slight flaw observable during the study and noticeable in the last question of the SUS questionnaire being dominantly answered with only the second-best option was the need for some guidance before achieving full efficiency with the system. Some effort should be concentrated on addressing this, especially if such systems are to grow in functionality they can provide. Even so, this work can further answer RQ2, by showing that chat based natural language interactions enabled by LLMs provide a very good user experience for BPM tasks.

6.2 Limitations and Future work

Although the system achieves its core objectives, this work is not without its limitations. Apart from the smaller deficits outlined in the discussion, this work was developed and written in the scope of Master's work project spanning three months, and some prioritization decisions had to be made to achieve the results presented. As the key aim of this work was to show the possibility of the PoC systems architecture and simulation capabilities, two core limitations are identified: First, evaluations have been done on a rather small scale, with the number and variety of test cases as well as the number of study participants being further expandable. Second, to demonstrate the functionality of the overall system over the process simulation functionalities, several simplifications have been made in the custom simulation engine. Notably, the process simulation models do not contain gateways and resources are modeled as employees only.

To address the identified limitations and further drive the vision of AI-driven BPM systems, this work proposes three possibilities for future work: First, further validation of the system and its functionalities should be done by expanding the number of test cases and queries and more importantly conducting more extensive user studies or case studies. Second, as the robustness to adding more functionality has been shown (5.4) the scope of the system should be extended by adding further process analytics tools, and leverage LangChain capabilities of chaining and invoking multiple actions to make use of their synergies. Lastly, proper integration into real world scenarios should be developed and researched, by, e.g., utilizing real-world process data and building interfaces into BPM-related IS as proposed by research (Kampik et al. 2023).

6.3 Practical implications

This work highlights the potential of AI-driven BPM systems addressing complex process management tasks. By integrating LLMs with prescriptive analytics, the system and its possible evolutions can be leveraged in practice, promising a variety of benefits.

Considering the P2P example followed throughout this work: An analyst responsible for the P2P process is trying to decide where to allocate two seasonal workers for the coming holiday season. Instead of manually collecting process-relevant data and modeling scenarios in a spreadsheet based on assumptions, they use the AI-driven BPM system. By defining available options in a query, and asking the system to simulate results, they get data-driven insights and recommendations in a few minutes. With further integration and extensions, the system could be adapted to handle more processes and more complex adjustment scenarios, like regulatory changes, supply chain disruptions, or epidemics affecting the workforce. Managers could ask the system about the implications of such changes and test different mitigation strategies.

These scenarios showcase the potential benefits of implementing and extending such a system in BPM practices: Their efficiency, simulation functionalities and laymen-friendly user interface can boost speed and quality of decision making. However, the use of LLMs inside the system poses some risks, as generated chains of actions and responses can be confidently false, leading to misinterpretations, wrong conclusions, or false simulation parameters. Some risks can be easily detected and mitigated by users, such as for wrong tools picked (as discussed in 6.1), users can adjust the query and retry. Others, such as the passing of wrong parameters for simulation, while visible in the system, might not be noticed without close inspection. Therefore, following the responses and recommendations blindly can lead to false decisions. Managers and users need to always be aware of the limitations and risks of the system and implement mitigation processes or functions. By addressing these risks, the system can become a robust, efficient, and reliable decision-making and support tool for BPM practitioners.

7 Conclusion

This work aimed to bridge a gap in the integration of LLMs with prescriptive analytics via the example of simulation, for BPM practices. By addressing research questions of how LLMs can be effectively used for system architectures capable of executing deterministic tools (RQ1) and how LLMs can further support systems in providing a user-friendly way to conduct process analytics (RQ2), this study has designed and developed a PoC system evaluated against three key objectives: a modular LLM-based architecture capable of using other tools, an integrated process simulation functionality, and a chat-based interface.

An LLM-based framework helps achieve the first objective, putting an AI agent at the core of a system architecture creating chains of actions including using a custom process simulation engine. By integrating this engine, internally utilizing further analytics such as process mining to derive a simulation model, the system achieves its second objective but with some areas of improvement regarding handling complex or incorrect parameters. Finally, a chat-based interface enables users to easily access the tools integrated into the system. The very good results of the user study show the achievement of the third objective.

This work lays the foundation for extendable, AI-driven BPM systems that could further transform how organizations conduct process optimization and decision-making. To address limitations relating to the limited scale of the evaluation as wells as the functional scope of the system, this work concludes with a call for further practical and academic work, to build on the presented results and the PoC system design by 1) validating the shown results on a broader scope, 2) enhancing amount and capabilities of the process analytics tools, and 3) developing and implementing AI-driven BPM systems in real-world scenarios.

Bibliography

Aalst, Wil M. P. van der, Martin Bichler, and Armin Heinzl. 2018. "Robotic Process Automation." *Business & Information Systems Engineering* 60 (4): 269–72.

Armistead, Colin, Jean-Philip Pritchard, and Simon Machin. 1999. "Strategic Business Process Management for Organisational Effectiveness." *Long Range Planning* 32 (1): 96–106.

Badakhshan, Peyman, Bastian Wurm, Thomas Grisold, Jerome Geyer-Klingeberg, Jan Mendling, and Jan vom Brocke. 2022. "Creating Business Value with Process Mining." *The Journal of Strategic Information Systems* 31 (4): 101745.

Bandara, Wasana, Guy Gable, and Michael Rosemann. 2005. "Factors and Measures of Business Process Modelling: Model Building through a Multiple Case Study." *European Journal of Information Systems* 14 (4): 347-360.

Bangor, Aaron, Philip T. Kortum, and James T. Miller. 2008. "An Empirical Evaluation of the System Usability Scale." *International Journal of Human–Computer Interaction* 24 (6): 574–594.

Berti, Alessandro, and Mahnaz Sadat Qafari. 2023. "Leveraging Large Language Models (LLMs) for Process Mining (Technical Report)." *arXiv*: 2307.12701.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. "Language Models Are Few-Shot Learners." *arXiv*: 2005.14165.

Chen, Hsinchun, Roger H. L. Chiang, and Veda C. Storey. 2012. "Business Intelligence and Analytics: From Big Data to Big Impact." *MIS Quarterly* 36 (4): 1165–1188.

Davenport, T.H., and J.E. Short. 1990. "The New Industrial-Engineering - Information Technology and Business Process Redesign." *Sloan Management Review* 31 (4): 11–27.

Dumas, Marlon, Fabiana Fournier, Lior Limonad, Andrea Marrella, Marco Montali, Jana-Rebecca Rehse, Rafael Accorsi, et al. 2023. "AI-Augmented Business Process Management Systems: A Research Manifesto." *ACM Trans. Manage. Inf. Syst.* 14 (1): Article 11.

Dwivedi, Yogesh K., Nir Kshetri, Laurie Hughes, Emma Louise Slade, Anand Jeyaraj, Arpan Kumar Kar, Abdullah M. Baabdullah, et al. 2023. "Opinion Paper: 'So What If ChatGPT Wrote It?' Multidisciplinary Perspectives on Opportunities, Challenges and Implications of Generative Conversational AI for Research, Practice and Policy." *International Journal of Information Management* 71: 102642.

Fiedler, K.D., V. Grover, and J.T.C. Teng. 1995. "An Empirical-Study of Information Technology Enabled Business Process Redesign and Corporate Competitive Strategy." *European Journal of Information Systems* 4 (1): 17–30.

Greasley, Andrew. 2019. *Simulating Business Processes for Descriptive, Predictive, and Prescriptive Analytics*. Walter de Gruyter GmbH & Co KG.

Gröger, Christoph, Holger Schwarz, and Bernhard Mitschang. 2014. "Prescriptive Analytics for Recommendation-Based Business Process Optimization." In *Business Information Systems*, edited by Witold Abramowicz and Angelika Kokkinaki, 25–37. Springer International Publishing.

Grohs, Michael, Luka Abb, Nourhan Elsayed, and Jana-Rebecca Rehse. 2023. "Large Language Models Can Accomplish Business Process Management Tasks." *arXiv*: 2307.09923.

Kaftantzis, Savvas, Alexandros Bousdekis, Georgia Theodoropoulou, and Georgios Miaoulis. 2024. "Predictive Business Process Monitoring with AutoML for next Activity Prediction." *Intelligent Decision Technologies* 18 (3): 1965–1980.

Kampik, Timotheus, Christian Warmuth, Adrian Rebmann, Ron Agam, Lukas N. P. Egger, Andreas Gerber, Johannes Hoffart, et al. 2023. "Large Process Models: Business Process Management in the Age of Generative AI." *arXiv*: 2309.00900.

Khan, Asjad, Aditya Ghose, Hoa Dam, and Arsal Syed. 2023. "A Survey of Process-Oriented Data Science and Analytics for Supporting Business Process Management." *arXiv*: 2301.10398.

Kim, Jeong Hyun, Jungkeun Kim, Jooyoung Park, Changju Kim, Jihoon Jhang, and Brian King. 2023. "When ChatGPT Gives Incorrect Answers: The Impact of Inaccurate Information by Generative AI on Tourism Decision-Making." *Journal of Travel Research* 64 (1): 51-71.

LangChain. 2023. "Benchmarking Agent Tool Use". *LangChain Blog December 2023*. Accessed December 13, 2024. https://blog.langchain.dev/benchmarking-agent-tool-use/.

Lewis, James R. 2018. "The System Usability Scale: Past, Present, and Future." International *Journal of Human-Computer Interaction* 34 (7): 577–590.

Margherita, Alessandro. 2014. "Business Process Management System and Activities." *Business Process Management Journal* 20 (5): 642–662.

McKinsey & Company. 2024 "The State of AI in Early 2024". *McKinsey & Company Insights.* Accessed December 7, 2024. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai.

McCormack, Kevin, and B Johnson. 2001. "Business Process Orientation, Supply Chain Management, and the e-Corporation." *IIE Solutions* 33 (10):33–37.

Melão, Nuno, and Michael Pidd. 2000. "A Conceptual Framework for Understanding Business Processes and Business Process Modelling." *Information Systems Journal* 10 (2): 105–129.

Mikalef, Patrick, and John Krogstie. 2020. "Examining the Interplay between Big Data Analytics and Contextual Factors in Driving Process Innovation Capabilities." *European Journal of Information Systems* 29 (3): 260–287.

Oijen, Joost van, and Pieter de Marez Oyens. 2023. "Empowering Military Decision Support through the Synergy of AI and Simulation." *Paper presented at The 2023 NATO NMSG Symposium, Monterey, CA, October.*

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, et al. 2024. "GPT-4 Technical Report." *arXiv*: 2303.08774.

Peffers, Ken, Tuure Tuunanen, Marcus Rothenberger, and S. Chatterjee. 2007. "A Design Science Research Methodology for Information Systems Research." *Journal of Management Information Systems* 24 (01):45–77.

Reinkemeyer, Lars. 2020. "Process Mining in a Nutshell." In *Process Mining in Action: Principles, Use Cases and Outlook*, edited by Lars Reinkemeyer, 3–10. Springer International Publishing.

Sallam, Malik. 2023. "ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns." *Healthcare* 11 (6): 887.

Smart, P. A., H. Maddern, and R. S. Maull. 2009. "Understanding Business Process Management: Implications for Theory and Practice." *British Journal of Management* 20 (4): 491–507.

Sun, Wei, Scott McFaddin, Linh Ha Tran, Shivaram Subramanian, Kristjan Greenewald, Yeshi Tenzin, Zack Xue, Youssef Drissi, and Markus Ettl. 2024. "PresAIse, a Prescriptive AI Solution for Enterprise." *INFOR: Information Systems and Operational Research* 62 (4): 629–645.

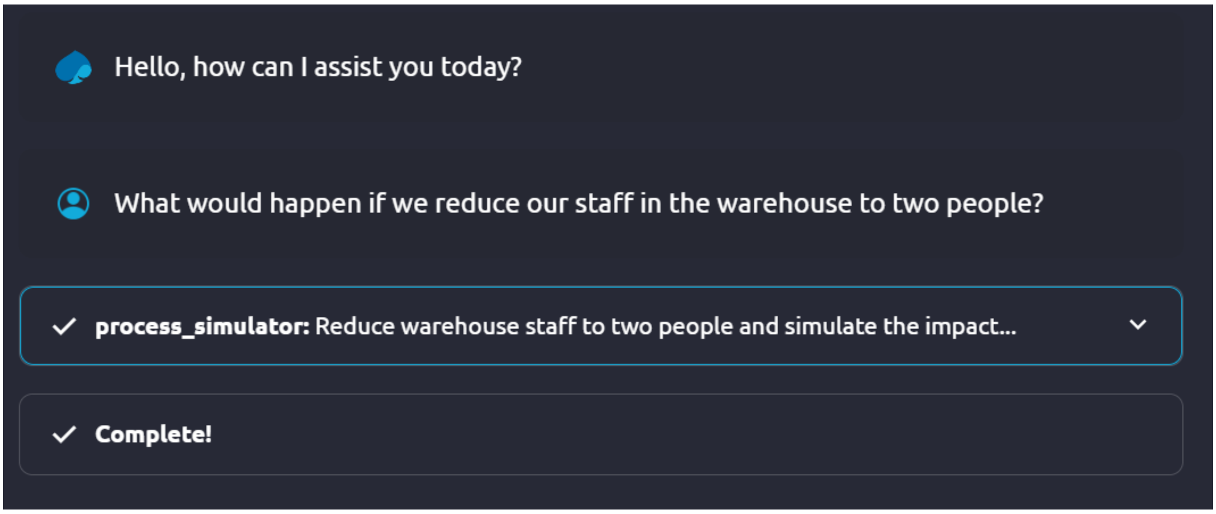
Trkman, Peter. 2010. "The Critical Success Factors of Business Process Management." *International Journal of Information Management* 30 (2): 125–34.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. "Attention Is All You Need." *arXiv*: 1706.03762.

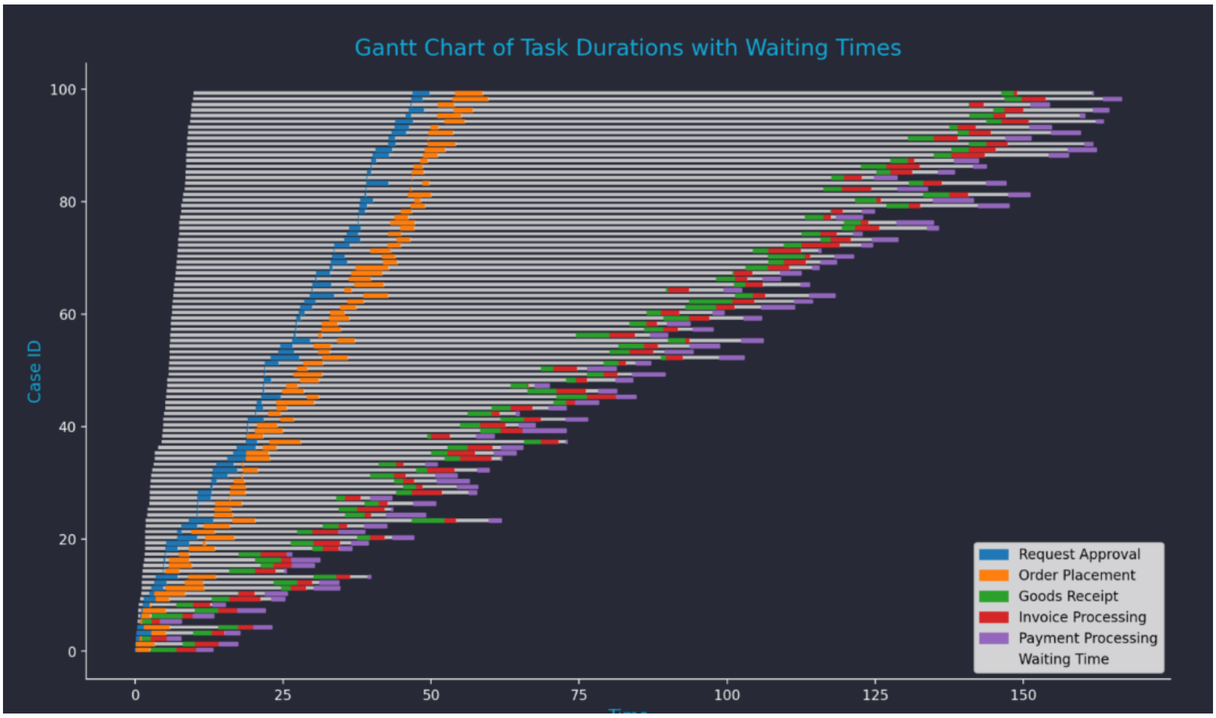
Appendix

A: Screenshots of the PoC System

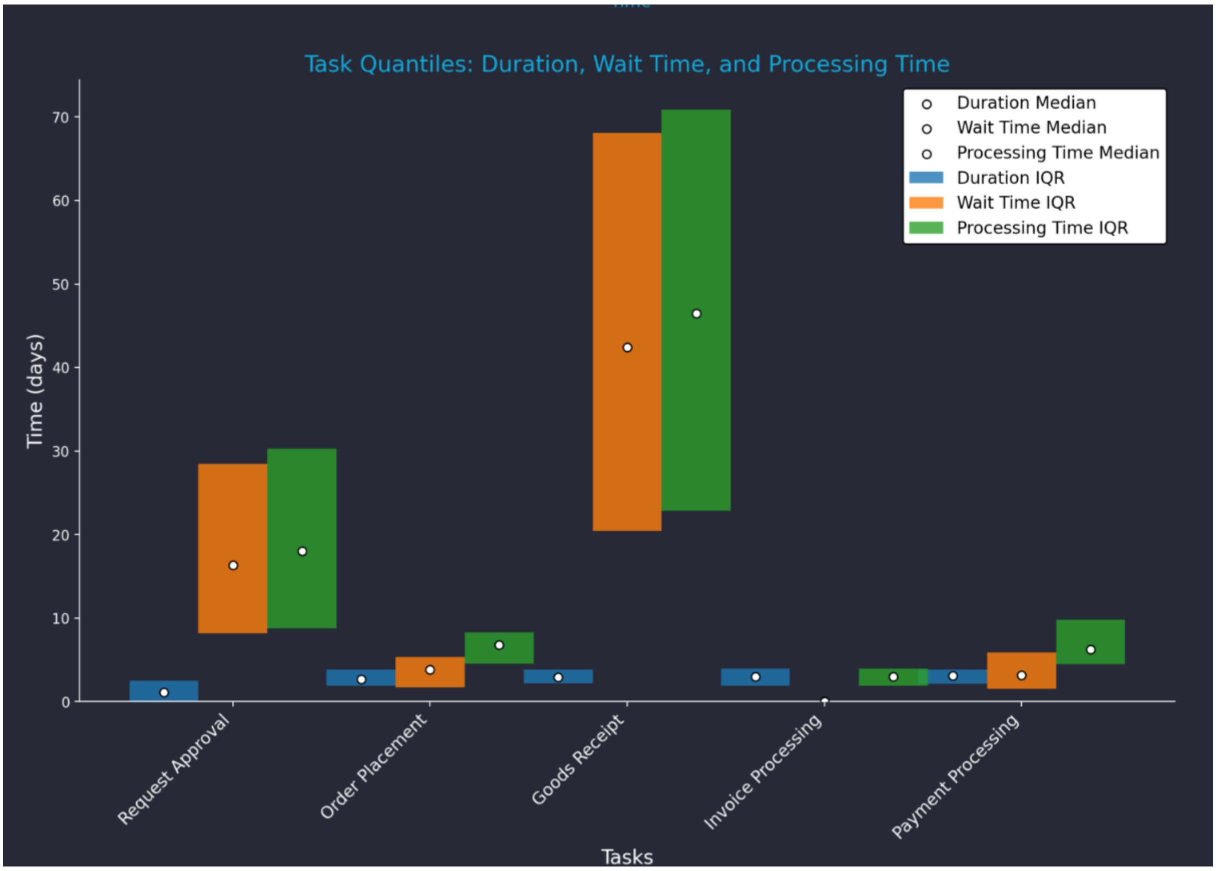
The following is a cut-up series of screenshot displaying the same interaction and response, after the user has entered the query *"What would happen if we reduce our staff in the warehouse to two people?"*.



Appendix figure 1: Start of conversation (1/6), the user entered a query, and the system starts executing it.



Appendix Figure 2: A Gantt chart is display (2/6) illustrating the wait times and processing times (x-axis) for each task (colored bars) for each case that was simulated (one bar on the y-axis).



Appendix Figure 3: A chart is displayed (3/6) illustrating the quantiles of wait and processing times are displayed. Here the user can e.g. see that the reduction in warehouse staff has made the task they are responsible for, "Goods Receipt", much longer then all other tasks, even the other bottleneck "Request approval".

The simulation results for reducing the warehouse staff to two people in the Purchase-to-Pay process are as follows:

Simulation Results Summary:

- **Total Cases Run**: 100
- **Total Tasks Completed**: 500
- **Number of Unique Tasks**: 5
- **Average Case Duration**: 82.58 days

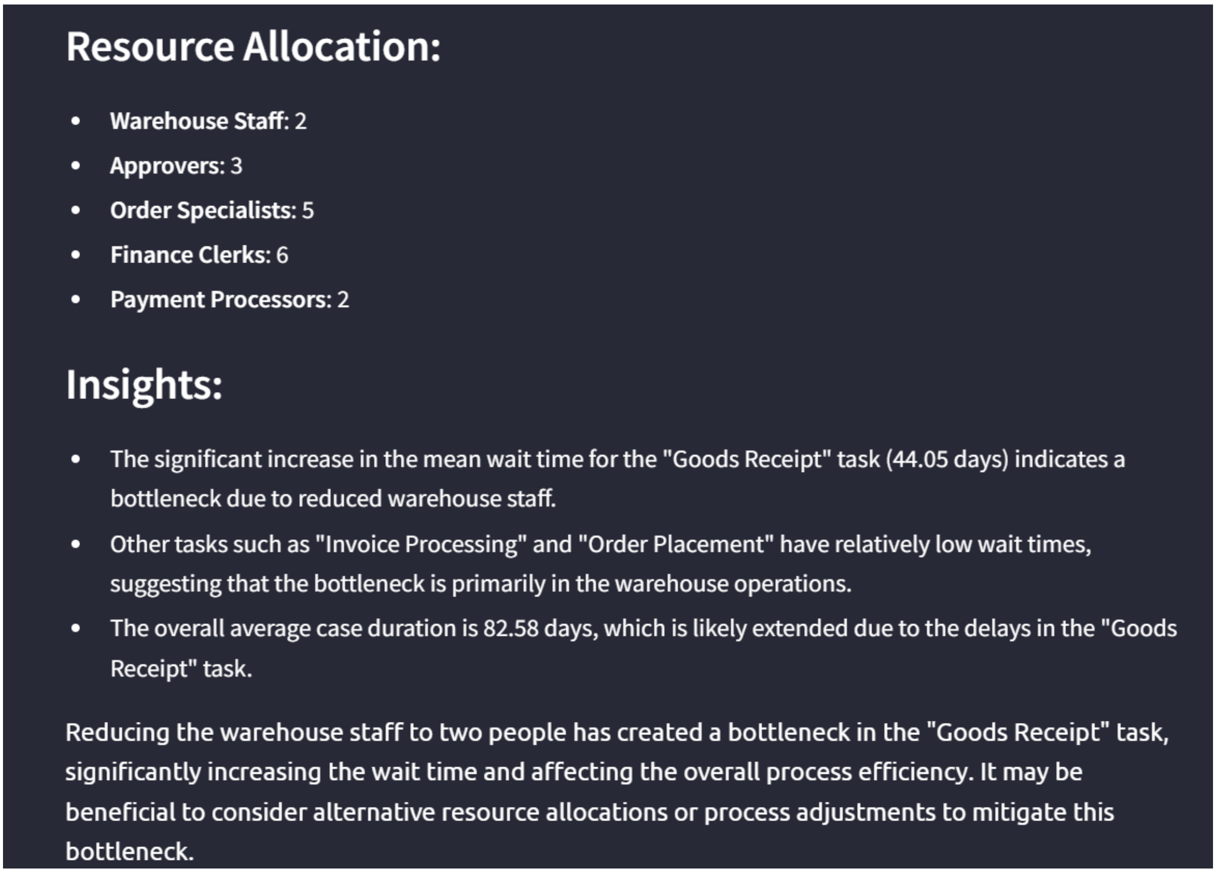
Task-Specific Metrics:

1. **Goods Receipt**:
   - Mean Duration: 2.96 days
   - Mean Wait Time: 44.05 days
   - Total Occurrences: 100
2. **Invoice Processing**:
   - Mean Duration: 2.95 days
   - Mean Wait Time: 0.00 days
   - Total Occurrences: 100

Appendix Figure 4: Start of the simulation summaries provided (4/6)

3. **Order Placement**:
   - Mean Duration: 2.80 days
   - Mean Wait Time: 3.53 days
   - Total Occurrences: 100
4. **Payment Processing**:
   - Mean Duration: 3.11 days
   - Mean Wait Time: 4.08 days
   - Total Occurrences: 100
5. **Request Approval**:
   - Mean Duration: 1.46 days
   - Mean Wait Time: 17.64 days
   - Total Occurrences: 100

Appendix Figure 5: End of the simulation summaries provided (5/6)



Resource Allocation:

- **Warehouse Staff:** 2
- **Approvers:** 3
- **Order Specialists:** 5
- **Finance Clerks:** 6
- **Payment Processors:** 2

Insights:

- The significant increase in the mean wait time for the "Goods Receipt" task (44.05 days) indicates a bottleneck due to reduced warehouse staff.
- Other tasks such as "Invoice Processing" and "Order Placement" have relatively low wait times, suggesting that the bottleneck is primarily in the warehouse operations.
- The overall average case duration is 82.58 days, which is likely extended due to the delays in the "Goods Receipt" task.

Reducing the warehouse staff to two people has created a bottleneck in the "Goods Receipt" task, significantly increasing the wait time and affecting the overall process efficiency. It may be beneficial to consider alternative resource allocations or process adjustments to mitigate this bottleneck.

Appendix Figure 6: Resource allocation information for verification is provided as well with final insights (6/6). The system detects that a new bottleneck is created and passes that information onto the user.

B: Demonstration Data

This section outlines the data used for demonstration purposes in the system. The data aims to simulate a Purchase-to-Pay process, as described in B.1.

B.1: Simplified Purchase-to-Pay Process

To demonstrate the system capabilities, we use the following simplification of the purchase to pay process:

Appendix Figure 7: Simplified Purchase-to-Pay process used for demonstration

B.2: Extract from the event log used

The following is an abstract of the 1000 rows event log generated to demonstrate the capabilities of the system:

Appendix Table 1: Excerpt from the P2P event log used for demonstration

Case ID	Activity	Timestamp
105	Order Placement	2024-11-20
148	Goods Receipt	2024-11-19
149	Request Approval	2024-11-01
133	Request Approval	2024-11-25
83	Order Placement	Not yet concluded
136	Invoice Processing	2024-12-04
Continuing for 996 further rows.		

B.3: Resource excel used

The following is the same data that is used in the resources excel file used as demonstration data in the system:

Resource name	Task supported	Capacity	Description
Approvers	Request Approval	3	Team of approvers handling requests.
Order Specialists	Order Placement	5	Specialists handling the placement of orders.
Warehouse Staff	Goods Receipt	4	Staff managing goods receipt in the warehouse.
Finance Clerks	Invoice Processing	6	Clerks processing invoices in finance.
Payment Processors	Payment Processing	2	Staff or system processing payments.

C: Evaluation results:

C.1: Evaluation system objective 1

C.1.1: Methodology to evaluate chain of actions

Example for evaluation of the chain of actions:

The following example illustrates the evaluation procedure: When running the query: *"Simulate the process with only 3 people in the warehouse"*, the callback handler returns the following simplified chain of events: *1. Chain start, 2. LLM start, 3. LLM end, 4. Agent action, 5. Tool start, 6. Tool End, 7. LLM start, 8. LLM end, 9. Agent finish, 10. Chain End*. The evaluation then checks for the name of the tool started in step 6 and finds that it is the "process_simulator" tool, as expected when formulating the query.

C.1.2: Descriptions of dummy tools

Tool name	Description
process_simulator	The same description and input parameters as the ProcessSimulation tool, with no results to evaluate more quickly.
bottleneck_analysis	Dummy tool described as being capable of identifying and analyzing bottlenecks in a process.
business_process_modeling	Dummy tool described as being capable of modeling a business process and its sub processes.
generate_optimization_scenarios	Dummy tool described as being capable of creating optimization scenarios based for processes.
process_mining_analysis	Dummy tool described as being capable of utilizing process mining for various analytical goals.

C.1.3: Evaluation results of running the queries

The following table shows the results of running the queries with the dummy tools.

Appendix Table 2: Queries and tool invocation results against expected tool invocation

ID	Query	Invoked Tool	Expected Tool
1	Simulate the process with only 3 people in the warehouse	process_simulator	process_simulator

2	What happens if we reduce workers by 2 in the warehouse?	process_simulator	process_simulator
3	Run a simulation for 5 workers across all tasks	process_simulator	process_simulator
4	Simulate with only 1 person approving requests	process_simulator	process_simulator
5	What happens if we double the warehouse staff?	process_simulator	process_simulator
6	Simulate the workflow with 2 fewer finance clerks	process_simulator	process_simulator
7	What is the impact of increasing payment processors by 3?	process_simulator	process_simulator
8	Simulate reducing invoice processing time by 20%	process_simulator	process_simulator
9	What happens if payment processors are reduced to 1?	process_simulator	process_simulator
10	Run a simulation where only one person works on all tasks	process_simulator	process_simulator
11	Identify the bottlenecks in the approval workflow	bottleneck_analysis	bottleneck_analysis
12	Which steps in the process have the highest wait times?	bottleneck_analysis	bottleneck_analysis
13	Analyze bottlenecks in the current warehouse process	bottleneck_analysis	bottleneck_analysis
14	What are the major process bottlenecks in our workflow?	bottleneck_analysis	bottleneck_analysis
15	Determine bottlenecks in the invoice approval process	bottleneck_analysis	bottleneck_analysis
16	What are the causes of delays in the approval workflow?	bottleneck_analysis	bottleneck_analysis
17	Which task in invoice processing has the most wait time?	bottleneck_analysis	bottleneck_analysis
18	Analyze bottlenecks caused by limited warehouse staff	bottleneck_analysis	bottleneck_analysis
19	Which approval tasks create the largest delays?	bottleneck_analysis	bottleneck_analysis
20	Determine the most significant bottleneck in our workflow	bottleneck_analysis	bottleneck_analysis
21	Create a process model for the current P2P workflow	business_process_modeling	business_process_modeling
22	Generate a process model for invoice processing	business_process_modeling	business_process_modeling
23	Model the process flow for the approval workflow	business_process_modeling	business_process_modeling
24	What is the business process model for order placement?	business_process_modeling	business_process_modeling
25	Create a model of the warehouse operations workflow	business_process_modeling	business_process_modeling
26	Generate a detailed process model for payment processing	business_process_modeling	business_process_modeling
27	Can you model the current invoice approval process?	business_process_modeling	business_process_modeling
28	Create a visual process model for our procurement workflow	business_process_modeling	business_process_modeling
29	Generate a workflow model for all tasks in P2P	business_process_modeling	business_process_modeling

30	What does the current process map look like for claims workflow?	business_process_modeling	business_process_modeling
31	Generate optimization scenarios for warehouse staff	generate_optimization_scenarios	generate_optimization_scenarios
32	What are the best optimization scenarios for approvals?	generate_optimization_scenarios	generate_optimization_scenarios
33	Provide optimization options for payment processors	generate_optimization_scenarios	generate_optimization_scenarios
34	Can you suggest optimization scenarios for invoice processing?	generate_optimization_scenarios	generate_optimization_scenarios
35	Generate scenarios for optimizing the P2P workflow	generate_optimization_scenarios	generate_optimization_scenarios
36	Suggest optimization scenarios for reducing wait times	generate_optimization_scenarios	generate_optimization_scenarios
37	Provide scenarios for balancing resource utilization	generate_optimization_scenarios	generate_optimization_scenarios
38	Generate optimization ideas for improving throughput times	generate_optimization_scenarios	generate_optimization_scenarios
39	What are the best ways to optimize warehouse operations?	generate_optimization_scenarios	None, standard GPT-4o answer.
40	Suggest optimization scenarios for minimizing bottlenecks	generate_optimization_scenarios	generate_optimization_scenarios
41	Analyze process mining logs for bottlenecks in approvals	process_mining_analysis	process_mining_analysis
42	What do the process mining logs say about warehouse delays?	process_mining_analysis	process_mining_analysis
43	Analyze compliance issues in our invoice approval process	process_mining_analysis	process_mining_analysis
44	Use process mining to find inefficiencies in order placement	process_mining_analysis	process_mining_analysis
45	What does process mining reveal about payment processing?	process_mining_analysis	None, standard GPT-4o answer.
46	Analyze event logs for compliance issues in P2P	process_mining_analysis	process_mining_analysis
47	What inefficiencies can we find using process mining?	process_mining_analysis	None, standard GPT-4o answer.
48	Use process mining to detect delays in our workflows	process_mining_analysis	process_mining_analysis
49	What do process mining logs say about task handoffs?	process_mining_analysis	None, standard GPT-4o answer.
50	Analyze event logs to find process deviations	process_mining_analysis	process_mining_analysis

C.2: Evaluation system objective 2

The following table shows the results of running the resource parameters tests. *Pass* means the expected resource was changed, *fail* means the resource change was successful but not the expected change, and *error* means the parsing tool threw an expected or unexpected error.

Appendix Table 3: Result of running resource parsing tests

ID	Instruction	Result	Behavior	Reasoning
1	Reduce warehouse staff by 2	Pass	Expected	
2	Increase finance clerks by 3	Pass	Expected	
3	Reduce payment processors by 1	Pass	Expected	
4	Increase order specialists by 2	Pass	Expected	
5	Add 1 approver	Pass	Expected	
6	Reduce warehouse staff by 1, increase finance clerks by 2	Pass	Expected	
7	Reduce order specialists by 2, add 1 payment processor	Pass	Expected	
8	Increase all resources by 1	Pass	Expected	
9	Double the warehouse staff	Error	Unexpected	Unexpected error parsing
10	Halve the finance clerks	Error	Unexpected	Unexpected error parsing
11	Remove all payment processors	Error	Expected	Expected error
12	Assign fractional capacity: set warehouse staff to 1.5	Pass	Expected	
13	Set negative capacity: reduce order specialists to -1	Error	Expected	Expected error, capacity must be greater than 0
14	Distribute workers unevenly: set approvers to 10, warehouse staff to 0	Error	Expected	Expected error, capacity must be greater than 0
15	Assign all resources to one team: set finance clerks to 20	Pass	Expected	
16	Reduce approvers by 1	Pass	Expected	
17	Add 3 order specialists	Pass	Expected	
18	Set payment processors to 5	Pass	Expected	
19	Increase warehouse staff by 3	Pass	Expected	
20	Reduce approvers by 2 and finance clerks by 1	Pass	Expected	
21	Double the order specialists and reduce warehouse staff by 2	Fail	Unexpected	Order specialists unchanged
22	Increase finance clerks to 12	Pass	Expected	
23	Reduce all resources by 1	Error	Unexpected	Unexpected error parsing
24	Halve all resources	Fail	Unexpected	Set all resources to 0.5 instead of halving them
25	Set all resources to 10	Pass	Expected	

26	Remove all approvers	Error	Expected	Expected error, capacity must be greater then 0
27	Reduce warehouse staff to 1 and finance clerks to 3	Pass	Expected	
28	Increase payment processors by 2	Pass	Expected	
29	Set order specialists to 0 and increase warehouse staff by 3	Error	Expected	Expected error, capacity must be greater then 0
30	Reduce all resources to 1	Pass	Expected	

C.3: Evaluation system objective 3

C.3.1: The study overview given to the users for the study

System context:

- *The shown system is a prototype of an AI-driven Process Management Assistant. Compared to similar AI-Assistants, such as ChatGPT, this specific prototype can call a process simulation tool.*
- *The tool is capable of testing out different scenarios for the process, currently this refers to the adjustment of resources i.e. employees available for each step in the process.*

Process and resources context:

*This simplified Purchase-to-Pay (P2P) process and resource allocations are available inside the AI-assistant for direct reference in queries. Over **10 days**, the process usually receives **100 purchase requests**.*

Appendix Figure 8: The illustration of process steps and resources for the P2P process as given in the study overview

Your role:

You are a BPM practitioner e.g. a Process Analyst Owner or similar role. Your objective is to manage and optimize your process (P2P).

Your tasks:

1. *Use the assistant to simulate the process with* ***no resource adjustments*** *to understand the current average throughput times of each process step.*
2. *Your manager wants to* ***reduce the staff in the warehouse to two people****. Analyze the impact of such a reduction on the process using the assistant.*
3. *Your company is significantly scaling, and you might be able to negotiate* ***2 more people*** *in either (A) the warehouse, (B) as request approvers or (C) as financial clerks. Use the assistant to determine which scenario you should push for.*

Instructing the agent to simulate the process as well as always providing the resource adjustment can help with reliability. Scenarios can also be written in one query.

C.3.2: Participants of user study

The following table shows the participants of the user study that determined the System Usability Scale (SUS) score:

Appendix Table 4: Table describing the participants of the user study

ID	Job level
P1	Junior grade level consultant
P2	Junior grade level consultant
P3	Senior grade level consultant
P4	Senior grade level consultant
P5	Management grade level consultant

C.3.3: Results of System Usability Scale (SUS) study

The following table shows the results of the participant for each of the SUS questionnaire statements, with the 1-5 representing the used Likert scale (1: Strongly disagree, 2: partly disagree, 3: neither agree nor disagree, 4: partly agree, 5: strongly agree):

Appendix Table 5: Results for each statement from each participant

Statement | Participant	P1	P2	P3	P4	P5
I think that I would like to use this AI assistant frequently.	5	4	5	5	5
I found the AI assistant unnecessarily complex.	1	1	1	1	2
I thought the AI assistant was easy to use.	4	4	5	5	5
I think that I would need the support of a technical person to be able to use this AI assistant.	1	1	2	1	1
I found the various functions in this AI assistant well integrated.	5	4	5	5	5
I thought there was too much inconsistency in this AI assistant.	1	1	1	2	1
I imagine that most people would learn to use this AI assistant very quickly.	5	5	5	4	5
I found the AI assistant very cumbersome to use.	1	2	1	1	1
I felt very confident using the AI assistant.	4	5	5	5	4
I need to learn a lot of things before I could get going with the AI assistant.	2	1	2	1	2

C.3.4: SUS Score results

The following table shows the resulting SUS scores for each participant after applying the score calculation (For odd questions 1 from value given, for even questions subtract 4 from value given, add it all up and multiply by 2.5):

Appendix Table 6: Average SUS score for each participant

Participant	P1	P2	P3	P4	P5
Score	92.5	90	95	95	92.5

Calculating the total mean results in 93.

C.4: Evaluation of overall system performance

C.4.1: Dummy tool descriptions for overall evaluation

The following table summarizes the dummy tool descriptions used for the overall system evaluation. The first five were used for all scenarios, with the last 5 only used for the *10 dummy tools* test category.

Appendix Table 7: Dummy tools for overall performance evaluations

Tool name	Description
process_documentation	Dummy tool described as being capable of generating detailed process documentation based on an input process.
kpi_reporting	Dummy tool described as being capable of generating KPI reports to track performance metrics for a given process.
compliance_check	Dummy tool described as being capable of performing a compliance check with known regulations of an input process.
workflow_visualization	Dummy tool described as being capable of creating workflow visualizations based on natural language descriptions.
task_priorization	Dummy tool described as being capable of giving task prioritization strategies based on process diagrams and resource descriptions.
process_variant_analysis	Dummy tool described as being capable of analyzing and comparing multiple process variants from an event log.
benchmark_process_performance	Dummy tool described as being capable of benchmarking a process performance against industry standards.
evaluate_automation_feasibility	Dummy tool described as being capable of evaluating the automation feasibility of all tasks in a given process.
risk_assesment	Dummy tool described as being capable of assessing potential risks of a given process and providing mitigation advice.
resource_allocation_optimization	Dummy tool described as being capable of describing optimal resource allocations strategies to reduce costs in a process.

C.4: Summary of query results

The following table shows if a query got an overall correct response from the system for all six testing scenarios (A: No additional tools, B: No additional tools with prompt, C: Five additional tools, D: Five additional tools with prompt, E: Ten additional tools, D: Ten additional tools no prompt):

Appendix Table 8: Results of the queries run for the overall system evaluation

Category	Query	A	B	C	D	E	F
Basic	Simulate the P2P process with 2 people in the warehouse.	✓	✓	✓	✓	✓	✓
Basic	What would happen to our P2P process if we add one more person to the requesters?	✓	✓	✓	✓	✓	✓
Basic	Simulate the P2P process with no changes.	✓	✓	✓	✓	✓	✓
Basic	What happens if we reduce warehouse staff by one person?	✓	✓	✓	✓	✓	✓
Basic	Simulate the impact of adding two workers in all departments	✓	✓	✓	✓	✓	✓
Basic	Simulate the P2P process with one fewer finance clerk.	✓	✓	✓	✓	✓	✓
Basic	What would happen if we double the warehouse staff?	✓					
Basic	Simulate the process with only one request approver available.	✓	✓	✓	✓	✓	✓
Basic	How does removing one approval staff member impact process time?	✓	✓		✓		✓
Basic	What happens if warehouse staff is reduced to three?	✓	✓	✓	✓	✓	✓
Basic	Simulate the process with one fewer staff in the finance team.	✓	✓	✓	✓	✓	✓
Basic	Simulate the P2P process with 5 people in the warehouse.	✓	✓	✓	✓	✓	✓
Basic	What happens if we add 20 people as requesters?	✓			✓	✓	✓
Basic	Simulate the process with a only half an staff person as order specialist.	✓	✓	✓	✓	✓	✓
Basic	How does doubling all staff affect process efficiency?						
Basic	Simulate the P2P process with 20 warehouse staff.	✓	✓	✓	✓	✓	✓
Basic	What happens if the approval team is now 12 people?	✓	✓		✓	✓	✓
Basic	Simulate reducing finance clerks to 1	✓	✓	✓	✓	✓	✓
Basic	What will happen if we add two more payment processers in the P2P process	✓	✓	✓	✓	✓	✓
Basic	What happens if we have 3 fewer people in the warehouse	✓	✓	✓	✓	✓	✓

Subtotal		**19**	**17**	**15**	**18**	**17**	**18**
Advanced	Simulate the P2P process with warehouse staff reduced by 2 and finance clerks increased by 1.	✓	✓	✓	✓	✓	✓
Advanced	Compare Scenario A: Warehouse staff reduced by 1 and Scenario B: Approvers increased by 2. Analyze which scenario reduces bottlenecks.		✓	✓	✓		✓
Advanced	Evaluate the impact of reducing order specialists by 2 on average case durations and wait times.	✓	✓	✓	✓	✓	✓
Advanced	Simulate the process with one additional payment processor and one fewer finance clerk. Analyze resource utilization changes.	✓	✓	✓	✓	✓	✓
Advanced	Run a simulation with warehouse staff now at 10. What is the effect on bottlenecks in the process?	✓	✓	✓	✓	✓	✓
Advanced	Analyze the impact of adding one employee to all the tasks. Does that make sense or should we allocated differently?	✓		✓	✓		✓
Advanced	Compare resource utilization when finance clerks are increased by 2 versus when payment processors are reduced by 1.	✓	✓	✓	✓	✓	✓
Advanced	Simulate the process with no adjustments as a benchmark. Now simulate the impact of adding 2 warehouse staff and reducing order specialists by 1. Evaluate bottleneck shifts.	✓	✓	✓	✓	✓	✓
Advanced	Evaluate how reducing warehouse staff by 3 affects the wait times for the goods receipt step.	✓	✓	✓	✓	✓	✓
Advanced	Simulate the P2P process to if we reduce all teams by one person.	✓	✓	✓	✓	✓	✓
Subtotal		**9**	**9**	**10**	**10**	**8**	**10**
Analytical	What is the average case duration when warehouse staff are reduced by 2?	✓	✓	✓	✓	✓	✓
Analytical	Calculate the total wait time for the 'Goods Receipt' task when order specialists are set to 4.	✓	✓	✓	✓	✓	✓
Analytical	What is the percentage change in wait time for 'Invoice Processing' when payment processors are increased by 1?	✓	✓	✓	✓		✓
Analytical	What is the average wait time for tasks when warehouse staff are set to 5?	✓	✓	✓	✓	✓	✓
Analytical	How does the average task duration change when finance clerks are reduced by 2?	✓	✓	✓	✓	✓	✓
Analytical	What is the overall case throughput when approvers are reduced by 1?	✓	✓	✓	✓	✓	✓
Analytical	Calculate the total bottleneck time when the 'Order Placement' task has 1 fewer available resources.				✓	✓	
Analytical	Determine how the average case duration changes when payment processors are increased by 3.	✓	✓	✓	✓	✓	✓
Analytical	What is the maximum wait time for any task when approvers are increased by 2?	✓	✓	✓	✓	✓	✓
Analytical	Identify the task with the smallest duration change when warehouse staff are set to 2.	✓	✓	✓	✓	✓	✓
Analytical	What happens to the average case duration when warehouse staff are reduced by 2?	✓	✓	✓	✓	✓	✓
Analytical	Identify the task with the highest wait time when finance clerks are reduced by 1.	✓	✓	✓	✓	✓	✓
Analytical	What is the most delayed task when warehouse staff are set to 3?	✓	✓	✓	✓	✓	✓
Analytical	Evaluate the bottlenecks caused by reducing payment processors by 1.	✓	✓	✓	✓	✓	✓

Analytical	Which task has the longest average duration when the order specialists are reduced by 2?	✓	✓	✓	✓	✓	✓
Analytical	Identify the most delayed step when approvers are increased by 2.	✓	✓		✓	✓	✓
Analytical	Explore how average case duration changes with warehouse staff adjustments reaching from 1 to 5 people in the warehouse.		✓		✓		✓
Analytical	What happens to task wait times when approvers are decreased to 1?	✓	✓	✓	✓	✓	✓
Analytical	Analyze the effect of adding 2 payment processors on average case duration.	✓	✓	✓	✓	✓	✓
Analytical	Which task is least impacted when finance clerks are increased by 2?	✓	✓		✓	✓	✓
Subtotal		**18**	**19**	**16**	**20**	**18**	**19**
Toal correct %		**94**	**90**	**82**	**96**	**86**	**94**