Leiden University

ICT in Business and the Public Sector

Designing A Framework For Evaluating An AI Agent From Business Managers Perspective

Name:	Anshul Mendhekar
Student ID:	s4042239
Date:	23/03/2026
1st supervisor:	Joost Visser
2nd supervisor:	Natalia Amat Lefort
Company supervisor:	Erik Weller (FrieslandCampina)

MASTER'S THESIS

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Einsteinweg 55
2333 CC Leiden
The Netherlands

Abstract

Background: The rapid advancement of low-code development platforms and generative artificial intelligence has enabled the swift development and deployment of AI agents across various industries. Organizations are poised to rely on these agents to lower operational expenses, enhance service quality, and increase service availability for the customers. Companies are undergoing digital transformation, leading to growing interest in AI agents to support customers and internal operations. However, business managers seek to evaluate AI agents beyond the technical performance metrics and need a comprehensive framework that enables assessment from a managerial perspective. By establishing a framework, organizations can gain structured insights into the value and risks of AI agents across financial, user experience, learning and growth, and business process perspectives.

Aim: The aim of this research is to develop and validate a comprehensive evaluation framework for business managers. This framework allows them to structurally assess AI agents getting deployed in customer operations. It supports early decision-making by evaluating AI agents across operational, experiential, technical, and organizational dimensions before scaling and production.

Method: The study adopts a design science research methodology. The framework was created based on an extensive review of existing literature on AI agent evaluation frameworks, the evolution of AI in business applications, human-AI interaction, and AI agents in customer operations, including associated AI-related risks. In addition to the literature review, we conducted semi-structured interviews. We studied the project template and held workshops to understand business processes. These activities helped us identify the dimensions important for evaluating an AI agent from a managerial lens.

Results: A final scorecard, mapping key dimensions to the balanced scorecard's perspectives, enabled business managers to evaluate an AI agent and make clear go/no-go decisions for the project.

The developed artifact supported systematic assessment of an AI agent across four perspectives: financial, user, business process, and learning and growth. This was intended for business managers. The design and demonstration of the framework involved 11 participants.

The survey results showed that an AI agent added business value, proved cost-efficient, and could be scaled financially. From the user perspective, managers indicated they trusted AI when accuracy was high, which led to user adoption and a preference for automated work. From a business-process perspective, an AI agent demonstrated faster document handling and reduced manual work. Lastly, the learning and growth indicator showed that the agent is viewed as a novel approach to help employees take on higher-value tasks and requires limited training when working with AI-generated outputs.

Conclusion: This study addresses the current difficulties managers face when evaluating an AI agent beyond technical metrics. Using the balanced scorecard framework, business managers identified areas for improvement and could offload repetitive tasks to an AI agent. This inspired them to use AI in current business processes, saving time and costs while increasing productivity. Ultimately, this work advances AI agent evaluation by tailoring assessment methods to the specific context. This better aligns with the needs and perspectives of business managers.

Acknowledgements

The thesis was written under the supervision of, Joost Visser, who I would like to thank for the inspiration and guidance to start the research, and his intensive support throughout the project. The numerous brainstorm and constructive feedback sessions have helped me tremendously to arrive at this point. Your guidance has undoubtedly helped me grow throughout this period, and you consistently challenged me to think in terms of opportunities and possibilities rather than limitations.

I would also like to thank my second supervisor, Natalia Amat Lefort. Your research paper inspired me to begin my work on AI agents, and your guidance helped me tie together the loose threads of my ideas into a coherent whole. I believe that your guidance has enabled me to elevate my thesis to a higher level.

A special thanks to FrieslandCampina for hosting my thesis internship. I am especially grateful to my supervisors there, Erik Weller. Thank you for your time, support, and for sharing your expertise with me. I also want to extend my thanks to Customer Operations Manager Jolien Kroon and her team Linda Blaak, Fleur van Breda, Paul Rensen van de Berg and IT Manager Jasper Leenstra for providing the necessary support both in terms of giving access to tools and sharing the AI agent use cases.

Lastly, I am deeply grateful to my family for their unwavering encouragement and support throughout this research. I could not have achieved this without you guys.

Contents

Chapter 1

Introduction

The rapid rise of technology like Artificial Intelligence (AI) is fundamentally redefining the way we work (changing nature of work) [23]. Customer and employee experience with the digital technologies and service ecosystems are shifting towards autonomous ways of working [23, 45]. For example, service robots are being used to greet customers instead of employees [23]. Currently, artificial intelligence capabilities are evolving at a fast pace such that AI has already mastered different service intelligence layer such as mechanical and analytical but now moving towards intuitive and empathetic intelligences [23].

With increasing adoption of these technologies business managers must embrace this transformative shift by integrating machine learning, intelligent digital technologies, and natural language processing into their service departments to serve the customer better [63, 45]. However, business managers must understand how these technologies can enhance customer experiences, in what scenarios we need to use advanced tools vs when a human touch is required [46].

Authors Waelbers et al. [63] mentioned previously the focus was on replacing human labor with the help of an automation system. Another study by Bitner et al. [7] talks about self service technologies that focused solely on reducing costs and increased convenience for customers. Now, the focus has now shifted from replacing humans to co-create value with technology augmentation, use human empathy and judgment as strengths that would complement the technologies [63].

When technologies like AI, augment human capabilities employees take on four suggested roles such as innovators, differentiators, enablers and coordinators [45]. However, when intelligent technologies are used to standardize the processes and employees are made to follow those rules to be efficient, this creates a challenging situation where in business managers must maintain the balance between autonomy, how much technology automation is needed and give employee the flexibility to make decisions, thereby reducing strain on them [45, 63].

Evaluating these intelligent agents requires to move beyond traditional machine learning parameters such as F1 score, accuracy, recall and precision [59], that help with model centric analysis of the system or through traditional service quality measurement tools like the SERVQUAL framework that measures service quality based on expected and perceived service [48] but we need to have an evaluation dimensions which is relevant to business managers based on the problem we are trying to solve, so context based criteria in explainable AI is very relevant [33].

From an academic perspective, the evaluation process of AI has always been focused on technical performance indicators relevant for conducting experiments in controlled settings [59]. The results used are highly relevant for a data scientist or people with technical background for evaluating AI capabilities and algorithms [59].

However, business managers are more concerned about whether AI agent should be deployed, scaled or restricted for a particular business case that are characterized by financial risks, environmental risks and potential risk of misuse or failure when deployed within the organization [6].

The paper by Shukla [57] highlights standard benchmarks such as MMLU, HELM that are available to assess the AI agent but it comes with limitations such as it measures only static question and answers, and fail to measure multi-agent setup. Additionally, Shukla [57] systematically reviewed 84 studies between

2023 and 2025 which shows, about 83% of the benchmarks primarily address technical performance, while the remaining 30% concentrate on human-centered aspects or economic impacts. The paper by [57] identified an important research gap that mentioned AI systems scores well on the benchmarks as they are done in controlled settings but when real world scenarios are tested they might fail due to inconsideration of sociotechnical factors such as poor workflow integrations, lack of human trust. To address the gap the author [57] proposed a multidimensional framework with parameters such as task completion rates (capability and efficiency), robustness and adaptability, safety and ethics, human centered interactions and economic sustainability to evaluate AI agents in a systematic and structured way.

Similarly, we also draw motivation from [57] to move away from technical benchmarks to asses the AI agent and shift towards holistic perspective that would incorporate the different needs of the stakeholder [33]. For example, the study by Langer et al. [33], supports this by mentioning different stakeholders have different kind of expectation from explainable AI. Another example, business stakeholders evaluates AI on reliability, trust and acceptance based on the problem we are trying to solve [33]. These insights are missing when evaluating an AI agent from a business manager perspective.

The objective of this thesis is to provide business managers with a multi dimensional framework to evaluate an AI agent in its inception stage, this framework helps business manager make clear go/no-go decision for further devolvement and deployment. The proposed framework incorporates some technical performance indicators, socio-technical parameters and organizational needs.

1.1 Problem statement

The usage of generative artificial intelligence in the form of low-code tools, such as Microsoft Copilot, has helped the employees to be more efficient and improve their performance [50]. The use of such technologies could lead to potential economic impact for the organizations in terms of cost savings and productivity [50].

According, to the study conducted by Brynjolfsson et al. [10] identified that less experienced employees showed increase in productivity when support staff worked together with AI services, this data came from an AI based customer service support software company.

Similarly, study conducted by Khneyzer et al. [29] mentioned when AI chatbots are deployed to answer standard queries and provide 24x7 customer support, it reduced operational cost and the waiting time.

On the contrary, another research by Ferraro et al. [16] stated that deployment of such technologies introduced risks such as erroneous decisions and data privacy issues that could be a concern for the customers.

Situation at FrieslandCampina Within FrieslandCampina managers are developing AI agent use cases to automate repetitive tasks, time intensive tasks such as extracting, summarizing and gathering comments from vendor PDFs.

However, the business managers are currently struggling to evaluate such use cases that make use of agentic AI, as majority of the time technical parameters such as accuracy, precision, recall and confidence score along with some financial measures such as (ROI) return on investment, cost savings, development cost and maintenance cost. These parameters form the primary evaluation technique for an AI agent.

Business managers want a structured way of evaluating an AI agent project that considers holistic view of themes such as end user community, integration with business process and organizational readiness to be included with financial measures.

General Research Problem Accordingly, the research problem focuses on the absence of a comprehensive framework that allows business managers to assess AI agents early in their life cycle, so they can decide whether to proceed with the project.

The Problem Statement

Problem Statement: Business managers need a structured, comprehensive framework to evaluate an AI agent in its inception stage and drive a clear go/no-go decision for further development and deployment of the AI project.

Figure 1.1: Overview of the Problem Statement

1.2 Research questions

Our problem statement helped us to propose and investigate three research questions below:

1. What key dimensions should be considered when evaluating AI agents used in customer operations from a managerial perspective?
2. How can these dimensions be integrated into a structured framework that enables business managers to systematically evaluate AI agents in real-world contexts?
3. How effective is the proposed framework when applied in an organizational setting, such as FrieslandCampina's customer operations team?

To answer these research questions, we develop an AI agent for a specific case at FrieslandCampina. The research assesses what methods are currently available to managers to evaluate an AI agent, and then designs a comprehensive evaluation. Later, validate the usefulness through by setting up interviews with business managers.

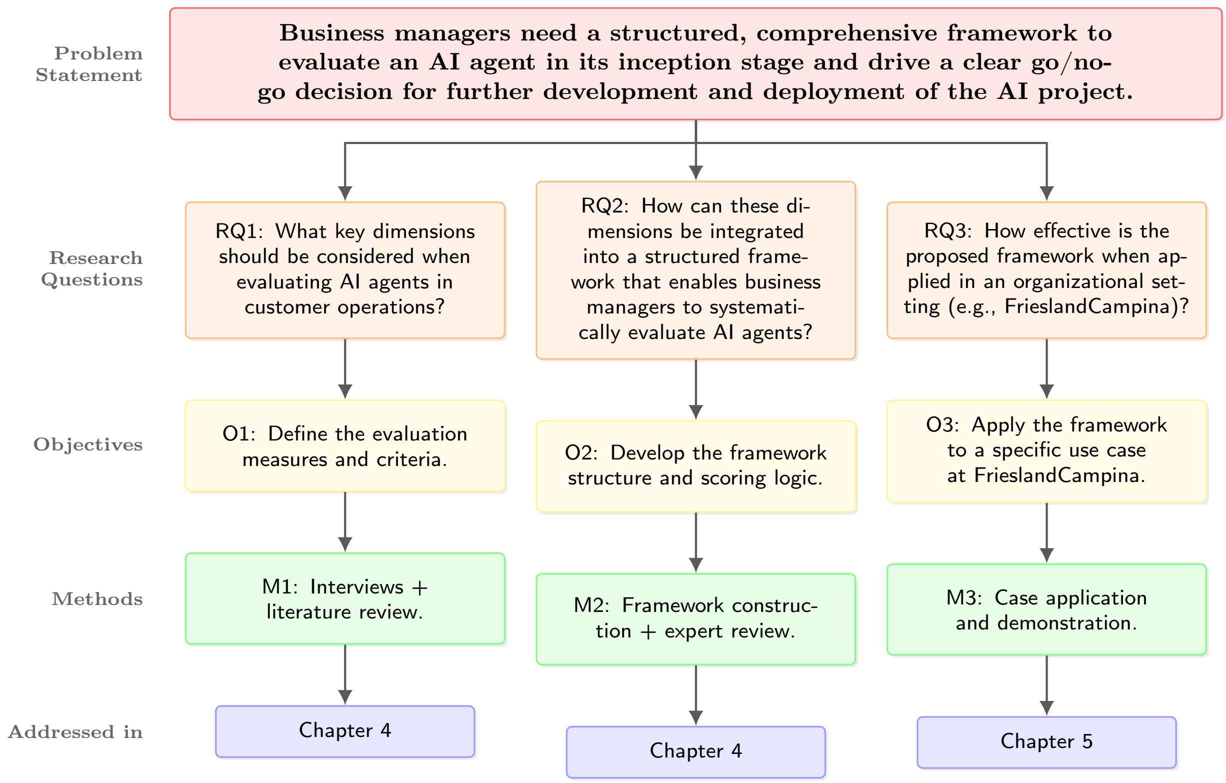


Figure 1.2: Research question overview: Mapping Problem Statement to Research Questions, Objectives, Methods, and Chapters.

1.3 Overview of the thesis

In Chapter 1, we introduced the adoption of artificial intelligence technology, AI agents and how it influences the business landscape, change in role of human workforce, evaluation of AI technologies. We identified a research gap that form the basis for our problem statement and is later concluded by designing the three research questions.

The remainder of the thesis is structured as follows. Chapter 2 covers the related work and theoretical background that covers definition of artificial intelligence (AI) and AI agents, evolution of artificial intelligence (AI) in business applications along with role of AI in customer operations. We also discuss and give an overview on the evaluation frameworks, evolution of performance management system and later conclude the chapter with research gap and opportunities.

In Chapter 3, we discuss about the methodology used to conduct our research, describe the data sources used, and conclude the Chapter 3 with evaluation measures for our study.

For Chapter 4, we focus on design process of the framework, the key scorecard dimensions and how we integrate the dimensions into a structured, comprehensive framework. We conclude this chapter with a final scorecard that contains key dimensions for the evaluation of AI agent from managers perspective.

The discussion in Chapter 5, starts with describing the evaluation setup and conclude it with evaluation results.

For Chapter 6, we discuss the research questions, framework objectives framed in Chapter 4. Later, conclude the chapter with limitations and future work.

Finally, for Chapter 7, we provide a short conclusion for our research project and way forward for the research.

Chapter 2

Background and related work

In this chapter related research regarding the concepts of Artificial Intelligence (AI), AI agents and characteristics of AI agents are described in section 2.1. Apart from this, we also give a brief overview on the usage of AI agents in organizational context stated in section 2.2. Moreover, we studied AI-Productivity paradox mentioned in section 2.3, brief overview on technical parameters of AI systems stated in section 2.4, and an overview of evaluation frameworks for LLM agents mentioned in section 2.5. Additionally, we studied different scorecard approaches mentioned in section 2.6. Lastly, we concluded the Chapter 2 with research gaps and framework opportunities stated in section 2.7.

2.1 Understanding Artificial Intelligence(AI) and AI agents

Artificial Intelligence is the study of systems that can perceive their environment with help of sensors and take actions to achieve their goal via actuators [54].

Actuators are known as intelligent agents whose foundational principle is rationality, that is to maximize the probability of success based on the information it receives [54].

Another study by Baird and Maruping [4] describes two information systems (IS) that are stated below:

- The agentic information system has autonomy i.e. can act without much human intervention, is goal oriented i.e. pursue outcomes under uncertainty and has adaptability i.e. change behavior based on environments [4].
- The traditional information system for example decision support systems where human retains control over the final action and system is only used for data processing [4].

Authors Baird and Maruping [4] mentioned the delegation mechanism which are appraisal, distribution and coordination .

To better understand this delegation framework [4], we consider an example of Tesla or Waymo cars where in the appraisal that is the first mechanism where both human and AI evaluate each other before delegating. For example human may have doubts about fully trusting the self driving capability and the AI will check human if they are awake by sending vibrations to the steering wheel. In the distribution mechanism, the human instructs the car to take him to nearest cafe but use a specific route. Last, in the coordination mechanism the human watches whether the car is steering correctly or not to prevent any accidents.

The paper by Wooldridge and Jennings [66] helped us to understand the characteristics or properties of an agent in terms of hardware or software systems, that has properties such as autonomy (no human intervention required), social ability (interacting with multiple agents), reactivity (respond or take action as per the environment) and pro-activeness (depending on the specific goal they can take action).

The paper by Lacity and Willcocks [32] talked about Robotic Process Automation (RPA) that is deterministic/rule based and work on structured data, this helped us to automate repetitive, routine tasks and rule driven work by copying human interactions with the existing systems.

Authors Lacity and Willcocks [32] highlighted most organizations start with RPA and then involve human-robot interactions by letting robot do the heavy lifting of repetitive work. On the other hand, human does cognitive analysis like making decisions with unstructured data, infer when RPA has failed [32].

2.2 Understanding AI agents in organizational context

In this section we studied literature on the role of AI in business context to understand how AI technologies is evolving overtime stated in section 2.2.1. We also studied different intelligence layers where AI plays a key role in service sector mentioned in section 2.2.2. Lastly, we concluded the section 2.2 by giving an overview on increased usage of autonomous systems and its effects on society, organizations and individuals mentioned in section 2.2.3.

2.2.1 Evolution in the role of AI agents in business applications

We see AI moving through distinct stages of maturity, moving from basic process automation to analysis and engaging with the customers to solve problems as business landscape expands, acquire new markets and bring new products [13].

The hype around artificial intelligence is real and most organizations are often swayed by it [13]. By investing into AI transformative moonshot projects that often fail to deliver any value [13].

Similar, case was discussed by the authors Davenport and Ronanki [13] were in IBM Watson invested 62 million dollars to cure certain forms of cancer but in the end it failed to even treat a patient. Another project started by chief information officer (CIO) to help customers with invoicing, billing and help the staff with IT problems, the project was used to improve patient satisfaction, operational efficiency, and report financial return [13].

The authors Davenport and Ronanki [13] also surveyed 152 technology projects. Later, categorized the projects as "shown in figure 2.1", here are the categories as follows: (1) Robotics process automation (routine, repetitive and administrative tasks); (2) Cognitive insights (analytics on structured data for insights); (3) Cognitive engagement (using deep learning to engage with employees/customers) [13].

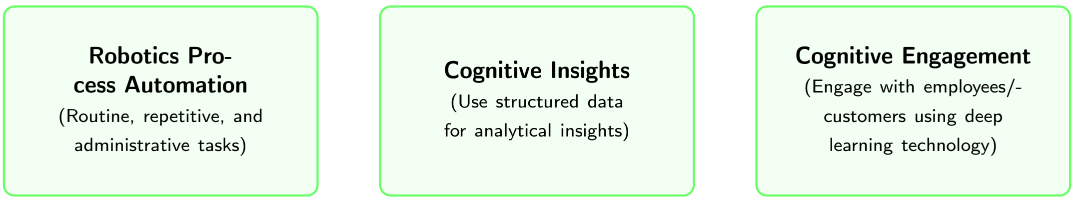


Figure 2.1: These three stages are adapted and defined the role of AI in different business needs adopted from Davenport and Ronanki [13]

The role of AI is moving from cost cutting, optimization, efficiency that was viewed as providing economic value towards autonomous executions [13].

Recent advancements in technologies like generative artificial intelligence (AI) and large language models (LLMs) that have matured in cognitive engagement by executing complex tasks autonomously without much human intervention [13]. This has created a shift toward algorithmic management where in tasks such as resource allocation, real time feedback and performance monitoring that previously done by humans have now been taken over by advanced algorithm and assume the role of managerial functions [28].

With advancements in technology, authors Kellogg et al. [28] noticed managerial complexity is introduced since certain tasks are being taken over due to algorithm control. Thus, the locus of agency shifts, this shift makes it difficult for organizations to know who is in control [4]. With Agentic IS, that have the power to take autonomous decisions, execute tasks without human supervision [4]. However, we need to keep human in the loop [11] so authority is shared and organizations establish a dual agency setup [4].

2.2.2 AI agents in customer operations

Customer operations is a strategic business unit, responsible for managing, designing and optimizing end to end business process to provide a positive experience for customers. This function of the organization is the first to get affected by recent technological developments in artificial intelligence because it has high volumes, repetitive/routine tasks and the availability of structured data since all the data is stored in customer relationship management (CRM) software [23]. This makes it easier for the service robots to access the data and give personalized experience to the customers [65].

The paper by Huang and Rust [23] mentioned how artificial intelligence plays a key role in the service sector and argues how AI is replacing the humans based on tasks complexity and emotional intelligence. The author Huang and Rust [23] argues that the service sector was traditionally dominated by humans as they could understand the feelings of the customer/user, understand issues and tailor the services but now technology is capable of doing these tasks at a cheaper rate and are slowly getting better than humans.

The main contribution of paper [23] is the four intelligence layer namely mechanical, analytical, intuitive and empathetic [23]. The four stages of intelligence are also "shown in the figure 2.2".

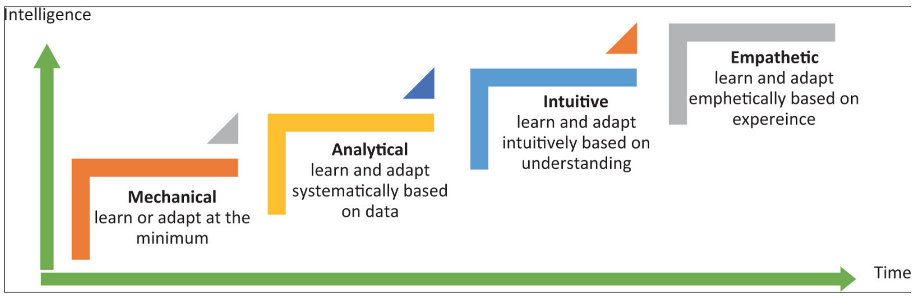


Figure 2.2: The Four Intelligences (Adopted from [23])

The authors Huang and Rust [23] discussed each stage one by one and examine how AI excels at a stage and pushes humans to upgrade their skills for the subsequent stages [23].

1. Mechanical Intelligence: The first stage, where AI performs routine, repetitive, transactional tasks via standardization and automation [23]. An example for this is self service bots or virtual bots that act as frontline customer service to answer some repetitive questions.
2. Analytical Intelligence: The Second stage, where in AI learns and adapts from historical data, it uses rules and logic to solve a problem [23]. For example IBM's Watson helps Human and resource for tax block preparation.
3. Intuitive intelligence: The third stage, where AI learns through artificial neural network or deep learning via statistics model [23]. For example, use of copilot agents to interpret data from power bi reports and summarize it or an LLM able to design a logo for you.
4. Empathetic Intelligence : The fourth stage, where in AI learns and adapt through experiences, it also learns from emotions, communication style, learning and affective computing [23]. An example is Sophia humanoid robot created by Hanson robotics [65].

The systems that are built using Artificial Intelligence have already started to dominate in the mechanical and analytical intelligences [23] that are within the customer operations department.

For example in FrieslandCampina we see usage of a ServiceNow chatbot named as Bonnie is able create tickets on behalf of the employees based on the conversations it has and is able to route it to correct team. This shows that chatbots are now able to have complex conversations that previously required either humans [65] to solve their problem.

However, the chatbots or virtual assistants that are deployed as a frontline agents to interact or have a complex dialogue with a customer [65] are operationally very critical as hallucinations or the interactions with AI are insensitive i.e. the responses it generates may damage the brand image and affect customer satisfaction.

A qualitative study was conducted by Kim and Lee [30] that mentioned generative AI is prone to hallucinations but how AI responds to user or customer affect trusts and satisfaction levels.

Authors Kim and Lee [30] found three major findings which are as follows:

1. The user self efficacy levels were boosted when AI made mistakes and responded with an acknowledgment saying thank you for noticing [30].
2. Users were forgiving when AI blamed external sources rather than itself [30].
3. User wants to know why AI hallucinated rather than a simple sorry statement [30].

Concluding from the paper [30], we need to transfer some of the findings into simple metrics when evaluating the AI agent on social appropriateness, service consistency, transparency and reliability.

2.2.3 Increasing usage of Autonomous Systems and Effects

New technologies and tools use generative artificial intelligence, these AI systems are moving beyond the normal support systems to autonomous roles where agents are present in the organizational processes.

Authors Raisch and Krakowski [52] highlights two concepts that are automation and augmentation.

1. Automation refers to tasks that are performed by machines to replace humans to achieve cost efficiencies [52].
2. Augmentation refers to human and machines working together to achieve a certain task [52].

Overly, relying on automation could have detrimental effects on business and society such as humans loose the specific skill for the work being performed by machines or humans loosing their job because of too much automation [52].

Authors Raisch and Krakowski [52] suggested, organizations must move towards a more virtuous cycle wherein automation takes over repetitive and mundane tasks, so that humans focus on tasks that bring value to the organizations. Thereby, helping humans to up skill in recent technologies and build their augmentation skill to work with AI [52].

Another study by Wirtz et al. [65], introduces the concept of service robots, that are highly autonomous and deliver tailored services to the customers. These robots are already present in different shapes, form and contain different attributes such as Alexa (virtual), Sophia (humanoid) and Roomba (vacuum cleaner bot) [65].

Authors Wirtz et al. [65] mentioned, depending on the task complexity based on cognitive and emotional tasks, these service task gets distributed between humans and service robots .

For example the heavy backend work that requires complex cognitive analysis but doesn't require social interactions moves to robots whereas when complex emotions and social interactions are required moves to human [65].

In future, the organization need to involve human-robot coordination when tasks gets highly complex both in terms of emotional and cognitive ability [65]. Such rapid adoption of highly autonomous systems would lead to certain challenges such as privacy issues, mass unemployment, and dehumanization of workforce [65].

Despite the rapid adoption and deployment of such AI technologies, it's becoming difficult for organizations to create balance between augmentation and automation [52]. Also, how to distribute tasks based on emotional and cognitive complexity [65].

Therefore, we need a framework grounded on human-AI interactions [2] and understand stakeholders desiderata based on explainable AI [33] that includes evaluating AI agents based on its usefulness, trust, usability, performance and accuracy.

2.3 The productivity paradox with AI

Authors Brynjolfsson et al. [9] discussed on artificial intelligence and modern day productivity paradox discusses about two contradictory trends around AI technologies that are technology optimism and the economic statistics relating to labor productivity [9].

Simply put, the technological transformation around AI has been massive such as machine learning algorithm improving over time in terms of perception and recognition and the error rate is going down significantly compared to human who are better at the task [9].

However, this doesn't reflect in the labor productivity growth statistical numbers for the US market. Moreover, this slowdown is also relevant to other OECD countries [9]. To resolve this paradox the authors Brynjolfsson et al. [9] introduce the concept of J-curve that is under measuring capital.

This curve helps us to understand that initially the productivity will go down as we are in the process of training new employee, restructuring the processes and cleaning the datasets to get intangible benefits [9].

Brynjolfsson et al. [9] mentioned that steam engines main purpose in the nineteenth century was to pump water out of coal mines but later on it supported innovations in other sectors such as railroads and transportation like steam engine and steam ships respectively [9]. The topic discussed by Brynjolfsson et al. [9] is relevant to our study as it helped us understand how AI is being used as general purpose technologies to increase productivity and help with innovations in other sectors.

So, business managers should evaluate AI from different perspective rather than replacing employees, we should think how can it provide value across business groups by solving different kind of problems.

Another paradox that mentioned by Autor [3], cognitive tasks performed by humans, come intuitively to them but it cannot be explicitly stated on how its done either via code or verbalization. The research paper by Brynjolfsson et al. [10] overcomes this paradox by highlighting that Generative AI can learn from examples with the help of tacit knowledge.

The researchers Brynjolfsson et al. [10] studied the introduction of generative AI conversational chatbot from 5,172 customer support agents based in Philippines who worked for their US clients to handle technical queries.

The support staff used the agents to help customers with real time suggestions for responses. The key findings from the paper [10] was it increased the staff productivity by 15% on average when we consider the metric issues resolved per hour.

Apart from this, the authors Brynjolfsson et al. [10] noticed that support staff was able to handle more chats at the same time, the average time taken to resolve a ticket reduced. Moreover, the AI agent significantly helped low skilled workers as compared to high skilled worker [10]. This happened because of the tacit knowledge that the AI was trained on came from experts who knew how to resolve a case [10].

Additionally, the authors Brynjolfsson et al. [10] observed the data from service outages with AI agents where they checked if agents retained the knowledge thereby helped them gain new knowledge and skills.

Another set of results was highlighted by Brynjolfsson et al. [10] AI agents helped with fewer escalations, lowered the attrition rate with newer staff.

2.4 Managerial decision making and AI systems

There is increase in AI agents deployment across organizations to either enhance the business process or for cost efficiency. Because of this business managers are shifting their focus purely from evaluating an AI agent form technical parameters to managerial perspectives.

Although, AI agents are technically developed and optimized by IT specialist but the decisions about scaling, productionizing, adoption by users/teams are made by business managers upon seeing the results from AI proof of concept (POC). They carry the burden for operational performance, strategic alignment and risk exposure for the deployed solution. The managerial decision making is not concerned with technical performance of the solution but how the solution aligns with business objectives due to which governance of the solution is important [64].

The paper by Weill and Ross [64] mentioned that IT governance helps with overseeing the accountability, alignment and decision rights of the AI agents that is being currently developed so that it aligns with business objectives rather than on technical parameters [64].

The paper by Simon [58] about decision research gave two terms that are bounded rationality and satisfying terms adapted from [58], that means managers often make decisions based on incomplete information, limited time and cognitive constraints and they look for solutions that are good enough to solve their problem rather than best solutions to deploy.

When using applied machine learning we mostly use parameters such as accuracy, precision, f1-score, recall [59]. Each of these parameters are technical that allow us to capture model performance and fine tune the model to improve the performance, accuracy [59].

However, these parameters are not useful for business managers as it doesn't translate to business objective or strategic alignment of the business unit. The research papers on human centered AI [2] and explainable AI [33] helps us to understand that stakeholders need to evaluate AI agents on accountability, transparency, interpretability, reliability and ability to understand the black box models so that user understand the underlying logic behind the model and why certain output was delivered [33, 21].

From a managerial standpoint in customer operations a high confidence score doesn't mean much as it doesn't help manager to check if AI agent is showing consistent output across cases or it is creating additional work for the team. For instance, if we use ChatGpt 5.2 an LLM model to analyze an invoice containing handwritten remarks from customers there are chances that the model assigns high confidence score to incorrect interpretations due to which it creates a significant problem for managers and teams, who need to manually intervene to correct errors that system falsely flagged as accurate, thereby undermining the efficiency gains, reliability of the LLM as expected from automation.

As observed from above example, there is rarely a direct correlation between technical parameters and the business performance criteria that can be used by business managers. While the statistical parameters such as precision, recall, F1 score, etc are important for data science teams but business stakeholders evaluate success based on dimensions such as long term value, operational costs and user satisfactions.

When AI system results do not map effectively with organizational objectives it leads to a alignment problem between statistically optimized model and organizational utility [38]. The alignment gap worsens when we evaluate autonomous agent as they execute end to end flow in seconds and give us output that could impact the operational performance of the team.

The authors Seif et al. [56] proposes a multi fidelity tests that would help managers to do a tradeoff between confidence score and costs by running few high to low fidelity tests. This helps the manager to do cost and risk coverage tradeoff by providing scalable mechanism to test and evaluate an AI solution [56].

2.5 Overview of evaluation frameworks for LLM agents

In Section 2.4 we observed certain technical parameters are used to evaluate an AI system. The research conducted by Blagec et al. [8] across 2,298 benchmark data sets revealed that accuracy is the commonly used metric followed by precision, recall and F-1 score to measure model performance but it can lead to ambiguity and obscure results when datasets are imbalanced. The study [8] also reports inconsistencies in metric naming conventions and varying formulae definitions to measure a model performance that would lead to obscure results [8].

Enterprises need multidimensional metrics such as security, reliability, compliance and maintainability. Therefore, we focused on structured two dimensional framework that are, first being what needs to be measured (evaluation targets) and second, how can we measure (evaluation methods) it the right way [42].

The authors Mohammadi et al. [42] discussed the following sub dimensions:

1. To evaluate targets such as user expectations, reliability, ethical compliance and technical capability [42].
2. The evaluation process consists of type of interactions with LLM agents, nature of datasets, metrics for calculations, use of technical or public frameworks and environmental settings such as sandboxes or simulation modes [42].

An important aspect discussed under evaluation process is how evaluation metrics are computed, the papers [42, 68] highlighted methods such as using another LLM as a judge for multi agent setup or evaluating [42] via humans which can be used to grade agent's output in terms of correctness, helpfulness and adherence to instructions [68] which are also important for managers to help with adoption and confidence to use the AI agent.

The study [42] also reveals that when AI agents move from demos to real implementation project new challenges emerge which are as follows:

- Agents need to be reliable show deterministic behavior and should have audit trails [42].
- Access control protocols should be incorporated in agents and it must followed [42].
- Adherence to legal regulations such as GDPR or HIPAA [42].
- Long term interaction simulations needs to be run to check whether agent drifts away from context that could affect goal alignment, adaptability and reliability [42].

For managers this paper [42] provides some key dimensions such as reliability, consistency, maintainability and compliance to evaluate an AI agent based on two dimensional taxonomy [42]. Driving inspiration from the paper [42], we try to connect few evaluation objectives with our study as agent behavior in terms of behavioral metric such as cost and latency it allow managers to calculate ROI of agent vs the humans.

Similarly, managers can measure success rate [42] to evaluate predefined goals related to business objectives. The manager will also get to know about the reliability guarantees when agent is measured on pass@k metric [42] to evaluate consistency over same tasks with repetitive runs.

The research paper by Mohammadi et al. [42] gave some key dimensions such as task success, reliability, security and compliance and mentioned the importance of having a structured framework to evaluate an AI agent.

However, AI agents are built on LLM and its behavior/output is decided by underlying LLM so this paper [47] focused on evaluating LLMs based on the use of appropriate benchmarks and metrics which are as follows:

- How well does the model perform on tasks (performance and task accuracy) [47].
- Does the model give similar output when working with adversarial prompts (robustness) [47].
- Consider evaluation based on deployment cost and economic sustainability [47].
- How transparent, traceable and explainable the LLM output is (explainability and interpretability) [47].
- Check on bias, toxic behavior and misinformation (fairness and bias) [47].

The paper [47] highlighted four methodological approaches to test the dimensions. For our thesis, few concepts are relevant that are: (1)the human centered evaluation (the outputs are evaluated by humans); (2)adversarial/simulation testing (stress testing the model with adversarial prompts and scenario simulations) [47]. These are important tests that managers can rely on to make decisions on which LLM to go for model selection prior to developing an agent [47].

To use the above multi dimensional assessment the author proposes a five later hybrid evaluation framework [47]. First is measure of task completion using benchmark such as MMLU, BIG Bench, SuperGlue paired with metric like precision, recall, accuracy [47]. Second, testing the LLMs on adversarial prompts and edge cases using indicators like failure rate [47]. Third, ethical and safety assessment testing based on toxicity score, hallucinations and biases [47]. Fourth, tests on things that are relevant to the problem and the domain in which problem exists [47]. Finally, testing with humans to evaluate the outputs and refine the LLMs [47].

The paper [47] was introduced as it helped us with early selection of LLMs based on outputs that closely aligns with the business objectives and how it can be measured with suitable metrics.

Moreover, it gives a practical approach to look at the overall cost based on compute,latency and deployment cost [47]. This will help to compare LLMs based on cost efficiency. When looking at model performance based on the benchmarking score and metrics it will let manager know about failure rate that could translate to operational cost spent on correcting the LLMs/agent mistake that could indirectly help them to know if they can scale the agent [47].

The framework mentioned in [47] would also help managers to evaluate on reliability and consistency as we put test LLMs through stress tests that could help with dimensions such accuracy and trust.

2.6 Understanding scorecard approach and some key dimensions

In the early 1980's most of the performance management systems were focused on calculating the accounting parameters to measure the business performance and were grounded on accounting principles [31].

According to Gomes et al. [19], the evolution of performance management happened in two phases. First one in the late 1980's and the second phase was in late 1990's. The first phase was focused on financial measures such as profit and return on investments but the second phase focused not only on financial measures but also on the non financial measures that was tied with business strategy [19].

The literature survey conducted by Kurien and Qureshi [31] mentioned most of the model/framework development happened after 1980's. The study [31] also reveals that most models have gone through empirical testing and some are in theoretical development phase.

The paper [31] also acknowledges that most cited works in performance management system are SMART (1988), Performance Measurement Matrix (1989), Balanced Scorecard (1992) and integrated performance measurement system (1997) [31]. Table 2.3 highlights a list of models that were developed since the early 1980's till 2007.

Table 2. List of Performance Measurement Models (Taticchi et al., 2010 and Morgan, 2007)

Name of the model	Period of introduction
The ROI, ROE, ROCE and derivates	Before 1980s
The economic value added model (EVA) The activity based costing (ABC) – the activity based management (ABM,1988) The strategic measurement analysis and reporting technique (SMART,1988) The supportive performance measures (SPA,1989) The customer value analysis (CVA,1990) The performance measurement questionnaire (PMQ,1990)	1980-1990
The results and determinants framework (RDF,1991) The balanced scorecard (BSC,1992) The service-profit chain (SPC,1994) The return on quality approach (ROQ,1995)	1991-1995
The Cambridge performance measurement framework (CPMF,1996) The consistent performance measurement system (CPMS,1996) The integrated performance measurement system (IPMS,1997) The comparative business scorecard (CBS) The integrated performance measurement framework (IPMF,1998) The business excellence model (BEM,1999) The dynamic performance measurement system (DPMS,2000)	1996-2000
The action-profit linkage model (APL,2001) The manufacturing system design decomposition (MSDD,2001) The performance prism (PP,2001) The performance planning value chain (PPVC,2004) The capability economic value of intangible and tangible assets model (CEVITA,20041)	2001-2004
The performance, development, growth benchmarking system (PDGBS,2006) The unused capacity decomposition framework (UCDF,2007)	2006-2007

Figure 2.3: List of Performance Management Models/Frameworks adopted from Taticchi et al. [60] Source: Kurien and Qureshi [31]

The study conducted by Neely [43] using citation and co-citation analysis of papers published between 1981 and 2005, found out that balanced scorecard developed by Kaplan and Norton was the most cited work.

Similarly, research conducted by Alhaddi [1] on the evolution of performance management from 2005 till 2020 acknowledges that balanced scorecard has been the most cited references, citation referenced amounted to 296 citations.

We move away from citations analysis to understand how BSC was applied practically to different organizations.

Authors Hasan and Chyi [22] studied balanced scorecard application across industries including healthcare, education, agriculture and manufacturing. The paper highlights [22] BSC was used to identify improvement opportunities with internal business processes, areas of improvement, focus on evaluation based on user needs, and measure the qualitative insights gathered from employees feedback. We draw inspiration from the paper [22] to tailor our approach and create a framework based on the problem addressed.

In section 2.4 we discussed, technical metrics such as precision, recall, F-1 score and confusion matrix [59] that are relevant for model performance and helped us verify results of an agent but managers cannot use it to evaluate an AI agent holistically.

Furthermore, studies reviewed in section 2.5 helped us to know about the dimensions important to evaluate an AI agent and LLMs. Moreover, we became aware why structured way of evaluating an AI agent from the section 2.5 and this helped us with some relevant metrics for business managers to make decisions for the AI project.

However, we still need a framework relevant for managers to map strategic objectives as per organization vision and strategy. Later, look from cost perspective and non financial measures such as user perception, organization capabilities and integration with business process [22].

The decision to scale, productionize and further investments for an AI agent developed by IT Team needs to taken care by business manager. For this, we introduced another framework by Van Grembergen [61], that is built by cascading different balanced scorecard from IT level to company/organization level. Figure 2.4 illustrates how the cascading effect works and how it is all related to the foundational balanced scorecard and understand how IT creates value for business.

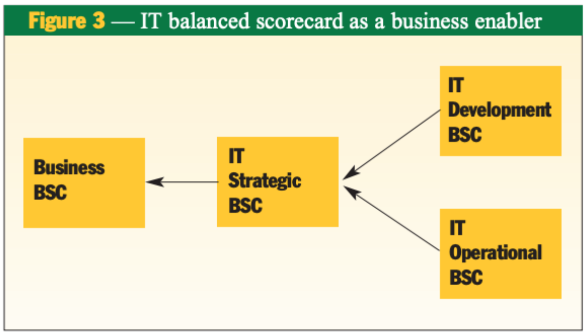


Figure 2.4: The cascading effect of IT scorecards and how it relates to business scorecard adopted from Van Grembergen [61]

So in simple terms, we first understand the business objectives i.e mission for each view and then map it into the four perspectives namely financial, user, internal business process and learning and growth [27].

Later on, map the above business objectives to IT strategic balanced scorecard and IT development balanced scorecard [61]. Lets expand this concept via an example lets say FrieslandCampina wants to combine Europe's various customer support operations teams into a singe cluster headquartered in Netherlands, as the organizations move towards digitization and automation.

To map this into balanced scorecard, we use financial and customer perspective where the goal is to reduce the operational cost by 20% but maintain the customer satisfaction score to 3.9 out of 5.

The IT strategic balanced scorecard [61] will have a goal for the IT team to build an LLM agent that handles simple questions like order information, delivery transport updates, etc. The metric here will be how many tickets were successfully resolved without customer staff support, if this is low this directly affects our business balanced score card.

The IT development balanced scorecard [61] goal would be to focus on developing an AI agent that will not hallucinate, give accurate answers based on user ask. So, if these metrics fail then it affects our IT strategic business balanced scorecard [61]. We will be using a similar approach to map the key dimensions as they are IT and business related to the balanced score card so that managers can evaluate the AI agents.

As seen earlier, we have a foundational level balanced scorecard that measure the performance of business at an enterprise level that was given by Kaplan and Norton [27]. Later on we moved to an IT balanced score card to understand on how IT can create value for business and the cascading effect on how IT and business strategies are connected to each otherVan Grembergen [61].

Now, we will make use of another balanced scorecard that helped us to evaluate IT applications projects from different lenses such as business value, user orientation, internal process and future readiness that are highly relevant Martinsons et al. [36] for our case study.

Since this framework [36] is an extension of the foundational business scorecard [27], the financial perspective [27] is measured by return on investments, payback period can now be evaluated via business value [36] when working with IT projects or IT applications to check on cost control measures [36] such as cost of an employee doing the work.

Apart from that we can also use value categories such as value linking to know about benefits gained from system, generated savings when using value acceleration [36].

This framework [36] gave direction to understand the business value by linking to evaluation categories of value and risks [36]. This helped us to address some of the dimensions that will be relevant for financial perspectives [36].

Moreover, the user orientation perspective captures some of the key dimensions such as credibility that will help us to evaluate trust and also help us to measure this in terms of user satisfaction for our survey. This perspective also capture the value by evaluating the user experience in terms of how helpful the IS system [36].

The third perspective is about internal processes that focuses on delivering the services at low cost but high quality and measured via time metrics such as fine tuning applications [36].

Finally, future readiness perspective focused on future challenges and continuous improvements as well having specific budget for to train on new technologies so that internal employees are prepared for the future work [36]. It also focuses on the user satisfaction and technical performance of systems captured under portfolio management [36].

2.7 Gaps and opportunities for framework

The reviews of existing literature looked at the role of AI in customer operations and how it has evolved overtime, understand the productivity paradox with AI, model centric evaluation frameworks and LLM agents evaluation frameworks. This helped to understand why evaluation of AI agents should also be done at managerial level not just using technical parameters.

Each of these topics aided us to get an understanding on what are the necessary measures that needs to be considered during AI agent evaluation and provided opportunities to develop a comprehensive structured framework.

Moreover, as highlighted in the paper by Sokolova and Lapalme [59] where technical parameters are considered more important when evaluating an AI agent but an important measure such as business objectives or organization utility [38] is missed that creates a disconnect with technical parameters and organizational strategy.

Furthermore, research on evaluation framework for LLM agents mentioned in section 2.5 suggested that some attempt has been made by researchers to measure the results based on performance, robustness,safety, reliability and maintainability. These parameters are focused on the sociotechnical dimensions which gave a starting point to form a more comprehensive framework for business managers [57].

Even though some perspective regarding future readiness for the organizations and how it can continuously improve to better align with the future technology of Agentic system is missing. Hence, we need a perspective that helps us to understand the organizational readiness.

Additionally, the focus has been on how to select certain dimensions and the method to evaluate these dimensions the right way [42] but often forgetting about how it can be useful for internal business process such as how well AI agent integrate with the existing business processes.

Most of the research paper on the evaluation frameworks of LLM agents or for models that use AI are good for measuring the models or agent itself but doesn't look at the broader perspective from managerial lens.

To summarize the sections covered in literature review of existing evaluation frameworks for LLM agents, we would summarize our findings and understanding in the table 2.1.

Table 2.1: Literature review of few sections summarized

Frameworks Discussed	**Primary Focus**	**Metrics Used**	**Research Paper**
Model Centric	Model Performance	Accuracy, F-1 Score, Recall, Precision	Sokolova and Lapalme [59]
Balanced LLM evaluation	LLM's performance	Robustness, Safety, Bias, Fairness, Reliability,Toxicity, Task performance and Accuracy	Shukla [57], Pandhare [47], Mohammadi et al. [42]
Financial Analysis (Before 1980's)	Cost/Benefit	ROI, NPV	Kurien and Qureshi [31]
Balanced Scorecard	Holistic Strategy	Financial + Non-Financial measures	Kaplan and Norton [27]

During my internship, we observed the measurement system that is used to evaluate an AI agent is based on the lagging parameters as seen from Table 2.3 that concentrate on accounting principles or financial measures such as ROI, activity based costing [31] but we moved beyond that and focused on non financial measures as well.

Simply, put how much money needs to be invested in AI vs the benefit it can bring to us in the coming years or the number of hours saved by using the solution. We need to move beyond the financial measures and also focus on the measures such as user experience that tells us how comfortable the employees are while using the solution.

In conclusion, the literature on balanced scorecard approach demonstrates established practices and how it can be used to fulfill a gap rather than working with a siloed AI evaluation frameworks. Each section explored above identified opportunities for a structured framework that would strike a balance between dimension measures from the evaluation frameworks and how to relate them to the BSC structure.

A clear, upfront framework will enable the business managers to evaluate an AI agent and make early decision whether to productionize, scale or drop the project.

Chapter 3

Method

In this chapter, we outlined the structured approach taken to develop the framework. First, we discussed about the Design Science Research Methodology (DSRM), followed by the six step process. Later, we understand the steps used to design a balanced scorecard, and finally, discussed the techniques used to generate the questions for respective dimensions.

3.1 Design Science Research Methodology

We have adopted the Design Science Research Methodology (DSRM), that is a problem solving model which helps to create innovative artifacts and the design knowledge through novel solutions for the real world problems thereby enhancing human knowledge [62]. These artifacts can be models, design principles, frameworks, algorithms and methodologies.

To bridges the gap between academic work and practical world, we develop an artifact and test it with the relevant stakeholders. The framework created by Peffers et al. [49] divides the research process into six activities as shown in Figure 3.1. We have explained all the steps below:

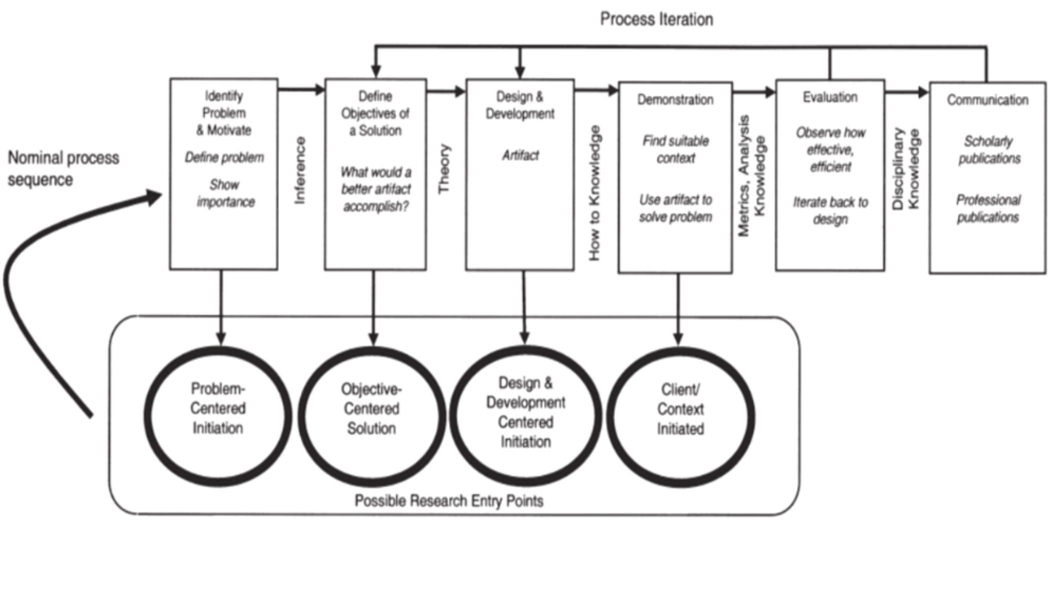


Figure 3.1: DSRM Methodology Process Model (Adopted from Peffers et al. [49])

1. Activity 1: Problem Identification and Motivation:

   "Identify, define the specific research problem and justify value of the solution"

The research problem identification began by identifying how managers are currently evaluating an AI agent and what parameters are being used to give a go/no-go decision for further investments, deployment, and, productionize at scale. We gathered this information via interviews with different stakeholders (business managers, team members and operations leads) and looked at the past projects.

Insights from literature also revealed key gaps on how AI agents are evaluated from technical perspectives and absence of structured framework for managers. This phase helped us to define the core problem statement stated in section 1.1 and revealed why a framework is needed to guide managers to evaluate an AI agents in structured way with different perspectives and help them with a clear g0/no-go decision.

2. Activity 2: Define the objectives of the solution

   "Develop the objectives of solution from the problem statement and gain knowledge on what is possible and feasible"

   After the problem identification stage, discussions from the stakeholders and relevant literature review, this study proposes a framework objectives in 4.1 to be fulfilled. Each objective helped us to answer research questions developed in section 1.2.

3. Activity 3: Design and Development

   "Creation of an artifact"

   During this phase, we determine the artifact's desired functionality and architecture, followed by the creation of an actual artifact. To create the artifact and it's dimension we used balanced scorecard framework. The output of this phase is structured, comprehensive framework along with key dimensions, this framework is the result of literature review insights, team feedback and internal pain points.

4. Activity 4: Demonstration

   "Use the artifact to solve a particular problem"

   In this phase, we applied the artifact to solve a particular problem. The business problem –in this case freight document solution built on AI agent capabilities– to assess the fit, gap identification, and check how it helps the manager for decision making and investment decisions.

5. Activity 5: Evaluation

   "Test how well the artifact supports the solution to problem"

   In this phase, we checked on the usability of the framework and how well it support our objectives by comparing it to the actual results. We check back on team members whether a particular dimension is effective or not and move on to the next phase that is communication.

6. Activity 6: Communication

   "Communicate the problem and design of the artifact"

   During this phase, a final tailored framework is shared to the team to improve the current way of working and motivate them for adoption of the framework, followed by applying the framework to other business project related to AI agents.

3.2 Scorecard Design Process

To asses an AI agent form managers perspective, we used the steps mentioned in section 3.1. This helped us to generate the dimensions and the relevant structured questions for the framework. We used different data sources as shown in table 3.1, that helped shape our final scorecard.

Table 3.1: Overview of the data sources and its purpose

Data Sources	Purpose
Company project template	Understand the current process on how current IT projects are evaluated
Semi structure interviews (n = 5)	Identify the core dimensions and key challenges
Workshop	Identify how the team evaluates an AI solution and the user perspectives
Steps involved in designing a balanced scorecard	Help us to design mission and objectives of the four perspectives

3.2.1 Interviews

During the initial phase of this research, semi-structured interviews were conducted with the team members of the customer operations team at FrieslandCampina. This setup of semi structured interview allowed us to explore how a project is evaluated, the knowledge of AI agents, what perspectives are important for managers to evaluate an AI agents and know more about the customer operations process.

Each member is a specialist in supply chain management with varied skill set in different areas such as customer service, improving the customer operations process, little familiarity with AI agents and artificial intelligence, application specialist for EDI orders, case management and order processing. All 6 members across the department, including CIS specialist for customer operations, customer operations managers, senior manager and EDI specialist were interviewed using a questionnaire.

In total, six in depth interviews were conducted, each session lasted for 25 minutes, and during the interview some open ended questions listed in A were asked.

This allowed us to know more about the strategic goals of the company and the vision they foresee for a particular department and necessity of using AI in non value added tasks. This technique directly helped us to form a problem statement and work towards few solution objectives that are the activities 1 and 2 of the DSRM methodology which established main issues, problem identification and goals.

Furthermore, we also setup some interview question with the operational leads who would be using the AI solution and how would they evaluate an AI agent as well as exploring their core competencies if they require a deep explanation on how AI agent is working, how results are measured and the ease of use.

3.2.2 Stakeholder Identification

Apart from conducting the interviews, we needed to map the key stakeholders who will be using, developing/building and evaluating the solution in FrieslandCampina. The same is covered in the figure 3.2.

This approach also addresses the first two steps of DSRM and was very important to understand how AI agents are developed, used by the teams and who manages it. This will help us to identify who would be using our artifact either directly or indirectly and who it will serve ultimately.

The stakeholder mapping is a core step to help us with the practical realities of the business scenario and how to align the framework structure with it. The job roles and responsibilities, interviews, job titles were used to categorize the stakeholders and the distance to the developed product/solution.

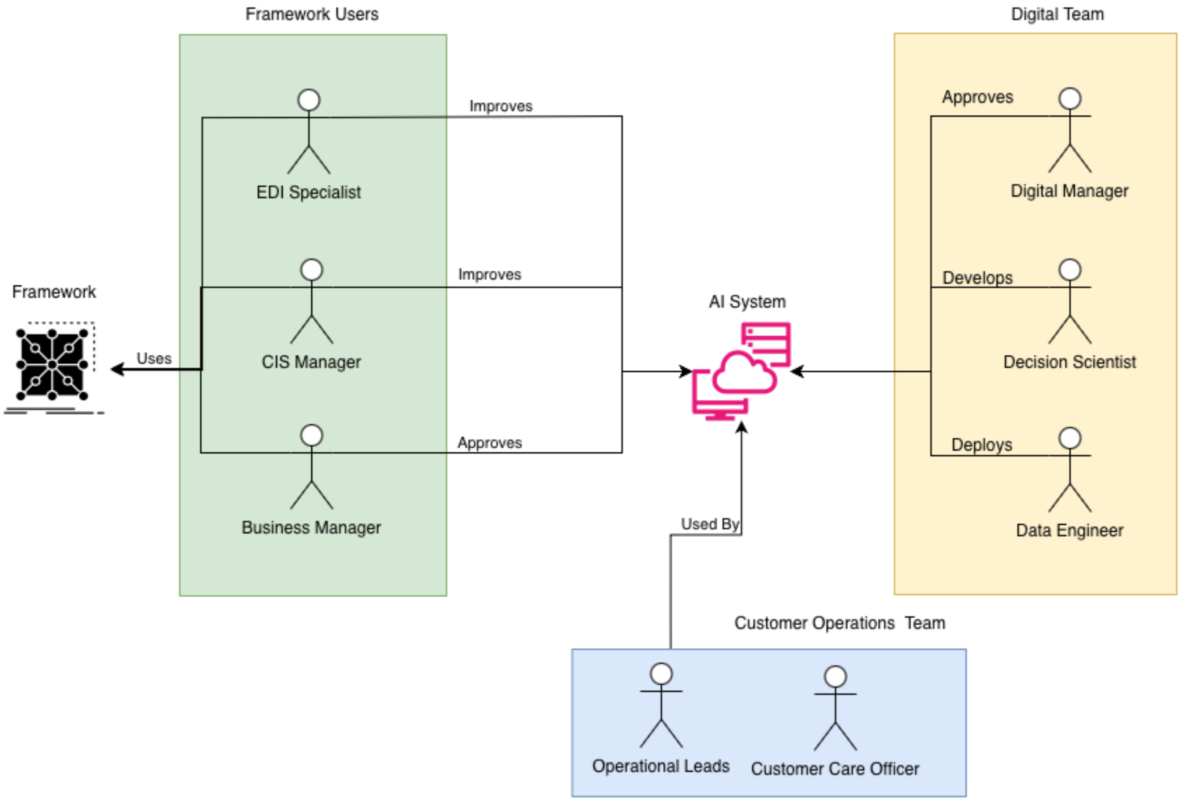


Figure 3.2: Stakeholders at FrieslandCampina

We identified four primary stakeholder groups that are as follows:

1. **Business Managers**: These roles are important as they help identify the gaps, opportunities on where we can use an AI solution to enhance the customer service departments by bringing in new ideas and key pain points. Apart from this, they are also responsible for approving the AI solution and have a say in go/no-go decisions. They will validate the results of an AI solution initially during proof of concept and will be key users who would be using the framework.

2. **Operational Leads and Customer Care Officer**: These individuals serve as the frontline support staff to address the customer queries, questions and the backbone for order processing, maintaining customer relationship and processing credit notes at FrieslandCampina. They are very important stakeholder as their inputs helped us to know where we can use AI agent to eliminate non value added activities and understand the customer operations process related to our case study. They will be primary users of the developed solution and interact with it on a daily basis. They will help align the user perspectives of using an AI solution and other factors.

3. **Data Scientist and Data Engineer**: These individuals help us with the solution development and deployment for the suggested problem. They help with suggestions on what the most cost effective way to develop and maintain a solution. They decide on what key metrics/parameters are required to evaluate a solution but not strictly technically and how it will fit in the FrieslandCampina architecture.

4. **Digital Manager**: The digital manager is responsible for development and deployment of AI solutions within the customer operations department. They play a key role in translating the customer operations strategy to business group objectives into scalable AI driven capabilities that drive operational performance and elevate customer experiences. They also guide their team on which technologies to use as well as are the final approver of the solution before it goes to business manager.

5. **Continuous Improvement Service Managers and EDI Specialist**: These individuals are involved in projects to help with change management and report key improvements for managers and user perspective on what the solution is missing and how can we introduce it to the team.

They help the users to understand the technical part of the solution from a business perspective and look at the bigger picture on what skills are required by the employees.

3.2.3 Workshop

A workshop session was conducted for 60 minutes in which 5 people where invited from Europe and Retail Americas team. The invitees held different titles such as Continuous Improvement Service Manager, EDI Application Specialist, two Customer Care Officer and a Senior Manager. Before, we go to the next sentence let us first understand what Power Automate and SharePoint are.

Power Automate, is a Microsoft tool to design workflows so that we can automate the routine work, that results in streamlining the business process. For the business problem that is addressed, we make use of cloud flows [40]. For our research we used Power Automate to extract key data from freight documents, this was done by using LLMs and ML model that are available as connectors by Microsoft.

SharePoint, is a (software as a service) SAAS tool that is used by FrieslandCampina employees to store, access and create sites [41]. We are using SharePoint to store our results that is generated from Power Automate flow.

The context of the study focuses on an AI agent driven automation flow where freight documents from different vendors are processed using Power Automate, and some key value pairs are extracted and stored on the SharePoint list view.

The Power Automate tool helped to know if the flow was executed successfully and measured the success when looked at SharePoint list view but this failed to capture the business value that is perceived by business managers when evaluating an AI agent. The workshop aims to bridge the gap between the technical parameters and the parameters a business manager might use.

The workshop was conducted in three phases that are listed below:

- **Business Process Mapping:**
  - Objective: Map the process of customer operations related to freight document case and in which part of the process we can use AI solution/agent. This workshop activity was conducted after the approval and validation from customer operations manager.
  - Activity 1: Participants were shown the detailed process and asked to validate the findings. Furthermore, if the use of AI targeted the correct process and part. We also showcased the business value the solution will generate.
  - Activity 2: A proof of concept demo was shown to the participants using actual PDFs that arrive on email and how the AI agent extracts the key value pair from the PDFs and loads it to SharePoint list view.
  - Discussion: Some key points were discussed such as speed of the document processor vs manual handling, automation level of the setup, integration with the current processes and human collaboration once the AI output is generated.
- **Business Value Creation and Trust with AI**
  - Objective: Let participants understand the case volume load currently handled by customer care officers. Identify the user trust, Return on Investment and Cost required to develop the solution.
  - Activity: Participants were asked to validate the potential of the solution in terms of financial gains done by business managers and another set of participants who used the solution, were asked to check if the solution output can be trusted and how accurate it is.
  - Discussion: Some key perspectives of balanced scorecard were discussed such Financial perspective and User perspective. From the financial perspective we discussed on time saved, manual effort, cost reduction and scalability of the solution. In terms of User perspective, we discussed on how easy it is to use the solution, errors on the output produced and the accuracy of the extracted data on SharePoint list view.
- **Future state of Working and Organizational Readiness**

– Objective: To evaluate the new way of working and the shift from manual entry to an AI driven approach.

– Activity: Participants were asked to do some retrospectives on the solution and the current ways of working with freight documents. A practical scenario was asked were we compared the current ways of working vs when AI would do the same work with human intervention.

– Discussion: The discussions were centered around how AI agent helped the team in terms of user experience, productivity and efficiency of the solution. We also touched upon how this solution can be integrated with existing tools and business process. Apart from this, we interacted with managers on how can they validate the solution and understand why AI made an error. When discussed with senior manager, the focus was on how can AI add value to the team, what skill sets are required and does this solution actually help them with routine tasks. We also touched base upon the change management to bring the productionized version to the team and how can other business units benefit from the solution.

3.2.4 Steps Involved In Building a Balanced Scorecard

In this section, we will discuss the steps that are taken to build a specific balanced scorecard that is inspired from the research paper "The balanced scorecard: a foundation for the strategic management of information systems" by Martinsons et al. [36]:

1. Create an awareness regarding the concept of the balanced scorecard among the business managers and other stakeholders [36].
2. Get the data on corporate strategy, business strategy and the digital strategy; specifically on their objectives and goals related to their business department [36].
3. Define the goals and objective of the business group from the balanced scorecard four perspectives [36].
4. Develop an initial balanced scorecard on the data gathered on goals and objectives of the business department [36].
5. Check with the management for feedback and comments on the designed balanced scorecard and make changes if possible [36].
6. Reach an informal agreement on the balanced scorecard that can be used by the business managers [36].
7. Represent the scorecard and the logic behind the dimensions used to the respective stakeholders [36].

For our study, we will also follow the points mentioned above as a guideline to design the balanced scorecard to evaluate an AI agent for business managers.

3.2.5 Project Template Assessment

For our business scorecard design process we got the initial inspiration from the template showcased in the figure 3.3 and consider some key fields such as business value and cost for our framework design. We studied this template to understand some key dimensions and integrated into our balanced scorecard.

This is a very practical solution or framework that is currently practiced in FrieslandCampina to evaluate an IT project and give an overview on the business problem, solution development, product key features, project timeline and evaluation in terms of business value and cost.

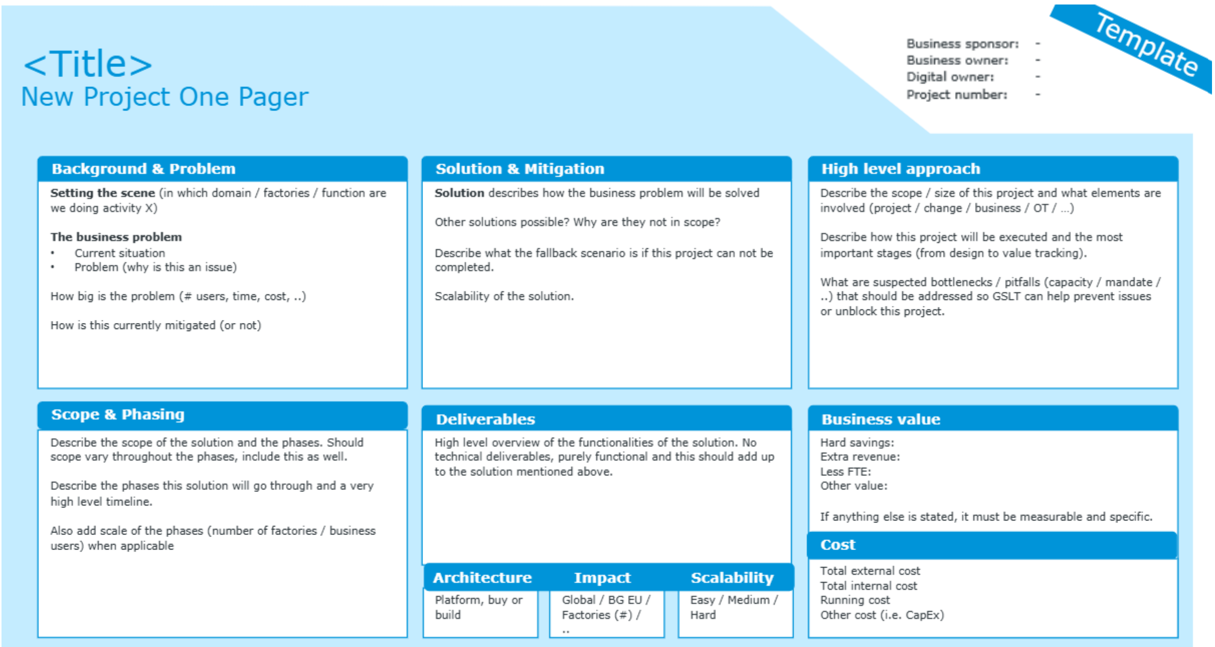


Figure 3.3: FrieslandCampina Project Template

Below is the summary of the key fields:

1. Background and Problem: This field is used to give an overview of the problem that IT/Digital End to End team is trying to solve.
2. Solution and Mitigation: In this we need to describe what kind of solution the business managers want and the solution scope, also the scalability of the given solution. This can be technical or non technical.
3. High level approach: This is a key field as it describes the scope and size of the problem and what are the important elements of the projects. It includes important stages of solution development from design to value tracking. Apart from this, we also need to fill in common pitfalls or bottlenecks that may come along the way as we develop the solution.
4. Scope and Phasing: This field describes the scope of the solution. Simply put, it includes how do we solve the problem and develop a solution in staggered phases. Also, the timeline of the project from start to end along with how many users will be there for testing, the scale of phases when solution is built for factories.
5. Deliverable: This gives the IT or Digital team an overview of the key features in the solution i.e. the functional scope for it. Along with what technology stack we need to use, the impact for it and the scalability of the solution to different teams/factories.
6. Business value: This is a key field for business managers to be filled in as it tracks how much value would be generated from the business solution in terms of cost, revenue generation and savings in FTE count. This is purely a quantitative field that can be measured.
7. Cost: This field is used by IT or Digital team to give an estimate on the development cost, resources allocation/cost, and the running cost once the solution is deployed (support and maintenance activities). This is also a quantitative field that needs to be measured.

3.2.6 Define Scoring Mechanism

A 1-5 Likert scale is used to measure each question for our case study approach that helps us to evaluate AI agents from managers perspective for each dimension followed by statements that needs to be scored by the relevant stakeholders. The questions formed were related to literature insights, interviews and workshop that help us to address the dimensions that are in considered in BSC.

Since we are in an organizational setting and the perspectives are subjective in nature therefore using Likert scale [34] helps us in following ways:

1. Reliability : The Likert scale allows the managers to measure the intensity of their experience with the developed solution with AI, it helps to show the agreement level with new technology [25].
2. Midpoint: The Likert scale allows the respondents/participants to actually select a midpoint that is neutral indicated by number/response (3) suggests that participants can express their neutrality or indecisiveness rather than forcing a bias when working with a new technology like AI [51].
3. Assessment: The research paper on attitudes towards AI at work further validates the use of Likert scale to measure the dimensions such as usability that would be one of the dimensions of the balanced scorecard [55].

Furthermore, to support our usage of Likert scale we studied additional research papers on different topics that uses balanced scorecard approach and design questionnaire using Likert scale. The study conducted on employee and patient satisfaction in healthcare states that a five point Likert scale was used to to measure employee satisfaction [35].

Another research on a postal services evaluate company performance using balance scorecard and to measure the performance for four different perspectives they develop the questionnaire using five point Likert scale [24]. Moreover, research conducted to evaluate financial performance of the company in retail sector used a five point Likert scale questionnaire to gather data from 75 retail firms [26].

In conclusion, we have seen usage of Likert scale type questionnaire to gather data when using balanced scorecard method to assess organizational performance, financial performance or satisfaction scores in different industries, its quite common practice to integrate the Likert type questionnaire to be used for survey.

3.2.7 Evaluation Measures

In this section we covered the different evaluation measures used to validate the survey statements and the dimensions used for the balanced scorecard perspectives to check on internal consistency and reliability of the survey.

Furthermore, we checked the mean and standard deviation of each perspective that helped to know about the participants sentiment and consensus of the balanced scorecard perspective.

Additionally, we covered the method used to establish relationship between each perspective via Pearson's correlation and how do we measure the balanced scorecard. The statistical evaluation measures such as calculation for internal consistency of items construct, standard deviation, mean, Pearson's correlation tests are inspired by the research paper [26] that used similar approach.

- **Measures for Internal Consistency of Items Construct**:

  In Section 3.2.6, we have seen why it is important to use the right scoring mechanism to measure the dimensions for balanced scorecard questionnaire.

  In this section, we covered an equally important topic that is to check on the measure of internal consistency reliability of the survey statements that have been developed for each perspective for respective dimensions.

  Simply put, it helps to know what we set out to measure (perspectives/dimensions in our balanced scorecard) via a survey is actually being measured in the right way and the survey is reliable. We will use two methods to check the internal consistency and reliability of the survey which are Cronbach's Alpha and Composite Reliability.

  Although, the most common measure of reliability of the constructs is done via Cronbach's Alpha due to its popularity with researchers as well as the number of citations within scientific articles [39]. The author McNeish [39] discussed certain flaws of this method such as Tau's equivalence, items being normally distributed and the randomness error or noise that we are unable to measure (errors in the items do no covary) [67], [20], [18], [44], [53].

  To overcome these drawbacks/assumptions of Cronbach's Alpha, the author McNeish [39] suggested a conceptually similar method that has been previously tested on various studies and performed better [39]. This method doesn't make strict assumptions like Cronbach's Alpha, therefore author suggested to use Composite Reliability test [39].

Lets see the mathematical formula for Cronbach's Alpha and Composite Reliability (Omega) below:

1. Cronbach's Alpha: We use the formula of Cronbach's Alpha from the paper [12]. This is used to check the internal consistency and reliability [39] of the perspectives of the BSC.
2. Composite Reliability: The formulae for composite reliability was mentioned in paper [37]. We used this formulae to check the internal consistency and reliability of the dimensions of BSC for our research as well as to address Cronbach's Alpha assumptions [39]. The usage of the above formulae is seen in section 5.3.

- **Average and Standard Deviation**

  We also need to calculate the average/mean of the perspectives for the sentiment analysis on how our balanced scorecard is doing and the standard deviation on the general consensus of the participants for the dimensions used regarding the evaluation of an AI agent. The formula for this is adopted from [17] covering mean and standard deviations.

- **Pearson's Correlation Method**

  To calculate the relationship between the four perspectives, we used Pearson product moment corelation coefficient mathematical formula as outlined in the research paper [17], this was used to establish the inherent relationship between the balanced scorecard perspectives as mentioned in [27].

Chapter 4

Design

The chapter covers the design of the balanced scorecard approach for this research. First, we frame our framework objectives that gave shape and foundation to the scorecard. Second, we discussed about process of designing a balanced scorecard, this is followed by expanding each perspective and its related dimension. Later, we revealed the balanced scorecard to be used by business managers to make a clear go/no-go decision for the AI agent project.

4.1 Framework objectives

Historically, the evaluation of artificial intelligence has been done through technical parameters but now we look at certain guidelines when humans interact with AI systems at different life cycles [2].

Initial interaction, we need to understand the capability of AI system with the help of metrics [2]. Working with an AI system, should show information relevant to the problem we are trying to solve [2]. When AI system gives wrong information, ability to understand why it gave incorrect information and what led to the wrong recommendation [2]. Lastly, when AI systems become part of work routines, users should provide feedback to improve the system and notify users if there are any changes [2].

Moving away from AI system interactions from user perspectives [2] to understand the interpretability of AI such as machine learning models [15].

The author Doshi-Velez and Kim [15] highlights the definition of interpretability that is understand or explain the machine learning (ML) systems in a format that is understood by humans.

To evaluate the interpretability of technologies that use AI, the author Doshi-Velez and Kim [15] mentioned evaluation methods such as asking humans what AI output and explain what AI does (human-grounded), using mathematical formulae such as decision trees to understand the interpretability of AI (functionally-grounded), and finally involving domain experts to verify if AI has fulfilled its objective (application-grounded) [15]. This approach gives business manager an understanding on how to understand the logic and test the AI system.

Another research by Chaturvedi and Dasgupta [11] presents a framework for using AI in strategic decision making that takes into account factors such as social, organizational and individual and make a decision to accept or reject an AI solution.

Additionally, authors Chaturvedi and Dasgupta [11] mentioned key themes such as AI helps with the analysis of data to find trends and patterns but humans are needed to make the final decision, establish trust by bridging the communication gap between manager and employee, and replacement of humans due to AI capability [11]. In short, these are some of the factors that contribute to an organization's acceptance of an AI system.

Demonstrated above there are lot of parameters that guided managers to evaluate an AI system ranging from human-AI interactions [2], understanding AI systems in terms of interpretability [15] and drivers for accepting AI in organizations [11].

This helps us with a research gap on how can we provide managers with a structured, comprehensive framework to accommodate some key parameters to evaluate an AI agent and help managers make decision. By using and adapting BSC's multidimensional perspective [27], the research helps to fulfill the shortcomings into actionable insights to make decision whether to continue with development and deployment of project.

To realize this goal, the Design Science Research Methodology (DSRM) described by Peffers et al. [49] is used to guide the development of this framework. Given its systematic structured approach to problem solving, it follows a six-step process: (1) problem identification; (2) definition of solution objectives; (3) design and development; (4) demonstration; (5) evaluation; and (6) communication.

To help us with our research questions, we design some key framework objectives stated below:

1. **Shift evaluation focus from code level metric to business value:** We need to re-evaluate the AI agents in terms of strategic value rather than solely on code level metrics.
2. **Design a structured comprehensive assessment instrument:** Drive a go no-go decision by providing clear matrix formatted dimension criteria for all the four perspectives namely financial, user experience, internal business process and learning and growth areas. This will create a actionable, measurable and viable balanced scorecard consisting of questions.
3. **Evidence based adoption and scaling decisions:** Help business managers by creating a decision support structure that ensures managers can evaluate an AI agent in the early life cycle stages. As part of responsible AI we need to integrate "Human in the loop" so that AI has the ability to learn from human correction and integrate well into the existing workforce [11]. Apart from this, we also need to showcase that AI agents help the support staff of customer operations with routine tasks and that is done through an AI agent proof of concept.
4. **Strategic alignment with the organizational goals:** Finally, the research will aim to validate the frameworks effectiveness with interviews from stakeholders in practical world setting, specifically examining the application within FrieslandCampina's customer operations. The objective is to showcase how a generic management framework can be used to evaluate an AI agent for a specific industry and a use case ensuring that AI agents acts as an extended workforce of the corporate strategy rather than an isolated technology tools [14].

Framework Objectives

1. Shift focus from code level metric to business value.
2. Design a structured and comprehensive assessment instrument.
3. Enable an evidence-based adoption and scaling decisions.
4. Ensure strategic alignment with organizational goals.

Figure 4.1: Overview of the Framework Objectives

4.2 Case Assessment Scorecard Design

This section will cover the results of each analysis method that were represented in Chapter 3 and how they eventually led to the formation of dimensions and their related question for the scorecard.

1. **Business Group Strategy Data**

   From Section 3.2.4, we will understand the specific business group vision, mission, strategic focus points that are relevant to the customer operations business group in the Netherlands. To understand this, we will list the key elements that focuses on driving process and performance excellence, and build a future proof customer centric supply chain, below are the listed elements:

   - Vision: To be number one in customer centric global FMCG dairy supply chain and be data driven as well as future ready customer experience.
   - Mission: Deliver the agreed services to the customer everyday, in a sustainable and cost efficient manner.

- Aspiration: Serve the customers and consumers 24x7, ensure perfect service for our customers.
- Strategic Focus Points:
  Drive performance and process excellence
  - Ensure continuous improvement programs on main KPI's when these are rolled out.
  - Create end to end operational control while making use of our system landscape.

  Build future proof customer centric supply chain
  - Increase end to end performance to the customers by means of digitization.
  - Drive organizational efficiency.
- Enablers: Build digital capabilities that fit the current and future supply chain. Furthermore, build a diverse environment focused on individual development needs.

Based on the gathered data from management team, we will try to incorporate these pointers to arrive at its mission and objectives for each perspectives. We will now discuss each perspective's mission and its objective below

(a) Financial Perspective
  - Mission: The AI Agent should show meaningful value to the business [36]
  - Objectives:
    Ensure the demo project provides business value [36]
    Control the cost of the project [36]
    Be as cost efficient as possible to fulfill your goal with the project
    Cross collaborate with Digital and Business team to maintain and establish good image and reputation with the management [36]
    Establish clear cost benefits in terms of savings shown in Euros saved [36]

(b) User Perspective
  - Mission: Ensure the product or services built for the business team adds value [36]
  - Objectives:
    Ensure you consider end user inputs while building the solution [36]
    Fulfill the end user requirements
    Utilize the IT opportunities [36]
    Maintain and establish relationship with end user community [36]

(c) Internal Business Perspective
  - Mission: Ensure delivery of new services/product in an effective and efficient manner [36]
  - Objectives:
    Anticipate new requirements from the end user community [36]
    Provide cost effective training for the IT product/services [36]
    Ensure IT applications are developed without disrupting the current business activities [36]

(d) Learning and Growth Perspective
  - Mission: Continuous improvements and prepare for new ways of working [36]
  - Objectives:
    Upgrade your employee skills through training and development [36]
    Invest/Research into upcoming technologies and analyze potential use cases for business in a cost effective way [36]

2. **Initial Diagnosis** Initial analysis during one-on-one interactions with teams and business managers, revealed some key insights and findings about the current state of evaluation of IT projects.

   We discussed each point that reveals an "area of limited visibility" for managers. Because of limited visibility managers are unable to evaluate an AI agent holistically, and suggested to design a balanced scorecard with modified dimensions to diagnose the problems [22].

(a) **Value Creation Gap** Managers often used lagging indicators such as ROI, payback period and development cost as the primary indicator for the financial analysis, that is not enough when dealing with AI agent because there are hidden cost that needs to be considered such as scalability cost, the number of hours the AI agent will save.

This needs to be converted to an employee cost of work, savings generated by the agent. Furthermore, scalability for which we need to consider the processing cost, and additional resources to be scaled to different departments/business groups.

(b) **Employee-AI friction** Managers need to think about how can they trust, integrate and consider organizational level challenges when an AI agent is introduced for their daily workloads without disrupting the current business processes.

Managers want to see a short demo on how an AI agent behaves with few examples to gauge on AI output so that they can check the error rate, accuracy and how easy it is to work with an agent.

Furthermore, this demo also needs to be shared with their respective teams so that actual end users can experience working with AI agent and give valuable feedback with the interaction, experience and ease of use with the setup to know if it actually helps or is it creating more work for them.

(c) **Manager's Organizational Challenge** As the word of AI agent solution spread through the work floor, employees become concerned about their jobs and the new way of working which gave rise to another challenge that is skill development such as new training program for employees and new skill developments.

For senior managers it provokes them to think in terms of innovation capability of the business group to not be cost efficient but also how to better serve the customers by reducing the employee cognitive workload.

To summarize the above pointers, we created a table to link some of the problems that where identified and the diagnostic need as well as some resulting dimensions from it.

Table 4.1: Overview of problems, diagnostic need and connection with scorecard

Problems	**Diagnostic Need**	**Linked Scorecard Dimension**
Shadow Costs	Understand the marginal cost that needs to be considered when scaling or increasing the processing capability of AI Agent	Scalability (Financial Perspective)
Ease of use	Identify the experience and interaction when working with AI agent	User experience (User perspective)
Lack of Trust	Check on error rates, accuracy of AI agent output	Accuracy (User perspective)
Passive Dependence on AI	Ensure the agent acts as an extended colleague or an enabler that shifts the cognitive load to itself	Skill Development (Learning and Growth Perspective)
AI Demo Trap	Make sure the AI agent is demoed to different business groups to validate the novel approach and new ways of thinking	Innovativeness (Learning and Growth Perspective)

3. **Review Of Company Project Template and Workshop Setup**

As discussed in the section 3.2.5, we can observe that template concentrates more on financial parameters such as ROI and how much savings we can generate from the given idea or solution. This helps us to give rise to our next dimension which is Cost Efficiency and ROI/Value Creation under the financial perspective.

In the section 3.2.3, we conducted a small workshop setup where in we discussed about business process mapping, value creation and trust with AI as well as future way of working and organization readiness, these discussion helped us to gather dimensions based on Efficiency and Automation, User experience in regards to trust, and adoption towards AI, skill development and innovative approach to problem solving.

We will now combine the points discussed regarding the project template and workshop setup in a tabular format as seen in table 4.2 to see how these insights fit with the dimensions.

Table 4.2: Connecting project template and workshop insights with scorecard

Source	**What was identified?**	**What does this mean?**	**Linked Scorecard Dimension**
IT Project Template FrieslandCampina	Less FTE and Hard Savings	AI agent has lowered the manual effort cost and the overall cost of document processing activities, Operational savings in terms of hours saved (converted to cost)	Cost Efficiency (Financial Perspective)
IT Project Template FrieslandCampina	Extra Revenue, Other Value	AI agent help to free up employee time, benefits of using AI agent outweigh the implementation cost, increased productivity	Value Creation (Financial Perspective)
Workshop Activity (Business process mapping)	Time taken by document processor, automation level setup	Compared the time taken by human vs the document processor to know about efficiency and the level of automation(whether its an overkill or appropriate for the task)	Efficiency (Internal Business Process Perspective)
Workshop Activity (Business process mapping)	Potential integration with existing process, tools	Demo shows the integration of the workflow and how the collaboration between human and AI agent will happen	Integration with Existing Tools (Internal Business Process Perspective)
Workshop Activity (Business value creation and Trust with AI)	End user interaction with setup	Customer care officer gave a feedback on the experience, interaction and what they understood by confidence score, error rate	User Experience and Accuracy (User Perspective)
Workshop Activity (Future state of working and Organizational Readiness)	Novel approach, Innovative thinking, Task reduction	Managers encouraged new ways of working and explore, demo new ideas to different department, Employees mentioned this would reduce cognitive workload and they can focus on other tasks	Skill Development, Innovativeness, and Employee Experience (Learning and Growth)

4. **Insights from Literature Study**

From the sections 2.4, we understood that when working with LLMs we need to account for technical parameters such as accuracy, precision and recall scores so that users can estimate how accurate the data is and help them with error rates by looking at the AI output.

These parameters helped us to establish trust with the AI system by looking at the scores present in the output. It gave them confidence when the error rates are less to rely on the AI Agent which gave rise to two dimensions that is Accuracy and User Adoption.

Another section 2.5, this helped us to look at the sociotechnical parameters such as task success, reliability, and consistency that are highlighted by different authors such as Pandhare [47], Mohammadi et al. [42].

The below table 4.3 highlights how the technical parameters can relate to the scorecard dimension:

Table 4.3: Connecting some technical and sociotechnical parameters to scorecard dimensions

Parameters From Literature Review	**What this means in practice for our AI solution**	**Linked Scorecard Dimension**
Accuracy, Precision, Recall and Confidence Score	These parameters are embedded in the AI solution and help end user to associate how the solution output correlate to the numbers of the mentioned parameters; it helps them to check whether data is accurate for their use, establish trust based on statistical parameters by comparing the output and parameter scores, Check/calculate error rates if they are more or manageable	Accuracy and Adoption (User Perspective)
Reliability and Consistency	Check how well AI agent performs based on performance and task accuracy, check the robustness by adversarial prompts, interaction simulations to check if AI Agent drifts away from context	Consistency (Internal business Process Perspective)
Task Success (Pass@k Metric)	Do multiple runs of AI Agent on similar tasks and also during peak workload hours to ensure it gives consistent and reliable outputs	Consistency (Internal business Process Perspective)

4.3 Final Scorecard

As a key output of this chapter, the scorecard was created to support structured evaluation of AI agents for business managers at FrieslandCampina. This section represents the final form of the scorecard, grounded in the iterative process defined by Design Science Research Methodology (DSRM) as mentioned in the methods Chapter 3 under the section 3.2. The balanced scorecard contains four perspectives and captured key dimensions, as showcased in figure 4.2 and what each dimension measures is captured in the table 4.4:

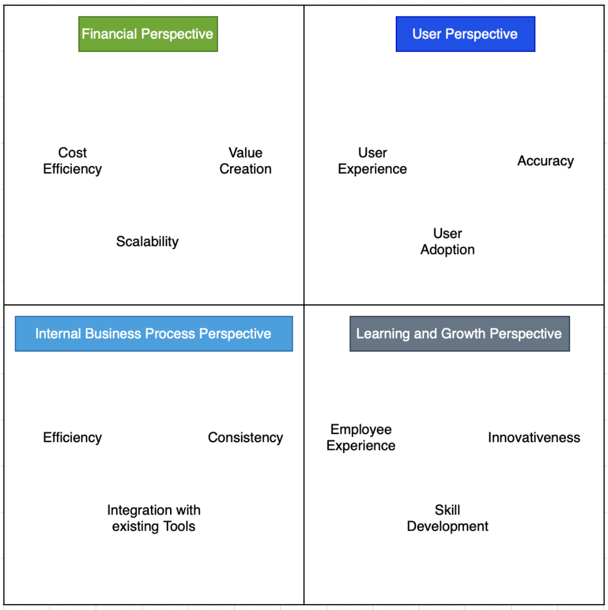


Figure 4.2: Final Scorecard

Table 4.4: Final Balanced Scorecard

Perspectives	Dimensions	Dimensions Measure
Financial Perspective	Cost Efficiency, Value Creation, and Scalability	• Cost Efficiency: Helps to measure the operational cost compared to support staff. • Value Creation: Measures the business value AI Agent brings as compared to the relative cost (implementation cost). • Scalability: Measure if the AI Agent can be scaled with a marginal cost increase.
Customer/User Perspective	User Experience, Accuracy, and User Adoption	• User Experience: Helps the manager to understand how easy an AI Agent is to use and how it fits with daily routine. • Accuracy: Helps us to know if the AI agent is producing correct results for operational use. • User Adoption: Helps us to know if the end user community has agreed to work with the agent and can it work without supervision.
Internal Business Process Perspective	Efficiency, Consistency, and Integration with existing tools	• Efficiency: Helps us to provide information on an AI Agent if it speeds up the document processing time and does it bring optimization. • Consistency: Helps us to understand if the Agent is reliable and consistent (operational stability) enough during high workloads and similar cases. • Integration with existing tools: Helps us to know if other system requires technical architectural change or AI Agent integrates smoothly with effective human AI collaboration.
Learning and Growth	Employee Experience, Skill Development, and Innovativeness	• Employee Experience: Measures if an AI Agent fulfills the end user objectives by taking away repetitive task and how the employee would feel about job satisfaction when AI onboard their work. • Innovativeness: Measures if an organization is ready to adopt and embrace new technology and be future ready. • Skill Development: Measures if an employee requires additional training for an AI solution or current trainings are enough to work with it.

Chapter 5

Evaluation Setup

In this chapter, the scorecard illustrated in figure 4.2 is evaluated to test its quality and usability at FrieslandCampina for a selective case. The evaluation is performed first by showing a demonstration of an AI agent and later by sharing a survey with participants in different roles, that will be the potential end-users of the framework, who will interact either directly or indirectly with it. Later, we discussed the statistical method as described in 3.2.7 to check the how the (balanced scorecard) BSC perspectives and dimensions which were part of the survey, evaluated for this research. The evaluation is done in three phases and listed below:

1. **Phase 1**: It consists of setting up the Power Automate workflow to showcase the technical ability on how AI Agent helps with the document extraction process for the customer operations team.
2. **Phase 2**: This consists of sharing a digital self administered questionnaire that is made with the help of Qualtrics system.
3. **Phase 3**: Evaluated the survey results via statistical methods mentioned in 3.2.6.

For the participants to effectively answer the questionnaire, we have executed the phases right after each other. For more detailed setup, screenshots and the digital survey containing all the questions and questionnaire contents can be found in the appendix section C.

5.1 Phase 1: Setting Up Power Automate Flow and Showcasing Result

In the first phase, the participants where showcased the Power Automate Flow that is built using Microsoft AI model and LLM powered instructions to extract data from the documents that is received on the shared mailbox of customer care operations team. The participants were asked to observe the AI Agent setup and it's results. This phase is performed in person, as it allows the participants to interact, understand, and ask non-standardized questions about the setup as well as share their opinion freely on the working of the agent. To know how the setup looks like and how results were showcased, the content is present in appendix section B.

5.2 Phase 2: Survey Results

After, phase-1 was conducted we launched the self administered survey to our participants. The appendix C contains all the survey questions. Below we share the survey results.

5.2.1 Survey Results

The second phase of the evaluation consist of a digital survey questionnaire mailed to participants after the Phase-1, this setup is self-administered, allowing the participants for a swift execution. Here, we present the participants with four perspective with each containing three dimensions and each dimension containing a survey statement relevant to measure the dimensions respectfully. The survey setup, consent form, questions, and questionnaire statements can be found in appendix section C.

1. **Financial Perspective**
   - Cost Efficiency: As seen from figure 5.1, the participants were in agreement that AI agent saves time, allows manager to re-purpose its employee time and has cost advantage when compared to an (full time employee) FTE.

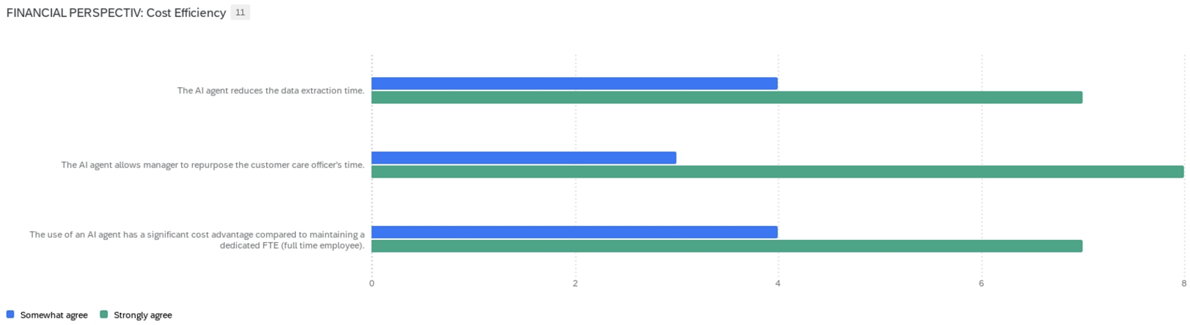


Figure 5.1: Cost Efficiency Dimension

   - Value Creation: From the figure 5.2, we observe some minor neutrality between the participants about the survey statements AI agent reduces employee burnout and increase productivity and the output generated by the AI agent but these are overshadowed when we combine the agreements from other participants. Lastly, we see participants agreed for the survey statement that AI agent improves the processing throughput.

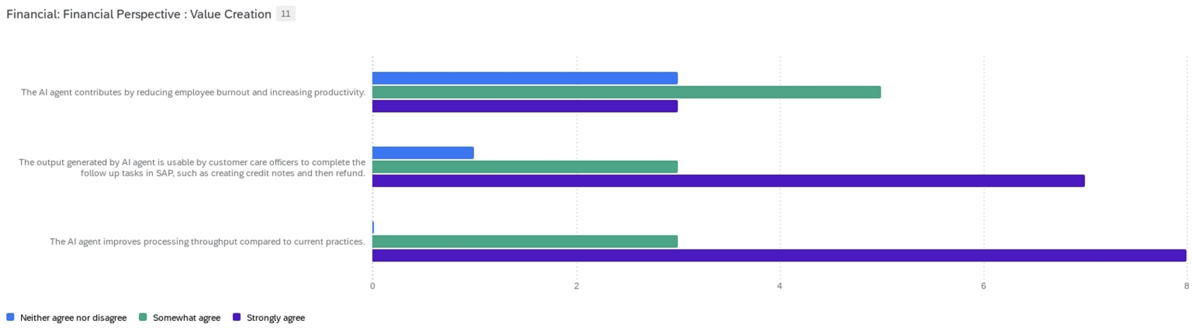


Figure 5.2: Value Creation Dimension

   - Scalability: From the survey figure 5.3, majority of the participants agree that AI can process documents compared to humans. There was also strong agreement that AI agent can adapt to new documents without much training and scaled to different business groups.

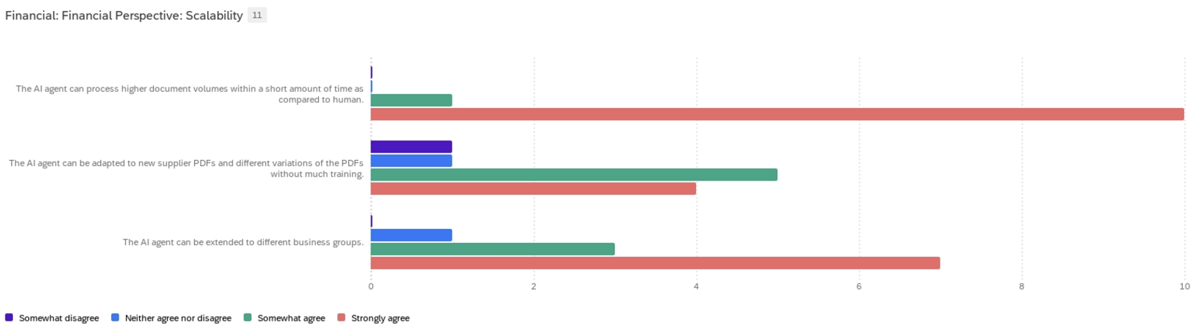


Figure 5.3: Scalability Dimension

2. **User Perspective**

- User Experience: From the figure 5.4, The first survey statement has mixed output wherein majority agreed but some are neutral and at least one doesn't agree with the Power Automate setup. For the second statement, there a high agreement between the participants about opening of the PDF when AI agent is used but few have a neutral stance. We see that majority of the participants agreed AI generates an output understood by the teams whereas only one took a neutral stand.

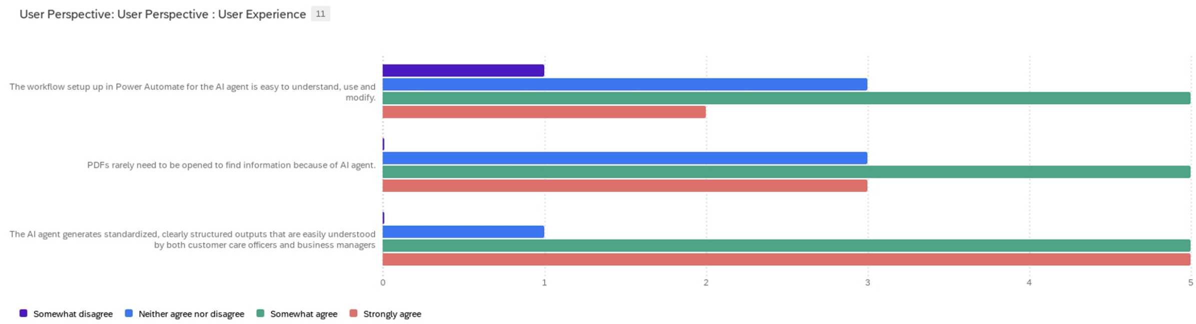


Figure 5.4: User Experience Dimension

- Accuracy: From the figure 5.5, we see that majority of the participants agreed that AI agent showed accurate data, fields extracted are usable and agent is likely to generate accurate output even for different suppliers whereas for similar survey statements observe only 2 participants took a neutral stance.

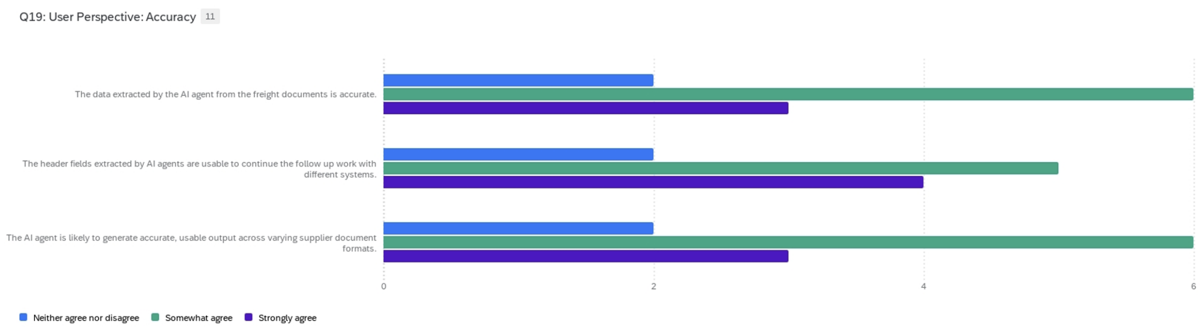


Figure 5.5: Accuracy Dimension

- Adoption: From the figure 5.6, we observe majority of the participant in agreement with the survey statements that measure user adoption. However just for the last survey statement only 2 participants tool a neutral stand.

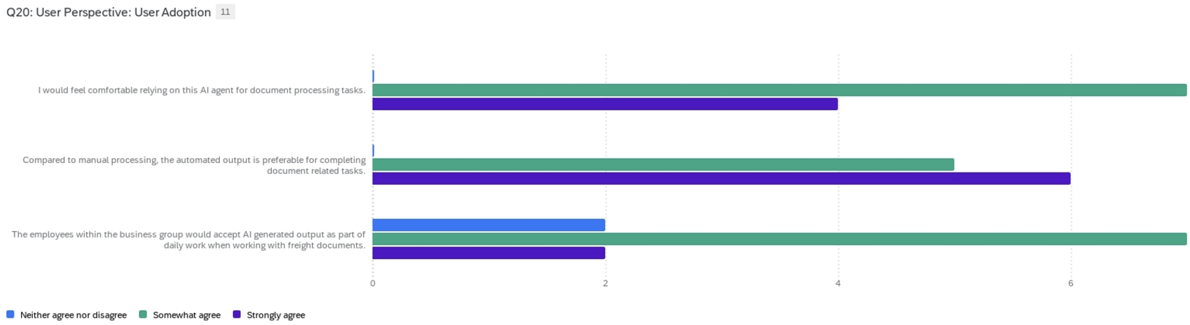


Figure 5.6: User Adoption Dimension

3. **Internal Business Process Perspective**

- Efficiency: From the figure 5.7, for the survey statements measuring efficiency that states AI has the potential to reduce manual steps, shorter processing time and reduce turnaround time for complaint handling, we noticed all participant agreed.

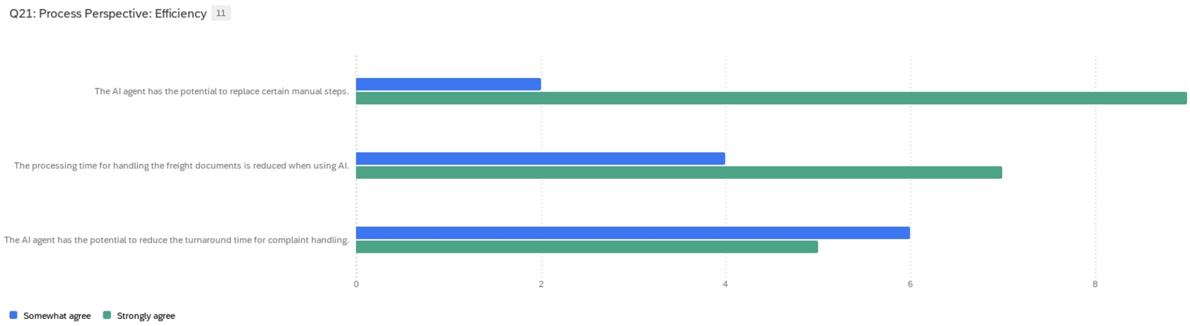


Figure 5.7: Efficiency Dimension

- Consistency: From the figure 5.8, only few participants took neutral stance that measured consistency from the three survey statements that are AI agent produced consistent results, output remained consistent and when trained on new documents the results were usable but majority were in agreement.

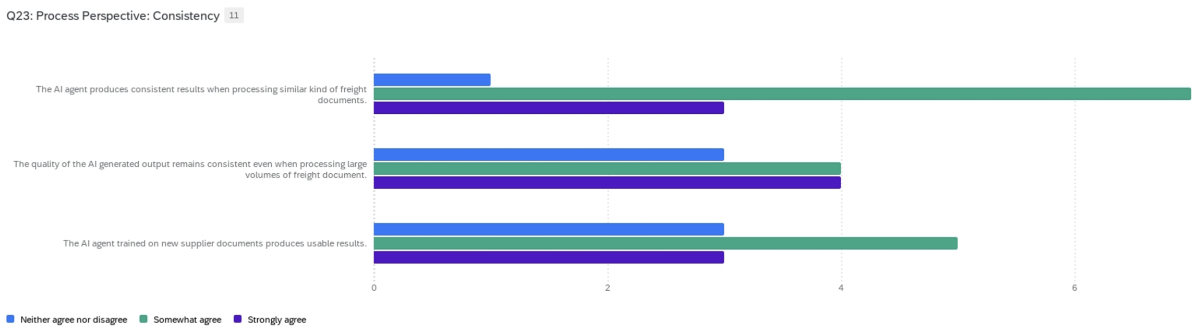


Figure 5.8: Consistency Dimension

- Integration with Existing Tools: As per figure 5.9, for the first survey statement we observe majority of the participants agreed that AI agent output can be used with existing Microsoft product. For the second survey statement, there was a single disagreement, four participants in agreement and others response were equally distributed between neutral stance and agreement to integrate the agent with SAP and Copilot Studio. Lastly, majority were in agreement and only three participants took neutral stance when they responded for the survey statement that states AI agent fits with current technology stack of the company.

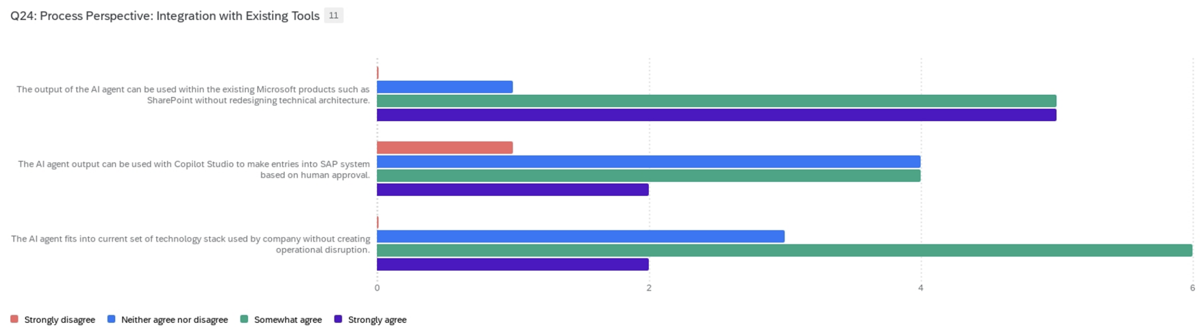


Figure 5.9: Integration with Existing Tools Dimension

4. **Learning and Growth Perspective**

   - Employee Experience: From the figure 5.10, we observe that majority participants agreed that AI agent reduces repetitive work. Similar trend was followed for the second survey statement that states AI agent allows employees to focus on high value tasks. Lastly, majority of the participants disagree that AI will have a negative impact to employee job satisfaction.

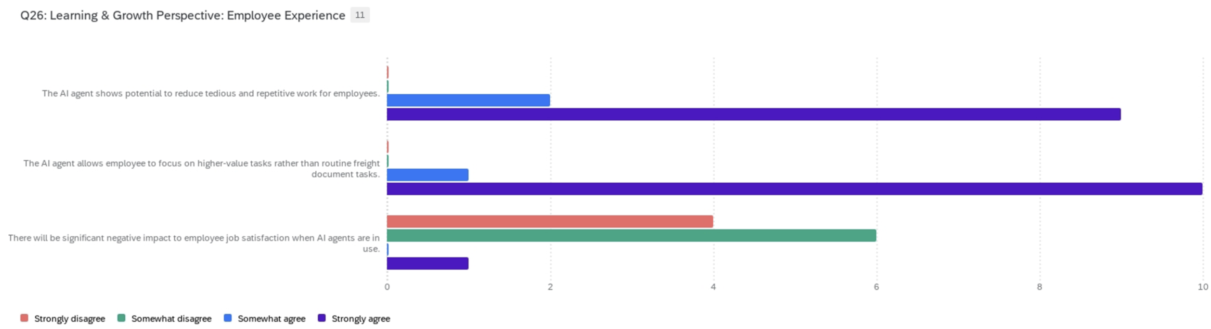


Figure 5.10: Employee Experience Dimension

   - Innovativeness: From the figure 5.11, we observe that majority of the participants agreed with all the three survey statement that measures Innovativeness.

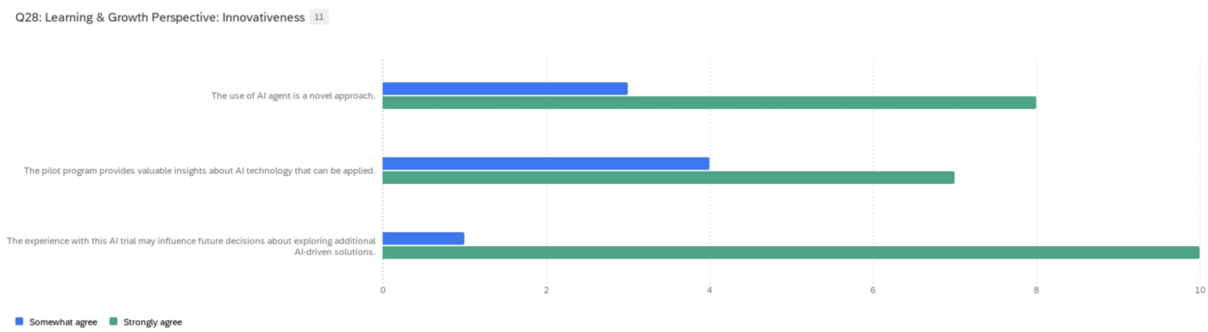


Figure 5.11: Innovativeness Dimension

   - Skill Development: From the figure 5.12, we observe for the first survey statement that participants are in agreement that employees are able to understand AI agent output on SharePoint.

We see a mixed response type with participants when we checked the second survey statement that states employees poses necessary skills to work with the AI tools. Lastly, majority of the participant agreed that it not much training is required to use the output of AI but two participants response were equally divided between neutral stance and disagreement for the last survey statement.

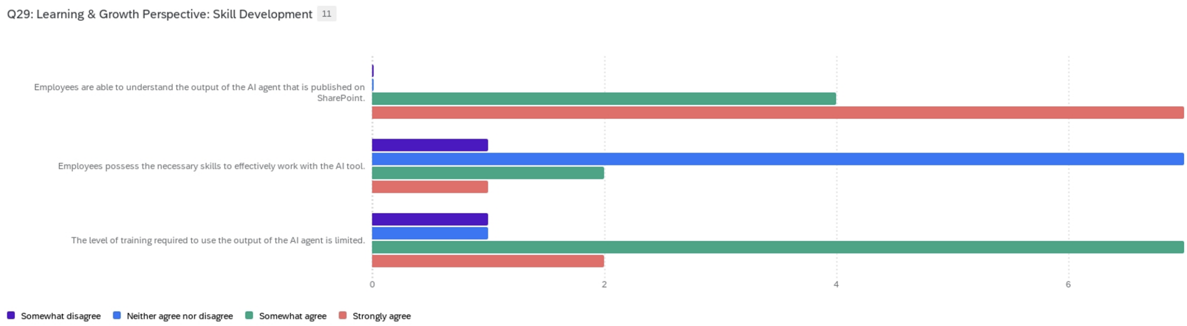


Figure 5.12: Skill Development Dimension

5.3 Phase 3: Evaluation Results

The evaluation was conducted with eleven participants, each holding a different job role within the same business group that is Europe and Retail Americas, who interacts with customer operations team either directly or indirectly. An overview of the participants can be found in the table 5.1.

The evaluation results will be divided into two phases that will cover reliability test of the survey results to measure its internal consistency. The second phase, will display the results from each perspective and the insights/observations gathered from it.

Participant	Framework Role	Formal Role
P1	CIS Manager	Continuous Improvement Specialist
P2	CIS Manager	Continuous Improvement Specialist
P3	Business Manager	Project Manager Commercialization
P4	Decision Scientist	Senior Data Scientist
P5	Business Manager	IT SAP Solution Specialist
P6	EDI Specialist	Distribution Specialist
P7	Business Manager	Customer Operations Manager
P8	Business Manager	Customer Management Officer
P9	Business Manager	Key User SAP
P10	EDI Specialist	Application Specialist
P11	Business Manager	Manager Customer Operations

Table 5.1: Evaluation Participants

Reliability Test Results

We discuss an important topic that is the reliability test of the item constructs to validate the survey questionnaire gathered from the survey results. As discussed in section 3.2.7, we showcased the practical use of the Cronbach's Alpha and Composite Reliability formula.

- **Cronbach's Alpha**: We conducted an analysis on the four perspectives to check if there's an internal consistency and reliability for the items and does it measure the each perspective correctly when using Likert-scale questionnaire.

  The results indicate, that all the Cronbach's alpha values are above the industry acceptable threshold level that is above 0.70, that signifies acceptable reliability of the four perspective: Financial Perspective(0.78), User Perspective (0.80), Internal Business Process Perspective (0.83), and, Learning and Growth Perspective (0.78).

These findings further strengthens that the Likert-scale questionnaire is reliable to measure an AI Agent with the help of a balanced scorecard approach, and, the numbers further bolstered that data collected is dependable and consistent for further analysis. The results are showcased in the table 5.2.

Table 5.2: Cronbach's Alpha Result For Four Perspectives

Perspective Name	**Survey Statements Per Perspectives**	**Cronbach's Alpha Score**	**Composite Reliability (CR)**
Financial	9	0.787	0.837
User	9	0.809	0.887
Internal Business Process	9	0.833	0.885
Learning and Growth	9	0.789	0.879

Even though, the numbers look satisfactory as per academic perspective but there are drawbacks when using this method such as it considers all the items/constructs used to evaluate each perspective contributes equally, doesn't account for the random noises/errors that can happen when a human is taking the survey such as survey fatigue, not understanding the underlying trait of the construct or similar wordings of the statements.

This is the reason why Cronbach's Alpha is flawed and we checked the results into more detail by using Composite Reliability at item/construct level.

- **Composite Reliability**: To overcome the assumptions/drawbacks of Cronbach's Alpha, we conducted a Composite Reliability test on dimension level (12 dimensions) to ensure the reliability of the survey instrument. This calculation was conducted alongside Cronbach's Alpha score and Factor loading, that helps to identify the reason behind low Alpha score levels.

  Having a small sample size (n=11), this makes Cronbach's Alpha highly sensitive and due to tau equivalence (respective survey items equally contribute to respective perspective). As shown in the table 5.3, we can observe Cronbach's Alpha score are below 0.70 for cost efficiency, User Experience and Integration with Existing tools as well as Employee Experience.

  As observed from the last column that is Factor Loading, the individual data of the survey statement reveals why the Cronbach's Alpha scores are low. It also helps us to establish that not every survey statement equally contributes to the dimension measure. This can also be observed from the table 5.3 that have low factor loading score.

Table 5.3: Composite Reliability Result For The Dimensions

Dimension Name	**Survey Statements Per Dimension**	**Cronbach's Alpha Score**	**Composite Reliability (CR)**	**Factor Loading**
Cost Efficiency	3	0.409	0.728	[0.554, 0.922, 0.554]
Value Creation	3	0.699	0.858	[0.584, 0.934, 0.903]
Scalability	3	0.684	0.878	[0.861, 0.731, 0.920]
User Experience	3	0.456	0.744	[0.266, 0.854, 0.895]
Accuracy	3	0.866	0.920	[0.876, 0.841, 0.950]
User Adoption	3	0.677	0.835	[0.880, 0.876, 0.600]
Efficiency	3	0.805	0.886	[0.796, 0.917, 0.834]
Consistency	3	0.619	0.777	[-0.038, 0.983, 0.982]
Integration with existing tools	3	0.377	0.722	[0.702, 0.561, 0.773]
Employee Experience	3	0.186	0.734	[0.913, 0.912, 0.090]
Innovativeness	3	0.806	0.888	[0.928, 0.893, 0.722]
Skill Development	3	0.803	0.885	[0.721, 0.841, 0.970]

We were curious to know what lead to the low Cronbach's Alpha score for the dimension **Employee Experience** an important factor under **Learning and Growth Perspective**, we checked the raw data and observed that most respondents were able to identify the negative construct used in statement 3 ("Negative Impact on Job Satisfaction") and associate the statement 1 and 2 were about positive experience with AI (reduces tedious work and Shift focus to higher value task).

Another observation regarding the dimension **Consistency** under **Internal Business Process Perspective**, has a negative factor loading (-0.038) for the survey statement 1 ("AI can easily handle similar documents") but it has high factor loading (0.983, 0.982) for statements 2 ("AI can handle large volumes") and 3 ("AI can handle new supplier format PDFs easily").

We again, looked at the raw data of the survey and observed that respondents ("Manager Customer Operation (P11) and CIS Manager (P2)") gave survey item one a higher score (5) but survey items 2 and 3 were lowered (3) but then the opposite happened with ("Customer Management Officer(P8) and CIS Manager (P1)"), they scored lower for survey item 1 but higher (5) for survey items 2 and 3 due to which it established a negative corelation with survey item 1 as per statistics it moved in the opposite direction and PCA analysis looks for the best fit.

In conclusion, we can mathematically say that our survey instruments is reliable and shows internal consistency when we use Cronbach's Alpha score at the four perspective level since the score is greater than 0.70 threshold value.

Furthermore, it also shows similar consistency and reliability when we use it at dimensions level as the Composite Reliability score is greater than the threshold value of 0.70 for each dimension.

Descriptive Statistics and Correlation Result

We did a descriptive analysis of the self administered digital survey results, the mean/average was calculated to measure the central tendency of participants sentiments across the four perspective of the balanced scorecard approach, while standard deviation helped with the spread of the scores across the dimensions by rating the survey statements using Likert-scale, thereby letting us to know the consensus of the participants whether they have agreed on the dimensions stated in the form of survey statements.

This result data help us to justify, this framework allows the managers to take a look beyond financial perspectives ($M = 4.54$) such as value creation, costs. It helps the management to take a look at other perspectives to help oversee the efficiency, integration and optimization of internal process perspectives ($M = 4.22$), also consider/monitor the end user community regarding their experience and adoption with an AI agent through user perspective ($M = 4.15$), and strategically manage the teams Learning and Growth ($M = 4.43$) by ensuring future workforce is ready to adapt new technologies to eliminate the repetitive tasks and focus on higher value objectives.

Additionally, we also conducted a standard deviation calculation on the four perspectives of the balanced scorecard to know about the general consensus if the dimensions support the perspectives and help identify any high polarized view among the respondent.

The standard deviation score ranged from (0.38 to 0.47) for the perspectives that demonstrates high managerial consensus within the participant group about the evaluation of AI agent via a balanced scorecard approach.

The results of the descriptive analysis containing the data for mean and standard deviation is mentioned on the table 5.4 for each perspective.

Table 5.4: Descriptive Statistics Result

Perspective Name	**Mean/Average Score**	**Standard Deviation**
Financial	4.54	0.38
User	4.15	0.44
Internal Business Process	4.22	0.47
Learning and Growth	4.43	0.39

Additionally, a correlation test is conducted to evaluate the relationship between the four perspectives namely: Financial, User, Internal Business Process, and Learning and Growth. The results show that all the perspective have positive relationship with each perspective validated the interconnected nature of the balanced scorecard.

A strong relationship between **Learning and Growth** and **User perspective** *(r=0.81)* is seen. This indicates that if AI agent shows higher accuracy and correctly extracts data from the freight documents, managers believe it can actually reduce the repetitive and tedious tasks, which in-turn would help the employee to move to higher value tasks thereby reducing the burnout and adding value to the company.

Additionally, moderate to strong relationship was noticed between **Internal Process**, **Financial** *(0.54)*, and **Learning and Growth perspective** *(r= 0.59)* outcomes. This helps the business managers that to achieve value creation (Financial Perspective) we need to integrate the workflows in our existing tools (Internal Business Perspective) that would eventually help with daily operations and embrace technology (Learning and Growth Perspective).

Surprisingly, weakest correlation relationship was established between **User perspective** and **Internal Business Process perspective** *(r=0.19)*, that indicates that though for the dimension Accuracy (dimension under User perspective) scores were high but the autonomous integration with Copilot Studio making entries into SAP mentioned under Integration with Existing Tools (dimension under Internal Business Process perspective) scored low, that means managers want human in the loop connection rather than making automatic entries in the ERP system.

The results of the correlation analysis for each perspective is mentioned in the table 5.5.

Table 5.5: Pearson Correlation Result

Perspective Name	**Financial**	**User**	**Internal Process**	**Learning and Growth**
Financial	1	0.38	**0.54**	0.37
User	0.387	1	0.19	**0.81**
Internal Process	0.54	**0.19**	1	0.59
Learning and Growth	0.37	0.81	**0.59**	1

Chapter 6

Discussion

In this section, we interpreted and put the results in context from the section 5.3 and discussed the implications in terms of how well the research was able to answer the key research questions and how the results supported the framework objectives that were mentioned in section 4.1. Followed by limitations of the framework and research then concluded the discussion with suggestions for future work.

6.1 Research Questions

We initially defined three research questions mentioned in section 1.2 to address our problem statement mentioned in section 1.1. The first two research questions (RQ1, RQ2) were discussed in the same section due to its similarity and the research approach taken to answer it. Lastly, we answered the research question three (RQ3).

1. What key dimensions should be considered when evaluating AI agents used in customer operations from a managerial perspective?
2. How can these dimensions be integrated into a structured framework that enables business managers to systematically evaluate AI agents in real-world contexts?
3. How effective is the proposed framework when applied in an organizational setting, such as FrieslandCampina's customer operations team?

6.1.1 RQ1 and RQ2

The main goal of the research was to explore how can business managers be supported by structured multidimensional framework that guides them to evaluate an AI agent which is currently in its early development stages, enhancing the manager's ability to assess the business value and feasibility for further development.

We addressed a specific problem: currently managers often make decisions based on just two parameters that are cost of development and savings that the project will bring but often ignore other perspectives such as value creation, user experience, user adoption and integration with other business processes. Evaluating on these parameters is quite critical as these decisions help with go/no-go decisions for the AI agent project.

To address the gap, we first conducted a literature review that is covered in section 2.7. This helped us to understand how AI agents are currently evaluated, we got insights into technical parameters such as F-1 score, accuracy, recall, precision and sociotechnical parameters such as robustness, safety, fairness, reliability and task performance but all these were siloed parameters that were derived from different AI evaluation frameworks.

Moving away from literature review, we conducted mini interviews, weekly one on one's with business manager and having spent few months with the customer operations team, we learned that evaluating an AI agent requires a multidimensional approach that goes beyond the financial measures. While values creation, cost efficiency and scalability measures in terms of financial perspective remains important for a business manager but this study helps us to look at different perspectives such as user, internal business process, learning and growth that also plays a key role in evaluating an AI agent.

Based on literature review, surveys, mini interviews and workshops we were able to get all the necessary dimensions and group them in respective perspectives for a balanced scorecard framework as discussed in Chapter 4, where we cover all the dimensions that needs to be evaluated and then group it into respective scorecard perspective which helps the business manager to evaluate an AI agent.

We need to structure all the dimensions identified in RQ1, this study used a design science research methodology (DSRM) approach to arrive at a final scorecard as illustrated in figure 4.2. We translated the theoretical framework of the scorecard into measurable constructs like designing a qualitative survey containing survey statements for each dimensions for respective perspectives, allowing the us to capture specific data points for the balanced scorecard perspective performance.

We used a Likert-scale 5 point method to evaluate the dimensions for each perspective, the questionnaire helped the managers to evaluate an AI agent and make a decision for further development of the project. The validity of the survey was measured through statistical analysis such as cronbach's alpha and composite reliability scores, both test were used to check the internal consistency and reliability of the survey basically helping us to measure what we set out to measure.

When we discussed the final scorecard and results from the survey with different participants such as customer operations manager, customer operations lead and senior managers. They all had different perspectives for the final scorecard, for example customer operations manager focused more on the potential of the AI agent to reduce the tedious and rote task for the employees and later delve into financial aspect such as the cost savings the project might bring in, understand the cost in scaling the solution to different teams.

On the other hand senior managers focused more towards the organizational readiness such as experimentation with new AI technologies and learn the impact on employees whether the workforce is ready to work with AI basically helping them plan learning modules for AI, responsible use of it and focus on new skill development for the workforce.

Furthermore, the customer operations lead would focus more on the user perspectives as this solution would be used by their team members so the focus on ease of use, accuracy and trust (User perspective) became more concerning for them rather than other perspectives in play.

6.1.2 RQ3

The effectiveness of the proposed framework was tested first statistically via standard deviation, mean, Pearson's correlation tests and performance of the balanced scorecard that confirmed the overall success of the framework being highly effective in evaluating an AI agent and helped managers make decision for the project.

We saw a positive sentiment and engagement with the framework survey questions as the mean scores were greater than 4.15 and low standard deviations ($SD < 0.50$) across the four perspectives that gave us a quantitative proof that the participants were in consensus that AI agent can benefit the teams, employees and provides value. Furthermore, we did a correlation test to show that all the dimensions were interconnected with each other and had a positive relationship with other perspectives.

A particular observation was made between the relationship of user and internal business process perspective where the score was 0.19, a very low score. Upon discussions with IT manager and business manager we realized the reason for low score, the company is going through SAP upgrade from ECC to S/4HANA due to which the integration into SAP via tools like copilot studio was not possible despite the AI agent scoring high in accuracy. Along with this, the company focused on "Human In the Loop" where in a direct autonomous entry will not be allowed because of responsible use of AI.

Even though we gave the business managers the BSC framework to evaluate an AI agent but the decision for further development was still made on financial savings, cost of development and productivity value it brings to the employee but this experiment opened new doors for the company to experiment more with new emerging technologies such as AI, Gen-AI and tools like copilot studio, M365 copilot. This made the managers realize on how to incorporate AI in their business processes and offload tedious, repetitive and manual tasks to AI agents.

This scorecard inspired them to consider end user community inputs in terms of ease of use, display of confidence level so that support staff can make meaningful decisions when working with an AI solution. Apart from this it motivated senior management to improve the current way of working when AI comes into picture and what skills are needed for the workforce to work with new/upcoming technologies

In conclusion, the framework statistically succeeded in the tests but the practical implementation could be tough considering the presence of an already existing IT project template decided by management and a strict focus on savings generated by the IT project. However, on the bright side it gave managers new insights on the possible use cases for AI agents that can be grouped together to be further developed by the digital team.

6.2 Framework Objectives

We defined four framework objectives in the section 4.1, now we will discuss to what extent these have been achieved.

- Shift evaluation focus from code level metric to business value: We have successfully achieved this framework objective, we have captured necessary dimensions from a managers perspective that add value to the business. Moreover, the dimension like "Cost Efficiency" wherein AI agent showcases significant cost advantage compared to full time employee (FTE) captured a key metric important for business managers.
- Design a structured comprehensive assessment instrument: We have successfully achieved the objective by designing a balanced scorecard for business managers to evaluate an AI agent, this was supported by designing a detailed survey questionnaire to capture the overall performance of the scorecard and help them make go/no-go decisions.
- Evidence based adoption and scaling decisions: The framework was partially successful as it gave a structured decision support structure to evaluate an AI agent early in its development stage and showcase potential to reduce the tedious or repetitive task performed by support staff. However, due to ongoing SAP upgrade to S/4HANA, we were unable to scale the presented solution to other teams and integrate well into existing business processes/tools, this is also indicated with a lower score for the internal business perspective. However, it inspired the senior management and business manager to develop use cases for AI and upgrade their workforce with skills, training and be ready to work with new/upcoming technologies in the AI/Gen-AI space to provide excellent customer service and motivated them to look at cases that creates strategic value for the organization.
- Strategic alignment with the organizational goals: We were partially successful to achieve this framework objective as one of the key strategic focus for the department was the use of AI and ensure continuous improvement with a business process and drive organizational efficiency but the practical usage of the framework in a practical setting would be difficult considering the tedious task of developing questions and getting survey response from participants. Furthermore, the framework inspired them to look for areas within the department on were to use AI and how to make use of AI agent as an extended colleague so that it takes care of repetitive and manual tasks.

6.3 Limitations

This section covers the limitations in relation to the research process. It first examined the process of framework creation and questionnaire followed by the limitations of the evaluation processes. Finally, it examined the theoretical limitations of the proposed framework.

6.3.1 Limitations in questionnaire and scorecard

The development of the balanced scorecard was inspired via a specific business problem addressing credit note handling for customers in the operational environment of FrieslandCampina's customer operations team. The evaluation setup, designing of questionnaire specific to the problem we addressed, the interviews, workshops and the way of working/business processes were all developed with a specific objective in mind related to FrieslandCampina's freight document case.

Although, this ensured good case and contextual fit, it gave a limitation in generalization of scorecard, survey setup and overall performance of the balance scorecard that may not directly be transferrable to a new problem statement without modification to the dimensions.

Additionally, the company is going through digital transformation and cost cutting due to which stakeholders were busy with high workloads and competing business priorities. This led to a restricted sample size, and scheduling interviews was bit challenging. While the data gathered provided valuable insights but a larger sample size or extended study of the topic would have yielded a more detailed understanding of the dimensions for the scorecard.

Due to restricted sample size, the survey statement showed some polarization of views especially for the integration of the solution with copilot studio, which affected our internal business process perspective score, we tried to address this by conducting an experimentation of the setup before answering the survey questions.

Finally, the length of the survey statement consisted of thirty six questions, completing a digital survey could induce cognitive load or respondent fatigue on participants. This could deplete cognitive resources, potentially leading to a less thoughtful response towards the end of the survey. This was done to ensure data completeness as it was useful for us to run quantitative analysis on the results, we tried our best to split the survey into four perspectives to reduce the cognitive load.

6.3.2 Theoretical limitations of the framework

In this section, we covered the basic inherent quality carried over by balanced scorecard, basically theoretical limitation of the framework.

Firstly, the study inherits a particular quality of the balanced scorecard developed by Kaplan and Norton [27], that mentioned the sequential, causal linkage between the perspectives, which states improving on one perspective such as Learning and Growth perspective will bring process improvements and help to create financial value. The case study at FrieslandCampina tells a different story, if we invest in employee skill development to get them ready to work with AI might not have a direct impact on process improvements and would not create a financial value for the company.

Secondly, the challenge of capturing the perspectives of different types of users. Accommodating the diverse viewpoints of various user groups necessitates adaptations or modifications to the framework. Ideally, this framework tries to capture all the essential needs from the different users ranging from EDI specialist, Continuous Improvement Service Manager to Senior Managers.

For Customer Operations Managers were more concerned on the cost factor and savings the solution brings whereas Senior Managers were focused on organizational readiness parameters, opposed to EDI specialist or CIS Managers who were concerned on the dimensions of the user perspectives. This raises a limitation on whether we were able to address all the requirements from the users with the help of this framework.

6.4 Future Work

While the study successfully demonstrated the use of multidimensional framework for evaluating an AI agent that helped business managers to make go/no-go decisions, this section covers the following framework design refinements that can be applied beyond the current setup and suggestions for future research.

6.4.1 Framework design refinements

This section outline two specific areas for refining the framework : generalization of the balanced scorecard and taking a new approach for the balanced scorecard perspectives.

Firstly, generalization of the balanced scorecard so it can be applied beyond FrieslandCampina's setup. The key dimensions and the survey statements for each of the dimensions needs to changed as per different project types and the organization objectives [22]. Adjustment's to scorecard to accommodate the following dimensions as stated by Shukla [57] in the research paper include: tool usage, communication between agents,human in the loop trust scoring and economic sustainability parameters such as cost per outcome, energy consumption needs to be added to the relevant perspective for future research.

Regarding different sectors, the framework aims to encourage organizations to look beyond mere financial measures, considering not only their business interests but also the interests of employees, end user community and integration with the existing technology stack [22]. Future research could investigate which additional dimensions, beyond those associated with the current applications, must be adapted to ensure the effectiveness of the framework across different sectors [22].

Another option, the framework was inspired from the proof of concept that was developed via Power Automate tool but if the solution is more information system oriented then we need to take a new approach based on the research paper by Martinsons et al. [36], that is an extension of the Kaplan and Norton [27] balanced scorecard, wherein the author Martinsons et al. [36] refine the perspectives and add different type of missions and objectives related to information systems, this approach could be more suited for AI solutions that are developed through Microsoft's Azure setup.

6.4.2 Suggestions for future research

Two suggestions for the future research are mentioned, focusing on framework utilization and organizations sizes.

Firstly, framework evaluation can be done at smaller organizations, particularly those that use similar setup but have different tools like n8n (kind of similar to Power Automate) to extract data from freight documents/invoices/billings with different sectors other than (fast moving consumer goods) FMCG. These organizations could benefit a lot from this type of approach to evaluate a tool that uses AI agent and make decisions whether to further investigate or go for premium licensed versions.

Lastly, further research needs to be conducted to focus on how the framework is utilized in practical settings. For example, studies could examine at what stages of the development of AI agent this framework is used such as early stages and post-deployment. Also, notice how varying user needs are addressed and which type of users can make most use of it. The real world setting test would determine if the framework is measuring what it's supposed to.

Chapter 7

Conclusions

The research is aimed to design, validate, and operationalize a multidimensional structured framework to evaluate an AI agent during its early developmental phase so that business managers can make a clear go/no-go decisions for the project.

Many existing evaluation frameworks already exists that helped decision scientist to either fine tune the AI model or evaluate whether AI is technically viable such as measuring its algorithmic accuracy, precision, and latency. However, for business managers, these evaluation approaches/metrics are often tool technical and isolated.

The business managers needs a holistic framework that is able to measure if an AI solution can be trusted by employees, whether it smoothly integrates in current business process and existing tools, or it ultimately drives financial success, ensuring that the evaluation of the AI solution/project is directly linked to the organization's broader strategic business objectives.

Using freight document AI proof of concept as a case study at FrieslandCampina addresses the gap by bridging technical parameters with strategic objectives set by business manager. The framework adapts balanced scorecard approach that empowers business managers to evaluate an AI agent project during their early life cycle stages. By transforming the qualitative data of the employee consensus and sentiment into hard, actionable data, it guides the business managers beyond anecdotal feedback, enabling them to make clear, evidence based go/no-go decisions before moving to full scale enterprise integration.

To develop this framework, a design science research methodology was used, that led to three key components: (1) a multidimensional structured balanced scorecard, (2) a statistical validation of employee consensus and understanding relationship between different perspectives, and (3) data driven, approach based on overall performance of the balanced scorecard.

Towards answering our problem statement, we addressed three primary research that are as follows:

1. What key dimensions should be considered when evaluating AI agents used in customer operations from a managerial perspective?

   This was addressed by using previous literature study, interviews and conducting workshops to showcase the proof of concept and later mapping the dimensions to the respective balanced scorecard perspectives namely financial, user, business process and learning and growth. These dimensions prove that to evaluate an AI agent from business standpoint it needs a multidimensional approach and cannot be measured in isolation. The success depends entirely on capturing the financial measures such as cost efficiency, savings generated and non-financial measures such as integration with current business process/tools, employee experience, providing a consistent and holistic way to assess these digital AI solutions.

2. How can these dimensions be integrated into a structured framework that enables business managers to systematically evaluate AI agents in real-world contexts?

   We first mapped all the respective dimensions into balanced scorecard perspectives, and then translating these dimensions into a granular qualitative survey instrument, where a single perspective contained three dimensions were measured using three survey statements. By structuring the evaluation into specific dimensions (for example accuracy, user adoption, and user experience mapped to user perspective), the framework acts as a way to measure the employees consensus, sentiments into measurable, actionable data points regarding the AI agent.

3. How effective is the proposed framework when applied in an organizational setting, such as FrieslandCampina's customer operations team?

The statistical analysis of the framework of the survey statements data proved the framework to be highly successful that confirmed employees are ready to work with AI solutions, saves time, and management would like to explore new ways of working with AI but more importantly the correlation test revealed a weak correlation between perceived AI accuracy and integration with existing workflows/tools, the survey data allowed us to identify this relationship and also relate it with the digital transformation the company is going through.

Simply, put this means even though the accuracy scores are high but integration with other systems will be done after SAP S/4HANA transformations. This revealed a critical insight that allowed managers to avoid a risky, fully autonomous solutions and instead work with "Human in the loop" component. Moreover, the overall performance of the balanced scorecard (0.87) helped customer operations manager to go ahead and collaborate with digital team to build a solution that would address the freight document cases and eventually move away from proof of concept solution.

From a scientific perspective, this work bridges a critical gap that is merging few technical parameter and some sociotechnical measures when evaluating an AI agent by structuring the different constructs into respective perspectives of the balanced scorecard that includes financial perspective, user perspectives, business process perspectives and parameters related to organizational readiness. It gives a clear, interpretable connection between different perspectives (financial, user, business process and learning and growth) and how the structured way of evaluation helps business managers to make decisions. From practical standpoint, it offers ready to use comprehensive structured decision support system for business managers, designed to make decisions that results in clear measurable actions for a particular AI agent project.

The future research should focus on longitudinal study with multiple AI agent projects to see if the overall performance of the balanced scorecard. Additionally, we need to check if the relationship between different perspective would change once the company moves away from ECC setup to SAP S/4HANA, that would help us to understand the relationship change between user perspective and integration of solution with existing tools.

Finally, expanding this structured evaluation to other functional departments with some obvious change in survey statements to fit the context, this will help the study not only be academically relevant but also contribute to the real world scenarios when managing future AI solutions.

Bibliography

[1] Hanan Alhaddi. Evolution of performance measurement research: An update on research development from 2005 to 2020 and future outlook for the field. *Operations and Supply Chain Management: An International Journal*, 16(3):399–412, 2023. doi: 10.31387/oscm0540399.

[2] Saleema Amershi, Daniel Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, et al. Guidelines for human-ai interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019. doi: 10.1145/3290605.3300233.

[3] David Autor. Polanyi's paradox and the shape of employment growth. Working Paper w20485, National Bureau of Economic Research, September 2014.

[4] Aaron Baird and Likoebe M Maruping. The next generation of research on is use: A theoretical framework of delegation to and from agentic is artifacts. *MIS Quarterly*, 45(1):315–341, 2021. doi: 10.25300/MISQ/2021/15882.

[5] Imran Batada and Asmita Rahman. Measuring system performance & user satisfaction after implementation of ERP. *Informing Science: The International Journal of an Emerging Transdiscipline*, 15:603–611, 2012. doi: 10.28945/1679. URL `https://www.researchgate.net/publication/320655114_Measuring_System_Performance_User_Satisfaction_after_Implementation_of_ERP`.

[6] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623. ACM, 2021. doi: 10.1145/3442188.3445922.

[7] Mary Jo Bitner, Amy L Ostrom, and Matthew L Meuter. Implementing successful self-service technologies. *Academy of Management Executive*, 16(4):96–108, 2002. doi: 10.5465/ame.2002.8951333.

[8] Kathrin Blagec, Georg Dorffner, Milad Moradi, and Matthias Samwald. A critical analysis of metrics used for measuring progress in artificial intelligence, 2021. URL `https://arxiv.org/abs/2008.02577`.

[9] Erik Brynjolfsson, Daniel Rock, and Chad Syverson. Artificial intelligence and the modern productivity paradox: A clash of expectations and statistics. Working Paper 24001, National Bureau of Economic Research, Cambridge, MA, November 2017. URL `http://www.nber.org/papers/w24001`.

[10] Erik Brynjolfsson, Danielle Li, and Lindsey R. Raymond. Generative ai at work. *Quarterly Journal of Economics*, 140(2):889–942, 2025. doi: 10.1093/qje/qjae044. URL `https://academic.oup.com/qje/article/140/2/889/7990658`.

[11] A. Chaturvedi and M. Dasgupta. Managerial perception of AI in strategic decision-making. *Journal of Computer Information Systems*, pages 1–13, 2024. doi: 10.1080/08874417.2024.2418403.

[12] Lee J. Cronbach. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334, 1951. doi: 10.1007/BF02310555.

[13] Thomas H Davenport and Rajeev Ronanki. Artificial intelligence for the real world. *Harvard Business Review*, 96(1):108–116, 2018.

[14] Fabien De Geuser, Stella Mooraj, and Daniel Oyon. Does the balanced scorecard add value? Empirical evidence on its effect on performance. *European Accounting Review*, 18(1):93–122, 2009. doi: 10.1080/09638180802481698.

[15] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[16] Carla Ferraro, Vlad Demsar, Sean Sands, Mariluz Restrepo, and Colin Campbell. The paradoxes of generative ai-enabled customer service: A guide for managers. *Business Horizons*, 67(5):549–559, 2024. doi: 10.1016/j.bushor.2024.04.013.

[17] Andy Field. *Discovering Statistics Using IBM SPSS Statistics*. SAGE Publications, London, 5th edition, 2018.

[18] G. E. Gignac, T. C. Bates, and K. Jang. Implications relevant to CFA model misfit, reliability, and the five-factor model as measured by the NEO-FFI. *Personality and Individual Differences*, 43(5): 1051–1062, 2007. doi: 10.1016/j.paid.2007.02.024.

[19] Carlos F. Gomes, Mahmoud M. Yasin, and João V. Lisboa. A literature review of manufacturing performance measures and measurement in an organizational context: A framework and direction for future research. *Journal of Manufacturing Technology Management*, 15(6):511–530, 2004. doi: 10.1108/17410380410547906.

[20] J. M. Graham. Congeneric and (essentially) tau-equivalent estimates of score reliability: What they are and how to use them. *Educational and Psychological Measurement*, 66(6):930–944, 2006. doi: 10.1177/0013164406288165.

[21] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5): 1–42, 2018. doi: 10.1145/3236009.

[22] Rashedul Hasan and Tai Chyi. Practical application of balanced scorecard-a literature review. *Journal of Strategy and Performance Management*, 5:87–103, 09 2017.

[23] Ming-Hui Huang and Roland T. Rust. Artificial intelligence in service. *Journal of Service Research*, 21(2):155–172, 2018. ISSN 1094-6705. doi: 10.1177/1094670517752459.

[24] Soleyman Iranzadeh, Sadegheh Hosseinzadeh Nojehdeh, and Nahideh Najafi Emami. The impact of the implication of balanced scorecard model (BSC) in performance of the post company. *Problems and Perspectives in Management*, 15(4):188–196, 2017. doi: 10.21511/ppm.15(4-1).2017.02.

[25] Ankur Joshi, Saket Kale, Satish Chandel, and D. K. Pal. Likert scale: Explored and explained. *British Journal of Applied Science & Technology*, 7(4):396–403, 2015. doi: 10.9734/BJAST/2015/14975.

[26] Loso Judijanto, Meiske Wenno, Muhammad Faisal, Hasmia Melati Arifin, and Erfendi Regar. Evaluation of financial performance with the balanced scorecard approach in retail companies. *West Science Business and Management*, 2(3):1031–1037, September 2024. URL `https://wsj.westscience-press.com/index.php/wsbm`.

[27] Robert S. Kaplan and David P. Norton. The balanced scorecard—measures that drive performance. *Harvard Business Review*, 70(1):71–79, 1992.

[28] Katherine C Kellogg, Melissa A Valentine, and Angèle Christin. Algorithms at work: The new contested terrain of control. *Academy of Management Annals*, 14(1):366–410, 2020. doi: 10.5465/annals.2018.0174.

[29] Chadi Khneyzer, Zaher Boustany, and Jean Dagher. Ai-driven chatbots in crm: Economic and managerial implications across industries. *Administrative Sciences*, 14(8):182, 2024. doi: 10.3390/admsci14080182. URL `https://doi.org/10.3390/admsci14080182`.

[30] Hayoen Kim and Sang Woo Lee. Investigating the effects of generative-ai responses on user experience after ai hallucination. *PEOPLE: International Journal of Social Sciences*, 2024:92–101, 2024. doi: 10.20319/icssh.2024.92101.

[31] Georgy P. Kurien and M. N. Qureshi. Study of performance measurement practices in supply chain management. *International Journal of Business, Management and Social Sciences*, 2(4):19–34, 2011. URL `https://api.semanticscholar.org/CorpusID:18146255`.

[32] Mary C Lacity and Leslie P Willcocks. A new approach to automating services. *MIT Sloan Management Review*, 58(1):41–49, 2016.

[33] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. What do we want from explainable artificial intelligence (XAI)? A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary research. *Artificial Intelligence*, 296:103473, 2021. doi: 10.1016/j.artint.2021.103473.

[34] Rensis Likert. A technique for the measurement of attitudes. *Archives of Psychology*, 22(140):5–55, 1932.

[35] Andrea Lorden, Alberto Coustasse, and Karan P. Singh. The balanced scorecard framework: A case study of patient and employee satisfaction. *Health Care Management Review*, 33(2):145–155, 2008. doi: 10.1097/01.HMR.0000304502.65969.34.

[36] Maris Martinsons, Robert Davison, and Dennis Tse. The balanced scorecard: A foundation for the strategic management of information systems. *Decision Support Systems*, 25(1):71–88, 1999. doi: 10.1016/S0167-9236(98)00086-4. URL `https://www.sciencedirect.com/science/article/pii/S0167923698000864`.

[37] Roderick P. McDonald. *Test theory: A unified treatment*. Lawrence Erlbaum Associates, Mahwah, NJ, 1999.

[38] Jack McKinlay, Marina De Vos, Janina A. Hoffmann, and Andreas Theodorou. Understanding the process of human-ai value alignment. *arXiv preprint arXiv:2509.13854*, 2025. doi: 10.48550/arXiv.2509.13854. URL `https://doi.org/10.48550/arXiv.2509.13854`.

[39] Daniel McNeish. Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, 23(3):412–433, 2018. ISSN 1082-989X. doi: 10.1037/met0000144.

[40] Microsoft. What is Power Automate?, 2025. URL `https://learn.microsoft.com/en-us/power-automate/flow-types`. Accessed: 2026-02-22.

[41] Microsoft. What is SharePoint?, 2026. URL `https://support.microsoft.com/en-us/office/what-is-sharepoint-97b915e6-651b-43b2-827d-fb25777f446f`. Accessed: 2026-02-22.

[42] Mahmoud Mohammadi, Yipeng Li, Jane Lo, and Wendy Yip. Evaluation and benchmarking of llm agents: A survey. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, page 6129–6139, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400714542. doi: 10.1145/3711896.3736570. URL `https://doi.org/10.1145/3711896.3736570`.

[43] Andy Neely. The evolution of performance measurement research: Developments in the last decade and a research agenda for the next. *International Journal of Operations & Production Management*, 25(12):1264–1277, 2005. doi: 10.1108/01443570510633648.

[44] M. R. Novick and C. Lewis. Coefficient alpha and the reliability of composite measurements. *Psychometrika*, 32(1):1–13, 1967. doi: 10.1007/BF02289400.

[45] Amy L. Ostrom, A. Parasuraman, David E. Bowen, Liliana Patrício, and Christopher A. Voss. Service research priorities in a rapidly changing context. *Journal of Service Research*, 18(2):127–159, 2015. doi: 10.1177/1094670515576315.

[46] Amy L. Ostrom, James M. Field, Deborah Fotheringham, Mahesh Subramony, Anders Gustafsson, Katherine N. Lemon, Ming-Hui Huang, and Janet R. McColl-Kennedy. Service research priorities: Managing and delivering service in turbulent times. *Journal of Service Research*, 24(3):329–353, 2021. doi: 10.1177/10946705211021915.

[47] Harshad Vijay Pandhare. Evaluating large language models: Frameworks and methodologies for ai/ml system testing. *International Journal of Scientific Research and Management (IJSRM)*, 12(9):1467–1486, 2024. ISSN 2321-3418. doi: 10.18535/ijsrm/v12i09.ec08. URL `https://ijsrm.net`.

[48] A Parsu Parasuraman, Valarie Zeithaml, and Leonard Berry. Servqual a multiple-item scale for measuring consumer perceptions of service quality. *Journal of Retailing*, 64:12–40, 01 1988.

[49] Ken Peffers, Tuure Tuunanen, Marcus A Rothenberger, and Samir Chatterjee. A design science research methodology for information systems research. *Journal of management information systems*, 24(3):45–77, 2007.

[50] PICBE. Microsoft copilot: Transforming customer support and knowledge work. In *Proceedings of the International Conference on Business Excellence*, volume 18, pages 1448–1460, 2024. doi: 10.2478/picbe-2024-0153. URL https://sciendo.com/article/10.2478/picbe-2024-0153.

[51] Carolyn C. Preston and Andrew M. Colman. Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, 104 (1):1–15, 2000. doi: 10.1016/S0001-6918(99)00050-5.

[52] Sebastian Raisch and Sebastian Krakowski. Artificial intelligence and management: The automation–augmentation paradox. *Academy of Management Review*, 46(1):192–210, 2021. doi: 10.5465/amr.2018.0072.

[53] W. Revelle and R. E. Zinbarg. Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika*, 74(1):145–154, 2009. doi: 10.1007/s11336-008-9102-z.

[54] Stuart J Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Pearson, Hoboken, NJ, 4th edition, 2021.

[55] Astrid Schepman and Paul Rodway. The general attitudes towards artificial intelligence scale (GAAIS): Confirmatory validation. *International Journal of Human–Computer Interaction*, 39(13): 2724–2741, 2023. doi: 10.1080/10447318.2022.2085400.

[56] Robert Seif, Atharva Sonanis, Laura Freeman, Kristen Alexander, and Jitesh Panchal. Towards multi-fidelity test and evaluation of artificial intelligence and machine learning-based systems. *The ITEA Journal of Test and Evaluation*, 45(1):39–48, 2024. doi: 10.61278/itea.45.1.1001. URL https://doi.org/10.61278/itea.45.1.1001.

[57] Manish A. Shukla. Evaluating agentic ai systems: A balanced framework for performance, robustness, safety and beyond, 2025.

[58] Herbert A. Simon. *Administrative Behavior: A Study of Decision-Making Processes in Administrative Organization*. Free Press, 3 edition, 1976.

[59] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437, 2009. ISSN 0306-4573. doi: 10.1016/j.ipm.2009.03.002.

[60] Paolo Taticchi, Flavio Tonelli, and Luca Cagnazzo. Performance measurement and management: a literature review and a research agenda. *Measuring Business Excellence*, 14(1):4–18, 2010. doi: 10.1108/13683041011027418.

[61] Wim Van Grembergen. The balanced scorecard and it governance. *Information Systems Control Journal*, 2(1):40–43, 2000.

[62] Jan vom Brocke, Alan Hevner, and Alexander Maedche. *Introduction to Design Science Research*, pages 1–13. Springer International Publishing, Cham, 2020. ISBN 978-3-030-46781-4. doi: 10.1007/978-3-030-46781-4_1. URL https://doi.org/10.1007/978-3-030-46781-4_1.

[63] B. Waelbers, Alexander P. Henkel, and S. Bromuri. Augmenting human service agents with intelligent technologies. *Journal of Service Management*, 37(6):26–49, 2026. doi: 10.1108/JOSM-01-2025-0035.

[64] Peter Weill and Jeanne W. Ross. *IT Governance: How Top Performers Manage IT Decision Rights for Superior Results*. Harvard Business School Press, Boston, MA, 2004.

[65] Jochen Wirtz, Paul G Patterson, Werner H Kunz, Thorsten Gruber, Vinh Nhat Lu, Stefanie Paluch, and Antje Martins. Brave new world: service robots in the frontline. *Journal of Service Management*, 29(5):907–931, 2018. doi: 10.1108/JOSM-04-2018-0119.

[66] Michael Wooldridge and Nicholas R. Jennings. Intelligent agents: theory and practice. *The Knowledge Engineering Review*, 10:115–152, 1995. URL https://api.semanticscholar.org/CorpusID:221342993.

[67] Y. Yang and S. B. Green. Coefficient alpha: A reliability coefficient for the 21st century? *Journal of Psychoeducational Assessment*, 29(4):377–392, 2011. doi: 10.1177/0734282911406668.

[68] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36*, volume 36. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf.

Appendix A

Interview Questionnaire

Introduction

- Could you please introduce yourself and also help me with your roles and responsibilities ?
- How would you define an AI agent and base knowledge about it ?
- Currently, in customer operations do you interact with an AI solutions ?
- What are some of the key areas in customer operations process we can use AI agent?
- What is your success criteria when evaluating a new technology ?

Financial Perspective

- How do you measure return on investment on AI agents ?
- When considering cost efficiency, what metric or measure would you use to know if an AI agent is getting expensive ?
- How would you define scalability cost when scaling the solutions for other business units ?
- How much estimated time an AI solution would save you in the case of freight document that can be converted to employee cost ?

User Perspective

- What kind of measures do you think would help you to measure the trust dimension in AI solution ?
- Considering a hypothetical scenario for freight case, if AI agent takes more than 5 minutes to give the desired output, would that affect your employees productivity ?
- When evaluating accuracy of the AI solution what would be a threshold or tolerance level ?

Internal Business Perspective

- As you have seen the demo, in which part of the process will it save time for you ?
- If we scale this POC, will it cause additional work for your team to review the PDFs from the LLMs ?
- If the document processor processes 100 documents but 10 failed, would you consider it as an acceptable solution or not ?
- As seen from demo, do think if this solution is productionized it will help your team to bring down case resolution time ?
- In demo we have seen NedCargo documents that are much cleaner, structured and gives the right output. What other scenarios would you like to test before you can trust this solution ?
- Would this solution be a background activity that needs to be carried autonomously or would you require someone actively monitor and manage it ?

Learning and Growth Perspective

- Do you think this AI agent will free up your team's time spent actively on repetitive and monotonous task ?
- As the company is moving forward with AI solution in different areas, do you think additional training/certifications are required by the team ?
- With the introduction of AI agent, would this threaten peoples job or would they see it as an extended colleague ?
- As you have seen the POC, does this encourage you to think how we can use AI in other business processes or with other business groups ?
- Would your team be open to learn new skills to work with AI agents ?

Appendix B

Power Automate Setup

In this section, we cover the power automate flow setup for the freight document extraction case at FrieslandCampina. For the proof of concept and experimentation setup, we currently extract the data from shared customer care folder with specific subject line, later save these PDFs in a SharePoint list which is processed by a document processor that is an AI model which extracts predefined fields like Order Nummer, Reference Nummer and Handwritten note and then we pass the same PDF to two LLM agent like ChatGpt to further summarize the PDF data with the help of instructions and extract the category of compliance, then store all this extracted data to SharePoint.

Below is a screenshot that covers this setup:

- Power Automate Flow Setup

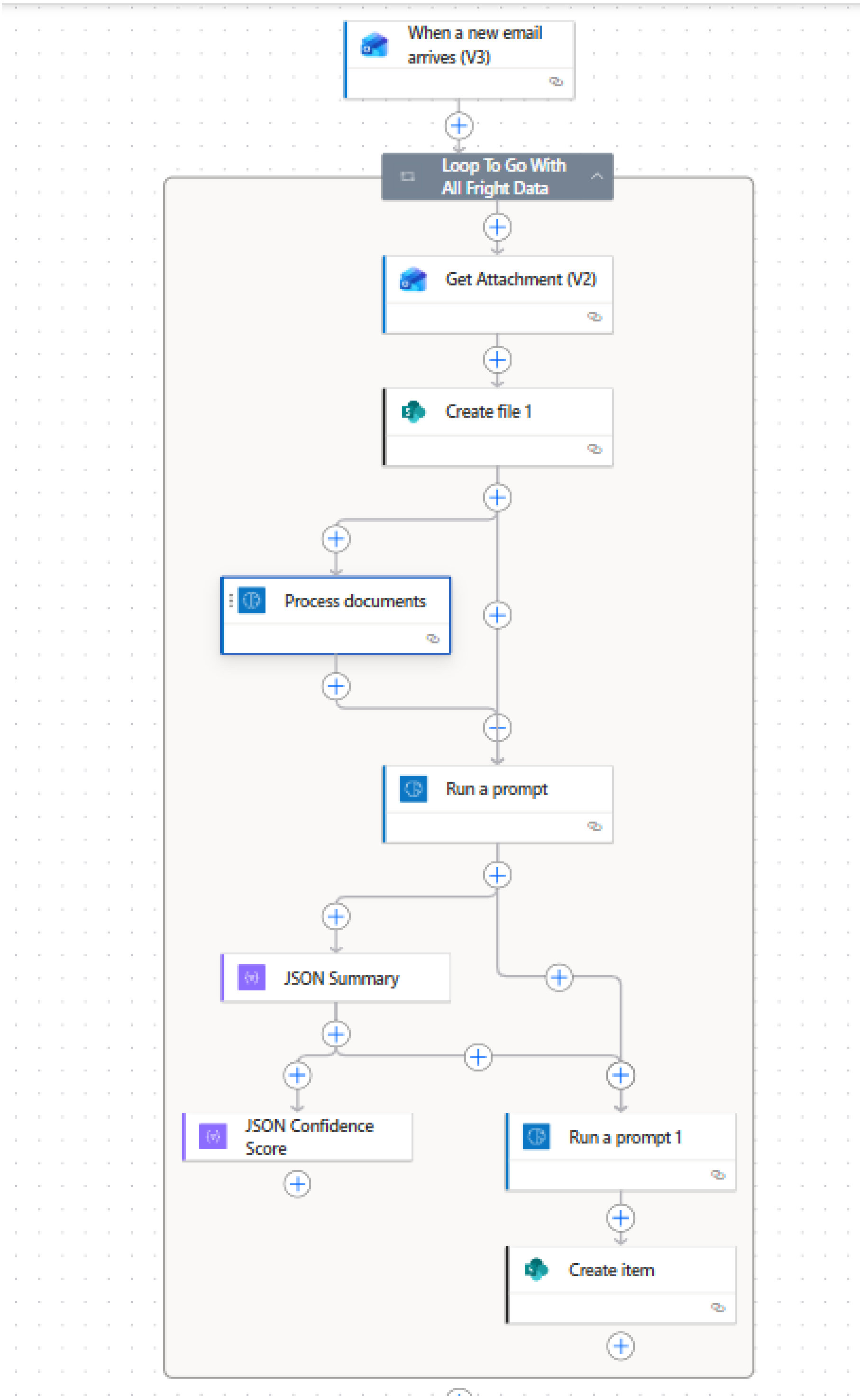


Figure B.1: Power Automate Setup

- Power Automate Flow Successful Run

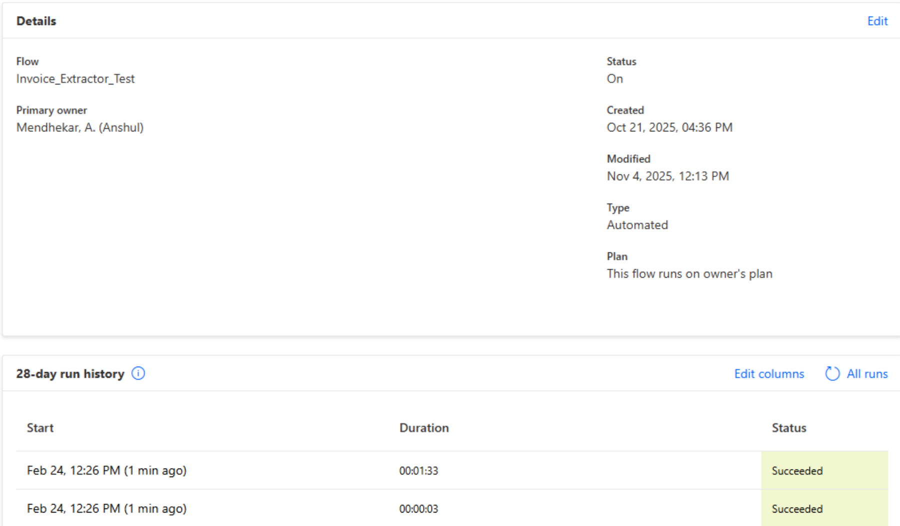


Figure B.2: Power Automate Flow Successful Run

- Artificial Intelligence Prompt-1 (To detect deviations from the PDFs)

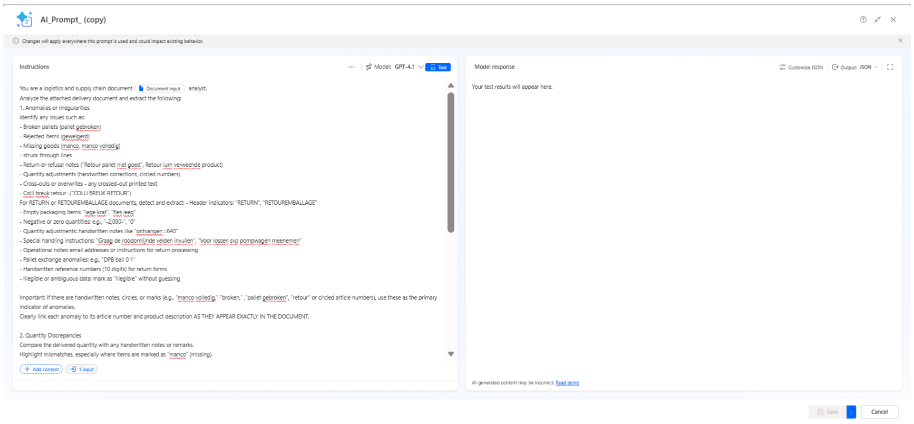


Figure B.3: Artificial Intelligence Prompt1

- Artificial Intelligence Prompt Continued

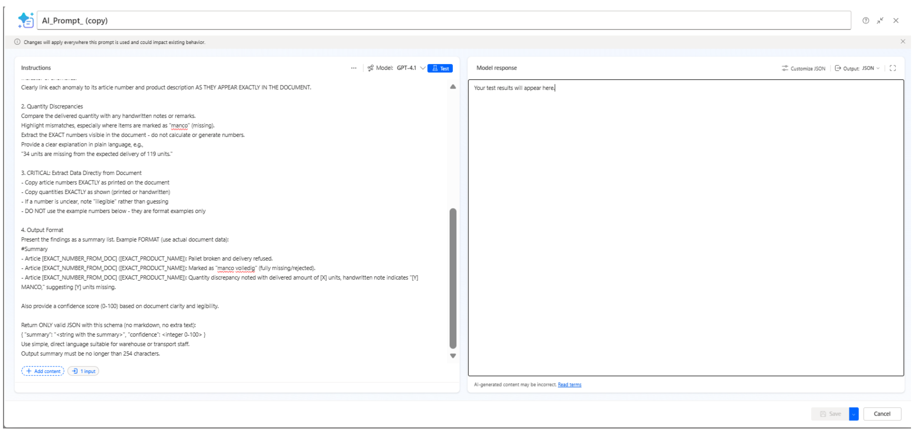


Figure B.4: Artificial Intelligence Prompt1

- Category Compliance

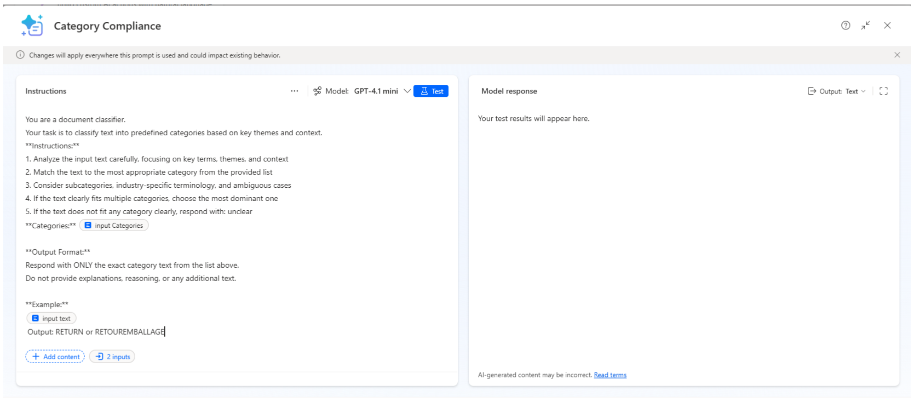


Figure B.5: Category Compliance

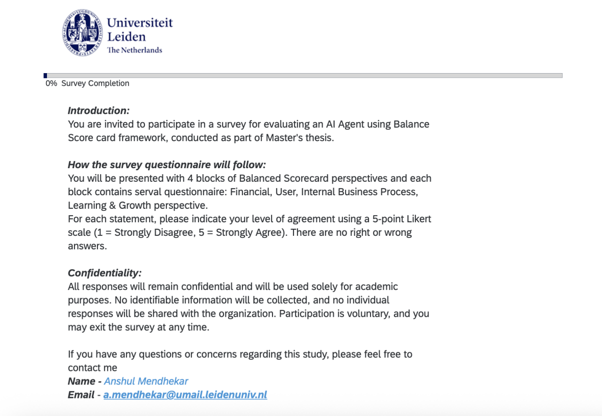
- Result SharePoint Here we share on what data was extracted from the same setup on to SharePoint. We can see the below screenshot that covers the result of this setup:

Figure B.6: Result SharePoint

Appendix C

Survey Questionnaire

The participant should first read and sign the consent form in C.1, C.2, before proceeding to the interaction with the questionnaire.



Universiteit Leiden
The Netherlands

0% Survey Completion

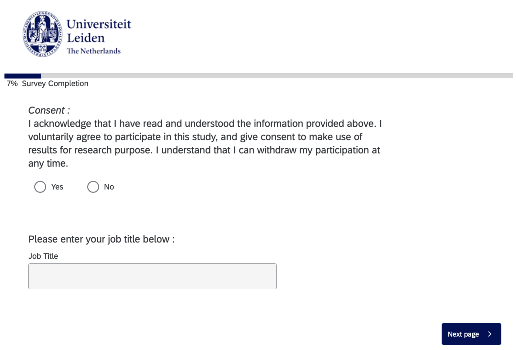
Introduction:
You are invited to participate in a survey for evaluating an AI Agent using Balance Score card framework, conducted as part of Master's thesis.

How the survey questionnaire will follow:
You will be presented with 4 blocks of Balanced Scorecard perspectives and each block contains serval questionnaire: Financial, User, Internal Business Process, Learning & Growth perspective.
For each statement, please indicate your level of agreement using a 5-point Likert scale (1 = Strongly Disagree, 5 = Strongly Agree). There are no right or wrong answers.

Confidentiality:
All responses will remain confidential and will be used solely for academic purposes. No identifiable information will be collected, and no individual responses will be shared with the organization. Participation is voluntary, and you may exit the survey at any time.

If you have any questions or concerns regarding this study, please feel free to contact me
Name - *Anshul Mendhekar*
Email - ***a.mendhekar@umail.leidenuniv.nl***

Figure C.1: Survey Start Intro



Universiteit Leiden
The Netherlands

7% Survey Completion

Consent :
I acknowledge that I have read and understood the information provided above. I voluntarily agree to participate in this study, and give consent to make use of results for research purpose. I understand that I can withdraw my participation at any time.

Yes No

Please enter your job title below :
Job Title

Next page

Figure C.2: Consent Form

Universiteit Leiden
The Netherlands

20% Survey Completion

Financial Perspective : Cost Efficiency

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
The AI agent reduces the data extraction time.	○	○	○	○	○
The AI agent allows manager to repurpose the customer care officer's time.	○	○	○	○	○
The use of an AI agent has a significant cost advantage compared to maintaining a dedicated FTE (full time employee).	○	○	○	○	○

Financial Perspective : Value Creation

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
The AI agent contributes by reducing employee burnout and increasing productivity.	○	○	○	○	○
The output generated by AI agent is usable by customer care officers to complete the follow up tasks in SAP, such as creating credit notes and then refund.	○	○	○	○	○
The AI agent improves processing throughput compared to current practices.	○	○	○	○	○

Figure C.3: Financial Perspective- Cost Efficiency and ROI Dimension

Financial Perspective: Scalability

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
The AI agent can process higher document volumes within a short amount of time as compared to human.	○	○	○	○	○
The AI agent can be adapted to new supplier PDFs and different variations of the PDFs without much training.	○	○	○	○	○
The AI agent can be extended to different business groups.	○	○	○	○	○

Next page >

Figure C.4: Financial Perspective- Scalability Dimension

Universiteit Leiden
The Netherlands

40% Survey Completion

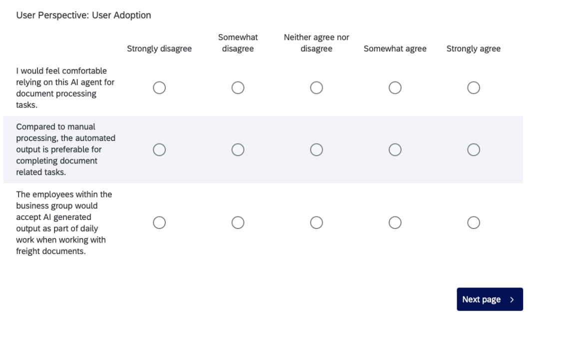
User Perspective : User Experience

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
The workflow setup up in Power Automate for the AI agent is easy to understand, use and modify.	○	○	○	○	○
PDFs rarely need to be opened to find information because of AI agent.	○	○	○	○	○
The AI agent generates standardized, clearly structured outputs that are easily understood by both customer care officers and business managers	○	○	○	○	○

User Perspective: Accuracy

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
The data extracted by the AI agent from the freight documents is accurate.	○	○	○	○	○
The header fields extracted by AI agents are usable to continue the follow up work with different systems.	○	○	○	○	○
The AI agent is likely to generate accurate, usable output across varying supplier document formats.	○	○	○	○	○

Figure C.5: User Perspective- User Experience and Accuracy Dimension

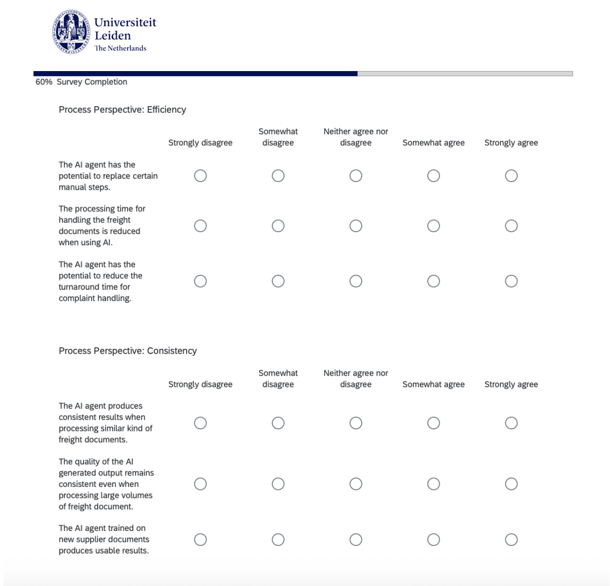


User Perspective: User Adoption

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
I would feel comfortable relying on this AI agent for document processing tasks.	○	○	○	○	○
Compared to manual processing, the automated output is preferable for completing document related tasks.	○	○	○	○	○
The employees within the business group would accept AI generated output as part of daily work when working with freight documents.	○	○	○	○	○

Next page >

Figure C.6: User Perspective- Adoption Dimension



Universiteit Leiden
The Netherlands

60% Survey Completion

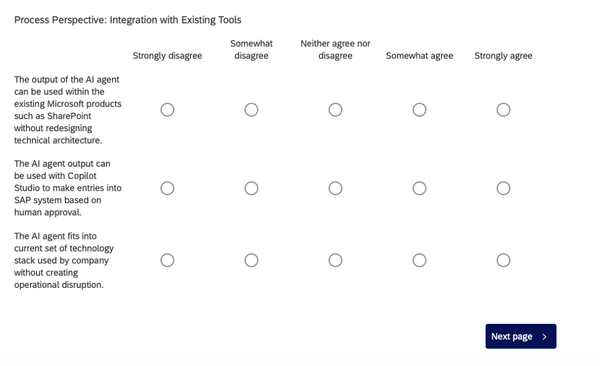
Process Perspective: Efficiency

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
The AI agent has the potential to replace certain manual steps.	○	○	○	○	○
The processing time for handling the freight documents is reduced when using AI.	○	○	○	○	○
The AI agent has the potential to reduce the turnaround time for complaint handling.	○	○	○	○	○

Process Perspective: Consistency

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
The AI agent produces consistent results when processing similar kind of freight documents.	○	○	○	○	○
The quality of the AI generated output remains consistent even when processing large volumes of freight document.	○	○	○	○	○
The AI agent trained on new supplier documents produces usable results.	○	○	○	○	○

Figure C.7: Internal Business Process Perspective -Efficiency and Consistency

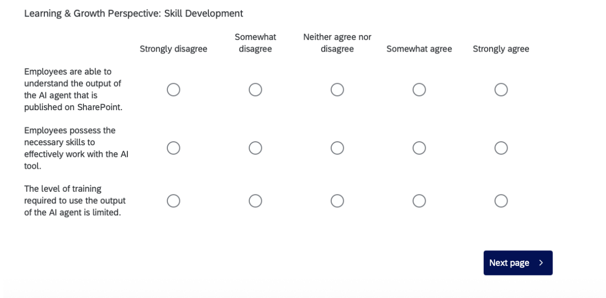


Process Perspective: Integration with Existing Tools

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
The output of the AI agent can be used within the existing Microsoft products such as SharePoint without redesigning technical architecture.	○	○	○	○	○
The AI agent output can be used with Copilot Studio to make entries into SAP system based on human approval.	○	○	○	○	○
The AI agent fits into current set of technology stack used by company without creating operational disruption.	○	○	○	○	○

Next page >

Figure C.8: Internal Business Process Perspective -Existing Tools

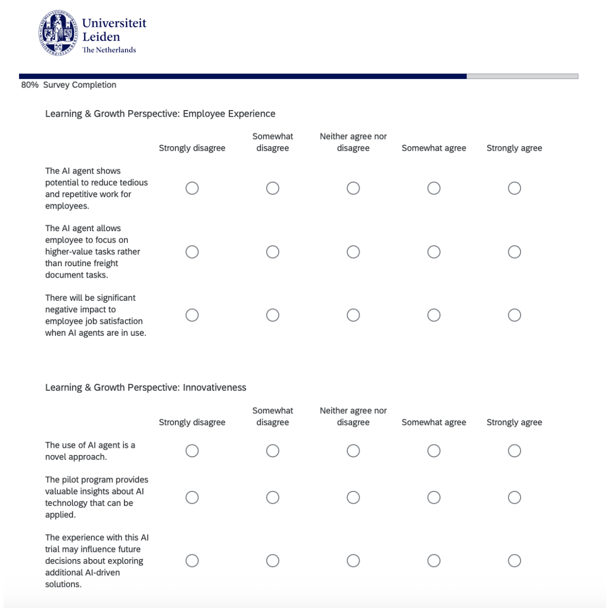


Learning & Growth Perspective: Skill Development

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
Employees are able to understand the output of the AI agent that is published on SharePoint.	○	○	○	○	○
Employees possess the necessary skills to effectively work with the AI tool.	○	○	○	○	○
The level of training required to use the output of the AI agent is limited.	○	○	○	○	○

Next page >

Figure C.10: Learning and Growth Perspective -Skill Development Dimension



Universiteit Leiden
The Netherlands

80% Survey Completion

Learning & Growth Perspective: Employee Experience

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
The AI agent shows potential to reduce tedious and repetitive work for employees.	○	○	○	○	○
The AI agent allows employee to focus on higher-value tasks rather than routine freight document tasks.	○	○	○	○	○
There will be significant negative impact to employee job satisfaction when AI agents are in use.	○	○	○	○	○

Learning & Growth Perspective: Innovativeness

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
The use of AI agent is a novel approach.	○	○	○	○	○
The pilot program provides valuable insights about AI technology that can be applied.	○	○	○	○	○
The experience with this AI trial may influence future decisions about exploring additional AI-driven solutions.	○	○	○	○	○

Figure C.9: Learning and Growth Perspective -Employee Experience and Innovativeness Dimension

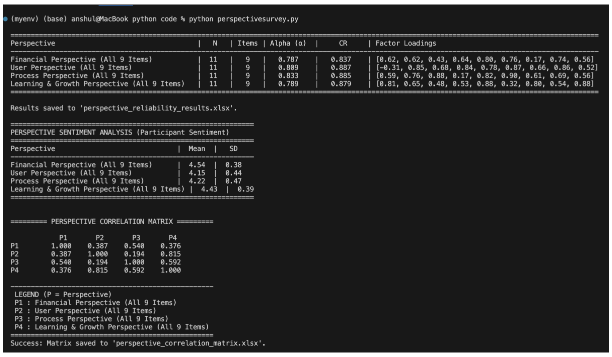
Appendix D

Python Code Snippets For Evaluation Measure

This section covers the code that was used to calculate the Factor Loadings, Pearson's Correlation, Standard Deviation, Mean, Composite Reliability and Cronbach's Alpha scores.

- Balanced Scorecard Perspective Code Snippet For Cronbach's Alpha, Composite Reliability, Factor Loadings and Correlation Score. The screenshots are available in figure D.1, D.2

Figure D.1: Balanced Scorecard Perspective Code Snippet For Cronbach's Alpha, Composite Reliability, and Factor Loadings ,Correlation Score



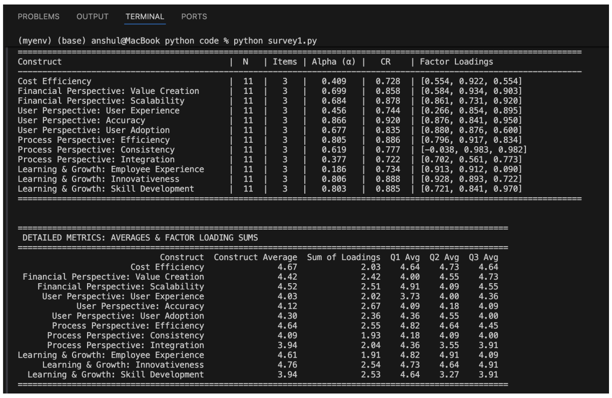
```
(myenv) (base) anshul@MacBook python code % python perspectivesurvey.py

==========================================================================================================================
Perspective                               |  N   | Items | Alpha (α)  |     CR     | Factor Loadings
--------------------------------------------------------------------------------------------------------------------------
Financial Perspective (All 9 Items)       |  11  |   9   |   0.787    |   0.837    | [0.62, 0.62, 0.43, 0.64, 0.80, 0.76, 0.17, 0.74, 0.56]
User Perspective (All 9 Items)            |  11  |   9   |   0.809    |   0.887    | [-0.31, 0.85, 0.68, 0.84, 0.78, 0.87, 0.66, 0.86, 0.52]
Process Perspective (All 9 Items)         |  11  |   9   |   0.833    |   0.885    | [0.59, 0.76, 0.88, 0.17, 0.82, 0.90, 0.61, 0.69, 0.56]
Learning & Growth Perspective (All 9 Items) |  11  |   9   |   0.789    |   0.879    | [0.81, 0.65, 0.48, 0.53, 0.88, 0.32, 0.80, 0.54, 0.88]
==========================================================================================================================

Results saved to 'perspective_reliability_results.xlsx'.

============================================================
PERSPECTIVE SENTIMENT ANALYSIS (Participant Sentiment)
============================================================
Perspective                              |  Mean  |   SD
------------------------------------------------------------
Financial Perspective (All 9 Items)      |  4.54  |  0.38
User Perspective (All 9 Items)           |  4.15  |  0.44
Process Perspective (All 9 Items)        |  4.22  |  0.47
Learning & Growth Perspective (All 9 Items) |  4.43  |  0.39
============================================================


========= PERSPECTIVE CORRELATION MATRIX =========

          P1       P2       P3       P4
P1       1.000    0.387    0.540    0.376
P2       0.387    1.000    0.194    0.815
P3       0.540    0.194    1.000    0.592
P4       0.376    0.815    0.592    1.000

--------------------------------------------------
 LEGEND (P = Perspective)
 P1 : Financial Perspective (All 9 Items)
 P2 : User Perspective (All 9 Items)
 P3 : Process Perspective (All 9 Items)
 P4 : Learning & Growth Perspective (All 9 Items)
==================================================
Success: Matrix saved to 'perspective_correlation_matrix.xlsx'.
```

Figure D.2: Result Of Balanced Scorecard Perspective Code Snippet For Cronbach's Alpha, Composite Reliability, Factor Loadings and Correlation Score

- Balanced Scorecard Dimensions Code Snippet For Cronbach's Alpha, Composite Reliability, Factor Loadings, Standard Deviation and Mean. The screenshots are available in figure D.3, D.4

```python
import pandas as pd
import numpy as np
from sklearn.decomposition import PCA

# ---------------------------------------------------------
# 1. Load the Survey Data
# ---------------------------------------------------------
# Ensure your Excel file is named 'survey_results_1.xlsx' and is in the same folder.
df_full = pd.read_excel('survey_results_1.xlsx')

# ---------------------------------------------------------
# 2. Reverse Code the Negative Question
# ---------------------------------------------------------
reverse_col = "Learning & Growth Perspective: Employee Experience - There will be significant negative impact to employee job satisfaction when AI agents are in use."

if reverse_col in df_full.columns:
    df_full[reverse_col] = pd.to_numeric(df_full[reverse_col], errors='coerce')
    df_full[reverse_col] = 6 - df_full[reverse_col]
else:
    print(f"WARNING: Could not find the column to reverse code. Please check the exact spelling.\n")

# ---------------------------------------------------------
# 3. Define Your Constructs
# ---------------------------------------------------------
constructs = {
    "Cost Efficiency": [
        "Cost Efficiency - The AI agent reduces the data extraction time.",
        "Cost Efficiency - The AI agent allows manager to repurpose the customer care officer's time.",
        "Cost Efficiency - The use of an AI agent has a significant cost advantage compared to maintaining a dedicated FTE (full time employee)."
    ],
    "Financial Perspective: Value Creation": [
        "Financial Perspective : Value Creation - The AI agent contributes by reducing employee burnout and increasing productivity.",
        "Financial Perspective : Value Creation - The output generated by AI agent is usable by customer care officers to complete the follow up tasks in SAP, such as creating credit notes and then refund.",
        "Financial Perspective : Value Creation - The AI agent improves processing throughput compared to current practices."
    ],
    "Financial Perspective: Scalability": [
        "Financial Perspective: Scalability - The AI agent can process higher document volumes within a short amount of time as compared to human.",
        "Financial Perspective: Scalability - The AI agent can be adapted to new supplier PDFs and different variations of the PDFs without much training.",
        "Financial Perspective: Scalability - The AI agent can be extended to different business groups."
    ],
    "User Perspective: User Experience": [
        "User Perspective : User Experience - The workflow setup up in Power Automate for the AI agent is easy to understand, use and modify.",
        "User Perspective : User Experience - PDFs rarely need to be opened to find information because of AI agent.",
        "User Perspective : User Experience - The AI agent generates standardized, clearly structured outputs that are easily understood by both customer care officers and business managers"
    ],
    "User Perspective: Accuracy": [
        "User Perspective: Accuracy - The data extracted by the AI agent from the freight documents is accurate.",
        "User Perspective: Accuracy - The header fields extracted by AI agents are usable to continue the follow up work with different systems.",
        "User Perspective: Accuracy - The AI agent is likely to generate accurate, usable output across varying supplier document formats."
    ],
    "User Perspective: User Adoption": [
        "User Perspective: User Adoption - I would feel comfortable relying on this AI agent for document processing tasks.",
        "User Perspective: User Adoption - Compared to manual processing, the automated output is preferable for completing document related tasks.",
        "User Perspective: User Adoption - The employees within the business group would accept AI generated output as part of daily work when working with freight documents."
    ],
    "Process Perspective: Efficiency": [
        "Process Perspective: Efficiency - The AI agent has the potential to replace certain manual steps.",
        "Process Perspective: Efficiency - The processing time for handling the freight documents is reduced when using AI.",
        "Process Perspective: Efficiency - The AI agent has the potential to reduce the turnaround time for complaint handling."
    ],
    "Process Perspective: Consistency": [
        "Process Perspective: Consistency - The AI agent produces consistent results when processing similar kind of freight documents.",
        "Process Perspective: Consistency - The quality of the AI generated output remains consistent even when processing large volumes of freight document.",
        "Process Perspective: Consistency - The AI agent trained on new supplier documents produces usable results."
    ],
    "Process Perspective: Integration": [
        "Process Perspective: Integration with Existing Tools - The output of the AI agent can be used within the existing Microsoft products such as SharePoint without redesigning technical architecture.",
        "Process Perspective: Integration with Existing Tools - The AI agent output can be used with Copilot Studio to make entries into SAP system based on human approval.",
        "Process Perspective: Integration with Existing Tools - The AI agent fits into current set of technology stack used by company without creating operational disruption."
    ],
    "Learning & Growth: Employee Experience": [
        "Learning & Growth Perspective: Employee Experience - The AI agent shows potential to reduce tedious and repetitive work for employees.",
        "Learning & Growth Perspective: Employee Experience - The AI agent allows employee to focus on higher-value tasks rather than routine freight document tasks.",
        "Learning & Growth Perspective: Employee Experience - There will be significant negative impact to employee job satisfaction when AI agents are in use."
    ],
    "Learning & Growth: Innovativeness": [
        "Learning & Growth Perspective: Innovativeness - The use of AI agent is a novel approach.",
        "Learning & Growth Perspective: Innovativeness - The pilot program provides valuable insights about AI technology that can be applied.",
        "Learning & Growth Perspective: Innovativeness - The experience with this AI trial may influence future decisions about exploring additional AI-driven solutions."
    ],
    "Learning & Growth: Skill Development": [
        "Learning & Growth Perspective: Skill Development - Employees are able to understand the output of the AI agent that is published on SharePoint.",
        "Learning & Growth Perspective: Skill Development - Employees possess the necessary skills to effectively work with the AI tool.",
        "Learning & Growth Perspective: Skill Development - The level of training required to use the output of the AI agent is limited."
    ]
}

# ---------------------------------------------------------
# 4. Math Function for Reliability
# ---------------------------------------------------------
def analyze_reliability(data, columns, construct_name):
    # Ensure data is numeric and drop NAs
    df_subset = data[columns].apply(pd.to_numeric, errors='coerce').dropna()
    n_size = len(df_subset)

    if n_size < 2:
        return {"Construct": construct_name, "N": n_size, "Error": "Not enough data"}

    # --- Cronbach's Alpha ---
    k = df_subset.shape[1]
    item_vars = df_subset.var(ddof=1)
    total_var = df_subset.sum(axis=1).var(ddof=1)

    if total_var == 0:
        alpha = 0.0
    else:
        alpha = (k / (k - 1)) * (1 - (item_vars.sum() / total_var))

    # --- Factor Loadings & Composite Reliability (CR) ---
    df_std = (df_subset - df_subset.mean()) / df_subset.std(ddof=1)
    df_std = df_std.fillna(0)

    pca = PCA(n_components=1)
    pca.fit(df_std)

    loadings = pca.components_[0] * np.sqrt(pca.explained_variance_[0])

    if sum(loadings) < 0:
        loadings = -loadings

    sum_loadings_sq = np.sum(loadings)**2
    sum_errors = np.sum(1 - loadings**2)

    if (sum_loadings_sq + sum_errors) == 0:
        cr = 0.0
    else:
        cr = sum_loadings_sq / (sum_loadings_sq + sum_errors)

    # Format loadings as a clean string (e.g., "[0.85, 0.91, 0.62]")
    loadings_str = "[" + ", ".join([f"{l:.3f}" for l in loadings]) + "]"

    return {
        "Construct": construct_name,
        "Sample Size (N)": n_size,
        "Items": k,
        "Cronbach's Alpha": round(alpha, 3),
        "Composite Reliability": round(cr, 3),
        "Factor Loadings": loadings_str
    }

# ---------------------------------------------------------
# 5. Loop Through and Output Results
# ---------------------------------------------------------
results = []

for name, cols in constructs.items():
    missing_cols = [c for c in cols if c not in df_full.columns]
    if missing_cols:
        print(f"\nWARNING: Skipping '{name}'. Missing columns in Excel file:")
        for missing in missing_cols:
            print(f" -> {missing}")
        continue

    res = analyze_reliability(df_full, cols, name)
    results.append(res)

# Output summary to Terminal with ENHANCED formatting
if results:
    results_df = pd.DataFrame(results)

    # 1. Print the Table Header
    print("\n" + "="*115)
    print(f"{'Construct':<42} | {'N':^4} | {'Items':^5} | {'Alpha (α)':^9} | {'CR':^7} | {'Factor Loadings'}")
    print("-" * 115)

    # 2. Print Each Row with Perfect Alignment
    for res in results:
        construct = res['Construct']
        n = res['Sample Size (N)']
        items = res['Items']

        # Extract the numbers first to avoid f-string quote errors
        raw_alpha = res["Cronbach's Alpha"]
        raw_cr = res["Composite Reliability"]

        # Now format them securely
        alpha = f"{raw_alpha:.3f}"
        cr = f"{raw_cr:.3f}"
        loadings = res['Factor Loadings']

        print(f"{construct:<42} | {n:^4} | {items:^5} | {alpha:^9} | {cr:^7} | {loadings}")

    print("="*115 + "\n")

    # Save results to Excel
    #results_df.to_excel("reliability_summary_results.xlsx", index=False)
    #print("Results successfully saved to 'reliability_summary_results.xlsx'.")
else:
    print("No valid data processed. Please check the warnings above.")

# ---------------------------------------------------------
# 6. Construct Averages & Sum of Factor Loadings
# ---------------------------------------------------------
detailed_metrics = []

for name, cols in constructs.items():
    # Skip if missing columns (handled previously, but good for safety)
    if any(c not in df_full.columns for c in cols):
        continue

    # Ensure data is numeric and drop NAs
    df_subset = df_full[cols].apply(pd.to_numeric, errors='coerce').dropna()

    if len(df_subset) < 2:
        continue

    # --- 1. Average for each construct survey statement ---
    item_means = df_subset.mean(axis=0)

    # --- 2. Average for the overall construct ---
    construct_average = item_means.mean()

    # --- 3. Sum of Factor Loadings ---
    # Standardize data and run PCA to match your earlier methodology
    df_std = (df_subset - df_subset.mean()) / df_subset.std(ddof=1)
    df_std = df_std.fillna(0)

    pca = PCA(n_components=1)
    pca.fit(df_std)

    loadings = pca.components_[0] * np.sqrt(pca.explained_variance_[0])
    if sum(loadings) < 0:
        loadings = -loadings

    sum_loadings = np.sum(loadings)

    # Build a dictionary for this construct's results
    construct_data = {
        "Construct": name,
        "Construct Average": round(construct_average, 2),
        "Sum of Loadings": round(sum_loadings, 2)
    }

    # Add the individual statement averages dynamically
    for i, col in enumerate(cols):
        # Naming them 'Q1 Avg', 'Q2 Avg', etc. to keep the printout clean
        construct_data[f"Q{i+1} Avg"] = round(item_means[col], 2)

    detailed_metrics.append(construct_data)

# Output summary to Terminal
if detailed_metrics:
    df_detailed = pd.DataFrame(detailed_metrics)

    print("\n" + "="*100)
    print(" DETAILED METRICS: AVERAGES & FACTOR LOADING SUMS")
    print("="*100)

    # Using pandas string formatting for a clean, dynamic table
    print(df_detailed.to_string(index=False))

    print("="*100 + "\n")

    # Optional: Save this new table to Excel alongside your other results
    # df_detailed.to_excel("detailed_construct_metrics.xlsx", index=False)
```

Figure D.3: Balanced Scorecard Dimensions Code Snippet For Cronbach's Alpha, Composite Reliability, Factor Loadings, Standard Deviation and Mean.



PROBLEMS OUTPUT TERMINAL PORTS

(myenv) (base) anshul@MacBook python code % python survey1.py

Construct	N	Items	Alpha (α)	CR	Factor Loadings
Cost Efficiency	11	3	0.409	0.728	[0.554, 0.922, 0.554]
Financial Perspective: Value Creation	11	3	0.699	0.858	[0.584, 0.934, 0.903]
Financial Perspective: Scalability	11	3	0.684	0.878	[0.861, 0.731, 0.920]
User Perspective: User Experience	11	3	0.456	0.744	[0.266, 0.854, 0.895]
User Perspective: Accuracy	11	3	0.866	0.920	[0.876, 0.841, 0.950]
User Perspective: User Adoption	11	3	0.677	0.835	[0.880, 0.876, 0.600]
Process Perspective: Efficiency	11	3	0.805	0.886	[0.796, 0.917, 0.834]
Process Perspective: Consistency	11	3	0.619	0.777	[−0.038, 0.983, 0.982]
Process Perspective: Integration	11	3	0.377	0.722	[0.702, 0.561, 0.773]
Learning & Growth: Employee Experience	11	3	0.186	0.734	[0.913, 0.912, 0.090]
Learning & Growth: Innovativeness	11	3	0.806	0.888	[0.928, 0.893, 0.722]
Learning & Growth: Skill Development	11	3	0.803	0.885	[0.721, 0.841, 0.970]

DETAILED METRICS: AVERAGES & FACTOR LOADING SUMS

Construct	Construct Average	Sum of Loadings	Q1 Avg	Q2 Avg	Q3 Avg
Cost Efficiency	4.67	2.03	4.64	4.73	4.64
Financial Perspective: Value Creation	4.42	2.42	4.00	4.55	4.73
Financial Perspective: Scalability	4.52	2.51	4.91	4.09	4.55
User Perspective: User Experience	4.03	2.02	3.73	4.00	4.36
User Perspective: Accuracy	4.12	2.67	4.09	4.18	4.09
User Perspective: User Adoption	4.30	2.36	4.36	4.55	4.00
Process Perspective: Efficiency	4.64	2.55	4.82	4.64	4.45
Process Perspective: Consistency	4.09	1.93	4.18	4.09	4.00
Process Perspective: Integration	3.94	2.04	4.36	3.55	3.91
Learning & Growth: Employee Experience	4.61	1.91	4.82	4.91	4.09
Learning & Growth: Innovativeness	4.76	2.54	4.73	4.64	4.91
Learning & Growth: Skill Development	3.94	2.53	4.64	3.27	3.91

Figure D.4: Results of Balanced Scorecard Dimensions Code Snippet For Cronbach's Alpha, Composite Reliability, Factor Loadings, Standard Deviation and Mean.

Appendix E

Cronbach's Alpha Score At Item Level

We report these results without analysis. Since they are based on just 11 respondents, these results should be seen as exploratory only. This section highlights the Cronbach's Alpha score at question level for each dimension is represented in the tables E.1, E.2, G.3, G.4. The python code snippet is represented in screenshot E.1 and the terminal output is represented in the screenshots E.2, E.3, E.4, E.5.

Table E.1: Cronbach's Alpha Result At The Lowest Level Of Data For Financial Perspective

Dimension Name	**Survey Statements Per Dimension**	**Cronbach's Alpha Score**	**Alpha Score If Particular Item Dropped**
Cost Efficiency	The AI Agent reduces data extraction time	0.11	0.556
Cost Efficiency	The AI agent allows manager to re-purpose the customer care officer's time	0.602	-0.435
Cost Efficiency	The use of an AI agent has a significant cost advantage compared to maintaining a dedicated FTE (full time employee)	0.111	0.556
Value Creation	The AI agent contributes by reducing employee burnout and increasing productivity	0.351	0.866
Value Creation	The output generated by AI agent is usable by customer care officers to complete the follow up tasks in SAP, such as creating credit notes and then refund	0.668	0.393
Value Creation	The AI agent improves processing throughput compared to current practices.	0.642	0.543
Scalability	The AI agent can process higher document volumes within a short amount of time as compared to human	0.61	0.673
Scalability	The AI agent can be adapted to new supplier PDFs and different variations of the PDFs without much training	0.515	0.708
Scalability	The AI agent can be extended to different business groups	0.664	0.364

Table E.2: Cronbach's Alpha Result At The Lowest Level Of Data For User Perspective

Dimension Name	**Survey Statements Per Dimension**	**Cronbach's Alpha Score**	**Alpha Score If Particular Item Dropped**
User Experience	The workflow setup up in Power Automate for the AI agent is easy to understand, use and modify	0.094	0.725
User Experience	PDFs rarely need to be opened to find information because of AI agent	0.317	0.293
User Experience	The AI agent generates standardized, clearly structured outputs that are easily understood by both customer care officers and business managers	0.51	0.0
Accuracy	The data extracted by the AI agent from the freight documents is accurate	0.712	0.84
Accuracy	The header fields extracted by AI agents are usable to continue the follow up work with different systems	0.666	0.887
Accuracy	The AI agent is likely to generate accurate, usable output across varying supplier document formats	0.867	0.697
User Adoption	I would feel comfortable relying on this AI agent for document processing tasks	0.598	0.458
User Adoption	Compared to manual processing, the automated output is preferable for completing document related tasks	0.584	0.468
User Adoption	The employees within the business group would accept AI generated output as part of daily work when working with freight documents	0.335	0.816

Table E.3: Cronbach's Alpha Result At The Lowest Level Of Data For Business Perspective

Dimension Name	**Survey Statements Per Dimension**	**Cronbach's Alpha Score**	**Alpha Score If Particular Item Dropped**
Efficiency	The AI agent has the potential to replace certain manual steps	0.571	0.816
Efficiency	The processing time for handling the freight documents is reduced when using AI	0.779	0.588
Efficiency	The AI agent has the potential to reduce the turnaround time for complaint handling	0.637	0.757
Consistency	The AI agent produces consistent results when processing similar kind of freight documents	-0.019	0.964
Consistency	The quality of the AI generated output remains consistent even when processing large volumes of freight document	0.713	0.0
Consistency	The AI agent trained on new supplier documents produces usable results	0.768	-0.071
Integration	The output of the AI agent can be used within the existing Microsoft products such as SharePoint without redesigning technical architecture	0.225	0.298
Integration	The AI agent output can be used with Copilot Studio to make entries into SAP system based on human approval	0.19	0.448
Integration	The AI agent fits into current set of technology stack used by company without creating operational disruption	0.302	0.173

Table E.4: Cronbach's Alpha Result At The Lowest Level Of Data For Learning and Growth Perspective

Dimension Name	**Survey Statements Per Dimension**	**Cronbach's Alpha Score**	**Alpha Score If Particular Item Dropped**
Employee Experience	The AI agent shows potential to reduce tedious and repetitive work for employees	0.209	0.026
Employee Experience	The AI agent allows employee to focus on higher-value tasks rather than routine freight document tasks	0.247	0.049
Employee Experience	There will be significant negative impact to employee job satisfaction when AI agents are in use	0.037	0.783
Innovativeness	The use of AI agent is a novel approach	0.821	0.538
Innovativeness	The pilot program provides valuable insights about AI technology that can be applied	0.748	0.64
Innovativeness	The experience with this AI trial may influence future decisions about exploring additional AI-driven solutions	0.489	0.894
Skill Development	Employees are able to understand the output of the AI agent that is published on SharePoint	0.481	0.892
Skill Development	Employees possess the necessary skills to effectively work with the AI tool	0.667	0.716
Skill Development	The level of training required to use the output of the AI agent is limited	0.911	0.4

Figure E.1: Cronbach's Alpha Score Calculation At Item Level via Python Code

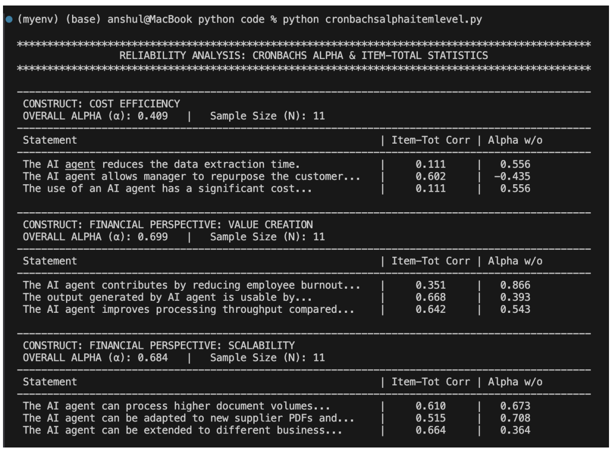
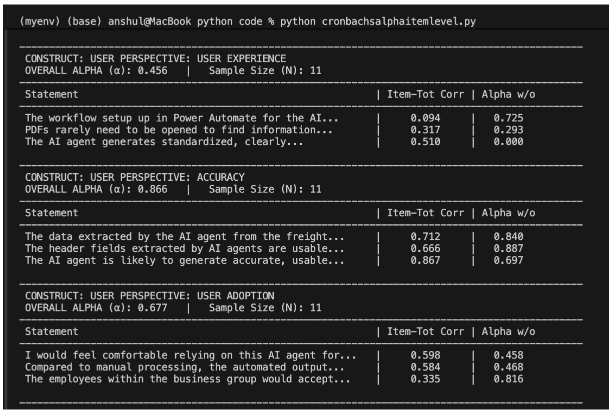
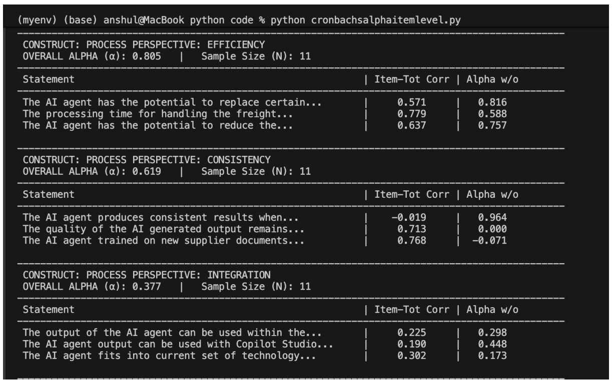


Figure E.2: Cronbach's Alpha Terminal Result At The Lowest Level Of Data For Financial Perspective



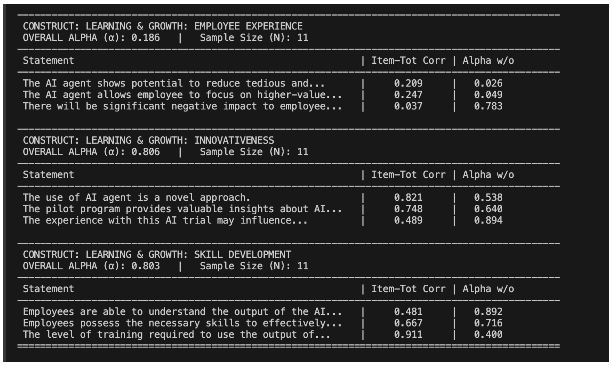
```
(myenv) (base) anshul@MacBook python code % python cronbachsalphaitemlevel.py
----------------------------------------------------------------------------------------------
CONSTRUCT: USER PERSPECTIVE: USER EXPERIENCE
OVERALL ALPHA (α): 0.456   |   Sample Size (N): 11
----------------------------------------------------------------------------------------------
Statement                                              | Item-Tot Corr | Alpha w/o
----------------------------------------------------------------------------------------------
The workflow setup up in Power Automate for the AI...  |     0.094     |   0.725
PDFs rarely need to be opened to find information...   |     0.317     |   0.293
The AI agent generates standardized, clearly...        |     0.510     |   0.000
----------------------------------------------------------------------------------------------
CONSTRUCT: USER PERSPECTIVE: ACCURACY
OVERALL ALPHA (α): 0.866   |   Sample Size (N): 11
----------------------------------------------------------------------------------------------
Statement                                              | Item-Tot Corr | Alpha w/o
----------------------------------------------------------------------------------------------
The data extracted by the AI agent from the freight... |     0.712     |   0.840
The header fields extracted by AI agents are usable... |     0.666     |   0.887
The AI agent is likely to generate accurate, usable... |     0.867     |   0.697
----------------------------------------------------------------------------------------------
CONSTRUCT: USER PERSPECTIVE: USER ADOPTION
OVERALL ALPHA (α): 0.677   |   Sample Size (N): 11
----------------------------------------------------------------------------------------------
Statement                                              | Item-Tot Corr | Alpha w/o
----------------------------------------------------------------------------------------------
I would feel comfortable relying on this AI agent for... |   0.598     |   0.458
Compared to manual processing, the automated output...  |    0.584     |   0.468
The employees within the business group would accept... |   0.335     |   0.816
----------------------------------------------------------------------------------------------
```

Figure E.3: Cronbach's Alpha Terminal Result At The Lowest Level Of Data For User Perspective



```
(myenv) (base) anshul@MacBook python code % python cronbachsalphaitemlevel.py
----------------------------------------------------------------------------------------------
CONSTRUCT: PROCESS PERSPECTIVE: EFFICIENCY
OVERALL ALPHA (α): 0.805   |   Sample Size (N): 11
----------------------------------------------------------------------------------------------
Statement                                              | Item-Tot Corr | Alpha w/o
----------------------------------------------------------------------------------------------
The AI agent has the potential to replace certain...   |     0.571     |   0.816
The processing time for handling the freight...        |     0.779     |   0.588
The AI agent has the potential to reduce the...        |     0.637     |   0.757
----------------------------------------------------------------------------------------------
CONSTRUCT: PROCESS PERSPECTIVE: CONSISTENCY
OVERALL ALPHA (α): 0.619   |   Sample Size (N): 11
----------------------------------------------------------------------------------------------
Statement                                              | Item-Tot Corr | Alpha w/o
----------------------------------------------------------------------------------------------
The AI agent produces consistent results when...       |    -0.019     |   0.964
The quality of the AI generated output remains...      |     0.713     |   0.000
The AI agent trained on new supplier documents...      |     0.768     |  -0.071
----------------------------------------------------------------------------------------------
CONSTRUCT: PROCESS PERSPECTIVE: INTEGRATION
OVERALL ALPHA (α): 0.377   |   Sample Size (N): 11
----------------------------------------------------------------------------------------------
Statement                                              | Item-Tot Corr | Alpha w/o
----------------------------------------------------------------------------------------------
The output of the AI agent can be used within the...   |     0.225     |   0.298
The AI agent output can be used with Copilot Studio... |     0.190     |   0.448
The AI agent fits into current set of technology...    |     0.302     |   0.173
----------------------------------------------------------------------------------------------
```

Figure E.4: Cronbach's Alpha Terminal Result At The Lowest Level Of Data For Business Perspective



```
----------------------------------------------------------------------------------------------
CONSTRUCT: LEARNING & GROWTH: EMPLOYEE EXPERIENCE
OVERALL ALPHA (α): 0.186   |   Sample Size (N): 11
----------------------------------------------------------------------------------------------
Statement                                              | Item-Tot Corr | Alpha w/o
----------------------------------------------------------------------------------------------
The AI agent shows potential to reduce tedious and...  |     0.209     |   0.026
The AI agent allows employee to focus on higher-value... |   0.247     |   0.049
There will be significant negative impact to employee... |   0.037     |   0.783
----------------------------------------------------------------------------------------------
CONSTRUCT: LEARNING & GROWTH: INNOVATIVENESS
OVERALL ALPHA (α): 0.806   |   Sample Size (N): 11
----------------------------------------------------------------------------------------------
Statement                                              | Item-Tot Corr | Alpha w/o
----------------------------------------------------------------------------------------------
The use of AI agent is a novel approach.               |     0.821     |   0.538
The pilot program provides valuable insights about AI... |   0.748     |   0.640
The experience with this AI trial may influence...     |     0.489     |   0.894
----------------------------------------------------------------------------------------------
CONSTRUCT: LEARNING & GROWTH: SKILL DEVELOPMENT
OVERALL ALPHA (α): 0.803   |   Sample Size (N): 11
----------------------------------------------------------------------------------------------
Statement                                              | Item-Tot Corr | Alpha w/o
----------------------------------------------------------------------------------------------
Employees are able to understand the output of the AI... |   0.481     |   0.892
Employees possess the necessary skills to effectively... |   0.667     |   0.716
The level of training required to use the output of... |     0.911     |   0.400
==============================================================================================
```

Figure E.5: Cronbach's Alpha Terminal Result At The Lowest Level Of Data For Learning and Growth Perspective

Appendix F

Formulae and Method to Evaluate the AI Agent with BSC

Balanced Scorecard Evaluation Measure In this section we will cover how to evaluate the performance of the balanced scored card that is designed for our case study. This approach is adapted from the research paper Batada and Rahman [5] that uses weighted average method and mentioned equal weight-age needs to be given to all the dimensions and then calculate the overall performance of the balanced scorecard.

The calculation follows a three step approach for each perspective that is listed below:

1. **Normalization Of Dimension Weights**: The factor loadings for each survey statements is divided by the sum of all the factor loadings for the given dimension to give a normalized weight that would sum to 1.0 as described in the research paper [5].
2. **Calculation of Weighted Dimension Score For A Perspective**: The average survey response of each survey statement will be multiplied by its normal weight. These are then summed and then divided by the maximum scale value that is 5 that helps to convert the matrix score that will range from 0.0 to 1.0, that is inline with the research paper approach [5].
3. **Summation Of The Perspective Scores**: We will now add all the performance score of the perspectives and divide it four since there are four perspective, to determine the overall performance score for our balanced scorecard [5]. In simple term, we are doing a average method to calculate the overall performance score of the balanced scorecard.

$$P_x = \frac{\sum_{k=1}^{m} M_k}{m}$$

P_x= The overall score for the perspectives x.

M_k= The standardized score for the dimension for that perspective k.

m= The total number of perspectives.

Appendix G

AI Agent Evaluation with BSC Approach

Balanced Scorecard Performance Evaluation

As mentioned in the section F, we followed and modified the approach to arrive at the overall performance of the balanced scorecard.

Let's proceed and see step by step calculation of the performance of AI agent for each perspective below:

- **Financial Perspective**

Table G.1: Performance Evaluation Score Of Financial Perspective

Dimension Name	**Normalized Weight [Q1,Q2, Q3]**	**Average Response[Q1, Q2, Q3]**	**Weighted Score[Q1, Q2, Q3]**	**Final Score** $[sum(weightedscore)/\mathbf{5}]$
Cost Efficiency	$[W_(q1) = 0.554/2.03]$, $[W_(q2) = 0.992/2.03]$, $[W_(q3) = 0.554/2.03]$	[4.64], [4.73], [4.64]	[0.273 * 4.64], [0.454 * 4.73], [0.273 * 4.64]	0.936
Value Creation	$[W_(q1) = 0.584/2.42]$, $[W_(q2) = 0.934/2.42]$, $[W_(q3) = 0.903/2.42]$	[4.00], [4.55], [4.73]	[0.241 * 4.00], [0.386 * 4.55], [0.373 * 4.74]	0.897
Scalability	$[W_(q1) = 0.861/2.51]$, $[W_(q2) = 0.731/2.51]$, $[W_(q3) = 0.920/2.51]$	[4.91], [4.09], [4.55]	[0.343 * 4.91], [0.291 * 4.09], [0.367 * 4.55]	0.909

$$Performance\ of\ the\ Financial\ Perspective = sum[0.936 + 0.897 + 0.909]/\ 3 = 0.914$$

- **User Perspective**

Table G.2: Performance Evaluation Score Of User Perspective

Dimension Name	**Normalized Weight [Q1,Q2, Q3]**	**Average Response[Q1, Q2, Q3]**	**Weighted Score[Q1, Q2, Q3]**	**Final Score [$sum(weightedscore)/$5]**
User Experience	$[W_(q1) = 0.266/2.02]$, $[W_(q2) = 0.854/2.02]$, $[W_(q3) = 0.895/2.02]$	[3.73], [4.00], [4.36]	[0.273 * 3.73], [0.454 * 4.00], [0.273 * 4.36]	0.823
Accuracy	$[W_(q1) = 0.876/2.67]$, $[W_(q2) = 0.841/2.67]$, $[W_(q3) = 0.950/2.67]$	[4.09], [4.18], [4.09]	[0.328 * 4.09], [0.315 * 4.18], [0.356 * 4.09]	0.823
User Adoption	$[W_(q1) = 0.861/2.36]$, $[W_(q2) = 0.731/2.36]$, $[W_(q3) = 0.920/2.36]$	[4.36], [4.55], [4.00]	[0.373 * 4.36], [0.371 * 4.55], [0.254 * 4.00]	0.866

Performance of the User Perspective $= sum[0.823 + 0.823 + 0.866]/\ 3 = 0.837$

- **Internal Business Process Perspective**

Table G.3: Performance Evaluation Score Of Internal Business Process Perspective

Dimension Name	**Normalized Weight [Q1,Q2, Q3]**	**Average Response[Q1, Q2, Q3]**	**Weighted Score[Q1, Q2, Q3]**	**Final Score [$sum(weightedscore)/$5]**
Efficiency	$[W_(q1) = 0.796/2.55]$, $[W_(q2) = 0.917/2.55]$, $[W_(q3) = 0.834/2.55]$	[4.82], [4.64], [4.45]	[0.312 * 4.82], [0.360 * 4.64], [0.327 * 4.45]	0.926
Consistency	$[W_(q1) = \text{-}0.038/1.93]$, $[W_(q2) = 0.983/1.93]$, $[W_(q3) = 0.982/1.93]$	[4.18], [4.09], [4.00]	[-0.020 * 4.00], [0.509 * 4.55], [0.509 * 4.74]	0.807
Integration with Existing Tools	$[W_(q1) = 0.702/2.04]$, $[W_(q2) = 0.561/2.04]$, $[W_(q3) = 0.773/2.04]$	[4.36], [3.55], [3.91]	[0.344 * 4.91], [0.275 * 3.55], [0.379 * 3.91]	0.842

Performance of the Internal Business Process Perspective $= sum[0.926 + 0.807 + 0.842]/\ 3 = 0.842$

- **Learning and Growth Perspective**

Table G.4: Performance Evaluation Score Of Learning and Growth Perspective

Dimension Name	**Normalized Weight [Q1,Q2, Q3]**	**Average Response[Q1, Q2, Q3]**	**Weighted Score[Q1, Q2, Q3]**	**Final Score [$sum(weightedscore)/$5]**
Employee Experience	$[W_(q1) = 0.913/1.91]$, $[W_(q2) = 0.912/1.91]$, $[W_(q3) = 0.090/1.91]$	[4.82], [4.91], [4.09]	[0.478 *4.82], [0.478*4.91], [0.047 * 4.09]	0.969
Innovativeness	$[W_(q1) = 0.928/2.54]$, $[W_(q2) = 0.893/2.54]$, $[W_(q3) = 0.722/2.54]$	[4.73], [4.64], [4.91]	[0.365 * 4.73], [0.352 * 4.64], [0.284 * 4.91]	0.951
Skill Development	$[W_(q1) = 0.721/2.53]$, $[W_(q2) = 0.841/2.53]$, $[W_(q3) = 0.970/2.53]$	[4.64], [3.27], [3.91]	[0.285 * 4.64], [0.332 * 3.27], [0.383 * 3.91]	0.781

Performance of the Learning and Growth Perspective $= sum[0.969 + 0.951 + 0.781]/\ 3 = 0.900$

Overall Performance of the Balanced Scorecard $sum[0.914 + 0.837 + 0.842 + 0.900]/\ 4 = 0.873$

Based on the research conducted by Batada and Rahman [5], we have seen the balanced scorecard performance scale for an excellent performance lies in the range of 0.81 and 1.0. This quantitative study successfully helps us to validate that balanced scorecard used for this particular case study of freight documents can be used for evaluating an AI agent for business managers.