HOCHSCHULE FÜR ANGEWANDTE WISSENSCHAFTEN

NEU ULM

MASTER DEGREE THESIS

Enhancing Process Management with AI: A Focus on Process Discovery and Analysis

to be awarded the degree of Master of Science in Digital Transformation and Global Entrepreneurship at the Hochschule Neu-Ulm

Supervised by
Professor Dr. Jörg-Oliver Vogt

Candidate Name
Mohammad Anees Raza
M. No. 329853
Neu-Ulm, 19/03/2025

Disclaimer: The content of this Thesis is Original. For parts, AI has been used only to draft sentences.

Table of Contents

List of Figures

List of Tables

Abstract

With the rise of artificial intelligence (AI), organizations are increasingly exploring how AI can enhance process management. One key area is process discovery and analysis, where AI models can help identify, map, and optimize business processes. This thesis examines the potential of AI, specifically ChatGPT, Gemini etc., to automate and improve process discovery and analysis.

The research focuses on evaluating how well these AI models can generate process models from textual descriptions and analyze process efficiency. A structured evaluation framework is developed to compare the performance of AI-generated solutions against traditional methods. Key metrics include accuracy, clarity, and efficiency in generating process maps and identifying bottlenecks. By using statistical analysis and a scoring system, this study assesses whether AI-driven process management can match or surpass human expertise. The findings highlight the strengths and limitations of AI models in understanding complex workflows, detecting inefficiencies, and suggesting improvements. While AI shows promise, challenges remain in handling domain-specific knowledge, process complexity, and consistency.

This research contributes to the ongoing discussion on integrating AI into business process management (BPM). It provides insights into how these models can be refined for better performance in real-world applications.

Chapter 1: Introduction

Since the mid-20th century, software engineering has undergone rapid transformation, driven by advances in computing technology, the increasing complexity of IT systems, and the growing demand for efficiency in business operations. As organizations expand and digital infrastructures become more complex, the need for structured process management has become more critical than ever (Mohammed, 2020). Process modeling plays a fundamental role in this context by providing a structured representation of business activities, facilitating workflow optimization, and ensuring that operations align with organizational goals. One widely used modeling standard is Business Process Model and Notation (BPMN), a graphical representation that helps organizations document, analyze, and improve their workflows. BPMN diagrams act as a bridge between business analysts and technical teams, enabling clear communication of processes and their execution within IT systems.

Traditionally, creating BPMN diagrams requires significant time, expertise, and iterative refinement, often involving multiple stakeholders such as business analysts, system architects, and process managers. The manual creation process is not only time-consuming but also prone to human errors, inconsistencies, and inefficiencies (Aguayo Publicidad, 2021). Even with careful validation, process models may contain inaccuracies that lead to misunderstandings, operational inefficiencies, or implementation issues. These challenges highlight the need for automation in process modeling, raising an important question:

Can artificial intelligence (AI) improve process discovery and analysis by making the creation of BPMN diagrams faster and more accurate?

If AI can successfully automate this process, it could significantly reduce time and effort while enhancing accuracy, consistency, and adaptability in business process management.

This study explores the potential of AI models— Gemini, and ChatGPT—to generate BPMN diagrams from textual descriptions of business processes. These AI models, powered by large language models (LLMs), have demonstrated remarkable capabilities in natural language understanding and content generation. However, their ability to comprehend, structure, and translate textual descriptions into formalized process models remains an open question (Neuberger et al., 2024). To determine their effectiveness, this research conducts an empirical evaluation comparing AI-generated BPMN diagrams with those created by human experts (Kourani et al., 2024). The study seeks to assess whether AI can match or even surpass human performance in generating syntactically correct and logically coherent BPMN diagrams. However, this is a challenging task, as AI models may struggle with domain-specific knowledge, graphical representation, and contextual accuracy—factors that are critical in business process modeling.

To address this issue, this research follows a structured approach to systematically evaluate AI-generated process models. A scoring system is developed to assess accuracy, syntax compliance, and logical consistency of the generated BPMN diagrams. Additionally, statistical methods, including T-tests and ANOVA, are applied to determine whether there is a statistically significant difference between human-generated and AI-generated process models. This study also examines AI's capability to analyze, interpret, and structure process information by working with XML-based process representations.

The thesis is structured to provide a comprehensive understanding of AI-driven process modeling. It begins with a background on process modeling methodologies and large language models (LLMs) to establish foundational concepts. This is followed by an overview of recent AI advancements and their applications in various industries, particularly field that benefits from business process management. The research methodology is then outlined, detailing the procedures used to generate and evaluate BPMN diagrams. The results of the empirical study are presented and analyzed to

determine the comparative performance of AI and human-created models. Finally, it provides a summary of findings, highlighting key insights into AI's strengths and limitations in process modeling, before concluding with recommendations for future research and potential improvements in AI-based process discovery.

The objective of this research is to contribute to the ongoing exploration of AI's role in business process discovery and analysis. By examining how well AI can automate BPMN diagram generation and evaluating its effectiveness compared to human expertise, this study aims to provide valuable insights into the current capabilities of AI in this domain. The findings will not only assess the feasibility of AI-driven process modeling but also highlight areas where AI needs further refinement to achieve greater accuracy, contextual understanding, and usability in real-world business applications. In doing so, this research lays the foundation for future advancements in AI-assisted business process management, potentially paving the way for more efficient and intelligent process automation solutions.

Chapter 2: Background

2.1. Definition of Business Process Modeling

A business process is a structured sequence of activities that, when executed in a defined order, transform specific inputs into desired outputs. It serves as a foundational element for understanding, analyzing, and optimizing business operations (Dumas et al., 2018). Before modeling a process, it is essential to identify the activities involved, their dependencies, and the sequence in which they occur. However, in complex workflows, distinguishing between individual activities can be challenging, particularly when tasks are interdependent or occur concurrently. In business environments, process modeling is a critical practice that involves creating structured visual representations of workflows to document, analyze, and improve operations (W. Van Der Aalst, 2016). One of the most widely adopted standards for this purpose is Business Process Model and Notation (BPMN), which provides a formalized method for representing business processes in a clear and standardized manner. BPMN enables organizations to create diagrams that illustrate the flow of activities, decisions, and interactions within a business process, ensuring alignment between technical and non-technical stakeholders.

The primary goal of BPMN is to offer a comprehensive yet intuitive visualization of business processes, facilitating communication among business analysts, process designers, and IT professionals. By providing a standardized notation system, BPMN helps stakeholders understand current workflows (as-is models) and design optimized future workflows (to-be models) (Von Rosing et al., 2014). This structured approach enables organizations to identify inefficiencies, streamline operations, and enhance overall business process management. BPMN diagrams consist of various graphical elements that represent different components of a process. These include tasks, which define specific actions within a workflow; gateways, which represent decision points and control the process flow; events, which signal the start, end, or intermediate

occurrences within a process; and connectors, which define the relationships between different elements. BPMN uses a set of standardized symbols and notations to ensure clarity and consistency across different industries and organizations.

As a business process modeling tool, BPMN is widely used in fields such as project management, enterprise resource planning (ERP), compliance auditing, and workflow automation. Its ability to bridge the gap between business and IT perspectives makes it a valuable framework for designing, analyzing, and optimizing business processes in a systematic and efficient manner (Weske, 2007).

2.1.1. Definition, Characteristics and Notions of Business Process Modeling

Processes are a fundamental component of any business, whether it provides a service or a product to its clients. The design and execution of these processes have a direct impact on service quality and operational efficiency. In a competitive market, an organization with well-structured and efficiently executed processes can gain a significant advantage over competitors offering similar services (Davenport, 1993). Business Process Modeling (BPM) is a formalized approach to representing business processes in a structured and visual manner. It enables organizations to analyze activities, interactions, and information flows within their operations, facilitating a better understanding of workflows (Hammer & Champy, 1993). By creating graphical representations of processes, BPM helps organizations identify areas for improvement, enhance communication among stakeholders, and foster collaboration. One of the key advantages of BPM is its ability to represent complex processes using standardized diagrams and symbols, making them easier to interpret and analyze (W. M. P. Van Der Aalst, 2013).

A core characteristic of BPM is its process-oriented nature, which focuses on capturing the sequence of activities and their relationships within a workflow. It helps stakeholders understand the flow of information and tasks without delving into implementation-specific details (Dumas et al., 2018). As processes become more

intricate, BPM models undergo iterative refinement based on stakeholder insights and feedback. This ensures that the models remain accurate, up-to-date, and effectively communicate the process structure (Weske, 2007).

The *first step* in business process modeling is identifying key organizational processes, understanding their limitations, and defining the activities involved. This step is crucial for establishing a clear framework for process analysis.

The *second step* involves creating visual representations of the processes, such as flowcharts or BPMN diagrams. These diagrams illustrate the sequence of activities, decision points, and information flows, enabling analysts to detect inefficiencies, bottlenecks, and opportunities for improvement.

The *third step* requires stakeholders to assess and optimize the process by identifying areas for enhancement. The goal is to improve overall efficiency and streamline operations (Dumas et al., 2013).

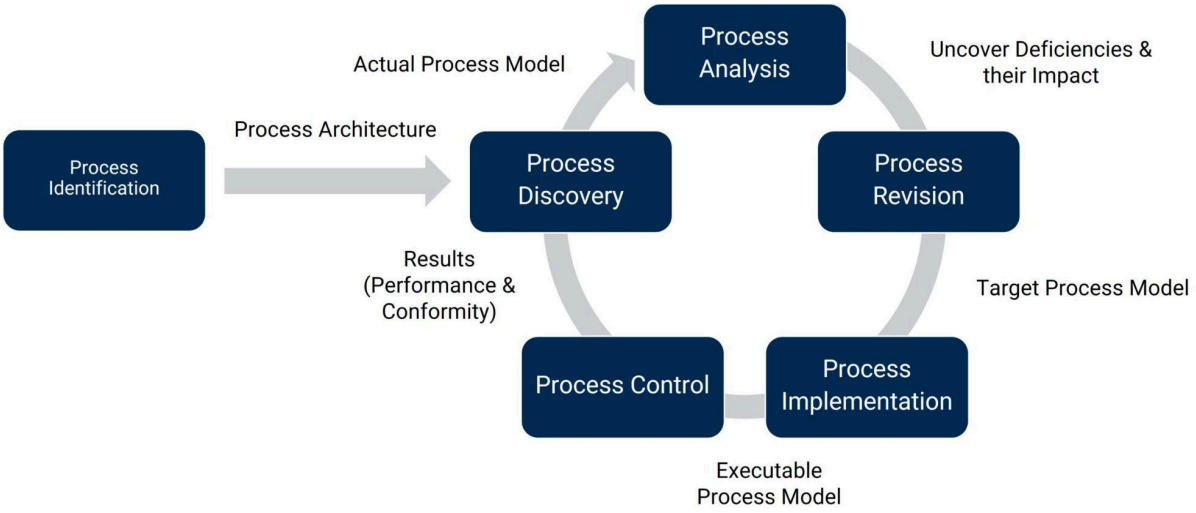

Figure 1 : Notions of business process modelling

Following this analysis, experts apply process simulation techniques to model different scenarios and evaluate process performance. By simulating workflows, organizations can identify inefficiencies and explore potential improvements before implementing changes. The *final step* focuses on process improvement, where experts propose

solutions to address inefficiencies. This may include redesigning workflows, adopting new technologies, or implementing best practices observed in similar cases (D'Ippolito, 2014).

BPMN was developed in response to the lack of a unified standard for representing business processes. Before its introduction, different organizations used varying notations, leading to inconsistencies in process documentation and communication. Recognizing this challenge, industry experts sought to create a standardized notation system to facilitate clear and uniform process representation across businesses (Chinosi & Trombetta, 2011). The first version, BPMN 1.0, was introduced in May 2004, providing a foundational framework for process modeling (Wong & Gibbons, 2008). Over the years, refinements were made to enhance its usability and effectiveness, culminating in the release of BPMN 2.0 in 2011. This version incorporated improvements that made BPMN an international standard, widely recognized as a best practice in business process modeling (Allweyer, 2016).

Business Process Model and Notation (BPMN) is a standardized graphical notation designed to model business processes in a way that is both intuitive for business users and precise for technical implementation. By utilizing simple and easily understandable symbols, BPMN facilitates communication between business analysts, process designers, and IT developers (Silver, 2011). As a business process modeling language, BPMN consists of specific notations and symbols used to represent different elements within a business process (Von Rosing et al., 2014). These include tasks, events, gateways, and connectors, each serving a distinct purpose in defining process structure and flow. The clear and standardized nature of BPMN makes it an essential tool for organizations seeking to optimize their operations, improve process efficiency, and ensure seamless communication between business and technical teams.

2.1.2. Available Formats for Business Process Modeling Diagrams

Business Process Model and Notation (BPMN) consists of several core components that serve as the foundation for visually representing business processes. These components help define the flow of activities, interactions, and decision-making logic within a process, ensuring clarity and standardization across different organizations.

One of the most fundamental components is **tasks**, which represent specific activities or actions carried out by individuals, departments, or systems involved in a business process. These tasks are depicted as rectangular boxes containing a brief description of the action being performed. Tasks can vary in complexity, from simple manual tasks to automated system-driven actions, and may include subcategories such as user tasks, service tasks, script tasks, and manual tasks, depending on their execution method (Bedwell et al., 2022).

Figure 2 : Types of tasks in BPMN

Another essential component is **gateways**, which represent decision points in a process and determine the flow of execution based on predefined conditions. Gateways are symbolized by a diamond shape and come in different types, each serving a distinct logical purpose. Exclusive gateways allow only one outgoing flow to continue, meaning that a specific condition determines which path is taken. Parallel gateways enable all outgoing flows to proceed simultaneously, converging later at a synchronization point. Event-based gateways determine the process flow based on external events, ensuring flexibility in process execution (Bedwell et al., 2022).

Figure 3 : Types of gateways in BPMN

Events are another integral part of BPMN and represent occurrences that influence the flow of a process. These events are depicted as circles and can be categorized into three main types: start events, intermediate events, and end events. A start event, represented by a thin circle, marks the initiation of a process. Intermediate events, shown as double circles, occur during the process and may involve elements such as timers, messages, or errors. End events, displayed as thick circles, signify the conclusion of a process or subprocess. Some events also include icons inside the circle to indicate their specific trigger, such as message receipt, timer expiration, or an error occurrence (Bedwell et al., 2022).

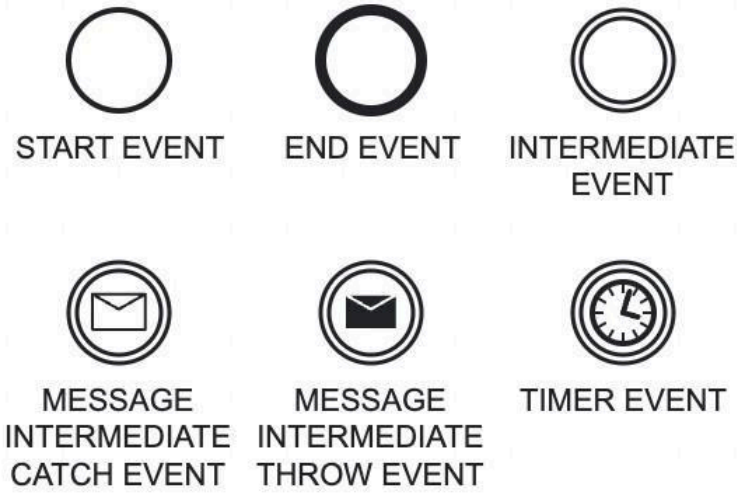

Figure 4: Types of events in BPMN

Flows are used to depict the movement and connections between different activities, ensuring a logical sequence within the process. There are two primary types of flows: sequence flows and message flows. Sequence flows, represented by a continuous arrow, establish the execution order of activities within the same participant's workflow.

Message flows, on the other hand, are shown as dashed arrows and illustrate the exchange of information between different participants or entities involved in the process. While sequence flows define the order of operations within a single organizational unit, message flows demonstrate how different stakeholders or systems communicate (Bedwell et al., 2022).

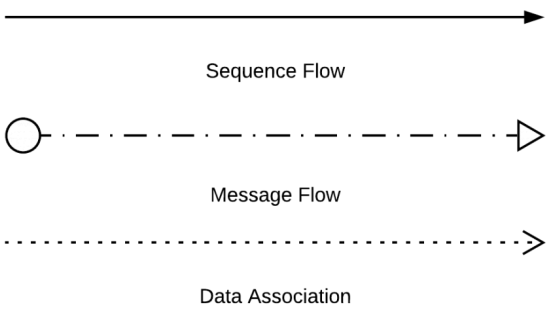

Figure 5 : Types of Sequence flow arrow

Pools and Lanes are used to define the roles, participants, and organizational structures involved in a process. They are depicted as large rectangular containers that categorize different process elements. A pool represents a high-level entity such as a company, department, or organization, while lanes subdivide pools to represent specific roles, teams, or subprocesses within that entity. This distinction ensures that responsibilities and interactions within a process are clearly defined (Bedwell et al., 2022).

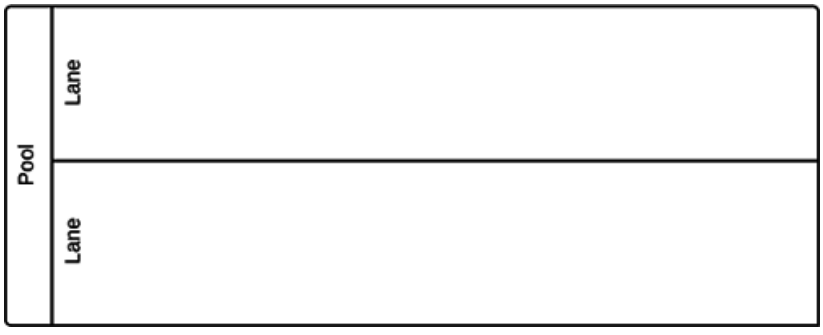

Figure 6 : Pools and Lanes in BPMN

These are just a few of the fundamental components used to construct BPMN diagrams. These diagrams are typically read from left to right, following the direction of the sequence flows. In cases where message flows connect different pools, the reading direction follows the flow of communication, which may be from top to bottom or vice versa. By adhering to these standardized elements, BPMN facilitates clear and structured process documentation, enabling businesses to analyze, optimize, and automate workflows efficiently.

2.2. Generative Pre-trained Transformers and Large Language Models

Generative Pre-trained Transformers (GPT) and Large Language Models (LLMs) have evolved to not only generate text but also to create diagrams, images, and even code. These multimodal models leverage the same foundational principles of large-scale pre-training on extensive datasets, but they are extended to handle not just language but also visual and code generation tasks. Like traditional LLMs, these advanced models are pre-trained using vast amounts of diverse data, including textual content, visual information, and structured data like code, to learn the underlying patterns and relationships in both language and visual domains (Fan et al., 2023).

The defining feature of such LLMs is their vast scale, both in terms of the size of the datasets they are trained on and the number of parameters they possess. These models are trained on massive datasets that combine text, images, diagrams, and even source code, often reaching petabyte-scale data processing. The term "Large" in LLMs refers to the sheer magnitude of data and the billions or trillions of parameters these models contain (Douglas, 2023). Parameters are internal variables that the model adjusts during training, enabling it to generate meaningful outputs, be it text, images, diagrams, or code, with contextual relevance. A larger number of parameters enables the model to better understand intricate relationships in text and images, enhancing its ability to perform a variety of tasks effectively (Douglas, 2023). These LLMs utilize transformer architecture, which is crucial for handling not only textual data but also

multimodal inputs. Transformers use self-attention mechanisms to evaluate relationships between elements in different modalities, whether it's words in a sentence, pixels in an image, or tokens in code (Uszkoreit, 2017). This architecture allows the model to build a rich understanding of complex inputs and outputs, enabling it to generate coherent text, meaningful images, accurate diagrams, and functional code based on a given prompt.

During the training phase, these models learn by predicting the next word in a sentence, generating the next pixel in an image, or suggesting the next line of code. Through iterative adjustments, they continuously improve their outputs by minimizing errors between their predictions and actual outcomes. This process allows the model to enhance its ability to generate human-like text, realistic images, informative diagrams, and syntactically correct code (Vaswani et al., 2017). Once pre-trained, these models can also be fine-tuned on specialized datasets to optimize their performance for specific tasks, such as generating code snippets, designing technical diagrams, or creating domain-specific images (Conneau & Lample, 2019).

One of the key strengths of such multimodal LLMs is their ability to perform tasks that combine text, visuals, and code generation. For instance, they can take a textual description and generate an image that matches the description, create a diagram that represents a process or system, or generate executable code based on specific programming instructions (Ramesh et al., 2021). This capability to generate multiple forms of output from a single input allows these models to be used in a wide array of applications, such as content creation, software development, data visualization, and design (Song & Xiong, 2021).

There are different variations of these models based on their specialization and use case. Some models are tuned to generate text, others to create images, while others are trained to generate or understand code. These multimodal models, however, can integrate multiple types of generation, such as producing a blog post accompanied by a relevant diagram or generating a coding solution along with visual explanations (Song &

Xiong, 2021). The versatility of these models is evident in their ability to be applied across diverse industries, including marketing, education, software engineering, and entertainment, making them invaluable tools for automating complex creative and technical tasks. The rapid advancement of multimodal LLMs has led to their adoption in numerous fields, where their ability to create integrated, cross-domain content enhances productivity and creativity. As these models continue to evolve, they are expected to improve not only in their technical accuracy but also in their understanding of context across different modalities, thus enabling more sophisticated and intelligent generation of text, images, diagrams, and code (Fan et al., 2023).

2.2.1. Introduction to Conversational Artificial Intelligence

Conversational Artificial Intelligence (AI) is a rapidly evolving field with a significant influence on various industries such as e-commerce, education, entertainment, healthcare, productivity, and journalism (Shawar & Atwell, 2007). As these technologies continue to advance, businesses, governments, and academic institutions are increasingly investing in conversational AI to enhance their operations and provide better user experiences. At its core, conversational AI relies on vast amounts of data and machine learning techniques to simulate human-like interactions (Song & Xiong, 2021). These systems can process both speech and text inputs, enabling them to understand user intent and respond in a manner that aligns with human conversation. It's important to distinguish between intents and entities in this context: intents help the AI system understand what the user is asking or seeking, while entities are used to extract specific information that can be used to provide relevant responses (Hassani & Silva, 2023).

Conversational AI is a branch of Artificial Intelligence focused on creating intelligent agents capable of replicating and automating human-like conversations. It leverages advances in natural language processing (NLP), which allows machines to comprehend and respond to human language, whether spoken or written (Hussain et al., 2019). This technology is being widely implemented across industries like aviation, tourism,

healthcare, and customer service, where it helps enhance user experience and streamline customer management through tools like chatbots and virtual assistants.

Although conversational AI is not yet widely adopted in sectors like architecture, engineering, and construction (AEC), there is great potential for it to revolutionize these industries. By improving productivity, assisting various tasks across project lifecycles, and enhancing user experiences, conversational AI can offer significant benefits in the AEC sector (Darko et al., 2020).

The process of conversational AI relies on several key steps in natural language processing (NLP). These steps include:

- Input Generation: Users provide input in the form of voice or text, usually through a website or an app.
- Input Analysis: The system deciphers the meaning of the input and identifies the user's intent through natural language understanding (NLU).
- Dialog Management: Based on the analyzed input, the system formulates a response that mimics human speech using natural language generation (NLG).
- Reinforcement Learning: Over time, the AI system refines its responses based on feedback and analysis, improving its performance with each interaction.

Conversational AI systems can be classified in various ways, based on their methodologies, communication channels, and intended goals. Common classifications include text-based and spoken dialogue systems, voice user interfaces, chatbots, embodied conversational agents, robots, and placed agents (Darko et al., 2020). The diversity of classification methods reflects the rapid development and growing interest in the field. As the field of conversational AI expands, companies are investing heavily in research and development to explore new possibilities and create innovative products. One of the leading companies in AI research is OpenAI, which focuses on the responsible development and deployment of artificial intelligence for the benefit of humanity (McTear, 2020).

2.2.2. Advantages and Disadvantages of Large Language Models

Large Language Models (LLMs) have proven to be groundbreaking in the field of artificial intelligence, offering a wide range of advantages while also presenting some notable disadvantages. These models, particularly those based on Generative Pre-trained Transformers (GPT), have demonstrated exceptional capabilities in understanding and generating human-like text. However, their application also comes with challenges and limitations that must be carefully considered.

One of the most significant ***advantages*** of LLMs is their enhanced productivity. LLMs can process and generate vast amounts of text much faster than humans, enabling the automation of repetitive tasks such as content generation, summarization, translation, and customer support. This high efficiency allows businesses and industries to save valuable time and resources while increasing overall productivity (Arrieta et al., 2019). LLMs also offer efficient parallelization of tasks, meaning they can handle multiple tasks simultaneously. This ability to perform numerous operations at once allows organizations to streamline workflows, reduce turnaround times, and improve overall effectiveness (Safari et al., 2020). For instance, LLMs can process data and generate insights from large datasets, all while executing various tasks at scale with minimal errors. The high success rate of LLMs in generating accurate, relevant, and coherent responses in a wide range of contexts is another key advantage. When properly trained, these models produce outputs that are often indistinguishable from human-written text. Furthermore, LLMs contribute to reduced error and defect rates in tasks that require complex data analysis, as they are less prone to human errors, such as biases or oversights. Another advantage is their ability to mitigate human error in decision-making processes. LLMs can analyze large datasets quickly and accurately, providing data-driven insights that help prevent costly mistakes. They can also operate 24/7, ensuring that businesses or systems relying on them remain operational around the clock. Lastly, LLMs accelerate decision-making by providing quick and reliable answers to complex questions. Their ability to analyze historical data, identify patterns, and make

predictions allows them to support high-speed decision-making in fields like finance, healthcare, and legal analysis (Khanzode & Sarode, 2023).

Despite their advantages, LLMs also present several ***disadvantages***. One of the primary concerns is their impact on the labor market. As these models become more capable, they may replace human workers in fields that involve routine, repetitive tasks, leading to job displacement and high unemployment rates in certain industries. LLMs also have substantial resource and time expenditure in terms of both training and operation (Arrieta et al., 2019). Training a large model requires massive amounts of computational power and energy, leading to high costs and environmental concerns. The process of fine-tuning and maintaining these models can also be resource-intensive, which may limit their accessibility for smaller organizations or startups. Another disadvantage is the absence of human emotional engagement. LLMs, like other AI systems, lack emotional intelligence and cannot truly understand human feelings, ethics, or social context. This limitation makes them ill-suited for tasks that require empathy, nuanced judgment, or moral decision-making. They are also prone to limiting creative problem-solving abilities since they are primarily based on patterns in data, which can restrict their ability to think outside the box or generate truly innovative ideas (Khanzode & Sarode, 2023).

Furthermore, LLMs can contribute to technological dependency, particularly in younger generations who may rely too heavily on these systems for decision-making and problem-solving. This reliance can undermine critical thinking skills and creativity, making it difficult for individuals to perform tasks independently without the assistance of AI. Another concern is the high development costs associated with creating and maintaining large language models (Schmager et al., 2025). The expertise, infrastructure, and time required to train these models contribute to their high price, making it a significant barrier for many organizations, especially those without substantial resources. Lastly, the potential for technological misuse is a significant drawback of LLMs. If not carefully controlled and monitored, these models can be used to spread misinformation, generate malicious content, or even manipulate public

opinion, which can lead to unintended consequences and harm (Khanzode & Sarode, 2023).

2.2.3. Ethics and Biases related to Large Language Models

Artificial intelligence is rapidly advancing, becoming increasingly integrated into daily life and offering significant benefits, particularly in enhancing productivity. As more people rely on AI technologies, concerns surrounding its ethical implications also grow. One major issue is the lack of clear legislation and regulation, which complicates discussions about how AI should be responsibly used and its potential impact on society (Blanchard & Taddeo, 2023).

Figure 7 : Ethical Concerns of AI technology

A critical ethical concern is algorithmic bias, which stems from the fact that AI systems are trained on large datasets. These datasets, while vast, are not always perfectly representative of the real world. As a result, the AI may make decisions that reflect the biases present in the data, leading to discriminatory outcomes based on factors such as race, gender, or socioeconomic status. In sensitive areas such as hiring practices, financial services, or judicial decisions, biased AI could perpetuate existing inequalities and exacerbate societal discrimination (Stahl et al., 2022). Another pressing ethical

issue is privacy and data protection. AI systems require vast amounts of personal data to function, raising concerns about how this data is collected, stored, and used. There is often a lack of transparency about how personal information is managed, which can erode public trust in these technologies. The opacity of AI systems also presents a challenge for accountability. These models are often described as "black boxes," meaning their decision-making processes are difficult to understand. This lack of transparency can undermine trust in AI outputs and make it harder to assign accountability when things go wrong, such as in cases of errors or damages caused by the system (Stahl et al., 2022).

Job displacement is another significant ethical concern related to AI. As automation and AI take over more routine tasks, workers in certain industries may face unemployment or have to adapt to new roles. This shift could exacerbate economic inequality, particularly for those without the skills to transition into new job markets created by AI technologies. While AI has the potential to improve efficiency and productivity, its rapid integration into the workforce may leave many people behind, leading to social and economic disruptions. The potential of AI for manipulation and misinformation also poses a significant ethical challenge. When used irresponsibly, AI tools can generate disinformation or spread false narratives, intentionally or unintentionally. This misuse of AI can undermine public trust, manipulate public opinion, and destabilize society. It is crucial that AI systems are designed to be fair, transparent, and impartial to prevent such negative consequences (Blanchard & Taddeo, 2023). Additionally, the environmental impact of AI is an emerging concern. The infrastructure required to power these systems consumes significant energy and contributes to carbon emissions. As AI technology continues to evolve, it is essential to consider its long-term environmental sustainability, ensuring that the benefits of AI do not come at the cost of damaging the planet.

Chapter 3: Applications of AI

Artificial Intelligence (AI) has significantly transformed the way organizations approach Business Process Modeling, particularly in the domain of process discovery. Process discovery, a crucial aspect of BPMN modeling, involves extracting process models from event logs to understand and improve business operations. AI-driven techniques enhance this process by automating the identification of process flows, detecting inefficiencies, and predicting potential optimizations. AI models such as ChatGPT and Gemini contribute to process discovery by analyzing vast amounts of unstructured data, learning from past workflows, and generating accurate BPMN diagrams with minimal human intervention. These advancements facilitate greater efficiency, reduce errors, and provide organizations with deeper insights into their business processes, ultimately leading to improved decision-making and operational efficiency.

3.1. Education and Academia

AI has significantly contributed to process discovery in education and academia by streamlining administrative processes and improving learning methodologies. Universities and educational institutions generate extensive event logs related to student enrollment, grading systems, course registration, and faculty management. AI-powered process discovery tools can analyze these logs to identify inefficiencies and suggest optimized workflows. For instance, AI can detect bottlenecks in student registration processes, recommend automation in grading systems, and enhance resource allocation for academic scheduling. By applying BPMN-based AI models, institutions can improve administrative decision-making and enhance the overall student experience (Van der Aalst, 2016).

3.2. Finance

In the financial sector, AI-driven process discovery has revolutionized compliance monitoring, fraud detection, and transaction management. Financial institutions deal with vast amounts of transactional data, requiring accurate and efficient process

modeling to maintain regulatory compliance and operational transparency. AI models like ChatGPT and Gemini assist in identifying process deviations, detecting fraudulent activities, and optimizing loan approval workflows by extracting insights from historical transaction logs. BPMN models enhanced with AI can automate risk assessments, reduce manual intervention, and improve customer service efficiency (Mendling et al., 2018).

3.3. Manufacturing

Manufacturing industries benefit from AI-powered process discovery by optimizing production lines, reducing waste, and enhancing supply chain management. AI models analyze real-time production data to identify inefficiencies in manufacturing workflows, detect machine failures before they occur, and optimize inventory management. By applying BPMN process modeling, manufacturers can automate decision-making, streamline quality control, and improve resource utilization. AI-driven predictive maintenance further reduces downtime and increases production efficiency, leading to higher profitability and sustainability in industrial operations (Weske, 2019).

3.4. Tourism

The tourism industry leverages AI-driven process discovery to enhance customer experiences, optimize booking systems, and improve service delivery. AI analyzes customer interactions, travel preferences, and booking trends to develop efficient BPMN models for travel agencies, hotels, and airlines. AI-driven chatbots, personalized itinerary planning, and automated customer service systems improve operational workflows and enhance user satisfaction. Moreover, AI models help detect travel fraud, optimize pricing strategies, and predict demand fluctuations, ensuring better decision-making in tourism management (Dumas et al., 2018).

Chapter 4: Methodology

This thesis explores the efficiency, accuracy, and potential of artificial intelligence in automating the generation of business process modeling diagrams. By leveraging AI-powered tools such as ChatGPT, and Gemini the research assesses how these models compare to human-generated Business Process Model and Notation (BPMN) diagrams. The objective is to establish a foundational understanding of AI's capability in both technical precision and creative adaptability within process modeling.

The methodology employed a sequential and iterative approach rather than a simple direct comparison. The research recognized that AI-generated outputs, particularly in early stages, might not always be optimal. However, through iterative refinement and improved prompt engineering, the quality of AI-produced diagrams significantly improved. This iterative process not only enhanced the AI's ability to generate more accurate and structured BPMN diagrams but also provided valuable insights into how these models learn, adapt, and refine their outputs over time.

By comparing AI-generated diagrams from ChatGPT and Gemini with those created by human with right answer, this study highlights the strengths and limitations of AI in process modeling. Key areas of analysis include the logical coherence of workflows, adherence to BPMN standards, and the ability of the model to understand complex process structures. Additionally, the research evaluates how different AI models interpret and visualize business processes, shedding light on their varying capabilities in structuring and optimizing workflows. The findings contribute to a broader understanding of AI's role in automating business process modeling. While AI offers significant advantages in speed and consistency, human expertise remains essential for ensuring contextual accuracy, nuanced decision-making, and creative problem-solving. This research underscores the importance of combining AI with human oversight to achieve optimal results in BPMN diagram generation and business process automation.

4.1. Background and Literature Review

Business Process Modeling and Notation (BPMN) diagrams play a crucial role in visually representing business processes, facilitating clear communication between business analysts, developers, and stakeholders. Ensuring the quality and effectiveness of BPMN diagrams is essential, as they serve as a bridge between conceptual process design and actual implementation (White, 2004). The challenge lies in evaluating these diagrams to determine their clarity, correctness, effectiveness, and completeness, especially when two structurally different diagrams may be behaviorally identical. Various approaches have been proposed to assess the equivalence and efficiency of BPMN diagrams, leveraging mathematical techniques and workflow analysis (Wohed et al., 2006).

One method for evaluating BPMN diagrams involves mathematical formalism to determine process equivalence. Even if two diagrams differ in structure, they can be analyzed for underlying behavioral similarities. This evaluation consists of defining specific criteria, establishing formal relationships, verifying transitivity and equivalence properties, and conducting equivalence checks (Vergidis et al., 2007). Mathematical models allow for precise assessment and comparison of business processes, ensuring that different BPMN representations remain functionally consistent (Safieddine & Nakhoul, 2018). Techniques such as automata theory, process algebra, and Petri Nets are widely used to validate equivalence by comparing different process structures and identifying discrepancies in behavior (Wohed et al., 2006). These mathematical evaluations contribute to maintaining coherence, reliability, and process optimization. Beyond mathematical techniques, evaluating BPMN diagrams based on workflow appropriateness provides another layer of analysis (Lopes & Guerreiro, 2023). Business processes involve control flow, data flow, and resource allocation, all of which impact workflow efficiency (Stravinskiene & Serafinas, 2020). The Workflow Pattern framework categorizes BPMN evaluation into three main perspectives: control flow, data flow, and resource management. The control flow perspective assesses the logical arrangement of process elements such as tasks, gateways, and event sequences, ensuring that the

overall workflow is effectively structured. The data perspective focuses on how information is handled within the process, including data visibility, transfer, and storage. This perspective ensures that BPMN models accurately represent data interactions and dependencies (Stravinskiene & Serafinas, 2020). Lastly, the resource perspective evaluates the allocation and management of resources within the process, particularly through BPMN elements like pools and lanes. Proper representation of resource interactions enhances process clarity and operational feasibility.

With the advent of artificial intelligence, the automation of BPMN diagram generation has introduced new possibilities for improving process modeling. This thesis incorporates AI models such as ChatGPT and Gemini to generate BPMN diagrams and compare them with human-created versions. The research follows an iterative approach, refining input prompts to enhance AI-generated outputs over time. Early AI-generated BPMN diagrams may lack precision or completeness, but continuous refinements improve accuracy, demonstrating how AI learns and adapts in technical tasks. Comparing AI-generated diagrams with human-created ones highlights strengths and limitations in AI's ability to interpret, structure, and optimize business processes. While AI accelerates process modeling and ensures consistency, human oversight remains essential for contextual understanding and nuanced decision-making. By integrating AI into BPMN diagram generation, this thesis aims to explore the potential for automating business process modeling while maintaining high standards of accuracy and clarity. The iterative refinement of AI models like ChatGPT, and Gemini shows how machine learning can enhance process visualization, reducing manual effort while improving efficiency. However, the need for structured evaluation methods—whether through mathematical validation or workflow assessment—remains crucial to ensuring that AI-generated BPMN diagrams align with industry standards and effectively represent real-world business processes (Segatto et al., 2013).

4.2. Study Design

This rthesis focuses on evaluating the capabilities of artificial intelligence models, specifically ChatGPT, and Gemini in generating Business Process Model and Notation (BPMN) diagrams compared to human-created solutions. The study is based on a set of assignements exercises taken from the Digital Process Management course within the Master's program in Artificial Intelligence and Data Analytics as well as Strategic Information Management at Hochschule Neu Ulm, spanning from 2023 to 2025. These assignements were carefully selected to encompass a range of difficulties, ensuring a comprehensive assessment of both AI-generated and human-created BPMN diagrams. The assignements were provided by the course instructor as well as answers were discussed in class, serving as a solid foundation for analyzing the modeling capabilities of both AI and human specialists.

The study was conducted using a structured approach, beginning with a deep understanding of BPMN elements such as processes, gateways, flows, pools, and lanes. The BPMN diagrams were created using BPMN.io as well as Signavio, a widely used tools for modeling business processes. The research initially involved manually constructing the diagrams, ensuring that the fundamental aspects of each process were accurately represented. Simultaneously, the same exercises were input into AI models using standardized prompts that closely mirrored the instructions provided to human participants. The outputs generated by ChatGPT, and Gemini were then systematically collected and compared to the human-developed solutions.

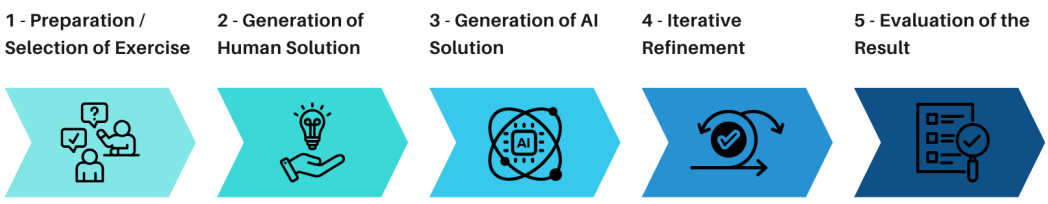

Figure 8 : Evaluation Process

To enhance the accuracy of AI-generated BPMN diagrams, an iterative refinement process was applied. This involved reviewing the initial AI outputs, identifying errors or inconsistencies, and adjusting the input prompts accordingly. Refinements included providing additional context, copy and pasting errors in XML file directly in the system, rephrasing instructions, and guiding the AI toward producing more precise and contextually relevant diagrams. Each iteration was carefully documented, noting the modifications made and their impact on the quality of the generated solutions. This iterative process allowed for a detailed examination of how AI models adapt and improve over multiple refinement cycles, demonstrating their ability to learn and enhance their performance in complex diagram-generation tasks. By comparing human and AI-generated BPMN diagrams, this research aims to highlight the potential efficiencies and challenges associated with using AI for business process modeling. The study provides insight into how AI models interpret and construct BPMN diagrams, emphasizing the role of iterative refinement in achieving optimal results. It also explores the extent to which AI can replicate human expertise in process modeling, shedding light on both the advantages and limitations of AI-driven automation in this domain. The research follows a structured methodology, beginning with exercise selection and preparation, followed by human and AI solution generation, and concluding with iterative refinement. These steps ensure a comprehensive analysis of AI's capabilities in BPMN diagram generation while demonstrating how AI models can be guided toward producing more accurate and refined outputs through continuous improvement.

4.3. Prompt Design

ChatGPT, and Gemini distinguish themselves by generating responses in a manner that mimics human-like reasoning and maintaining logical flow across multiple interactions. One of the primary challenges in leveraging these AI models for Business Process Modeling is ensuring that they provide accurate and contextually relevant answers rather than responses that may be technically incorrect or hallucinative. A crucial aspect of achieving this accuracy is the strategic application of prompt engineering, which

involves designing, refining, and optimizing prompts to guide AI models toward producing precise and meaningful outputs.

Prompt engineering plays a fundamental role in this study, as it determines how well AI models can generate BPMN diagrams using the BPMN 2.0 formalism. The process involved iterative adjustments to the input prompts, ensuring that the AI was not only capable of analyzing the problem logically but also generating a structured BPMN diagram with graphical representation through an XML file. Since the databases of these models, including ChatGPT, and Gemini, are trained on vast amounts of internet-based information, the key challenge was directing them to extract domain-specific knowledge relevant to business process modeling. Without precise prompts, the models could generate misleading or incorrect outputs, making prompt refinement essential to obtaining reliable results. To achieve better accuracy, prompts needed to be formatted correctly, explicitly stated, and structured with clarity. Providing contextual details and well-defined instructions improved the ability of AI to generate meaningful BPMN diagrams. Depending on the complexity of the query, different prompting strategies were tested, including few-shot and zero-shot prompting. Few-shot prompting involved providing the model with several examples to guide its response, making it more likely to produce relevant and structured outputs. In contrast, zero-shot prompting relied solely on the model's pre-existing knowledge without examples, requiring a more carefully crafted prompt to yield the desired results.

Through repeated testing and refinement, it became evident that prompt engineering is an iterative process. By continuously refining input queries, adjusting the specificity of instructions, and incorporating examples when necessary, the AI models demonstrated an improved ability to understand domain-specific requirements and generate more precise BPMN diagrams. The study highlights that while AI models such as ChatGPT, and Gemini are powerful tools for business process modeling, their effectiveness largely depends on how well prompts are structured. The continuous improvement of prompting strategies ensures that AI-generated outputs remain accurate, domain-specific, and aligned with the technical requirements of BPMN diagrams.

4..3.1. Prompt Development Process for BPMN Diagrams

The GPEI methodology, which stands for Goal, Prompt, Evaluation, and Iteration, was one of the key approaches used in this study to assess the ability of AI models to generate BPMN diagrams. This iterative process involved defining a clear objective, crafting a structured prompt, evaluating the AI-generated responses, and refining the input to improve accuracy (Velásquez-Henao et al., 2023). Since AI-generated outputs can be inconsistent or incomplete, especially when dealing with structured graphical representations, the study focused on how modifications to the prompt affected the quality and reliability of the responses from ChatGPT, Gemini.

Through multiple attempts, it became evident that the initial prompts were too vague, leading to responses that lacked structure and specificity. Early outputs often consisted of textual descriptions rather than a fully structured BPMN diagram. The models struggled to interpret the problem in a way that translated into a visual format, highlighting the need for more precise instructions. One of the key adjustments was to instruct the AI to generate responses in XML format, as BPMN modeling tools rely on this structure for creating diagrams. However, every model of ChatGPT as well as other model like Gemini, initially failed to produce complete XML outputs, preventing successful diagram generation.

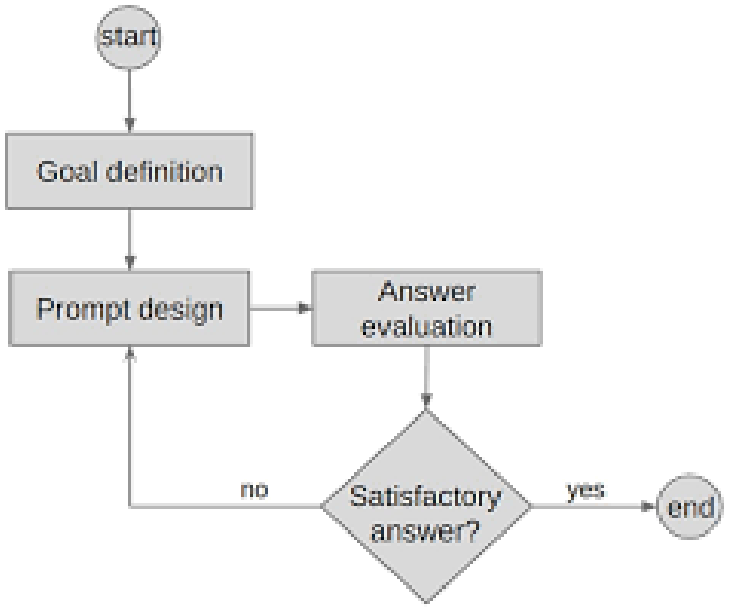

Figure 9 : GPEI Methodology

An alternative approach involved explicitly guiding the AI through a structured process. This included feeding the model a standardized BPMN diagram format, providing a detailed problem description, and then requesting a structured solution. This modification yielded slightly better results, with AI-generated outputs that could be partially visualized in BPMN tools. However, issues persisted, such as missing elements and inconsistencies in the generated diagrams. One method used to work around these limitations was breaking the response into multiple segments, requesting the AI to generate parts of the diagram separately. Despite this, maintaining continuity across sections proved difficult, revealing a fundamental limitation in AI models when handling graphical representations.

To ensure consistency in outputs, a complete shift in strategy was necessary. Instead of expecting AI models to generate fully functional BPMN diagrams, they were tasked with logically identifying key elements such as pools, lanes, processes, and gateways. Human intervention was then used to manually construct the diagrams based on AI-generated guidelines. This approach leveraged AI as an assistive tool rather than a

full automation solution, acknowledging the models' limitations in handling structured diagram formats.

The final phase of the study involved standardizing the prompts across multiple AI models to compare their performance objectively. By using the same prompt structure, it was possible to assess the differences in how ChatGPT, and Gemini approached BPMN diagram generation. ChatGPT demonstrated a stronger understanding of BPMN 2.0 formalism compared to Gemini, effectively incorporating event connections and maintaining a more logical sequence. This confirmed the hypothesis that models trained on larger datasets, like ChatGPT, would perform better in generating structured business process models.

Despite improvements, AI-generated diagrams were not always entirely accurate or complete, particularly when dealing with complex workflows. In cases where the problem description was intricate, additional instructions and refinements were necessary to guide the AI toward producing a more usable output. The primary expectation from AI was to generate a logical sequence that included all key BPMN elements, from start points to gateways, message flows, and endpoints. While AI models showed promise in assisting with BPMN diagram generation, human intervention remained necessary for refining and ensuring accuracy in the final outputs.

The study ultimately demonstrated that while AI models like ChatGPT, and Gemini can contribute to BPMN diagram generation, they are not yet fully capable of replacing human expertise in this domain. Instead, they serve as powerful tools for aiding business analysts and developers by providing structured insights, automating certain aspects of diagram creation, and streamlining the overall modeling process. The iterative nature of prompt engineering played a crucial role in optimizing AI-generated outputs, highlighting the importance of refining input structures to achieve more reliable and meaningful business process models.

4.4. Evaluation

4.4.1. Evaluation Criteria

The methodology for evaluating and interpreting results played a crucial role in this study. After refining the final prompt, ChatGPT, and Gemini were used to generate logical sequences for all the elements of the BPMN diagrams. The student conducting this research then constructed the diagrams based on these AI-generated instructions. Once the diagrams were completed, they were compared against the human-created versions to assess the differences and similarities between AI-generated and human-generated outputs.

The central objective of this study was to determine whether AI models could generate BPMN diagrams comparable to those created by humans and, if so, whether they could surpass human performance in terms of accuracy and structure. To carry out this evaluation effectively, it was necessary to establish specific assessment criteria. The study focused on two key aspects: syntactic evaluation and appropriateness evaluation. Syntactic evaluation examined the accuracy of the BPMN symbols, syntax, and overall structure used in the diagrams produced by AI and humans. Since BPMN 2.0 follows a formal notation, this type of evaluation was relatively straightforward and could be conducted objectively. Errors in symbol usage, misinterpretation of syntax, or incorrect application of BPMN language could be easily identified and analyzed. Unlike behavioral evaluation, where two structurally different diagrams could still be correct, syntactic evaluation strictly measured the adherence to BPMN 2.0 standards.

Appropriateness evaluation, on the other hand, focused on the overall logic and correctness of the diagrams based on workflow patterns. This assessment considered control flow and resource flow perspectives, ensuring that key BPMN components such as pools, lanes, and sequence flows were appropriately used. By evaluating how well the models adhered to standard workflow principles, it was possible to determine the effectiveness of AI-generated diagrams in representing business processes accurately.

Other evaluation criteria, such as semantic and practical assessments, were intentionally excluded from the study. Since the BPMN exercises were derived from academic exam papers rather than real-world business projects, it was challenging to determine the practical accuracy of a given diagram from a purely semantic standpoint. Additionally, the lack of specialized professional tools and expertise made it difficult to conduct an in-depth semantic analysis. Given these constraints, the study prioritized syntactic and appropriateness evaluations as the most reliable means of comparing AI-generated and human-generated BPMN diagrams. Through this methodology, the research aimed to provide insights into the potential of AI models like ChatGPT, and Gemini in assisting or even automating business process modeling. While AI demonstrated a capacity for generating structured BPMN diagrams, the study also highlighted areas where human intervention remained essential for refining and validating results.

4.4.2. Process of Evaluation

The evaluation process was structured into several distinct steps to ensure a comprehensive assessment. Establishing a directory for each problem, containing the professor's solution along with the solutions generated by three entities: a human participant, ChatGPT and Gemini.

- Defining a structured scoring system.
- Evaluating the appropriateness of all diagrams for each exercise and each participant.
- Conducting a syntactic analysis of all diagrams.
- Constructing the final results matrix—one to assess appropriateness and another to evaluate syntax errors.

To maintain clarity and efficiency, individual folders were created for each exam problem. Each folder housed the correct solution as provided by the professor and the corresponding solutions from the human participant, ChatGPT, and Gemini. This

organizational structure facilitated seamless comparison and evaluation. The next step involved defining the evaluation criteria for assessing the appropriateness of each diagram. This was determined by consensus, taking into account all possible cases. Appropriateness was assessed from two primary perspectives:

Resource Perspective: This criterion examined whether the pools and lanes in the solutions were correctly aligned with the actual problem requirements.
Control Flow Perspective: This focused on whether key elements of the diagram, including start events, tasks, gateways, timer events, and end events, were accurately represented. Furthermore, it examined the logical sequence of events to ensure adherence to the problem description.

To structure this assessment, a matrix was developed containing key evaluation criteria. Each diagram was scored based on a defined system:

Score	Definition
1	Correct
0.5	Partially Correct
0	Incorrect

Table 1: Structure of the appropriateness evaluation dataset for each exercise

A score of 1 indicated that the diagram component matched the professor's solution exactly. A score of 0.5 indicated that while the component did not fully align with the correct solution, it was close—either through a partially correct definition or assignment to a different pool. A score of 0 meant the component was either incorrect or entirely missing. The final matrix was structured such that for each exercise, each component was assessed for each participant (human, ChatGPT, and Gemini) and assigned a corresponding score. This evaluation process was carried out across a total of 19

exercises. The syntax evaluation process was straightforward and conducted using a specialized tool provided by the Department of Computer Engineering. This tool enabled an objective, quantitative assessment of syntax-related errors in each diagram.

A comprehensive table was used to document syntax errors, detailing the specific issues detected in each participant's solution. This analysis was repeated systematically across all 19 exercises, ensuring a thorough examination of both correctness and syntax compliance. The results were ultimately compiled in a structured format to facilitate comparison and draw meaningful insights regarding the performance of human-generated versus AI-generated solutions in Business Process Modeling.

4.4.4. Statistical Methods

A crucial component of this thesis involves the application of statistical methods to analyze the data collected from the evaluation of exam papers, where final scores were assigned. The data was organized into a summary table that presented the results, and the statistical analyses were conducted separately for two different types of evaluations: syntactic evaluation and adequacy evaluation. The process was divided into two main stages: data preparation and statistical analysis of the data.

For each actor (human, ChatGPT, Gemini), the percentage of correct, partially correct, and incorrect responses was calculated. However, the dataset was not ideal because analyzing each combination of response type (correct, partially correct, incorrect) and actor (human, ChatGPT, and Gemini) would have been overly complex and time-consuming. The focus of the thesis was on analyzing the overall performance of each actor, so to simplify this process, a weighted average score for each actor was computed. The formula used for calculating these scores is as follows:

- Weighted Human Score = (1 × Correct) + (0.5 × Partially Correct) + (0 × Incorrect)

- Weighted ChatGPT Score = (1 × Correct) + (0.5 × Partially Correct) + (0 × Incorrect)
- Weighted Gemini Score = (1 × Correct) + (0.5 × Partially Correct) + (0 × Incorrect)

This formula reflects the relative value of each type of response in the calculation of the total score. Once the data was organized and the weighted averages were calculated, the next step was to perform a syntactic analysis using statistical methods. Two types of analyses were conducted for this purpose: the t-test and Analysis of Variance (ANOVA).

The ***t-test*** is a statistical method used to determine if two samples from the same population have the same mean. Since the exercises used in the evaluation were common to all three actors (human, ChatGPT, and Gemini), a paired t-test was employed to compare the performance of each actor. Specifically, two t-tests were run: one comparing the performance of Human vs. ChatGPT and another comparing the performance of Human vs. Gemini. This test was used to evaluate whether the human performer or any of the AI models performed better, based on accuracy scores. ***ANOVA*** is a statistical method used to compare the means of three or more groups to determine if at least one group differs significantly from the others. It tests the null hypothesis, which states that all group means are the same, against the alternative hypothesis that at least one group mean is different. There are various types of ANOVA, depending on the variables being studied.

In this thesis, single-factor ***ANOVA*** was applied, focusing on a single independent variable (the exercises). The purpose was to evaluate whether there were significant differences in the mean scores across the three actors (human, ChatGPT, and Gemini). This test was used to analyze both syntactic performance and adequacy evaluation. The reason why ANOVA was performed after the t-tests was to ensure more reliable statistical results. The limitation of the t-test is that it requires pairwise comparisons, which increases the risk of type I error (false positive results). By conducting ANOVA after the t-tests, the risk of inaccurate results was reduced.

4.5.5. Limitations of used Evaluation Methodology

The selected methodology for evaluating the performance of different actors (ChatGPT, and Gemini Model) in Business Process Modeling (BPM) has some limitations, which can be categorized into two main types: Potential Biases and Constraints.

Potential Biases

The first type of limitation arises from observer biases, which can occur at any stage of the research process, including the initial research, the selection of BPMN exercises to be analyzed, and the evaluation phase. Since this study was mainly conducted by a single researcher (a student), there is an increased risk of such biases influencing the outcomes. The issue with observer bias is that it refers to cases where the researcher's own expectations or preconceived notions might influence their interpretation of the data, leading to skewed results.

To minimize the impact of this potential bias, multiple data sources were used to validate the results and ensure their accuracy. Furthermore, periodic reviews and consultations with research supervisors were implemented throughout the study to align the results with the agreed-upon methodology, helping to maintain consistency and coherence between the findings and the strategies followed during the course of the research.

Constraints

The second type of limitation involves various constraints that directly influence the evaluation process in BPM. These constraints include diagram complexity, the limited expertise of the human solver, and the scoring system used for adequacy assessment.

One such constraint is diagram complexity, which varies depending on the exercise. Simpler tasks tend to lead to higher accuracy rates across all the actors (ChatGPT, Gemini Model, and the human solver), whereas more complex tasks reveal differences

in problem-solving capabilities. The results from these exercises can therefore differ depending on the complexity level, making it harder to compare the actors directly in all cases. Another important factor is the competence of the human solver. In this study, the human actor was represented by a single student whose skills in solving BPMN modeling exercises are limited. This individual's performance may not be representative of a broader population, as other participants with different levels of expertise could produce vastly different results. The skill and experience of the human solver, therefore, have a significant impact on the outcome of the evaluation.

Finally, the scoring system used in the study is another limiting factor. It was based on three categories: correct, partially correct, and incorrect. However, this limited range could be expanded to offer a more detailed grading scale that accounts for a wider array of potential errors and anomalies that might occur in BPM diagrams. A more granular scale would provide a finer evaluation of the actors' performances, potentially leading to more accurate and nuanced results.The following table outlines the key limitations identified in the methodology, categorized into biases and constraints:

Type of Limitation	Description	Impact on Evaluation
Observer Biases	Bias from the researcher, influenced by personal expectations or perceptions.	Could skew data interpretation and result in biased outcomes.
Diagram Complexity	Varying complexity of BPMN exercises may affect performance across actors.	Simpler tasks may show less differentiation, while complex ones may highlight model limitations.
Human Solver Competence	The human solver’s limited BPMN expertise, based on a single student.	May lead to results not representative of a broader human capability range.

Scoring System	Use of three scores for adequacy assessment (correct, partially correct, incorrect).	More detailed scoring could have provided a finer analysis of performance.

Table 2: Identified Limitations and Their Impact

Chapter 5: Results and Discussion

5.1. Results

The statistical analysis in this study focuses on evaluating the performance differences between human evaluators and AI models, specifically ChatGPT and Gemini. Several key statistical tests were conducted to assess these differences, including paired sample t-tests, descriptive statistics, correlation analysis, one-way ANOVA, bootstrap resampling, and independent sample t-tests.

A paired sample t-test was conducted to compare the performance scores of ChatGPT and Gemini, as both AI models were evaluated on the same set of tasks. The analysis revealed that ChatGPT had a higher mean performance score (M = 0.525, SD = 0.3432) than Gemini (M = 0.325, SD = 0.3726).

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	ChatGPT Performance Score	.525	20	.3432	.0767
	Gemini Performance Score	.325	20	.3726	.0833

Table 3: ChatGPT vs. Gemini Performance Statistics

The paired correlation between these two models was found to be r = 0.653, which was statistically significant (p = 0.002), indicating a moderate to strong relationship between their performances. The mean difference between ChatGPT and Gemini was 0.200 (SD = 0.2991), and the paired t-test result was statistically significant (t = 2.990, df = 19, p = 0.008), confirming a meaningful performance difference. The effect size, measured using Cohen's d, was 0.669, which suggests a moderate to large effect, reinforcing the conclusion that ChatGPT performed consistently better than Gemini.

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	ChatGPT Performance Score & Gemini Performance Score	20	.653	.002

Table 4: Correlation Between ChatGPT and Gemini Performance Scores

To further examine the overall performance distribution, descriptive statistics were computed for human evaluators, ChatGPT, and Gemini. The results indicated that human evaluators consistently scored 1.000, with no variation (SD = 0.000). In contrast, ChatGPT exhibited a mean performance score of 0.525 with a standard deviation of 0.3432, while Gemini had a lower mean of 0.325 with a standard deviation of 0.3726. The variance for ChatGPT was calculated as 0.118, whereas Gemini exhibited a slightly higher variance of 0.139, suggesting greater variability in its performance. These descriptive statistics indicate that while both AI models demonstrated fluctuations in their performance, ChatGPT generally outperformed Gemini.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Variance
Human Performance Score	20	1.0	1.0	1.000	.0000	.000
ChatGPT Performance Score	20	.0	1.0	.525	.3432	.118
Gemini Performance Score	20	.0	1.0	.325	.3726	.139
Valid N (listwise)	20					

Table 4: Human, ChatGPT, and Gemini Performance Statistics

A correlation analysis was conducted to explore the relationships between human, ChatGPT, and Gemini performance scores. The Pearson correlation coefficient between ChatGPT and Gemini was found to be r = 0.653, with a p-value of 0.002, suggesting a significant positive correlation between their performances. However, the correlation

between human evaluators and the AI models could not be computed due to the constant human performance score of 1.000, preventing further relational analysis.

Correlations

		Human Performance Score	ChatGPT Performance Score	Gemini Performance Score
Human Performance Score	Pearson Correlation	.[a]	.[a]	.[a]
	Sig. (2-tailed)		.	.
	N	20	20	20
ChatGPT Performance Score	Pearson Correlation	.[a]	1	.653**
	Sig. (2-tailed)	.		.002
	N	20	20	20
Gemini Performance Score	Pearson Correlation	.[a]	.653**	1
	Sig. (2-tailed)	.	.002	
	N	20	20	20

**. Correlation is significant at the 0.01 level (2-tailed).

a. Cannot be computed because at least one of the variables is constant.

Table 5: Intercorrelation of Human, ChatGPT, and Gemini Performance

To determine whether significant differences existed among human evaluators, ChatGPT, and Gemini, a one-way ANOVA test was performed. The analysis yielded an F-value of 28.110 ($p = 0.000$), confirming that at least one of the groups had a significantly different mean score.

ANOVA

Human Performance Score

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	4.808	2	2.404	28.110	.000
Within Groups	4.875	57	.086		
Total	9.683	59			

Table 6: ANOVA Table for Human Performance Score

A post-hoc Tukey test was conducted to examine pairwise differences, revealing that human evaluators significantly outperformed both ChatGPT (mean difference = 0.475, p = 0.000) and Gemini (mean difference = 0.675, p = 0.000). However, the difference between ChatGPT and Gemini (mean difference = 0.200, p = 0.087) was not statistically significant. This result suggests that while ChatGPT tended to have higher performance scores than Gemini, the difference was not strong enough to be considered significant in this particular test.

Multiple Comparisons

Dependent Variable: Human Performance Score

Tukey HSD

(I) Labels	(J) Labels	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Human	ChatGPT	.4750*	.0925	.000	.252	.698
	Gemini	.6750*	.0925	.000	.452	.898
ChatGPT	Human	-.4750*	.0925	.000	-.698	-.252
	Gemini	.2000	.0925	.087	-.023	.423
Gemini	Human	-.6750*	.0925	.000	-.898	-.452
	ChatGPT	-.2000	.0925	.087	-.423	.023

*. The mean difference is significant at the 0.05 level.

Table 7: Post-Hoc Analysis: Human, ChatGPT, and Gemini Performance

To validate the robustness of these findings, a bootstrap resampling analysis with 100 samples was conducted. The confidence intervals for the mean differences aligned with the previously obtained values, reinforcing the conclusion that human evaluators significantly outperformed both AI models while the difference between ChatGPT and Gemini remained marginal.

Independent sample t-tests were also performed to compare the performance of each AI model against human evaluators separately. The comparison between human evaluators and ChatGPT yielded a t-value of 6.19 (p = 0.000), with a mean difference of

0.475, confirming a statistically significant disparity. Similarly, the comparison between human evaluators and Gemini produced an even larger t-value of 8.102 ($p = 0.000$), with a mean difference of 0.675. The effect sizes for these comparisons were large, with Cohen's d values of 1.958 and 2.562, respectively, indicating substantial performance differences. However, when comparing ChatGPT and Gemini using an independent samples t-test, the observed mean difference of 0.200 was not statistically significant ($t = 1.766$, $p = 0.085$). This aligns with the findings of the ANOVA post-hoc test, further supporting that although ChatGPT exhibited slightly better performance than Gemini, the variability in scores prevented a definitive conclusion regarding their relative superiority.

Overall, the results of these statistical tests consistently demonstrate that human evaluators significantly outperform both ChatGPT and Gemini in terms of performance scores. While ChatGPT generally performed better than Gemini, the statistical significance of this difference varied depending on the test applied. The correlation analysis further revealed a strong positive relationship between ChatGPT and Gemini's performances, indicating some consistency in how both models function. The application of bootstrap resampling validated these findings, confirming their reliability. Ultimately, these results suggest that while AI models are capable of performing evaluation tasks, they still exhibit notable limitations when compared to human evaluators, and the performance gap between ChatGPT and Gemini remains relatively small.

5.2. Discussion

Following an in-depth analysis of the collected data, it is evident that the performance of the three entities examined in this study—humans, ChatGPT, and Gemini—differed significantly. Initially, the research was grounded in a central null hypothesis:

"*The performance of the three actors (human, ChatGPT, and Gemini) in the study is equivalent, indicating no significant differences in performance.*"

However, statistical analysis demonstrated that human participants achieved higher accuracy scores than the AI models. Interestingly, the ANOVA analysis for syntax errors revealed an unexpected finding—ChatGPT exhibited a lower average syntax error rate than the human participant. Despite this, the null hypothesis was ultimately rejected as the p-value fell below the conventional alpha threshold of 0.05, confirming a statistically significant performance difference.

Given the experimental nature of this research, initial expectations assumed that human and AI-generated BPMN diagrams would yield comparable results. However, the findings contradicted this assumption, showing that AI models, at their current stage of development, cannot yet match the accuracy of human-generated diagrams, particularly when created by students or domain experts. With the rapid advancement of artificial intelligence, the automated generation of BPMN diagrams remains an open area of study. Notably, while prior research has explored the adaptation of AI for process modeling, no existing studies have directly addressed this specific challenge.

This study contributes to the field in two key ways. First, it serves as empirical evidence regarding AI-generated diagrams' viability and limitations. Second, it highlights important methodological constraints. One of the study's strengths lies in the diversity of its dataset. The problem descriptions used were taken from real exam questions designed to assess students' logical reasoning and understanding, making them inherently complex. This aspect enabled a rigorous evaluation of AI models’ ability to process and interpret varied problem statements. The iterative nature of the methodology, which allowed for continuous improvements, also strengthened the study. Nevertheless, several limitations must be acknowledged. The most critical limitation is that the human-generated diagrams were created by a single participant. Human performance is inherently subjective and influenced by individual expertise, which introduces a potential bias. A broader sample of human participants could provide a more comprehensive performance comparison. Another major limitation is that the AI models could not directly generate a graphical representation of BPMN diagrams.

Instead, they provided only the logical structure in textual form, requiring manual construction of the visual representation by the researcher. This raises an open question regarding the potential margin of human error in both graphical interpretation and subsequent evaluation of the AI-generated outputs.

Future research should explore multiple directions to address these limitations. From a technological perspective, one potential advancement is training AI models specifically for BPMN notation comprehension and graphical representation. This would involve three major development areas:

1. Enhancing AI understanding of BPMN notations
2. Training models to generate XML files that accurately represent the graphical structure of BPMN diagrams
3. Improving AI reasoning capabilities to ensure logical accuracy and compliance with BPMN standards

Beyond technological improvements, future research could also benefit from expanding the study to include a larger and more diverse set of human participants, reducing potential biases and increasing the reliability of the findings.

This study's significance lies in its attempt to empirically examine an area where direct scientific evidence remains scarce. BPMN diagrams play a crucial role in business process modeling, and this research evaluates AI’s potential as an assistive tool for domain experts. In the long run, AI automation could streamline BPMN diagram generation, potentially reducing manual effort and improving efficiency in process modeling. However, the integration of AI in BPMN automation also raises ethical concerns, particularly regarding data security and privacy. Business processes are often highly confidential, as they provide companies with a competitive edge. Allowing AI to analyze business processes and generate BPMN diagrams could pose risks if adequate data protection measures are not in place.

Additional concerns revolve around bias, fairness, explainability, and transparency in AI-generated results. Since AI models are trained on internet-based data, they may inadvertently reflect biases present in that data. Furthermore, AI algorithms are typically proprietary and inaccessible to end users, leading to concerns over transparency in the decision-making process. These factors contribute to skepticism about whether businesses can fully trust AI-generated BPMN diagrams. A significant challenge identified during this research, which was not initially anticipated, was the AI models' inability to produce coherent graphical representations of BPMN diagrams. The generated XML files lacked consistency and accuracy, necessitating the adoption of an alternative methodology. This limitation highlights a crucial area for future improvement in AI-driven process modeling.

Chapter 6: Evaluations

6.1. Evaluation of AI-Generated Diagrams

This section presents an in-depth analysis and interpretation of the quality and differences in BPMN diagrams generated by ChatGPT and Gemini models. The evaluation is conducted from three key perspectives: diagram quality, adherence to BPMN standards, and usability and interpretability. One of the main challenges encountered when assessing the quality of AI-generated diagrams was the models' inconsistent ability to capture key process components and relationships described in the problem statements. The results indicate that when the problems were complex or contained extensive details, the models often failed to generate an optimal response without additional user guidance. Providing this extra information required significant time and effort. Empirical evaluation of problem-solving accuracy showed that AI models successfully captured approximately 60% of the key elements in the given problem descriptions.

A notable difference emerged between the diagrams produced by ChatGPT and Gemini. In terms of complexity, Gemini-generated diagrams tended to be more detailed than those created by ChatGPT, which aligns with the fact that Gemini has access to a more extensive dataset and potentially superior domain-specific knowledge. However, contrary to expectations, ChatGPT’s diagrams exhibited greater syntactic accuracy. The final assessment revealed that ChatGPT produced a total of 33 errors across all exercises, whereas Gemini had 43 errors, marking a significant discrepancy in syntactic precision. It is important to clarify that when referring to diagrams "generated" by AI models, this does not mean they produced a fully visual BPMN representation. Instead, the AI models provided a textual representation outlining the logical sequence of activities. The actual graphical representation was constructed by humans, as the AI models lacked the capability to create an accurate visual format.

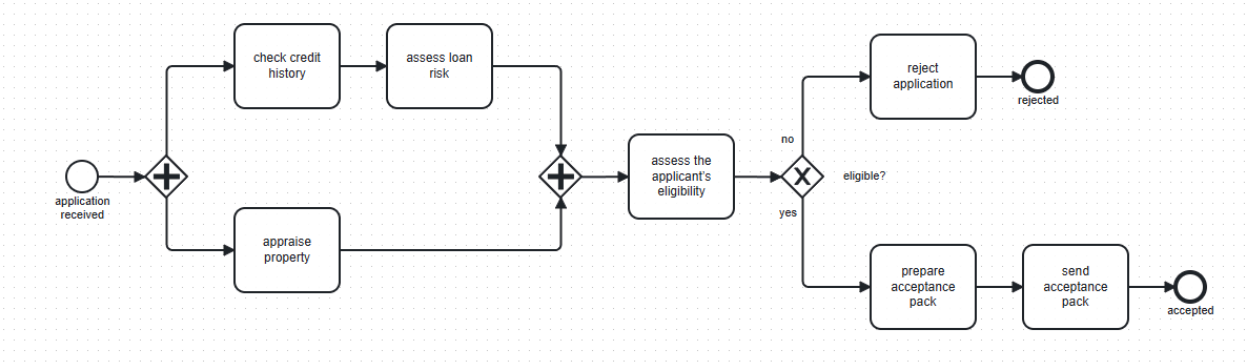

Figure 10 : Diagram generated from ChatGPT in BPMN XML Format

From a logical structure standpoint, both models demonstrated the ability to maintain a coherent sequence of tasks. ChatGPT successfully preserved logical task sequences in most cases, though its solutions often omitted key elements, preventing a fully complete BPMN representation. For instance, ChatGPT rarely identified multiple pools in a process and frequently overlooked message flows. Even when it recognized multiple lanes, it often failed to assign activities to specific lanes or pools, leading to inconsistencies within the generated solution. This limitation was observed frequently and contributed to errors in maintaining process integrity.

In contrast, Gemini struggled to maintain logical consistency between tasks, processes, and flows. Although it often identified message flows and pools more accurately than ChatGPT, it occasionally failed to specify the source and destination activities for these flows.

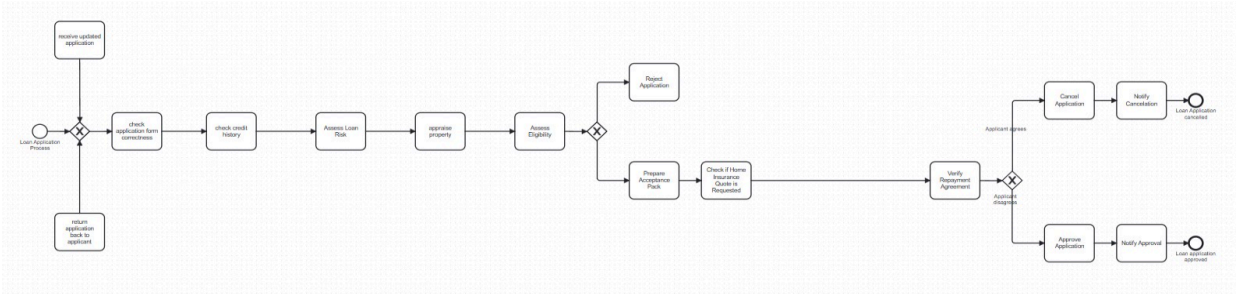

Figure 11 : Diagram generated from Gemini in BPMN XML Format

Furthermore, when prompted for clarification or additional details, Gemini sometimes lost continuity with its previous responses rather than seamlessly integrating new information. This issue further increased the time required to refine AI-generated diagrams and represented a significant limitation in Gemini's performance. An example of this can be seen in Figure 6.2, where a message collection event is present without a corresponding message generation event. With regard to adherence to BPMN 2.0 standards, both AI models followed the conventions to a certain extent. However, human intervention played a crucial role in ensuring compliance. In many instances, AI-generated solutions required additional user instructions to produce BPMN-compliant diagrams. The AI models occasionally overlooked key elements that are fundamental to proper BPMN modeling, requiring manual corrections to align the outputs with standard practices.

Some of the most frequently observed logical errors in AI-generated diagrams are summarized in the table below:

Error Type	ChatGPT	Gemini
Message Flow	Missing	Often Incomplete
Time Events	Missing	Rarely Included
Gateways	Incomplete in some cases	Often Incomplete
Continuity of Task in pool	Inconsistent	Inconsistent
Generation of other pools and lanes	Present but underutilised	No activities assigned

Table 8: Comparison of Model Errors

From an interpretability perspective, AI-generated diagrams were generally readable and understandable by individuals with a basic knowledge of BPMN modeling. However, due to frequent inaccuracies, they provided only a partial representation of the

underlying business processes. While these diagrams were not intended for use by a specific audience, their readability and logical flow remained sufficient for comparative analysis. Nonetheless, no dedicated evaluation was conducted to assess their semantic or pragmatic quality in real-world business process modeling contexts. While AI models demonstrate potential in BPMN diagram generation, their current capabilities remain limited. The results indicate that, while AI can assist in process discovery, human oversight is still required to refine outputs and ensure adherence to modeling standards. Future advancements in AI may improve accuracy and reliability, enabling greater integration of AI-generated BPMN diagrams into business process management.

6.2. Comparative Evaluation

The BPMN diagrams generated by artificial intelligence models were of higher quality than those produced by human actors. As outlined in the previous chapter, AI models are currently unable to generate more precise diagrams than humans in terms of overall quality.

Human-generated diagrams generally achieved better performance scores than AI-generated ones across nearly all evaluation criteria, including the identification of pools and lanes, the correct segmentation of process components, and the logical sequencing of events. However, an unexpected result emerged regarding syntactic accuracy. It was also revealed that ChatGPT produced diagrams with fewer syntactic errors than the human solver, with an error rate of only 30%, followed by the human at 32%, and Gemini at 38%. This suggests that, despite AI models generally underperforming in overall quality, they exhibited stronger syntactic accuracy in some cases. Another important factor to analyze is the consistency and variability between the BPMN diagrams generated by the three actors. For AI-generated diagrams, the same initial prompt was used to assess the degree of similarity between responses. Observations and data indicated that the diagrams generated by ChatGPT and Gemini demonstrated moderate variability. For the same problem statement, the BPMN diagrams produced by Gemini occasionally deviated significantly from those created by

ChatGPT. The evaluation metrics were based on how closely the AI-generated diagrams matched the expected solution. However, this approach had certain limitations, as Gemini sometimes generated diagrams that were close to the correct solution but did not fully conform to it. This raises the question of whether such outputs should be considered entirely incorrect or if they should be analyzed differently. Further research would be necessary to explore this aspect in greater detail.

The inconsistency in AI-generated BPMN diagrams can also be attributed to the evolving nature of artificial intelligence itself. This thesis was conducted over a period of approximately six months, and since AI models undergo continuous updates, their responses are not always consistent across different timeframes. Additionally, AI models inherently incorporate a degree of randomness in their outputs, which was evident in the variations observed. This characteristic of AI presents both advantages and disadvantages. On the positive side, AI's ability to introduce variability can foster creativity and innovation in BPMN modeling. Conversely, this inconsistency makes it difficult to achieve uniform results, which can be a drawback in business process modeling applications that require stability and accuracy. For human-generated BPMN diagrams, variability is influenced by experience and skill development. The more a person practices solving BPMN modeling exercises, the more proficient they become, leading to improved accuracy in subsequent diagrams. Additionally, human creativity plays a crucial role, as different individuals may approach the same process discovery task in unique ways, resulting in varying solutions and interpretations. This makes it difficult to establish a definitive distinction between AI- and human-generated BPMN diagrams solely based on this study.

From a statistical standpoint, the results do not provide sufficient evidence to suggest that AI models are currently capable of replacing humans in BPMN diagram generation. However, when viewed from a broader perspective, the findings indicate that AI-generated BPMN diagrams are evolving in quality. As AI models continue to improve, it will be interesting to explore their future role in software and business process modeling.

Chapter 7: Conclusions and Future Work

Artificial Intelligence is increasingly used in Business Process Modeling, particularly in process discovery. This study evaluates AI models like ChatGPT and Gemini in generating BPMN diagrams, assessing their accuracy and limitations. Findings show that AI-generated models do not perform equivalently to human-created ones, as statistical analysis rejected the null hypothesis. AI models struggle to generate complete and coherent BPMN diagrams, especially for complex problems, highlighting their limitations. Future research should focus on refining AI methodologies, improving logical consistency, and assessing semantic accuracy. AI offers potential benefits in automating workflows and optimizing processes across industries, though further advancements are needed to enhance reliability and effectiveness.

7.1. Conclusions

The results obtained provide answers to the initial hypothesis posed at the outset of this research while also serving as an empirical test, offering statistical validation of an ongoing scientific question. The statistical analyses conducted, including the t-test and ANOVA, refuted the null hypothesis that AI and human-generated models perform equivalently in process discovery for Business Process Modeling. The empirical findings, along with the p-value, which was lower than the conventional alpha threshold, indicate that there is no statistical evidence to confirm the initial hypothesis. This outcome was somewhat unexpected, as the initial assumption was that AI models would generally perform better than they did. The methodology used in this study, which may be open to interpretation, could also have influenced these results.

Moreover, this study provides insight into an emerging and relatively unexplored field, as there is currently a lack of substantial research offering empirical evidence on this topic. Additionally, the findings highlight a fundamental limitation of AI models—their inability to generate complete and coherent BPMN diagrams in XML format, particularly

when dealing with lengthy and complex problems. This suggests that further research is necessary to fully understand the capabilities of AI models in this domain. It is also crucial to acknowledge certain limitations that could affect the interpretation and generalizability of the results. These limitations may be either technical or methodological in nature. The technical constraints pertain to the AI models' ability to deliver the expected outcomes, while the methodological constraints relate to the study design, which may be subject to varying interpretations.

7.2. Future Work

In conclusion, this research on the role of AI in the process discovery of Business Process Modeling provides answers to key research questions while also revealing new areas that require further investigation. One possible improvement could involve refining the methodology used, such as training AI models further or exploring different AI models. Since this study primarily focused on evaluating and comparing syntactic errors, an intriguing avenue for future research would be to assess whether the generated process models are also semantically and pragmatically accurate. Addressing this question would require more extensive involvement of human experts with specialized domain knowledge. Additionally, inconsistencies were observed in the AI-generated logical workflows. The performance of artificial intelligence varied at times, highlighting an area that could be improved. Understanding how model architectures and algorithms function in AI-based process discovery could help mitigate the variability and inconsistencies observed in AI-generated responses when dealing with different scenarios.

The application of artificial intelligence in process discovery offers numerous opportunities for technological advancements. For instance, the technology industry stands to benefit significantly as AI-generated process models can streamline workflow development. Industries focused on optimizing business processes and logistics can also leverage AI to enhance operational efficiency. The academic sector could see advancements through AI-driven tools that generate various process models and

structured diagrams for summarizing textual data. These findings also hold significant practical implications. The IT industry, in particular, could gain substantial advantages from AI-generated process models. This technological advancement would simplify process modeling, minimize errors, reduce time consumption, and enable automation and comprehensive documentation of workflows.

Bibliography

Aguayo Publicidad. (2021). *BPMN for UX: Integration and Benefits in User-Centered Design | Aguayo's blog*. https://aguayo.co/en/blog-aguayo-user-experience/business-process-model-notation-for-ux/

Allweyer, T. (2016). *BPMN 2.0: Introduction to the Standard for Business Process Modeling* (2nd Edition). https://dl.acm.org/citation.cfm?id=1841147

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2019). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Bedwell, K., Garaccione, G., Coppola, R., Ardito, L., & Morisio, M. (2022). BIPMIN: A Gamified Framework for process Modeling education. *Information*, *14*(1), 3. https://doi.org/10.3390/info14010003

Blanchard, A., & Taddeo, M. (2023). The Ethics of Artificial Intelligence for Intelligence Analysis: a Review of the Key Challenges with Recommendations. *Deleted Journal*, *2*(1). https://doi.org/10.1007/s44206-023-00036-4

Chinosi, M., & Trombetta, A. (2011). BPMN: An introduction to the standard. *Computer Standards & Interfaces*, *34*(1), 124–134. https://doi.org/10.1016/j.csi.2011.06.002

Conneau, A., & Lample, G. (2019). Cross-lingual language model pretraining. *Neural Information Processing Systems*, *32*, 7057–7067. https://papers.nips.cc/paper/8928-cross-lingual-language-model-pretraining.pdf

Darko, A., Chan, A. P., Adabre, M. A., Edwards, D. J., Hosseini, M. R., & Ameyaw, E. E. (2020). Artificial intelligence in the AEC industry: Scientometric analysis and visualization of research activities. *Automation in Construction*, *112*, 103081. https://doi.org/10.1016/j.autcon.2020.103081

Davenport, T. H. (1993). Process innovation: reengineering work through information technology. *Choice Reviews Online*, *30*(08), 30–4486. https://doi.org/10.5860/choice.30-4486

D'Ippolito, B. (2014). The importance of design for firms' competitiveness: A review of the literature. *Technovation*, *34*(11), 716–730. https://doi.org/10.1016/j.technovation.2014.01.007

Douglas, M. R. (2023). Large language models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2307.05782

Dumas, M., La Rosa, M., Mendling, J., & Reijers, H. A. (2013). Fundamentals of Business Process Management. In *Springer eBooks*. https://doi.org/10.1007/978-3-642-33143-5

Dumas, M., La Rosa, M., Mendling, J., & Reijers, H. A. (2018). Fundamentals of Business Process Management. In *Springer eBooks*. https://doi.org/10.1007/978-3-662-56509-4

Fan, L., Li, L., Ma, Z., Lee, S., Yu, H., & Hemphill, L. (2023). A Bibliometric Review of Large Language Models Research from 2017 to 2023. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2304.02020

GeeksforGeeks. (2023, November 29). *Evolution of Software Development | History, phases and future trends*. GeeksforGeeks. https://www.geeksforgeeks.org/evolution-of-software-development-history-phases-and-future-trends/

Hammer, M., & Champy, J. (1993). Reengineering the corporation: A manifesto for business revolution. *Business Horizons*, *36*(5), 90–91. https://doi.org/10.1016/s0007-6813(05)80064-3

Hassani, H., & Silva, E. S. (2023). The role of ChatGPT in Data Science: How AI-Assisted conversational Interfaces are revolutionizing the field. *Big Data and Cognitive Computing*, *7*(2), 62. https://doi.org/10.3390/bdcc7020062

Hussain, S., Sianaki, O. A., & Ababneh, N. (2019). A survey on Conversational Agents/Chatbots Classification and Design techniques. In *Advances in intelligent systems and computing* (pp. 946–956). https://doi.org/10.1007/978-3-030-15035-8_93

Khanzode, K. C. A., & Sarode, R. D. (2023). Advantages and Disadvantages of Artificial Intelligence and Machine Learning: A literature review. *International Journal of Library & Information Science (IJLIS)*, 3.

Kourani, H., Berti, A., Schuster, D., & Van Der Aalst, W. M. P. (2024). Process Modeling with Large Language Models. In *Lecture notes in business information processing* (pp. 229–244). https://doi.org/10.1007/978-3-031-61007-3_18

Lopes, T., & Guerreiro, S. (2023). Assessing business process models: a literature review on techniques for BPMN testing and formal verification. *Business Process Management Journal*, *29*(8), 133–162. https://doi.org/10.1108/bpmj-11-2022-0557

McTear, M. (2020). Conversational AI: dialogue systems, conversational agents, and chatbots. *Synthesis Lectures on Human Language Technologies*, *13*(3), 1–251. https://doi.org/10.2200/s01060ed1v01y202010hlt048

Mendling, J., Weber, I., Van Der Aalst, W., Brocke, J. V., Cabanillas, C., Daniel, F., Debois, S., Di Ciccio, C., Dumas, M., Dustdar, S., Gal, A., García-Bañuelos, L., Governatori, G., Hull, R., La Rosa, M., Leopold, H., Leymann, F., Recker, J., Reichert, M., . . . Zhu, L. (2018). Blockchains for Business Process Management - Challenges and opportunities. *ACM Transactions on Management Information Systems*, *9*(1), 1–16. https://doi.org/10.1145/3183367

Mohammed, I. A. (2020). Critical analysis on the impact of software engineering in the technological industry. *INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS*. https://www.researchgate.net/publication/377159384_Critical_Analysis_on_the_Impact_Of_Software_Engineering_in_the_Technological_Industry

Mungoli, N. (2023). Exploring the synergy of prompt engineering and reinforcement learning for enhanced control and responsiveness in Chat GPT. *Journal of Electrical Electronics Engineering*, *2*(3). https://doi.org/10.33140/jeee.02.03.02

Neuberger, J., Ackermann, L., Han, V. D. A., & Jablonski, S. (2024). A Universal Prompting Strategy for Extracting Process Model Information from Natural Language Text using Large Language Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2407.18540

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021). Zero-Shot Text-to-Image Generation. *International Conference on Machine Learning*, 8821–8831. http://proceedings.mlr.press/v139/ramesh21a/ramesh21a.pdf

Safari, P., India, M., & Hernando, J. (2020). Self-Attention encoding and pooling for speaker recognition. *Interspeech 2022*. https://doi.org/10.21437/interspeech.2020-1446

Safieddine, F., & Nakhoul, I. (2018). Generic Business process model for SMEs in M-Commerce based on Talabat's case study. In *Lecture notes in computer science* (pp. 264–278). https://doi.org/10.1007/978-3-030-02131-3_24

Schmager, S., Pappas, I. O., & Vassilakopoulou, P. (2025). Understanding Human-Centred AI: a review of its defining elements and a research agenda. *Behaviour and Information Technology*, 1–40. https://doi.org/10.1080/0144929x.2024.2448719

Segatto, M., De Pádua, S. I. D., & Martinelli, D. P. (2013). Business process management: a systemic approach? *Business Process Management Journal*, *19*(4), 698–714. https://doi.org/10.1108/bpmj-jun-2012-0064

Shawar, B. A., & Atwell, E. (2007). Chatbots: Are they Really Useful? *Deleted Journal*, *22*(1), 29–49. https://doi.org/10.21248/jlcl.22.2007.88

Silver, B. (2011). *BPMN Method and style : with BPMN implementer's guide.* https://lib.ugent.be/en/catalog/rug01:002037299

Song, X., & Xiong, T. (2021). A survey of published literature on conversational artificial intelligence. *2021 7th International Conference on Information Management (ICIM)*, 113–117. https://doi.org/10.1109/icim52229.2021.9417135

Stahl, B. C., Schroeder, D., & Rodrigues, R. (2022). The Ethics of Artificial Intelligence: A Conclusion. In *SpringerBriefs in research and innovation governance* (pp. 107–111). https://doi.org/10.1007/978-3-031-17040-9_9

Stravinskiene, I., & Serafinas, D. (2020). The Link between Business Process Management and Quality Management. *Journal of Risk and Financial Management*, *13*(10), 225. https://doi.org/10.3390/jrfm13100225

Uszkoreit, J. (2017). *Transformer: A novel neural network architecture for language Understanding*. https://research.google/blog/transformer-a-novel-neural-network-architecture-for-language-understanding/

Van Der Aalst, W. (2016). *Process Mining: Data Science in action*. https://vipa.wiwi.uni-saarland.de/wordpress/wp-content/uploads/2014/11/14WS_BWinfo_Schlueko_ProcessMining.pdf

Van Der Aalst, W. M. (2014). Process mining in the Large: a tutorial. In *Lecture notes in business information processing* (pp. 33–76). https://doi.org/10.1007/978-3-319-05461-2_2

Van Der Aalst, W. M. P. (2013). Business Process Management: A Comprehensive Survey. *ISRN Software Engineering*, *2013*, 1–37. https://doi.org/10.1155/2013/507984

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is All you Need. *arXiv (Cornell University)*, *30*, 5998–6008. https://arxiv.org/pdf/1706.03762v5

Velásquez-Henao, J. D., Franco-Cardona, C. J., & Cadavid-Higuita, L. (2023). Prompt Engineering: a methodology for optimizing interactions with AI-Language Models in the field of engineering. *DYNA*, *90*(230), 9–17. https://doi.org/10.15446/dyna.v90n230.111700

Vergidis, K., Tiwari, A., & Majeed, B. (2007). Business Process analysis and optimization: Beyond reengineering. *IEEE Transactions on Systems Man and Cybernetics Part C (Applications and Reviews)*, *38*(1), 69–82. https://doi.org/10.1109/tsmcc.2007.905812

Von Rosing, M., White, S., Cummins, F., & De Man, H. (2014). Business Process Model and Notation—BPMN. In *Elsevier eBooks* (pp. 433–457). https://doi.org/10.1016/b978-0-12-799959-3.00021-5

Weske, M. (2007). *Business Process Management: Concepts, languages, architectures*. http://dx.doi.org/10.1007/978-3-540-73522-9

White, S. A. (2004). Business Process Modeling Notation (BPMN), Version 1.0. *Journal for Business Process Management International Conference*. http://bpms.ru/fileadmin/pdf/bpmn-1.0.pdf

Wohed, P., Van Der Aalst, W. M. P., Dumas, M., Ter Hofstede, A. H. M., & Russell, N. (2006). On the Suitability of BPMN for Business Process Modelling. In *Lecture notes in computer science* (pp. 161–176). https://doi.org/10.1007/11841760_12

Wong, P. Y. H., & Gibbons, J. (2008). A process semantics for BPMN. In *Lecture notes in computer science* (pp. 355–374). https://doi.org/10.1007/978-3-540-88194-0_22

```
 GET
  FILE='C:\Users\anees\Downloads\First Data Set.sav'.

>Warning # 67.  Command name: GET FILE
>The document is already in use by another user or process.  If you make
>changes to the document they may overwrite changes made by others or your
>changes may be overwritten by others.
>File opened C:\Users\anees\Downloads\First Data Set.sav
DATASET NAME DataSet1 WINDOW=FRONT.
T-TEST PAIRS=ChatGPT WITH Gemini (PAIRED)
  /ES DISPLAY(TRUE) STANDARDIZER(SD)
  /CRITERIA=CI(.9500)
  /MISSING=ANALYSIS.
```

T-Test

Notes

Output Created		19-MAR-2025 22:14:45
Comments		
Input	Data	C:\Users\anees\Downloads\First Data Set.sav
	Active Dataset	DataSet1
	Filter	<none>
	Weight	<none>
	Split File	<none>
	N of Rows in Working Data File	20
Missing Value Handling	Definition of Missing	User defined missing values are treated as missing.
	Cases Used	Statistics for each analysis are based on the cases with no missing or out-of-range data for any variable in the analysis.
Syntax		T-TEST PAIRS=ChatGPT WITH Gemini (PAIRED) /ES DISPLAY(TRUE) STANDARDIZER(SD) /CRITERIA=CI(.9500) /MISSING=ANALYSIS.
Resources	Processor Time	00:00:00,02
	Elapsed Time	00:00:00,01

```
[DataSet1]
```

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	ChatGPT Performance Score	.525	20	.3432	.0767
	Gemini Performance Score	.325	20	.3726	.0833

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	ChatGPT Performance Score & Gemini Performance Score	20	.653	.002

Paired Samples Test

		Paired Differences			
					95% Confidence ...
		Mean	Std. Deviation	Std. Error Mean	Lower
Pair 1	ChatGPT Performance Score - Gemini Performance Score	.2000	.2991	.0669	.0600

Paired Samples Test

		Paired ...			
		95% Confidence Interval of the ...			
		Upper	t	df	Sig. (2-tailed)
Pair 1	ChatGPT Performance Score - Gemini Performance Score	.3400	2.990	19	.008

Paired Samples Effect Sizes

			Standardizer[a]	Point Estimate	95% ...
					Lower
Pair 1	ChatGPT Performance Score - Gemini Performance Score	Cohen's d	.2991	.669	.175
		Hedges' correction	.3052	.655	.172

Paired Samples Effect Sizes

			95% ...
			Upper
Pair 1	ChatGPT Performance Score - Gemini Performance Score	Cohen's d	1.148
		Hedges' correction	1.125

a. The denominator used in estimating the effect sizes.
Cohen's d uses the sample standard deviation of the mean difference.
Hedges' correction uses the sample standard deviation of the mean difference, plus a correction ...

Notes

Output Created		19-MAR-2025 22:16:18
Comments		
Input	Active Dataset	DataSet0
	Filter	<none>
	Weight	<none>
	Split File	<none>
	N of Rows in Working Data File	60
Missing Value Handling	Definition of Missing	User defined missing values are treated as missing.
	Cases Used	All non-missing data are used.
Syntax		DESCRIPTIVES VARIABLES=trans1 /STATISTICS=MEAN STDDEV VARIANCE MIN MAX.
Resources	Processor Time	00:00:00,00
	Elapsed Time	00:00:00,00

Descriptives

Notes

Output Created		19-MAR-2025 22:16:57
Comments		
Input	Data	C:\Users\anees\Downloads\First Data Set.sav
	Active Dataset	DataSet1
	Filter	<none>
	Weight	<none>
	Split File	<none>
	N of Rows in Working Data File	20
Missing Value Handling	Definition of Missing	User defined missing values are treated as missing.
	Cases Used	All non-missing data are used.
Syntax		DESCRIPTIVES VARIABLES=Human ChatGPT Gemini /STATISTICS=MEAN STDDEV VARIANCE MIN MAX.
Resources	Processor Time	00:00:00,02
	Elapsed Time	00:00:00,01

```
[DataSet1]
```

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Variance
Human Performance Score	20	1.0	1.0	1.000	.0000	.000
ChatGPT Performance Score	20	.0	1.0	.525	.3432	.118
Gemini Performance Score	20	.0	1.0	.325	.3726	.139
Valid N (listwise)	20					

```
CORRELATIONS
 /VARIABLES=Human ChatGPT Gemini
 /PRINT=TWOTAIL NOSIG FULL
 /MISSING=PAIRWISE.
```

Correlations

Notes

Output Created		19-MAR-2025 22:19:01
Comments		
Input	Data	C:\Users\anees\Downloads\First Data Set.sav
	Active Dataset	DataSet1
	Filter	<none>
	Weight	<none>
	Split File	<none>
	N of Rows in Working Data File	20
Missing Value Handling	Definition of Missing	User-defined missing values are treated as missing.
	Cases Used	Statistics for each pair of variables are based on all the cases with valid data for that pair.
Syntax		CORRELATIONS /VARIABLES=Human ChatGPT Gemini /PRINT=TWOTAIL NOSIG FULL /MISSING=PAIRWISE.
Resources	Processor Time	00:00:00,02
	Elapsed Time	00:00:00,02

Correlations

		Human Performance Score	ChatGPT Performance Score	Gemini Performance Score
Human Performance Score	Pearson Correlation	.[a]	.[a]	.[a]
	Sig. (2-tailed)		.	.
	N	20	20	20
ChatGPT Performance Score	Pearson Correlation	.[a]	1	.653**
	Sig. (2-tailed)	.		.002
	N	20	20	20
Gemini Performance Score	Pearson Correlation	.[a]	.653**	1
	Sig. (2-tailed)	.	.002	
	N	20	20	20

**. Correlation is significant at the 0.01 level (2-tailed).

a. Cannot be computed because at least one of the variables is constant.

```
ONEWAY trans1 BY Index1
 /MISSING ANALYSIS
 /CRITERIA=CILEVEL(0.95)
 /POSTHOC=TUKEY ALPHA(0.05).
```

Oneway

Notes

Output Created		19-MAR-2025 21:59:52
Comments		
Input	Active Dataset	DataSet0
	Filter	<none>
	Weight	<none>
	Split File	<none>
	N of Rows in Working Data File	60
Missing Value Handling	Definition of Missing	User-defined missing values are treated as missing.
	Cases Used	Statistics for each analysis are based on cases with no missing data for any variable in the analysis.
Syntax		ONEWAY trans1 BY Index1 /MISSING ANALYSIS /CRITERIA=CILEVEL (0.95) /POSTHOC=TUKEY ALPHA(0.05).
Resources	Processor Time	00:00:00,02
	Elapsed Time	00:00:00,01

ANOVA

Human Performance Score

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	4.808	2	2.404	28.110	.000
Within Groups	4.875	57	.086		
Total	9.683	59			

Post Hoc Tests

Multiple Comparisons

Dependent Variable: Human Performance Score

Tukey HSD

(I) Labels	(J) Labels	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Human	ChatGPT	.4750*	.0925	.000	.252	.698
	Gemini	.6750*	.0925	.000	.452	.898
ChatGPT	Human	-.4750*	.0925	.000	-.698	-.252
	Gemini	.2000	.0925	.087	-.023	.423
Gemini	Human	-.6750*	.0925	.000	-.898	-.452
	ChatGPT	-.2000	.0925	.087	-.423	.023

*. The mean difference is significant at the 0.05 level.

Homogeneous Subsets

Human Performance Score

Tukey HSD[a]

Labels	N	Subset for alpha = 0.05	
		1	2
Gemini	20	.325	
ChatGPT	20	.525	
Human	20		1.000
Sig.		.087	1.000

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 20.000.

```
BOOTSTRAP
 /SAMPLING METHOD=SIMPLE
 /VARIABLES TARGET=trans1 INPUT=Index1
 /CRITERIA CILEVEL=95 CITYPE=PERCENTILE  NSAMPLES=1000
 /MISSING USERMISSING=EXCLUDE.
```

Bootstrap

Notes

Output Created		19-MAR-2025 22:01:24
Comments		
Input	Active Dataset	DataSet0
	Filter	<none>
	Weight	<none>
	Split File	<none>
Syntax		BOOTSTRAP /SAMPLING METHOD=SIMPLE /VARIABLES TARGET=trans1 INPUT=Index1 /CRITERIA CILEVEL=95 CITYPE=PERCENTILE NSAMPLES=1000 /MISSING USERMISSING=EXCLUDE.
Resources	Processor Time	00:00:00,02
	Elapsed Time	00:00:00,02

Bootstrap Specifications

Sampling Method	Simple
Number of Samples	1000
Confidence Interval Level	95.0%
Confidence Interval Type	Percentile

```
T-TEST GROUPS=Index1(1 2)
 /MISSING=ANALYSIS
 /VARIABLES=trans1
 /ES DISPLAY(TRUE)
 /CRITERIA=CI(.95).
```

T-Test

Notes

Output Created		19-MAR-2025 22:01:24
Comments		
Input	Active Dataset	DataSet0
	Filter	<none>
	Weight	<none>
	Split File	<none>
	N of Rows in Working Data File	38175
Missing Value Handling	Definition of Missing	User defined missing values are treated as missing.
	Cases Used	Statistics for each analysis are based on the cases with no missing or out-of-range data for any variable in the analysis.
Syntax		T-TEST GROUPS=Index1 (1 2) /MISSING=ANALYSIS /VARIABLES=trans1 /ES DISPLAY(TRUE) /CRITERIA=CI(.95).
Resources	Processor Time	00:00:03,53
	Elapsed Time	00:00:10,64

Group Statistics

				Bootstrap[a]	
	Labels		Statistic	Bias	Std. Error
Human Performance Score	Human	N	20		
		Mean	1.000	.000	.000
		Std. Deviation	.0000	.0000	.0000
		Std. Error Mean	.0000		
	ChatGPT	N	20		
		Mean	.525	-.001	.074
		Std. Deviation	.3432	-.0127	.0451
		Std. Error Mean	.0767		

Group Statistics

	Labels		Bootstrap[a] 95% Confidence Interval Lower	Upper
Human Performance Score	Human	N		
		Mean	1.000	1.000
		Std. Deviation	.0000	.0000
		Std. Error Mean		
	ChatGPT	N		
		Mean	.375	.667
		Std. Deviation	.2388	.4162
		Std. Error Mean		

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

Independent Samples Test

		Levene's Test for Equality of Variances F	Sig.	t-test for Equality of . t
Human Performance Score	Equal variances assumed	19.321	.000	6.190
	Equal variances not assumed			6.190

Independent Samples Test

		t-test for Equality of Means df	Sig. (2-tailed)	Mean Difference
Human Performance Score	Equal variances assumed	38	.000	.4750
	Equal variances not assumed	19.000	.000	.4750

Independent Samples Test

		t-test for Equality of Means	
		Std. Error Difference	95% Confidence ... Lower
Human Performance Score	Equal variances assumed	.0767	.3197
	Equal variances not assumed	.0767	.3144

Independent Samples Test

		t-test for Equality of Means	
		95% Confidence Interval of the ...	
		Upper	
Human Performance Score	Equal variances assumed	.6303	
	Equal variances not assumed	.6356	

Bootstrap for Independent Samples Test

			Bootstrap[a]	
		Mean Difference	Bias	Std. Error
Human Performance Score	Equal variances assumed	.4750	.0013	.0743
	Equal variances not assumed	.4750	.0013	.0743

Bootstrap for Independent Samples Test

			Bootstrap[a]	
			95% Confidence Interval	
		Sig. (2-tailed)	Lower	Upper
Human Performance Score	Equal variances assumed	.001	.3333	.6250
	Equal variances not assumed	.001	.3333	.6250

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

Independent Samples Effect Sizes

		Standardizer[a]	Point Estimate	95% ... Lower
Human Performance Score	Cohen's d	.2427	1.958	1.190
	Hedges' correction	.2476	1.919	1.166
	Glass's delta	.3432	1.384	.614

Independent Samples Effect Sizes

		95% ... Upper
Human Performance Score	Cohen's d	2.708
	Hedges' correction	2.654
	Glass's delta	2.130

a. The denominator used in estimating the effect sizes.
Cohen's d uses the pooled standard deviation.
Hedges' correction uses the pooled standard deviation, plus a correction factor.
Glass's delta uses the sample standard deviation of the control group.

```
BOOTSTRAP
 /SAMPLING METHOD=SIMPLE
 /VARIABLES TARGET=trans1 INPUT=Index1
 /CRITERIA CILEVEL=95 CITYPE=PERCENTILE  NSAMPLES=1000
 /MISSING USERMISSING=EXCLUDE.
```

Bootstrap

Notes

Output Created		19-MAR-2025 22:02:44
Comments		
Input	Active Dataset	DataSet0
	Filter	<none>
	Weight	<none>
	Split File	<none>
Syntax		BOOTSTRAP /SAMPLING METHOD=SIMPLE /VARIABLES TARGET=trans1 INPUT=Index1 /CRITERIA CILEVEL=95 CITYPE=PERCENTILE NSAMPLES=1000 /MISSING USERMISSING=EXCLUDE.
Resources	Processor Time	00:00:00,03
	Elapsed Time	00:00:00,03

Bootstrap Specifications

Sampling Method	Simple
Number of Samples	1000
Confidence Interval Level	95.0%
Confidence Interval Type	Percentile

```
T-TEST GROUPS=Index1(1 3)
 /MISSING=ANALYSIS
 /VARIABLES=trans1
 /ES DISPLAY(TRUE)
 /CRITERIA=CI(.95).
```

T-Test

Notes

Output Created		19-MAR-2025 22:02:44
Comments		
Input	Active Dataset	DataSet0
	Filter	<none>
	Weight	<none>
	Split File	<none>
	N of Rows in Working Data File	38255
Missing Value Handling	Definition of Missing	User defined missing values are treated as missing.
	Cases Used	Statistics for each analysis are based on the cases with no missing or out-of-range data for any variable in the analysis.
Syntax		T-TEST GROUPS=Index1 (1 3) /MISSING=ANALYSIS /VARIABLES=trans1 /ES DISPLAY(TRUE) /CRITERIA=CI(.95).
Resources	Processor Time	00:00:03,25
	Elapsed Time	00:00:10,88

Group Statistics

				Bootstrap[a]	
	Labels		Statistic	Bias	Std. Error
Human Performance Score	Human	N	20		
		Mean	1.000	.000	.000
		Std. Deviation	.0000	.0000	.0000
		Std. Error Mean	.0000		
	Gemini	N	20		
		Mean	.325	-.004	.081
		Std. Deviation	.3726	-.0144	.0451
		Std. Error Mean	.0833		

Group Statistics

	Labels		Bootstrap[a] 95% Confidence Interval Lower	Upper
Human Performance Score	Human	N		
		Mean	1.000	1.000
		Std. Deviation	.0000	.0000
		Std. Error Mean		
	Gemini	N		
		Mean	.167	.475
		Std. Deviation	.2557	.4375
		Std. Error Mean		

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

Independent Samples Test

		Levene's Test for Equality of Variances F	Sig.	t-test for Equality of . t
Human Performance Score	Equal variances assumed	76.452	.000	8.102
	Equal variances not assumed			8.102

Independent Samples Test

		t-test for Equality of Means df	Sig. (2-tailed)	Mean Difference
Human Performance Score	Equal variances assumed	38	.000	.6750
	Equal variances not assumed	19.000	.000	.6750

Independent Samples Test

		t-test for Equality of Means	
		Std. Error Difference	95% Confidence ... Lower
Human Performance Score	Equal variances assumed	.0833	.5063
	Equal variances not assumed	.0833	.5006

Independent Samples Test

		t-test for Equality of Means	
		95% Confidence Interval of the ...	
		Upper	
Human Performance Score	Equal variances assumed	.8437	
	Equal variances not assumed	.8494	

Bootstrap for Independent Samples Test

			Bootstrap[a]	
		Mean Difference	Bias	Std. Error
Human Performance Score	Equal variances assumed	.6750	.0041	.0805
	Equal variances not assumed	.6750	.0041	.0805

Bootstrap for Independent Samples Test

			Bootstrap[a]	
			95% Confidence Interval	
		Sig. (2-tailed)	Lower	Upper
Human Performance Score	Equal variances assumed	.001	.5250	.8333
	Equal variances not assumed	.001	.5250	.8333

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

Independent Samples Effect Sizes

		Standardizer[a]	Point Estimate	95% ... Lower
Human Performance Score	Cohen's d	.2635	2.562	1.708
	Hedges' correction	.2688	2.511	1.674
	Glass's delta	.3726	1.812	.955

Independent Samples Effect Sizes

		95% ... Upper
Human Performance Score	Cohen's d	3.398
	Hedges' correction	3.330
	Glass's delta	2.642

a. The denominator used in estimating the effect sizes.
Cohen's d uses the pooled standard deviation.
Hedges' correction uses the pooled standard deviation, plus a correction factor.
Glass's delta uses the sample standard deviation of the control group.

```
BOOTSTRAP
 /SAMPLING METHOD=SIMPLE
 /VARIABLES TARGET=trans1 INPUT=Index1
 /CRITERIA CILEVEL=95 CITYPE=PERCENTILE  NSAMPLES=1000
 /MISSING USERMISSING=EXCLUDE.
```

Bootstrap

Notes

Output Created		19-MAR-2025 22:06:42
Comments		
Input	Active Dataset	DataSet0
	Filter	<none>
	Weight	<none>
	Split File	<none>
Syntax		BOOTSTRAP /SAMPLING METHOD=SIMPLE /VARIABLES TARGET=trans1 INPUT=Index1 /CRITERIA CILEVEL=95 CITYPE=PERCENTILE NSAMPLES=1000 /MISSING USERMISSING=EXCLUDE.
Resources	Processor Time	00:00:00,03
	Elapsed Time	00:00:00,03

Bootstrap Specifications

Sampling Method	Simple
Number of Samples	1000
Confidence Interval Level	95.0%
Confidence Interval Type	Percentile

```
T-TEST GROUPS=Index1(2 3)
 /MISSING=ANALYSIS
 /VARIABLES=trans1
 /ES DISPLAY(TRUE)
 /CRITERIA=CI(.95).
```

T-Test

Notes

Output Created		19-MAR-2025 22:06:42
Comments		
Input	Active Dataset	DataSet0
	Filter	<none>
	Weight	<none>
	Split File	<none>
	N of Rows in Working Data File	38078
Missing Value Handling	Definition of Missing	User defined missing values are treated as missing.
	Cases Used	Statistics for each analysis are based on the cases with no missing or out-of-range data for any variable in the analysis.
Syntax		T-TEST GROUPS=Index1(2 3) /MISSING=ANALYSIS /VARIABLES=trans1 /ES DISPLAY(TRUE) /CRITERIA=CI(.95).
Resources	Processor Time	00:00:04,13
	Elapsed Time	00:00:11,02

Group Statistics

				Bootstrap[a]	
	Labels		Statistic	Bias	Std. Error
Human Performance Score	ChatGPT	N	20		
		Mean	.525	.001	.079
		Std. Deviation	.3432	-.0116	.0459
		Std. Error Mean	.0767		
	Gemini	N	20		
		Mean	.325	.000	.084
		Std. Deviation	.3726	-.0120	.0460
		Std. Error Mean	.0833		

Group Statistics

	Labels		Bootstrap[a] 95% Confidence Interval Lower	Upper
Human Performance Score	ChatGPT	N		
		Mean	.368	.676
		Std. Deviation	.2352	.4143
		Std. Error Mean		
	Gemini	N		
		Mean	.174	.500
		Std. Deviation	.2545	.4412
		Std. Error Mean		

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

Independent Samples Test

		Levene's Test for Equality of Variances F	Sig.	t-test for Equality of . t
Human Performance Score	Equal variances assumed	1.780	.190	1.766
	Equal variances not assumed			1.766

Independent Samples Test

		t-test for Equality of Means df	Sig. (2-tailed)	Mean Difference
Human Performance Score	Equal variances assumed	38	.085	.2000
	Equal variances not assumed	37.746	.086	.2000

Independent Samples Test

		t-test for Equality of Means	
		Std. Error Difference	95% Confidence ... Lower
Human Performance Score	Equal variances assumed	.1133	-.0293
	Equal variances not assumed	.1133	-.0293

Independent Samples Test

		t-test for Equality of Means	
		95% Confidence Interval of the ...	
		Upper	
Human Performance Score	Equal variances assumed	.4293	
	Equal variances not assumed	.4293	

Bootstrap for Independent Samples Test

			Bootstrap[a]	
		Mean Difference	Bias	Std. Error
Human Performance Score	Equal variances assumed	.2000	.0004	.1137
	Equal variances not assumed	.2000	.0004	.1137

Bootstrap for Independent Samples Test

		Bootstrap[a]	
		95% Confidence Interval	
		Lower	Upper
Human Performance Score	Equal variances assumed	-.0354	.4197
	Equal variances not assumed	-.0354	.4197

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

Independent Samples Effect Sizes

		Standardizer[a]	Point Estimate	95% ... Lower
Human Performance Score	Cohen's d	.3582	.558	-.077
	Hedges' correction	.3654	.547	-.076
	Glass's delta	.3726	.537	-.112

Independent Samples Effect Sizes

		95% ... Upper
Human Performance Score	Cohen's d	1.187
	Hedges' correction	1.164
	Glass's delta	1.173

a. The denominator used in estimating the effect sizes.
Cohen's d uses the pooled standard deviation.
Hedges' correction uses the pooled standard deviation, plus a correction factor.
Glass's delta uses the sample standard deviation of the control group.