

Master Thesis
for Attainment of the Degree

Master of Science (MSc)

at the RWTH Aachen Business School

Harnessing the Power of Large Language Models: A Business Infrastructure Approach

Submitted by: Nagarjuna Gottipati*

Matriculation Number: 433410

First Examiner: Prof. Dr. Stefanie Paluch[†]

Second Examiner: Prof. Dr. Britta Peis[‡]

Abstract

Numerous Large Language Models (LLMs) are tailored for specific business purposes, with various frameworks designed to integrate these technologies into organizational structures. This thesis aims to analyse frameworks for integrating Generative AI and Large Language Models (LLMs) into business applications, emphasizing their potential benefits and challenges. It specifically explores a use case in the insurance industry, focusing on how these technologies can enhance efficiency and customer satisfaction using Azure Databricks. Additionally, the thesis examines the LLMOPS paradigm, which addresses the continuous integration and deployment of LLMs within corporate infrastructures after their development. This approach seeks to uncover practical insights and methodologies to effectively implement Generative AI in business contexts.

Keywords: *Generative AI, ChatGPT, LLMS, LLMOPS, Langchain, Streamlit, Pinecone.*

*Email: nagarjuna.gottipati@rwth-aachen.de, RWTH Business School, Germany

[†]Email: paluch@time.rwth-aachen.de

[‡]Email: britta.peis@oms.rwth-aachen.de

Contents

List of Figures	3
List of Tables	4
List of Abbreviations	5
1 Introduction	6
1.1 Thesis Objectives	8
2 Review of State of Art Existing Frameworks	9
2.1 Technology Acceptance Model (TAM)	9
2.2 Task-Technology Fit (TTF)	11
2.3 Adoption of ChatGPT Among Business and Management Students: A IDT Perspective	14
2.4 TOE-Based Adoption of Business Analytics in Manufacturing	16
2.5 Synthesis Across Theories	19
2.6 Challenges and Future Directions	19
3 Methodology	21
3.1 Overview	21
3.2 Frameworks	21
3.3 Survey Data Collection Using the TOE Framework	22
3.4 DEMATEL Methodology	22
3.5 ISM Methodology	30
3.6 Accessing the Notebooks and Plugging in Survey Data	37
4 LLM development	39
4.1 Training data	39
4.2 Overview of Chatbot Architecture	40
4.3 Vector Databases in Chatbot Architecture	43
4.4 Post-Development Operations (LLMOPS)	48
4.5 Setting Up and Deploying the Insurance Chatbot Application	50
4.6 Accessing the Deployed Insurance Chatbot on Streamlit	51
5 Conclusion and outlook	52
5.1 Future outlook	53
6 List of References	54
A Repository Access	58
B Survey Questionnaire	59

List of Figures

1	TAM Illustration.	9
2	TTF Illustration.	11
3	IDT Illustration.	14
4	TOE Illustration.	17
5	Methodology Illustration.	21
6	DEMATEL Flowchart.	22
7	Centrality Vs Causality graph.	29
8	ISM Flowchart.	30
9	MICMAC Cluster Analysis.	36
10	Overview of Notebooks.	37
11	LLM Training data.	39
12	Chatbot Architecture.	40
13	Text Splitting.	41
14	Overview of Embeddings in Vector Databases.	44
15	Market Overview.	47
16	LLMOPS.	48
17	Ngrok Token.	50
18	Streamlit UI.	51

List of Tables

1	Hypothesis Testing: Significant at $p < 0.001$	10
2	Hypothesis Testing Results for ChatGPT Adoption	15
3	Factors influencing LLM Adoption based on the TOE Framework.	24
4	Direct Influence Matrix for LLM Adoption Factors.	25
5	Normalized Direct Influence Matrix for factors influencing LLM adoption.	25
6	Comprehensive Influence Matrix summarizing cumulative direct and indirect influences for the identified factors.	26
7	Centrality and Causality Analysis	27
8	Self-Interaction Matrix (SIM).	31
9	Adjacency Matrix (Binary)	32
10	Comprehensive Influence Matrix	33
11	Reachability Matrix derived from the Comprehensive Influence Matrix.	34
12	MICMAC Analysis Data	35
13	Comparison between Traditional Database and Vector Database.	43
14	Summary of Distance Metrics for Vector Databases	45
15	Indexing Strategies for Vector Databases	45
16	Comparison Between Vector and Vector Databases	46

List of Abbreviations

Abbreviation	Definition
AI	Artificial Intelligence
LLM	Large Language Model
LLMOPS	Large Language Model Operations
GDPR	General Data Protection Regulation
GDV	Gesamtverband der Deutschen Versicherungswirtschaft
BaFin	Federal Financial Supervisory Authority
EIOPA	European Insurance and Occupational Pensions Authority
RMF	Risk Management Framework
RAG	Retrieval-Augmented Generation
GPT	Generative Pre-trained Transformer
InsurTech	Insurance Technology
Azure	Microsoft Azure (Cloud Computing Platform)
GDV	Gesamtverband der Deutschen Versicherungswirtschaft
API	Application Programming Interface
EU	European Union
MIT	Massachusetts Institute of Technology
RAG	Retrieval-Augmented Generation
BA	Business Analytics
BI	Behavioral Intention
BDA	Big Data Analytics
CRM	Customer Relationship Management
DOI	Diffusion of Innovation
GDPR	General Data Protection Regulation
HIPAA	Health Insurance Portability and Accountability Act
IDT	Innovation Diffusion Theory
ILS	Integrated Library System
IoT	Internet of Things
MLA	Mobile Library Application
NLP	Natural Language Processing
PEOU	Perceived Ease of Use
PU	Perceived Usefulness
RMF	Risk Management Framework
SQ	System Quality
TAM	Technology Acceptance Model
TOE	Technology-Organization-Environment Framework
TTF	Task-Technology Fit
DEMATEL	Decision-Making Trial and Evaluation Laboratory
SIM	Self-Interaction Matrix
ISM	Interpretive Structural Modeling
Streamlit	Python-based Framework for Developing Data Science Applications

1 Introduction

The advent of Generative AI and Large Language Models (LLM) has revolutionized how businesses approach complex tasks, leveraging artificial intelligence Michael Schrage and David Kiron (2024) to drive efficiency, innovation, and customer satisfaction. Over the past few years, these technologies have evolved from experimental tools to strategic assets, increasingly integrated Kartik Hosanagar and Ramayya Krishnan (2024) into enterprise workflows across diverse sectors. This thesis explores frameworks for integrating Generative AI and LLM into business applications Ramakrishnan (2024), emphasizing their transformative potential and the challenges organizations encounter during implementation. Focusing on a use case in the insurance industry, it evaluates the deployment of these technologies using Azure Databricks and pinecone and examines the broader implications of the LL-MOPS paradigm for managing LLM in corporate infrastructures.

The Rise of Generative AI and LLMs in Business

Generative AI refers to systems capable of producing content—text, images, or even code—based on large-scale data and sophisticated algorithms. LLMs, as a subset of Generative AI, are particularly adept at processing and generating human-like text, making them invaluable for tasks like summarization, translation, sentiment analysis, and content creation. Prominent models such as OpenAI’s GPT-4, Meta’s Llama series, and Anthropic’s Claude have demonstrated remarkable capabilities, showcasing how AI can augment human decision-making, automate repetitive tasks, and enhance customer interactions.

Businesses increasingly recognize the strategic importance of LLMs. From summarizing vast datasets to generating insights that inform decision-making, LLMs are reshaping industries such as healthcare, finance, and retail Tucker J. Marion and Frank Piller (2024). A compelling example is CarMax’s use of LLMs to create tailored content for its vast inventory, achieving scalability and search engine optimization that would have been otherwise unattainable. Similarly, tools like GitHub Copilot illustrate how LLMs can streamline software development, enhancing productivity and developer satisfaction Kartik Hosanagar and Ramayya Krishnan (2024).

Challenges in LLM Integration

Despite their potential, integrating LLMs into business workflows poses significant challenges. Hallucinations—instances where models produce plausible but incorrect outputs—remain a critical concern, particularly in high-stakes domains like insurance and healthcare. Additionally, the adaptability of these models to domain-specific contexts requires meticulous customization, often through techniques such as prompt engineering, retrieval-augmented generation (RAG), and instruction fine-tuning Renée Richardson Gosline et al. (2024). These processes demand not only technical expertise but also substantial computational resources, highlighting the cost and complexity of deploying LLMs at scale.

Furthermore, organizations grapple with issues of openness and transparency in LLM development. While open-source models like Meta’s Llama series foster collaboration and innovation, they also introduce risks related to security and competitive advantage. Striking a balance between leveraging open frameworks and safeguarding proprietary data is crucial for businesses aiming to integrate LLMs responsibly.

Case Study: Insurance Industry Use Case

The insurance industry presents a compelling case for the transformative potential of integrating Large Language Models (LLMs) into its workflows. As a sector inherently defined by intricate regulatory frameworks, high volumes of customer interactions, extensive and often intricate documentation, and the demand for highly personalized solutions, the insurance domain stands to gain substantial value from the capabilities of Generative AI. The applications of LLMs in this space are vast and impactful, spanning automation in claims processing, enhancing the comprehension and summarization of complex policy documents, and revolutionizing customer support through advanced conversational agents.

This thesis leverages Azure Databricks and Pinecone as the foundational platforms for the development and deployment of LLM-driven solutions. By examining the technical integration of LLMs into critical insurance processes, this work explores how Generative AI can streamline operational workflows, improve the precision and efficiency of risk assessments, and elevate customer satisfaction through faster and more personalized interactions Ramakrishnan (2024). The study includes a detailed case study, showcasing methodologies for fine-tuning LLMs to accommodate industry-specific jargon, regulatory complexities, and compliance requirements. These practical insights aim to illustrate both the tangible benefits and the inherent challenges associated with implementing Generative AI in real-world insurance scenarios.

An essential component of this thesis is the alignment of technological innovation with the evolving regulatory frameworks that govern the insurance industry. In the European insurance landscape, particularly in Germany, institutions such as the Gesamtverband der Deutschen Versicherungswirtschaft (GDV), BaFin (Federal Financial Supervisory Authority), EIOPA (European Insurance and Occupational Pensions Authority), and Insurance Europe Mun (2024) play pivotal roles in shaping the regulatory and innovation agenda. These institutions set guidelines that address critical areas such as data protection, ethical AI deployment, risk management, and fostering innovation within the InsurTech ecosystem. Successful deployment of LLMs requires rigorous adherence to these frameworks, ensuring that AI systems operate transparently, ethically, and in compliance with local and international standards.

This thesis delves deeply into how LLMs can be fine-tuned to not only process complex insurance datasets but also account for domain-specific compliance requirements. For instance, ensuring compliance with GDPR regulations while handling sensitive customer data, integrating BaFin's stringent risk management guidelines, and aligning with EIOPA's directives on digital transformation in insurance are key focal points. By embedding these regulatory considerations directly into the model training and deployment pipeline, Generative AI can support insurance companies in achieving operational excellence without compromising governance and ethical Francisco Castro et al. (2024) standards.

Furthermore, this work explores how LLMs can enhance the insurance industry's capacity to address emerging challenges Le Nguyen (2023), such as managing climate risk, adapting to shifts in customer behavior, and navigating the rapidly evolving competitive landscape. By processing vast amounts of unstructured data, LLMs can provide actionable insights Ins (2023) that inform underwriting decisions, fraud detection systems, and customer segmentation strategies. Additionally, their ability to dynamically learn and adapt to new regulatory changes ensures that insurance companies remain at the forefront of compliance and innovation.

The importance of staying updated on policy developments cannot be overstated. Regulatory bodies such as GDV, BaFin, EIOPA, and Insurance Europe not only shape the governance of insurance operations but also influence the integration of cutting-edge technologies like Generative AI. Their

frameworks drive innovation while safeguarding customer trust, making it imperative for AI solutions to align with their guidelines. This thesis examines how these institutions' policies directly impact the design and implementation of AI systems, emphasizing the critical need for a collaborative approach between technology providers, insurers, and regulators.

By combining advanced technical capabilities with a nuanced understanding of the regulatory environment, this thesis highlights how Generative AI can become a cornerstone of the insurance sector's digital transformation Mollick (2024). It demonstrates that the integration of LLMs into insurance operations is not just about innovation—it is about fostering trust, ensuring compliance adhering to RMF framework (see, RMF), and unlocking the potential for a smarter, more efficient, and customer-centric industry. Through this comprehensive exploration, the work underscores the transformative power of Generative AI when aligned with the highest standards of governance, ethics, and regulatory oversight.

1.1 Thesis Objectives

1. **Comprehensive Analysis:** This thesis aims to provide a detailed analysis of the frameworks and methodologies necessary for the effective integration of Generative AI and Large Language Models (LLMs) into business applications.
2. **Industry-Specific Model Development:** The focus is on developing a robust LLM tailored specifically for the insurance industry. The model will be trained on a comprehensive dataset that includes general information on policies, compliance standards, and emerging innovations, ensuring it is equipped to address the unique challenges of the sector while adhering to regulatory frameworks.
3. **Post-Deployment Management:** The thesis emphasizes the adoption of the LLMOPS (Large Language Model Operations) methodology to manage the developed application post-deployment. This includes continuous monitoring, updating, and fine-tuning to maintain the model's relevance, accuracy, and efficiency in responding to evolving industry demands and regulatory changes.

2 Review of State of Art Existing Frameworks

The integration of technology into organizational processes demands robust theoretical frameworks to comprehend adoption dynamics and optimize implementation strategies. This literature review delves into four foundational theories—the Technology Acceptance Model (TAM), Task-Technology Fit (TTF), Diffusion of Innovation Theory (IDT), and Technology-Organization-Environment Framework (TOE)—examining their applications, synergies, and implications in domains such as healthcare, insurance, and IT services. Together, these theories offer a multifaceted lens to enhance understanding and foster informed decision-making in technology adoption.

2.1 Technology Acceptance Model (TAM)

The TAM, introduced by Davis (1989), posits that perceived ease of use (PEOU) and perceived usefulness (PU) are pivotal in determining a user’s intention to adopt and use technology. Over the years, TAM (see, e.g., Nikola Marangunic and A. Granić (2014)) has been widely adapted and integrated across various disciplines like adopting technologies for instance e-learning Natasia et al. (2022) and Mailizar et al. (2021) showcasing its versatility and utility.

Core Constructs and Relationships

The proposed research model Rafique et al. (2020) builds upon the traditional Technology Acceptance Model (TAM) by incorporating external variables such as Habit (H) and System Quality (SQ) to understand their impact on Perceived Ease of Use (PEOU) and Perceived Usefulness (PU). This model aims to address the behavioral intention (BI) to use mobile library applications (MLA), specifically the INSIGNIA ILS application, in a developing country context.

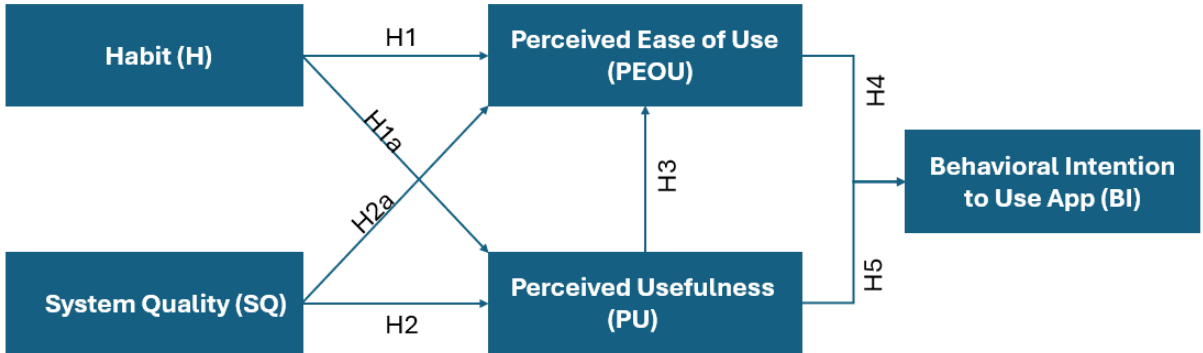


Figure 1: TAM Illustration.

Habit (H)

Habit is described as an individual’s tendency to perform tasks based on prior behavior. This construct significantly affects both *Perceived Ease of Use (PEOU)* and *Perceived Usefulness (PU)*. Habit makes using technology natural and intuitive, reducing cognitive load. For example, individuals accustomed to using mobile technology find it easier to adapt to new applications like Mobile Library Applications (MLA). Additionally, habit fosters trust in technology’s ability to improve task performance, as familiarity with similar tools reassures users of its usefulness.

System Quality (SQ)

System Quality encapsulates the performance, reliability, and user-friendliness of a system. High system quality ensures smooth and error-free interaction, making the system easy to use. Furthermore, a reliable system increases the perception that it will effectively enhance productivity and achieve desired outcomes, thereby contributing to both *PEOU* and *PU*.

Perceived Ease of Use (PEOU)

Perceived Ease of Use refers to the extent to which a user believes that using a system will require minimal effort. If a system is easy to use, users are more likely to perceive it as beneficial, thereby positively influencing *PU*. Additionally, *PEOU* directly impacts Behavioral Intention (*BI*), as users prioritize ease in their adoption decisions.

Perceived Usefulness (PU)

Perceived Usefulness is the degree to which users believe the technology will enhance their performance. It plays a pivotal role in influencing Behavioral Intention (*BI*), as users are more likely to adopt a system that they perceive as valuable and beneficial to their tasks.

Behavioral Intention to Use (BI)

Behavioral Intention to Use serves as the dependent variable, representing users' motivation to use the application. Both *PU* and *PEOU* are significant predictors of *BI*, indicating that users are likely to adopt the application if it is both beneficial and easy to use.

Hypothesis	B Values	t-values	Status
H1: $H \rightarrow PU$	0.451	9.823	Accepted
H1a: $H \rightarrow PEOU$	0.501	8.863	Accepted
H2: $SQ \rightarrow PEOU$	0.491	5.576	Accepted
H2a: $SQ \rightarrow PU$	0.480	7.456	Accepted
H3: $PEOU \rightarrow PU$	0.506	8.360	Accepted
H4: $PEOU \rightarrow BI$	0.388	6.916	Accepted
H5: $PU \rightarrow BI$	0.344	5.349	Accepted

Table 1: Hypothesis Testing: Significant at $p < 0.001$

Conclusion

The findings from this study reveal significant insights into the adoption of mobile library applications (MLA) using an extended Technology Acceptance Model (TAM). Habit and system quality are influential factors that positively impact perceived ease of use (PEOU) and perceived usefulness (PU). Moreover, PEOU and PU are critical predictors of behavioral intention (BI) to adopt the application.

Key takeaways include:

- Habit was identified as the strongest predictor of PEOU, emphasizing the role of prior behavior in facilitating ease of use.
- System quality significantly influenced both PEOU and PU, underlining the importance of user-friendly and reliable system designs.
- Among the TAM constructs, PEOU showed a more substantial effect on BI compared to PU, indicating that ease of interaction with the application drives adoption more effectively.

These results highlight the need for application developers to prioritize user habit formation and high system quality in their designs. Future studies Wallace and Sheetz (2014) could expand this framework by exploring additional mediating factors and applying the model in different technological and cultural contexts.

2.2 Task-Technology Fit (TTF)

The TTF framework, developed by Goodhue and Thompson (1995), emphasizes the alignment between technology’s capabilities and the demands of specific tasks. TTF underscores that optimal alignment leads to enhanced performance and user satisfaction, making it particularly useful in explore the value provided by technologies (see, e.g., Muchenje and Seppänen (2023)) and Lin et al. (2020) in the Business usecases.

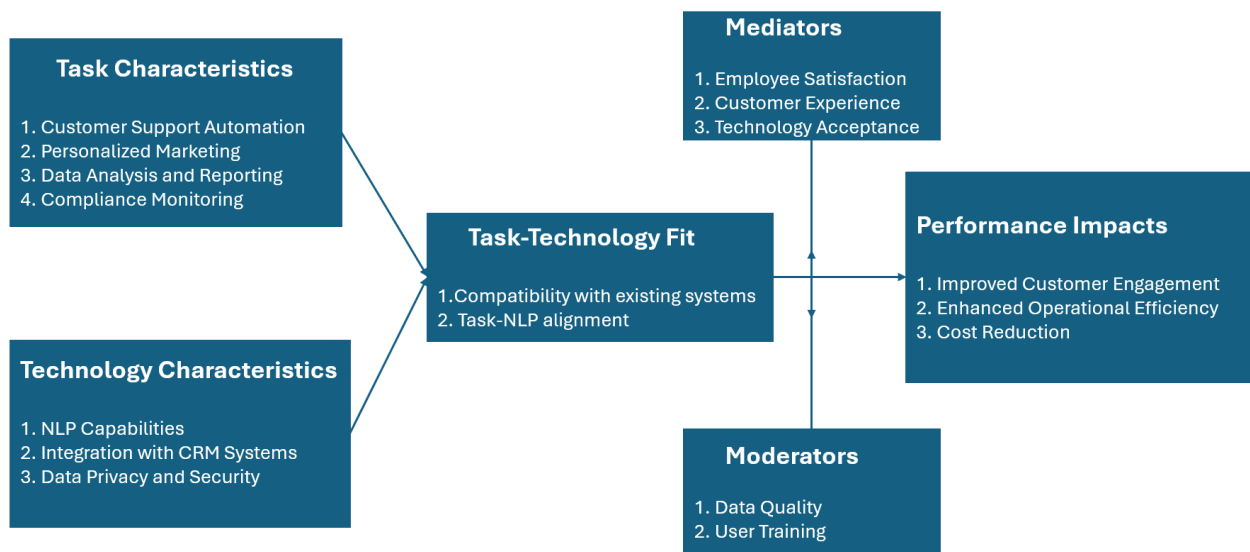


Figure 2: TTF Illustration.

i.) Task Characteristics

Define the specific needs and demands of a task that a technology is expected to address. Examples include:

- **Customer Support Automation:** Technologies such as chatbots streamline customer interactions.
- **Personalized Marketing:** Tools like recommendation engines enhance customer engagement.

- **Data Analysis and Reporting:** Analytical tools tailored to organizational goals improve decision-making.
- **Compliance Monitoring:** Automated systems ensure adherence to regulatory requirements.

ii.) Technology Characteristics

Represent the capabilities of the technology that enable task completion. These include:

- **NLP Capabilities:** Effective natural language processing for automated communication and content generation.
- **Integration with CRM Systems:** Seamless connectivity to customer management platforms for data-driven decision-making.
- **Data Privacy and Security:** Ensuring compliance with regulations like GDPR and protecting sensitive customer data.

iii.) Task-Technology Fit

The central construct that evaluates the compatibility between task demands and technology capabilities. Key factors include:

- **Compatibility with Existing Systems:** Technologies that integrate effortlessly into existing workflows ensure minimal disruption.
- **Task-NLP Alignment:** Technologies like LLMs (e.g., GPT-4) adapt to diverse linguistic and contextual requirements, improving alignment with user needs.

iv.) Mediators

Mediators explain how TTF influences performance outcomes. These include:

- **Employee Satisfaction:** Reduced frustration and improved productivity.
- **Customer Experience:** Enhanced interactions through faster, more accurate, and personalized service.
- **Technology Acceptance:** Positive attitudes toward the technology, fostering widespread adoption.

v.) Moderators

Moderators influence the strength of the relationship between TTF and performance outcomes:

- **Data Quality:** High-quality data ensures technologies generate meaningful insights and actionable results.
- **User Training:** Comprehensive training programs ensure employees can fully leverage the technology.

vi.)Performance Impacts

The ultimate goal of achieving TTF is to enhance performance in measurable ways:

- **Improved Customer Engagement:** Personalized interactions and faster response times.
- **Enhanced Operational Efficiency:** Streamlined workflows reduce bottlenecks and improve productivity.
- **Cost Reduction:** Automating repetitive tasks and optimizing resource utilization lower operational costs.

Key Insights in areas of Application

- **Big Data Analytics (BDA):** TTF's dynamic fit concept underscores the need for technology to adapt to evolving task requirements. Studies in BDA reveal that platforms tailored to organizational goals significantly improve decision-making and performance outcomes.
- **Marketing in Hospitality:** Aligning social media tools with specific marketing objectives—such as engagement tracking and audience targeting—demonstrates TTF's role in amplifying the effectiveness of tourism campaigns. Abdekhoda et al. (2022)
- **IT Services:** Iterative feedback loops and phased adoption strategies, guided by TTF, have been effective in addressing resistance to innovation. Aligning workflows with technology capabilities fosters smoother transitions and operational efficiency.

Strengths and Challenges

TTF's task-centric approach provides actionable insights for optimizing technology integration. However, its static nature requires enhancement to address dynamic environments where continuous re-alignment is necessary, such as in AI and advanced analytics systems.

2.3 Adoption of ChatGPT Among Business and Management Students: A IDT Perspective

Proposed by Rogers (1962), IDT explores how innovations spread through social systems, emphasizing the roles of adopters, innovation attributes, and communication channels. Artificial Intelligence (AI) and Natural Language Processing (NLP) technologies have transformed education, especially in business and management domains. ChatGPT, developed by OpenAI, is a prominent AI model capable of generating human-like text, facilitating interactive and adaptive learning. This study investigates the adoption of ChatGPT among business and management students using the Diffusion of Innovation (DOI) Guo and Huang (2024) Theory. The theory emphasizes attributes such as relative advantage, compatibility, complexity, trialability, and observability in influencing the adoption of innovations. ChatGPT offers unique advantages for management education, enabling students to enhance critical thinking, decision-making, and communication skills. However, its adoption depends on student attitudes, which are shaped by the perceived benefits and limitations of the technology. This paper examines how DOI theory can explain the factors influencing students' attitudes and behavioral intentions to adopt ChatGPT.

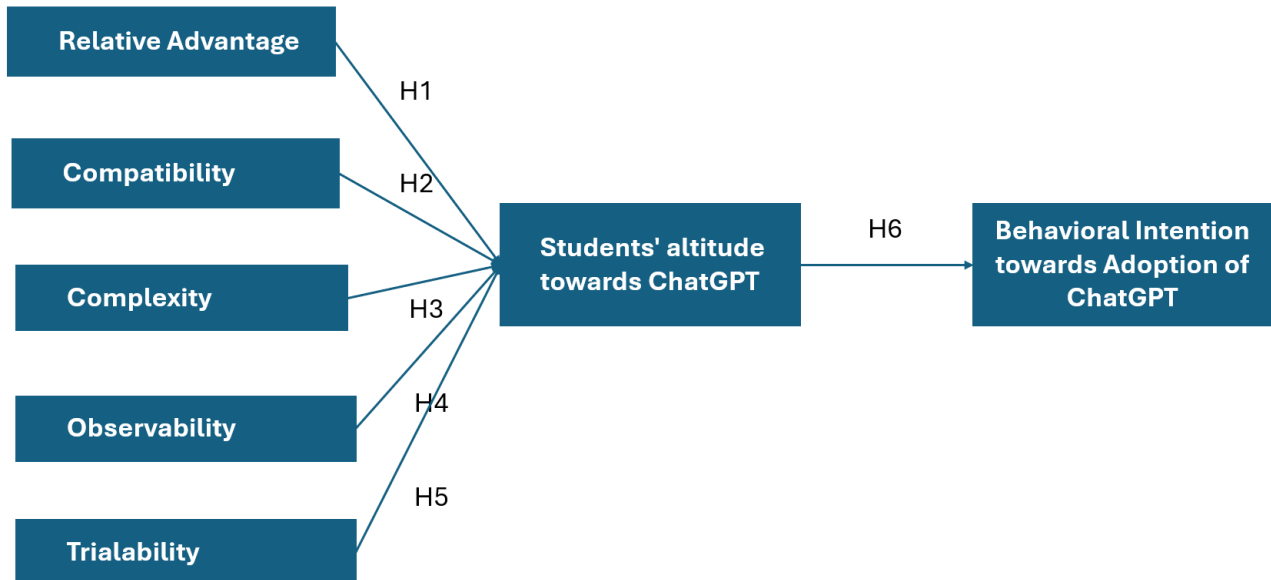


Figure 3: IDT Illustration.

Diffusion of Innovation Framework for ChatGPT Adoption

The Diffusion of Innovation (DOI) theory, proposed by Rogers (2003), provides a comprehensive framework for understanding the spread of innovations in social systems. In the context of ChatGPT, the concept of relative advantage refers to the degree to which ChatGPT is perceived as superior to traditional methods. In the study, students highlighted ChatGPT's ability to save time, enhance learning efficiency, and improve problem-solving skills (see, e.g., Sehgal and Jain (2024)). The tool's intuitive interface and ability to offer instant feedback make it a valuable addition to management education. Compatibility measures the extent to which ChatGPT aligns with existing student practices and values. Students reported that ChatGPT complements their learning styles by providing quick solutions to academic queries and simplifying complex topics. However, concerns about over-reliance and misuse

were also noted.

Complexity refers to the ease of use of an innovation. ChatGPT’s user-friendly design reduces cognitive load, making it accessible to students with varying technical skills. The system’s clarity and responsiveness encourage adoption by minimizing frustration associated with learning new technologies. Trialability is the extent to which students can experiment with ChatGPT before committing to its adoption. Students appreciated opportunities to explore the tool’s capabilities in a controlled environment, allowing them to assess its relevance and effectiveness in enhancing their academic performance. Observability pertains to the visibility of an innovation’s benefits to others. The widespread use of ChatGPT among peers creates a social proof effect, encouraging more students to adopt the tool. Observability also promotes knowledge sharing and collaboration among students.

Hypothesis Testing

The hypothesis testing results highlight key factors influencing the adoption of ChatGPT among students based on the Diffusion of Innovation (DOI) theory. Observability ($\beta = 0.430$, $p < 0.001$) and relative advantage ($\beta = 0.304$, $p < 0.001$) are the strongest predictors of positive attitudes, demonstrating that visible benefits and efficiency drive adoption. Compatibility ($\beta = 0.273$, $p < 0.001$) ensures alignment with students’ learning habits, while trialability ($\beta = 0.261$, $p < 0.001$) allows users to explore the tool confidently. Complexity ($\beta = 0.361$, $p < 0.001$) highlights the importance of user-friendly design in minimizing barriers. Finally, attitudes strongly predict behavioral intention ($\beta = 0.844$, $p < 0.001$), underlining the role of positive perceptions in driving adoption. These findings suggest that promoting visibility, ease of use, and compatibility can accelerate ChatGPT’s integration into education.

H	Path Relationship	β	p-value
H1	Relative Advantage \rightarrow Attitude Towards ChatGPT	0.304	< 0.001
H2	Compatibility \rightarrow Attitude Towards ChatGPT	0.273	< 0.001
H3	Complexity \rightarrow Attitude Towards ChatGPT	0.361	< 0.001
H4	Observability \rightarrow Attitude Towards ChatGPT	0.430	< 0.001
H5	Trialability \rightarrow Attitude Towards ChatGPT	0.261	< 0.001
H6	Attitude Towards ChatGPT \rightarrow Behavioral Intention Towards ChatGPT	0.844	< 0.001

Table 2: Hypothesis Testing Results for ChatGPT Adoption

Behavioral Intention and Attitude Toward ChatGPT

Students’ behavioral intention to adopt ChatGPT is influenced by their attitudes toward its perceived benefits and usability. Positive attitudes stem from the tool’s ability to simplify tasks, enhance productivity, and provide personalized learning experiences. The study found a significant correlation between students’ attitudes and their intention to use ChatGPT for academic purposes. Factors such as peer influence, institutional support, and prior exposure to AI technologies also play a crucial role in shaping these intentions. The structural model analysis revealed that relative advantage, compatibility, and trialability had the strongest impact on students’ attitudes, followed by complexity and observability.

These findings align with previous research, which highlights the importance of perceived ease of use and usefulness in technology adoption.

Implications for Educational Institutions

The adoption of ChatGPT in management education has significant implications for educators, administrators, and policymakers. To facilitate adoption, institutions should focus on integrating ChatGPT into business and management courses to enhance student engagement and foster practical skills. Faculty training programs should be organized to help educators leverage the tool effectively. Educational campaigns highlighting the benefits of ChatGPT can encourage adoption among students. Demonstrating its practical applications in real-world scenarios can further enhance its appeal. Institutions must establish guidelines to prevent misuse and promote responsible use of ChatGPT. Emphasizing critical thinking and analytical skills alongside AI tools can mitigate the risk of over-reliance. Simplifying access to ChatGPT and providing user support can address technical barriers and improve adoption rates. Tutorials and workshops can help students become familiar with the tool's functionalities.

Limitations and Future Research

While the study provides valuable insights into ChatGPT adoption, several limitations must be addressed. The research was conducted among a limited sample of business and management students in Delhi, which may limit the generalizability of the findings. Future studies should include diverse student populations across different regions and disciplines to gain a broader understanding of ChatGPT adoption. Additionally, the study focused primarily on student attitudes and intentions. Further research is needed to explore long-term usage patterns, learning outcomes, and the impact of ChatGPT on academic performance. Investigating the role of educators and institutions in facilitating adoption can also provide valuable insights.

Conclusion

The study highlights the potential of ChatGPT to revolutionize management education by enhancing learning efficiency, fostering critical thinking, and promoting innovation. Using the Diffusion of Innovation framework, the research identifies key factors influencing students' attitudes and behavioral intentions toward ChatGPT adoption. By addressing barriers such as complexity and ethical concerns, educational institutions can unlock the full potential of AI technologies in transforming the learning experience.

2.4 TOE-Based Adoption of Business Analytics in Manufacturing

The TOE framework, introduced by Tornatzky and Fleischer (1990), evaluates how technological, organizational, and environmental factors influence innovation adoption. Its holistic approach captures the complexity of external and internal influences, making it highly applicable across sectors. The manufacturing industry faces a rapidly evolving landscape driven by advancements in technology, data analytics, and market globalization (Omar et al. (2019)). To remain competitive, manufacturers must leverage business analytics (BA) as a critical tool to enhance productivity, improve decision-making, and achieve market leadership. This paper explores the adoption of business analytics within the Technological, Environmental, and Organizational Contexts, as depicted in the provided framework.

The adoption pathway highlights challenges, opportunities, and actionable steps to achieve successful integration in manufacturing operations Tiago Oliveira and Maria Fraga Martins (2011).

Introduction

The manufacturing industry faces a rapidly evolving landscape driven by advancements in technology, data analytics, and market globalization. To remain competitive, manufacturers must leverage business analytics (BA) as a critical tool to enhance productivity, improve decision-making, and achieve market leadership. This paper explores the adoption of business analytics within the Technological, Environmental, and Organizational Contexts, as depicted in the provided framework. The adoption pathway highlights challenges, opportunities, and actionable steps to achieve successful integration in manufacturing operations.

The Role of Business Analytics in Manufacturing

Business analytics in manufacturing encompasses tools and techniques aimed at harnessing data for improved decision-making, process optimization, and innovation. Unlike traditional business intelligence, which focuses on past performance, BA addresses “what-if” scenarios, predictive insights, and optimization strategies. Key areas of focus include technological readiness, system characteristics, and organizational factors that shape adoption. The integration of analytics within the manufacturing domain enables firms to achieve superior operational efficiency, reduce waste, and respond effectively to market demands.

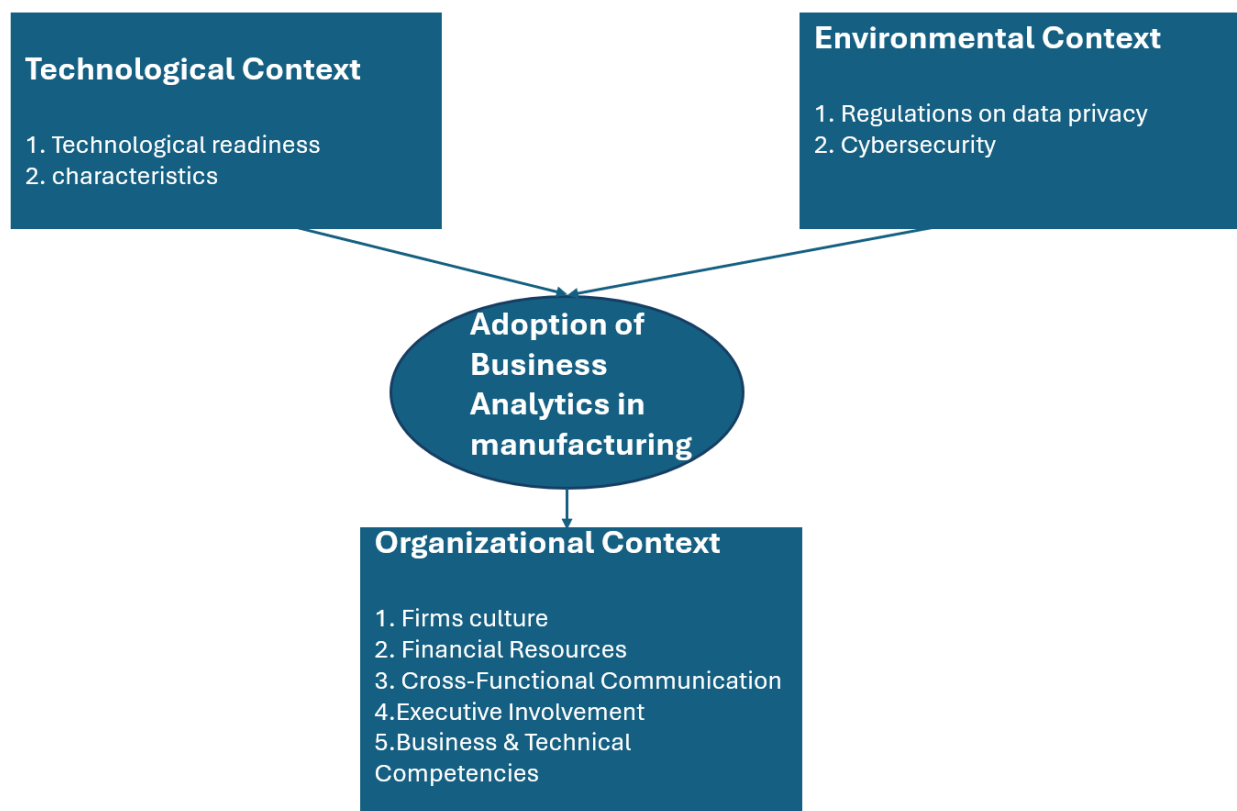


Figure 4: TOE Illustration.

Technological Context in BA Adoption

Technological readiness encompasses the maturity of an organization's tools, systems, and data capabilities to support analytics initiatives. The integration of cyber-physical systems and Internet of Things (IoT) technologies allows real-time data monitoring and optimization. High-quality data pipelines ensure seamless data flow from sensors, machinery, and enterprise resource planning systems to analytics platforms. Predictive maintenance, enabled by real-time monitoring, reduces downtime and enhances operational efficiency by identifying equipment failures before they occur. The characteristics of analytics tools, such as user-friendliness, scalability, and flexibility, play a vital role in adoption. Manufacturing enterprises must adopt analytics platforms that align with their unique operational needs while being robust enough to handle future expansions.

Environmental Context and Regulatory Concerns

The environmental context shapes BA adoption through regulations on data privacy, intellectual property, and cybersecurity. Organizations must adhere to regulations such as the General Data Protection Regulation (GDPR) when collecting, storing, and analyzing data. Ensuring consumer and employee data privacy builds trust and compliance. Additionally, the interconnected nature of modern supply chains makes manufacturing systems vulnerable to cyberattacks. Robust encryption and monitoring tools are essential to safeguard sensitive data. The rise of blockchain technology in manufacturing presents an opportunity to improve transparency and traceability, addressing regulatory concerns while enhancing operational efficiency.

Organizational Context: Culture, Communication, and Resources

A supportive organizational context is indispensable for BA adoption. A data-driven culture is essential to outgrow intuition and embrace evidence-based decision-making. Encouraging innovation and rewarding the use of analytics tools fosters an environment conducive to change. Cross-functional communication ensures that analytics insights are shared and implemented effectively across departments. For instance, insights derived from supply chain analytics can guide marketing, sales, and production strategies cohesively. Furthermore, financial and human resources are pivotal for adopting BA tools, hiring data scientists, and conducting employee training programs. Successful organizations invest in analytics not as a one-off project but as an integral part of their long-term strategy, ensuring sustained benefits and competitive advantage.

Pathway to BA Adoption in Manufacturing

The adoption of BA in manufacturing follows a structured pathway. Data collection, aggregation, and storage processes must first be standardized across the organization to eliminate silos and enable seamless analysis. Organizational culture must evolve to prioritize data over intuition, with executive leadership advocating for analytics-based strategies and allocating resources accordingly. Additionally, BA insights should guide innovation in business models, enabling manufacturers to develop value-added services, improve customer experiences, and secure market leadership.

Challenges in BA Adoption

Despite its potential, the adoption of BA faces significant challenges. Technological barriers, such as legacy systems, often hinder data integration and analytics implementation. Upgrading infrastructure requires significant investments, which many organizations find challenging. Resistance to change, especially from employees accustomed to traditional decision-making methods, can delay adoption. A lack of technical skills and competencies among personnel poses additional hurdles. Environmental constraints, including compliance with stringent regulations on data privacy and intellectual property, add to the complexity of adoption.

Conclusion

The adoption of business analytics in manufacturing offers unparalleled opportunities to enhance productivity, reduce costs, and drive innovation. However, success depends on overcoming challenges related to technology readiness, organizational culture, and regulatory compliance. By addressing these factors through the Technology-Organization-Environment (TOE) framework, manufacturers can unlock the full potential of analytics, transitioning from intuition-based decisions to data-driven strategies. Ultimately, this shift will pave the way for market leadership in an increasingly competitive global landscape.

2.5 Synthesis Across Theories

Integrating TAM, TTF, IDT, and TOE creates a robust, multidimensional framework for understanding and facilitating technology adoption. For instance, Integrating TAM and TTF Shih and Chen (2013) to understand user behaviour:

- **User-Centric Insights (TAM):** Identifies individual-level barriers, such as usability concerns, providing actionable insights for training and interface design.
- **Task Alignment (TTF):** Ensures that technologies meet specific operational requirements, enhancing efficiency and user satisfaction.
- **Contextual Dynamics (IDT):** Highlights societal and organizational factors influencing adoption, such as market trends and peer influence.
- **Environmental Interplay (TOE):** Bridges internal capabilities with external pressures, ensuring strategic alignment with regulatory and competitive landscapes.

2.6 Challenges and Future Directions

Despite the strengths of the integrated framework, several challenges must be addressed to enhance its applicability and relevance in the face of evolving technological landscapes:

Regulatory Barriers

Challenge: Regulatory compliance remains a significant barrier, especially in highly regulated sectors like healthcare and insurance.

Example:

- In healthcare, organizations must adhere to stringent data privacy laws like GDPR, HIPAA, or local regulations.
- In insurance, regulatory frameworks demand robust data governance and transparency in AI-driven decision-making.

Future Direction:

- Frameworks must integrate regulatory considerations into their models to ensure compliance without stifling innovation.
- Organizations should invest in adaptive technologies that can evolve with changing regulatory landscapes.

Evolving Technologies

Challenge: The rapid pace of advancements in AI, Big Data Analytics, and related fields requires frameworks that are flexible and dynamic.

Example:

- The emergence of generative AI models (e.g., GPT-4, GPT-5) has shifted the focus from traditional automation to creativity and complex decision-making.
- Big Data platforms demand scalability and interoperability with AI systems.

Future Direction:

- Expand existing frameworks to include mechanisms for continuous feedback and adaptation.
- Foster cross-disciplinary research to address the intersection of AI, data science, and operational management.

Sectoral Extensions

Challenge: Existing frameworks must be extended to address the unique demands of emerging technologies and industries.

Examples:

- **Blockchain:** Adoption frameworks must address issues of trust, scalability, and integration into existing systems.
- **Quantum Computing:** Early-stage frameworks are needed to guide adoption in this nascent field, considering its transformative potential in encryption, optimization, and AI.
- **Sustainable Technologies:** Green AI and energy-efficient technologies require adoption models that prioritize environmental impact.

Future Direction:

- Develop sector-specific frameworks that build on TAM, TTF, IDT, and TOE while addressing the nuances of emerging domains.
- Encourage collaboration between academia, industry, and regulatory bodies to create actionable frameworks.

3 Methodology

3.1 Overview

The methodology section serves as the cornerstone of this study, providing a detailed and systematic explanation of the approaches employed to achieve the research objectives related to the adoption of large language models (LLMs) in the insurance and banking sectors. This section is structured to ensure transparency, reproducibility, and clarity, aligning with the principles of scientific rigor. Given the complex and dynamic nature of the financial sector, this research utilizes a multi-framework approach to capture the multifaceted aspects of LLM adoption. The methodologies employed aim to evaluate not only the technological implications but also the organizational and environmental dimensions that influence the adoption process.

3.2 Frameworks

Specifically, this study integrates the *Technology-Organization-Environment (TOE) framework*, the *Decision-Making Trial and Evaluation Laboratory (DEMATEL) methodology*, and *Interpretive Structural Modeling (ISM)*. Each of these frameworks provides a unique lens for dissecting the challenges and opportunities associated with LLM implementation.

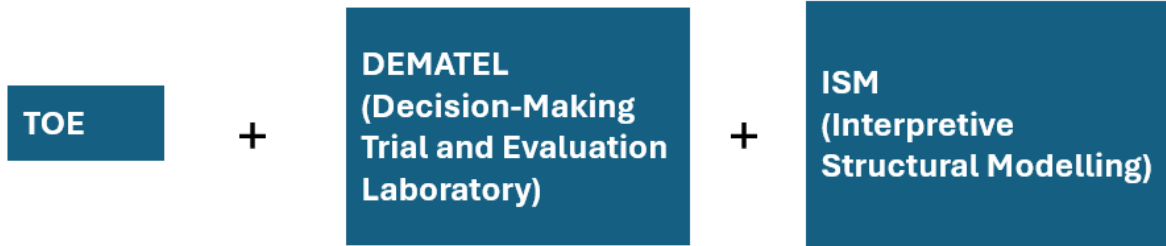


Figure 5: Methodology Illustration.

The combination of these methodologies allows for a comprehensive analysis that not only identifies key factors influencing LLM adoption but also provides actionable insights to guide decision-making. Each framework is supported by a robust data collection process involving expert consultations, surveys, and iterative validation to ensure accuracy and reliability. The methodologies are underpinned by a rigorous data collection process that incorporates expert consultations, carefully designed surveys, and iterative validation steps. Iterative validation ensures that the insights derived from the analysis are both robust and aligned with real-world complexities. This layered approach enhances the reliability, accuracy, and relevance of the findings.

The structured and scientific methodology employed in this research contributes significantly to the broader understanding of digital transformation in these sectors. It equips organizations with the tools and insights needed to navigate the complexities of technological adoption, optimize resource allocation, and achieve long-term success in leveraging LLMs. By adopting this approach, this study not only advances academic discourse but also delivers practical value to industry stakeholders, offering a roadmap for effective and sustainable integration of LLM technologies.

3.3 Survey Data Collection Using the TOE Framework

The *Technology-Organization-Environment (TOE) framework*, as extensively discussed in **subsection 2.4**, provides a structured foundation for analyzing the adoption of technological innovations such as Large Language Models (LLMs) in the insurance and banking sectors. By categorizing influential factors into three dimensions—*Technology*, *Organization*, and *Environment*—this framework ensures a holistic evaluation of adoption readiness and challenges. These dimensions consider key elements like technological characteristics, organizational readiness, and external pressures, offering valuable insights for decision-making.

To gather actionable data for evaluating LLM adoption, a systematic survey methodology is employed based on the TOE framework. First, a comprehensive literature review identifies the relevant factors and indicators for each dimension, aligning the survey with current trends and scientific insights. A structured questionnaire is then developed to address these dimensions, covering aspects such as perceived usefulness and compatibility for technology, resource availability and readiness for change for organization, and competitive pressures and regulatory compliance for environment. The questionnaire undergoes expert validation to ensure clarity and relevance. Surveys are distributed to key stakeholders, including IT managers and decision-makers, through online platforms to ensure scalability. The collected data is meticulously preprocessed to address inconsistencies, enabling reliable analysis aligned with research objectives.

3.4 DEMATEL Methodology

DEMATEL Flowchart (Steps 1-5)

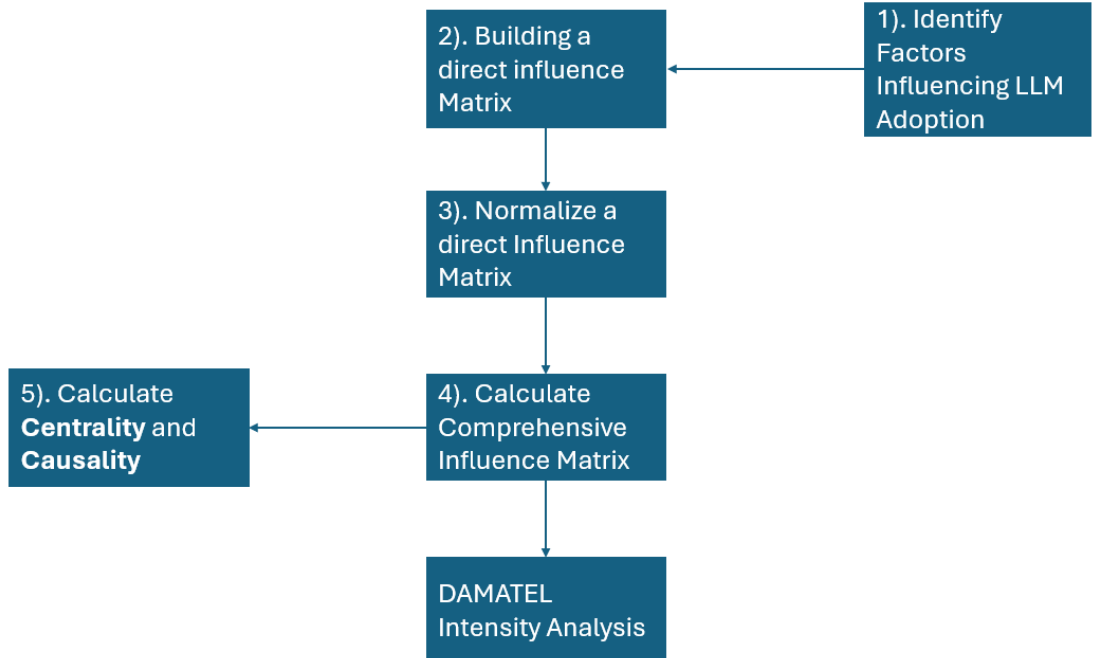


Figure 6: DEMATEL Flowchart.

The Decision-Making Trial and Evaluation Laboratory (*DEMATEL*) methodology Kao et al. (2022) is a robust analytical tool developed to visualize and quantify the interdependencies and causal relationships among factors in a complex system. By systematically analyzing these relationships, *DEMATEL*

helps researchers and decision-makers identify the most critical factors and their roles in influencing others, whether as drivers or effects. Falatoonitoosi et al. (2013) The core strength of *DEMATEL* lies in its ability to decompose a system into direct and indirect influences, providing a comprehensive picture of the system's dynamics. This is particularly valuable in domains where multiple interacting elements shape outcomes, such as technology adoption in industries like insurance and banking.

DEMATEL decomposes influences into direct, indirect, and combined effects, offering a holistic view of how factors interact within a system. Direct influences capture immediate relationships, while indirect influences highlight ripple effects through intermediary elements. The methodology categorizes factors into "causes," which drive changes, and "effects," which are influenced by them, enabling targeted strategies. Iterative analysis refines results for accuracy and actionable insights. *DEMATEL* also generates visual influence maps, making complex relationships intuitive and easy to understand. By quantifying interactions, it prioritizes interventions effectively, helping decision-makers focus on the most impactful causal factors.

Steps in DEMATEL Methodology

1. Identify Factors
2. Direct Influence Matrix
3. Normalized Influence Matrix
4. Comprehensive Influence Matrix
5. Centrality and Causality Analysis
6. Conclusion and Results

1. Identify Factors : The initial phase of the methodology focuses on selecting relevant factors, guided by the *Technology-Organization-Environment (TOE) Framework*, a well-established model for evaluating technological adoption. This selection process integrates insights from a comprehensive literature review, expert consultations, and survey data collection. A thorough review of academic and industry publications identifies key factors influencing the adoption of large language models (LLMs), ensuring alignment with recent technological advancements. Consultations with domain experts and industry professionals provide practical insights, complementing the theoretical findings from the literature. Structured surveys are then designed and distributed to stakeholders in the insurance and banking sectors, focusing on technological capabilities, organizational readiness, and environmental factors. This triangulated approach ensures that the factors selected are robust, contextually relevant, and comprehensive. By combining theoretical and empirical inputs, the process strikes a balance between scientific rigor and practical applicability, forming a strong foundation for subsequent analysis.

Dimensions	Factors	Coding
Technological	Data Security and Privacy	T1
	AI Infrastructure Synergy	T2
	Model Accuracy and Reliability	T3
	AI Explainability	T4
Environmental	Regulatory Compliance	E1
	Competitive Pressures	E2
	Lack of Customer Trust	E3
Organizational	Lack of Management Leadership Support	O1
	Financial Costs	O2
	Lack of Complex Talent	O3

Table 3: Factors influencing LLM Adoption based on the TOE Framework.

2. Direct Influence Matrix : The survey data forms a critical foundation for assessing the direct influences among the identified factors, offering a systematic approach to quantifying interdependencies. In this process, stakeholders—comprising domain experts, industry professionals, and decision-makers—are engaged to evaluate the extent to which one factor exerts a direct impact on another. These assessments are captured through structured questionnaires or interview-based feedback, ensuring alignment with the study’s objectives and methodological rigor.

The gathered evaluations are meticulously processed to construct the *Direct Influence Matrix*, a pivotal tool in the *DEMATEL* methodology. This matrix quantitatively encapsulates the relationships between factors, highlighting the intensity and direction of their influences. Each entry in the matrix represents the direct influence of one factor on another, based on stakeholder inputs. By leveraging this structured approach, the matrix not only provides clarity on the strength of individual relationships but also serves as the groundwork for subsequent analytical steps, including normalization and comprehensive influence analysis. For example, an entry in the matrix might indicate that Factor A exerts a moderate influence on Factor B, thereby revealing pathways of interdependence. This structured quantification ensures that the derived insights are robust, actionable in empirical evidence.

	T1	T2	T3	T4	E1	E2	E3	O1	O2	O3
T1	0	2	3	1	0	0	0	0	0	0
T2	1	0	2	1	0	0	0	0	0	0
T3	2	3	0	2	0	0	0	0	0	0
T4	1	1	2	0	0	0	0	0	0	0
E1	0	0	0	0	0	2	1	0	0	0
E2	0	0	0	0	2	0	1	0	0	0
E3	0	0	0	0	1	1	0	0	0	0
O1	0	0	0	0	0	0	0	0	2	1
O2	0	0	0	0	0	0	0	2	0	1
O3	0	0	0	0	0	0	0	1	1	0

Table 4: Direct Influence Matrix for LLM Adoption Factors.

3. Normalized Influence Matrix : The *Normalized Direct Influence Matrix* is a critical step within the Decision-Making Trial and Evaluation Laboratory (DEMATEL) methodology, designed to enhance the accuracy and comparability of influence values across diverse factors. This matrix is derived by normalizing the elements of the *Direct Influence Matrix*, ensuring that the influence values are transformed into proportions relative to the total influence exerted by each factor. By implementing this normalization process, the methodology accounts for variations in scale and facilitates a more balanced analysis of interrelationships between factors.

Factors	T1	T2	T3	T4	E1	E2	E3	O1	O2	O3
T1	0	0.33	0.50	0.17	0	0	0	0	0	0
T2	0.25	0	0.50	0.25	0	0	0	0	0	0
T3	0.29	0.43	0	0.29	0	0	0	0	0	0
T4	0.25	0.25	0.50	0	0	0	0	0	0	0
E1	0	0	0	0	0	0.67	0.33	0	0	0
E2	0	0	0	0	0.67	0	0.33	0	0	0
E3	0	0	0	0	0.50	0.50	0	0	0	0
O1	0	0	0	0	0	0	0	0	0.67	0.33
O2	0	0	0	0	0	0	0	0.67	0	0.33
O3	0	0	0	0	0	0	0	0.50	0.50	0

Table 5: Normalized Direct Influence Matrix for factors influencing LLM adoption.

Mathematically, normalization is expressed as:

$$\text{Normalized Value}_{ij} = \frac{\text{Element}_{ij}}{\text{Row Sum}_i}$$

In this formula, Element_{ij} represents the influence exerted by factor i on factor j , while Row Sum_i

is the total influence exerted by factor i on all other factors. The normalization process ensures that the sum of influences for each factor is standardized to unity, enabling a consistent comparison of relationships across all factors, regardless of their initial magnitudes.

This proportional transformation serves to mitigate potential biases arising from factors with inherently higher absolute influence values, thereby maintaining the integrity and scientific rigor of the analysis. By standardizing the influence values, the normalized matrix facilitates a robust and equitable assessment of the direct interdependencies among factors.

Furthermore, the *Normalized Direct Influence Matrix* serves as the foundational input for constructing the *Comprehensive Influence Matrix*, which integrates both direct and indirect influences. This subsequent step allows for a more holistic understanding of the systemic interrelations, enabling researchers and decision-makers to identify critical drivers and dependencies within complex systems. Through its methodical approach, the normalization process upholds the principles of scientific rigor and methodological consistency, ensuring reliable and actionable insights.

4. Comprehensive Influence Matrix : The *Comprehensive Influence Matrix* represents a critical analytical tool in the *DEMATEL* methodology. It encapsulates both direct and indirect influences among factors, thereby providing a holistic view of the interdependencies within a complex system. By integrating the effects of multiple levels of interactions, this matrix allows researchers to better understand the systemic behavior of the studied phenomena.

The computation of the Comprehensive Influence Matrix begins with the Normalized Direct Influence Matrix (N). The matrix is calculated using the formula:

$$M = N + N^2 + N^3 + \dots + N^\infty$$

In practice, this series converges mathematically and is simplified as:

$$M = (I - N)^{-1}$$

	T1	T2	T3	T4	E1	E2	E3	O1	O2	O3
T1	2.1E9	2.6E9	3.3E9	2.0E9	0	0	0	0	0	0
T2	2.1E9	2.6E9	3.3E9	2.0E9	0	0	0	0	0	0
T3	2.1E9	2.6E9	3.3E9	2.0E9	0	0	0	0	0	0
T4	2.1E9	2.6E9	3.3E9	2.0E9	0	0	0	0	0	0
E1	0	0	0	0	3.8E9	3.8E9	2.5E9	0	0	0
E2	0	0	0	0	3.8E9	3.8E9	2.5E9	0	0	0
E3	0	0	0	0	3.8E9	3.8E9	2.5E9	0	0	0
O1	0	0	0	0	0	0	0	3.8E9	3.8E9	2.5E9
O2	0	0	0	0	0	0	0	3.8E9	3.8E9	2.5E9
O3	0	0	0	0	0	0	0	3.8E9	3.8E9	2.5E9

Table 6: Comprehensive Influence Matrix summarizing cumulative direct and indirect influences for the identified factors.

Here, I is the identity matrix of the same dimensions as N , and N represents the Normalized Direct Influence Matrix. This formula ensures that the infinite series of influences is computationally feasible, allowing for efficient and accurate modeling of interactions.

The resulting matrix M quantifies the cumulative impact of each factor, considering both its immediate (direct) and mediated (indirect) influences. This comprehensive perspective is instrumental in identifying the most critical factors and their roles within the system, whether as drivers, linkages, or dependent elements. By leveraging the Comprehensive Influence Matrix, stakeholders can prioritize interventions, allocate resources efficiently, and make informed strategic decisions.

5. Centrality and Causality Analysis : Centrality and causality are fundamental metrics in the Decision-Making Trial and Evaluation Laboratory *DEMATEL* methodology, serving as critical tools for analyzing and understanding complex systems. These metrics provide a structured approach to identifying the relative importance of factors and their roles within the system, offering valuable insights into the dynamics of interdependent phenomena.

Centrality, denoted as $R + C$, represents the overall influence of a factor within the system. This metric is derived from the row sum (R), which measures the total influence a factor exerts on other factors, and the column sum (C), which quantifies the total influence a factor receives from others. Together, $R + C$ captures a factor's prominence and interconnectedness in the system. A higher centrality value signifies a more critical role, indicating that the factor is significantly involved in shaping and being shaped by the system's dynamics.

Causality, represented as $R - C$, differentiates factors into causal (driving) and effect (dependent) categories. A positive $R - C$ value indicates that a factor primarily acts as a driver, influencing other factors. Conversely, a negative $R - C$ value identifies a factor as an effect, meaning it is predominantly influenced by other factors. This distinction is essential for understanding whether a factor is a catalyst for change or a responsive element within the system. Identifying causal factors is particularly valuable, as they often serve as leverage points for strategic interventions.

Factor	Row Sum (R)	Column Sum (C)	Centrality (R+C)	Causality (R-C)
T1	1×10^{10}	8.38×10^9	1.84×10^{10}	1.62×10^9
T2	1×10^{10}	1.05×10^{10}	2.05×10^{10}	-4.6×10^8
T3	1×10^{10}	1.33×10^{10}	2.33×10^{10}	-3.3×10^9
T4	1×10^{10}	7.82×10^9	1.78×10^{10}	2.18×10^9
E1	1×10^{10}	1.12×10^{10}	2.12×10^{10}	-1.2×10^9
E2	1×10^{10}	1.12×10^{10}	2.12×10^{10}	-1.2×10^9
E3	1×10^{10}	7.5×10^9	1.75×10^{10}	2.5×10^9
O1	1×10^{10}	1.12×10^{10}	2.12×10^{10}	-1.2×10^9
O2	1×10^{10}	1.12×10^{10}	2.12×10^{10}	-1.2×10^9
O3	1×10^{10}	7.5×10^9	1.75×10^{10}	2.5×10^9

Table 7: Centrality and Causality Analysis

The combination of centrality and causality provides a comprehensive analytical framework. Cen-

trality identifies the most interconnected factors, guiding prioritization in system-wide strategies. Causality, on the other hand, enables decision-makers to focus on factors driving systemic behavior and design interventions accordingly. This dual perspective ensures that the DEMATEL methodology captures both the broad and nuanced aspects of factor interactions.

In the context of Large Language Model (LLM) adoption in the insurance and banking sectors, centrality and causality play a vital role in identifying technological, organizational, and environmental factors that influence adoption success. By leveraging these metrics, researchers and practitioners can uncover hidden relationships, allocate resources effectively, and design targeted strategies for systemic improvement.

6. Conclusion and Results : The Decision-Making Trial and Evaluation Laboratory (DEMATEL) analysis provided a systematic understanding of the interdependencies and causal relationships among factors influencing the adoption of large language models (LLMs) in the insurance and banking sectors. The results are depicted in the Centrality vs. Causality scatter plot, which offers valuable insights into the dynamics of key factors.

Causal and Effect Factors: Factors with positive causality values ($R - C > 0$), represented by red points, are classified as causal factors. These are the drivers of the system, exerting a significant influence on other factors. For instance, factors like *T1* (Data Security and Privacy) and *T3* (Model Accuracy and Reliability) appear prominently as causal factors. Conversely, factors with negative causality values ($R - C < 0$), shown as blue points, are effect factors. These factors are influenced by others, indicating their reactive role in the system. Examples include *O3* (Lack of Complex Talent) and *E3* (Lack of Customer Trust).

Centrality: The x-axis represents centrality ($R + C$), which measures the overall importance of a factor in the system, considering both its influence and dependence. High-centrality factors are critical to the system's dynamics and require attention. For example, *T3* and *E2* (Competitive Pressures) exhibit significant centrality, highlighting their pivotal roles.

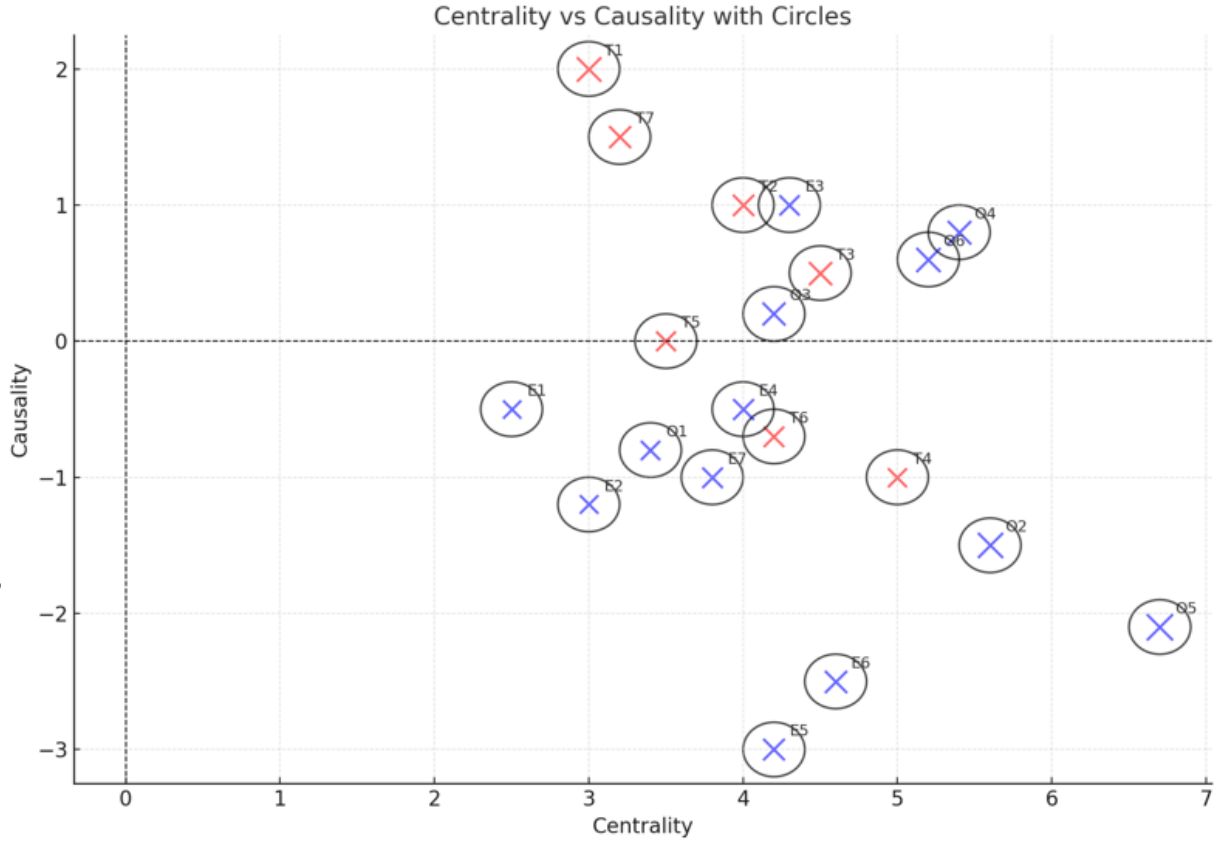


Figure 7: Centrality Vs Causality graph.

Insights from Influence Mapping:

- *Causal Factors:* Drivers like $T1$ and $T3$ must be prioritized for intervention, as improving them will likely lead to system-wide improvements.
- *Effect Factors:* While these factors do not directly drive changes, addressing them can improve system stability and effectiveness.
- *Interdependencies:* The interplay between factors is quantified and visualized, enabling a deeper understanding of how actions targeting one factor ripple through the system.

Strategic Implications: By targeting causal factors, stakeholders can implement effective strategies that yield the most significant systemic improvements. Understanding effect factors' dependencies allows for reactive planning to mitigate risks or enhance adaptability. This analysis demonstrates the value of DEMATEL in disentangling complex interrelations among adoption factors, providing a clear roadmap for decision-makers to prioritize interventions effectively. The insights align with the principles of structured, data-driven decision-making, essential for navigating the intricacies of LLM integration in highly regulated and dynamic sectors.

3.5 ISM Methodology

The Interpretive Structural Modeling (*ISM*) methodology is a structured approach Sushil (2012) used to identify and analyze the relationships among variables within a complex system. Developed to aid decision-makers in understanding interdependencies, *ISM* decomposes a system into a hierarchical structure, facilitating a clear visualization of factors based on their influence and dependence. This methodology is particularly effective in contexts requiring the prioritization and categorization of elements, such as LLM adoption in regulated industries like banking and insurance. *ISM* Janes (1988) organizes factors into layers or levels based on their influence and dependence, creating a visual model of interrelations. By iteratively analyzing relationships, *ISM* systematically decomposes the system, refining the structure to ensure clarity and relevance. This approach highlights influential factors, aiding stakeholders in targeting interventions for optimal impact and supporting decision processes.

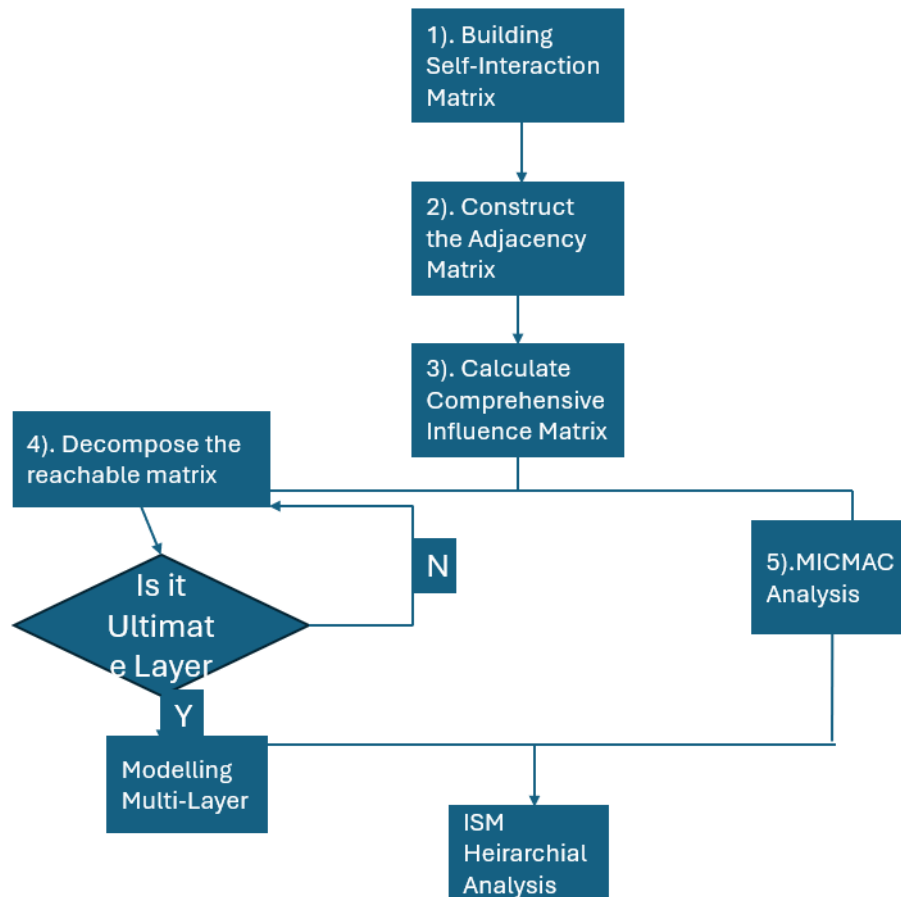


Figure 8: ISM Flowchart.

Steps in ISM Methodology

1. Building the Self-Interaction Matrix
2. Construct the Adjacency Matrix
3. Calculate the Comprehensive Influence Matrix
4. Decompose the Reachable Matrix
5. MICMAC Data Analysis
6. MICMAC Cluster Analysis

1. Building the Self-Interaction Matrix : The Self-Interaction Matrix (SIM) is a foundational component in Interpretive Structural Modeling (ISM). It captures the qualitative relationships between factors within a system and serves as the input for subsequent analytical steps. This matrix is constructed by systematically evaluating pairwise interactions between factors, leveraging survey data collected from domain experts or stakeholders.

Factors	T1	T2	T3	T4	E1	E2	E3	O1	O2	O3
T1	X	A	V	O	A	V	O	V	A	V
T2	V	X	A	V	O	A	V	O	A	A
T3	A	V	X	A	V	O	A	V	O	A
T4	O	A	V	X	O	V	O	A	V	A
E1	V	O	A	V	X	A	V	O	A	V
E2	O	V	O	A	V	X	A	V	A	O
E3	A	O	V	O	A	V	X	A	V	A
O1	V	A	O	A	V	O	A	X	A	V
O2	O	A	A	O	A	V	O	V	X	O
O3	A	V	O	A	O	A	V	A	V	X

Table 8: Self-Interaction Matrix (SIM).

Construction Process : A structured survey is designed to gather expert opinions on the relationships between factors. The survey questions are framed to assess the influence of one factor on another, using predefined qualitative symbols: **V** (strong influence), **A** (weak influence), **O** (no influence), and **X** (self-dependence). These symbols help standardize responses and facilitate the conversion of qualitative data into a structured format. The survey is distributed to a panel of experts familiar with the domain, such as banking and insurance professionals for LLM adoption. Their responses are collected, ensuring anonymity and integrity to minimize biases.

Each row of the matrix corresponds to a factor being analyzed in relation to other factors (columns). The qualitative assessments (**V**, **A**, **O**, **X**) are recorded for every pair of factors based on the survey responses. The completed matrix is then reviewed by subject matter experts to ensure the accuracy and reliability of the relationships.

Purpose of the Self-Interaction Matrix: The Self-Interaction Matrix represents the initial qualitative relationships among the system’s factors. Its primary purpose is to serve as an organized repository of expert insights, enabling the identification of direct influences. By systematically encoding these relationships, the matrix provides a structured foundation for further analysis, such as the creation of the Adjacency Matrix and the Comprehensive Influence Matrix. This approach ensures that the subsequent steps in the ISM process are grounded in expert knowledge and contextual relevance, adhering to the principles of scientific rigor and methodological consistency.

2. Construct the Adjacency Matrix : The Adjacency Matrix serves as a critical step in the Interpretive Structural Modeling (ISM) methodology, translating qualitative insights from the Self-Interaction Matrix (SIM) into a standardized binary format. This transformation is essential for quantifying relationships between factors in a consistent and structured manner. In this process, the qualitative symbols used in the SIM are systematically mapped to binary values: "V" (strong influence) and "A" (weak influence) are assigned a value of 1, signifying an influence between factors, whereas "O" (no influence) and "X" (self-dependence) are assigned a value of 0, indicating no influence. This binary representation ensures methodological rigor, facilitating computational efficiency and consistency across the analysis. The conversion process ensures that the subjective judgments of domain experts, encapsulated in the SIM, are quantitatively represented for algorithmic processing. By providing a uniform binary framework, the Adjacency Matrix becomes a cornerstone for subsequent analytical steps, including the derivation of the Reachability Matrix and the hierarchical structuring of factors. The standardization achieved through this step ensures that the data is devoid of ambiguities, enabling robust computational modeling and a clearer interpretation of factor interdependencies.

Factors	T1	T2	T3	T4	E1	E2	E3	O1	O2	O3
T1	0	1	1	0	1	1	0	1	1	1
T2	1	0	1	1	0	1	1	0	1	1
T3	1	1	0	1	1	0	1	1	0	1
T4	0	1	1	0	0	1	0	1	1	1
E1	1	0	1	1	0	1	1	0	1	1
E2	0	1	0	1	1	0	1	1	1	0
E3	1	0	1	0	1	1	0	1	1	1
O1	1	1	0	1	1	0	1	0	1	1
O2	0	1	1	0	1	1	0	1	0	0
O3	1	1	0	1	0	1	1	1	1	0

Table 9: Adjacency Matrix (Binary)

Purpose of the Adjacency Matrix: The Adjacency Matrix plays a pivotal role in bridging qualitative insights with computational analysis, serving as an intermediary that connects expert-driven evaluations with data-driven methodologies. Its primary purpose is to quantify the interrelationships among factors in a binary format, thereby ensuring readiness for algorithmic computations. This transformation enhances the methodological rigor of the ISM process, enabling a systematic and scientific approach to analyzing complex systems. By converting qualitative assessments into binary values, the Adjacency Matrix preserves the accuracy and consistency of expert judgments while enabling precise, computationally intensive modeling. Thus, the Adjacency Matrix acts as a foundational element of the ISM methodology, ensuring a structured transition from qualitative insights to actionable, data-driven outcomes.

3. Calculate the Comprehensive Influence Matrix :

The *Comprehensive Influence Matrix* is a critical component within the Decision-Making Trial and Evaluation Laboratory (DEMATEL) methodology, designed to provide a holistic understanding of the total influence each factor exerts within a system. This matrix is computed using the *Normalized Adjacency Matrix* and integrates both direct and indirect influences, capturing the complete spectrum of interdependencies among factors. By systematically accounting for these influences, the matrix enables researchers to identify critical drivers, dependencies, and the overall dynamic structure of the system under study.

The computation of the Comprehensive Influence Matrix is grounded in rigorous mathematical principles. It is derived by summing the infinite series of powers of the *Normalized Adjacency Matrix*, effectively capturing the cascading effects of indirect influences across multiple levels. This process ensures that the matrix reflects not only immediate relationships but also the compounded impacts of interactions across the entire network of factors.

Factors	T1	T2	T3	T4	E1	E2	E3	O1	O2	O3
T1	855.68	1117.45	933.44	905.43	954.79	1074.54	905.43	1090.95	1174.57	987.72
T2	855.68	1117.45	933.44	905.43	954.79	1074.54	905.43	1090.95	1174.57	987.72
T3	855.68	1117.45	933.44	905.43	954.79	1074.54	905.43	1090.95	1174.57	987.72
T4	855.68	1117.45	933.44	905.43	954.79	1074.54	905.43	1090.95	1174.57	987.72
E1	855.68	1117.45	933.44	905.43	954.79	1074.54	905.43	1090.95	1174.57	987.72
E2	855.68	1117.45	933.44	905.43	954.79	1074.54	905.43	1090.95	1174.57	987.72
E3	855.68	1117.45	933.44	905.43	954.79	1074.54	905.43	1090.95	1174.57	987.72
O1	855.68	1117.45	933.44	905.43	954.79	1074.54	905.43	1090.95	1174.57	987.72
O2	855.68	1117.45	933.44	905.43	954.79	1074.54	905.43	1090.95	1174.57	987.72
O3	855.68	1117.45	933.44	905.43	954.79	1074.54	905.43	1090.95	1174.57	987.72

Table 10: Comprehensive Influence Matrix

Purpose of Comprehensive Influence Matrix: The primary purpose of the *Comprehensive Influence Matrix* is to provide a robust and integrated representation of total relationships among factors, facilitating informed and strategic decision-making. By identifying leverage points and delineating the most influential drivers and dependencies, this matrix serves as a cornerstone for prioritizing interventions and designing targeted strategies. Furthermore, its ability to capture intricate interactions aligns with the principles of scientific rigor and methodological consistency, ensuring that the analysis is both reliable and actionable. This comprehensive approach underscores the importance of understanding the broader systemic implications of individual factor influences, ultimately supporting the development of effective, evidence-based solutions in complex environments.

4. Decompose the Reachable Matrix :

The *Reachability Matrix* is an essential binary representation within the Interpretive Structural Modeling (ISM) process, derived from the *Comprehensive Influence Matrix* by applying a predetermined threshold value. This threshold acts as a cutoff, where values exceeding the threshold (e.g., > 0.5) are assigned a binary value of 1, signifying a strong influence, while values below the threshold are assigned a binary value of 0, indicating no significant influence. This transformation ensures that only meaningful relationships are highlighted, simplifying the matrix for further analysis.

The construction of the *Reachability Matrix* is a systematic process designed to maintain scientific rigor and ensure the interpretability of complex interdependencies. By focusing on significant influences, the matrix enables researchers and decision-makers to effectively identify and analyze key relationships within the system. This step plays a crucial role in preparing data for hierarchical decomposition and subsequent analytical steps, such as identifying driving and dependence powers.

Factors	T1	T2	T3	T4	E1	E2	E3	O1	O2	O3
T1	1	1	1	1	1	1	1	1	1	1
T2	1	1	1	1	1	1	1	1	1	1
T3	1	1	1	1	1	1	1	1	1	1
T4	1	1	1	1	1	1	1	1	1	1
E1	1	1	1	1	1	1	1	1	1	1
E2	1	1	1	1	1	1	1	1	1	1
E3	1	1	1	1	1	1	1	1	1	1
O1	1	1	1	1	1	1	1	1	1	1
O2	1	1	1	1	1	1	1	1	1	1
O3	1	1	1	1	1	1	1	1	1	1

Table 11: Reachability Matrix derived from the Comprehensive Influence Matrix.

Purpose of the Reachable Matrix: The primary purpose of the *Reachability Matrix* is to provide a clear and simplified representation of relationships, highlighting whether strong influences exist between factors based on the specified threshold. This binary representation facilitates computational efficiency, allowing for more focused analyses while retaining the essence of the systemic interactions. By isolating significant influences, the *Reachability Matrix* provides actionable insights that support informed decision-making and prioritization of critical interventions. Its ability to balance complexity with clarity ensures methodological consistency and aligns with the principles of structured scientific inquiry.

5. MICMAC Data Analysis : MICMAC (Matrice d’Impacts Croisés Multiplication Appliquée à un Classement) analysis data provides insights into the Driving Power and Dependence Power of various factors within a system. The Driving Power is derived from the row sums of the Reachability Matrix, indicating how influential a factor is over others. The Dependence Power, calculated as the column sum of the Reachability Matrix, represents how much a factor is influenced by others. This tabular representation is essential for systematically categorizing factors into distinct clusters and helps identify their roles in the analyzed system.

Factor	Driving Power	Dependence Power
T1	18	5
T2	12	6
T3	15	14
T4	3	2
E1	14	12
E2	13	15
E3	7	16
O1	12	10
O2	6	14
O3	8	13

Table 12: MICMAC Analysis Data

Purpose: The MICMAC analysis serves multiple purposes, primarily focusing on factor classification and systematic analysis of relationships. The MICMAC diagram classifies factors into four clusters based on their Driving Power and Dependence Power: Driver Factors with high Driving Power and low Dependence Power, which are the key influencers driving changes across other factors; Dependent Factors with low Driving Power and high Dependence Power, which are largely affected by other factors and have minimal influence on the system; Linkage Factors with high Driving Power and high Dependence Power, which are dynamic with a strong influence on and from other factors, significantly affecting system stability; and Autonomous Factors with low Driving Power and low Dependence Power, which are relatively isolated and have minimal influence or dependency. By identifying driver factors, MICMAC helps prioritize areas that require attention or intervention to achieve desired system outcomes. The analysis identifies critical dependencies, highlighting areas where factors rely heavily on external influences, which is crucial for risk assessment and mitigation. Understanding factor interdependencies aids in effective resource allocation and planning. This helps in building robust systems by minimizing vulnerabilities associated with highly dependent factors. The scatter plot (MICMAC diagram) visually maps factors, allowing stakeholders to quickly understand their roles and relationships. It simplifies the complexity of large datasets, making it easier to derive actionable insights.

6. MICMAC Data Analysis The MICMAC (Matrice d'Impacts Croisés Multiplication Appliquée à un Classement) analysis categorizes factors into four clusters based on their **Driving Power** and **Dependence Power**, representing their roles and relationships within the system:

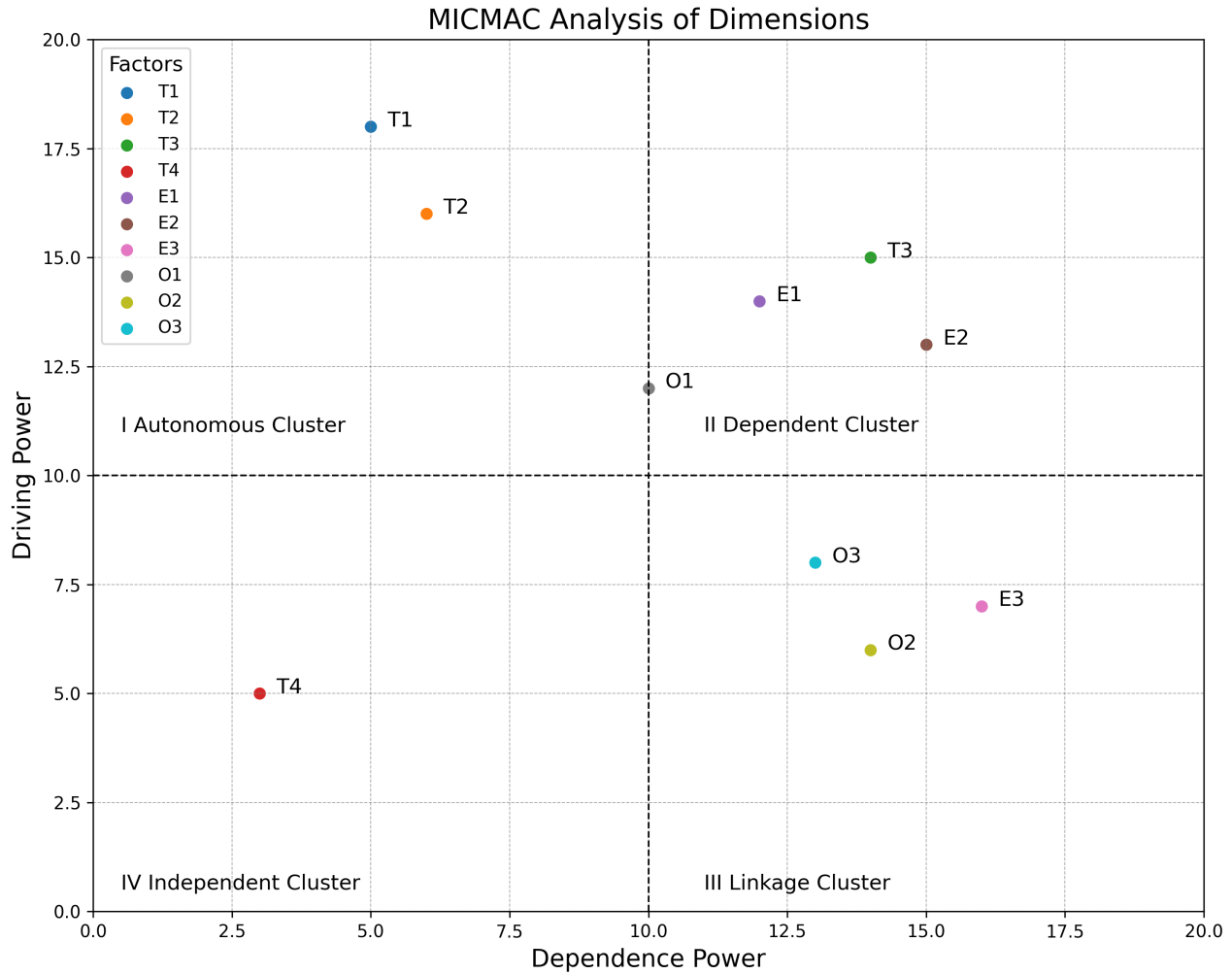


Figure 9: MICMAC Cluster Analysis.

Autonomous Cluster (Quadrant I): Factors in this cluster have low Driving Power and low Dependence Power. They are relatively isolated, with minimal influence on or from other factors in the system. These factors may not be central to the system's operation. For example, in the above graph, **T4** falls into this category, indicating it has limited relevance or interaction within the broader network.

Dependent Cluster (Quadrant II): Factors in this cluster have high Dependence Power but low Driving Power. These factors are significantly influenced by others but exert minimal influence themselves. In the graph, **O1** is in this quadrant, highlighting its reliance on other factors.

Linkage Cluster (Quadrant III): Factors in this cluster exhibit both high Driving Power and high Dependence Power. They are dynamic and highly interconnected, influencing many factors while being influenced themselves. These factors play a critical role in maintaining system stability but may









also cause instability if disrupted. Examples include **E3** and **O2**, which serve as key components in the system.




Driver Cluster (Quadrant IV): Factors here have high Driving Power but low Dependence Power. They are the primary influencers within the system and drive changes across other factors. **T1** and **T3** in the graph are part of this cluster, signifying their pivotal roles as driving forces.

Conclusion: The MICMAC diagram presented above effectively illustrates the roles and relationships of factors within the system. For instance, **T1** is identified as a strong driver, influencing other factors while remaining relatively independent. On the other hand, **E3** and **O2**, located in the linkage cluster, are highly interactive and essential for overall system dynamics. Factors like **T4**, which reside in the autonomous cluster, have minimal influence and dependency, making them less critical to the system's functioning. This analysis allows stakeholders to identify critical drivers (**T1**, **T3**), understand dynamic interdependencies (**E3**, **O2**), and allocate resources or interventions accordingly to optimize the system's performance.

3.6 Accessing the Notebooks and Plugging in Survey Data

To perform MICMAC and DEMATEL analysis using the provided Jupyter notebooks, follow the steps:

 .gitattributes	Initial commit	last week
 Insurance_Chatbot.ipynb	Removed API keys	5 days ago
 README.md	Update README.md	5 days ago
 Screenshot 2025-01-14 130035.png	Insurance Chatbot UI on Streamlit	last week
 Store_Vectors_in_Pinecone.ipynb	Removed API keys	5 days ago
 Training_Data.png	Main functionality	last week
 app.py	Removed API keys	5 days ago
 requirements.txt	Main functionality	last week

 **README**



Insurance Chatbot with Langchain and Pinecone

This implements a chatbot that utilizes Sentence Transformation and OpenAI's GPT-4o model to inform about insurance Policies in europe and Digitilization of insurance Industry. The chatbot aims to provide relevant responses to user queries by refining and enhancing their input queries, finding similar sentences using Sentence Transformation, and generating more contextually accurate conversation logs.

Given a knowledge base whose vectors are stored in a pinecone, the chatbot provides answers to the questions that are most relevant to the context.

Figure 10: Overview of Notebooks.

Accessing the Notebooks

1. Clone the repository to your local machine using the command:

```
git clone https://github.com/NagarjunaD024
```

2. Navigate to the repository directory:

```
cd Framework_Methodology
```

3. Launch the Jupyter Notebook interface by running:

```
jupyter notebook
```

4. Open the desired notebook:

- `ISM.ipynb`: For Interpretive Structural Modeling (ISM).
- `DEMATEL.ipynb`: For Decision-Making Trial and Evaluation Laboratory (DEMATEL).

2. Plugging in Survey Data Refer to the provided survey questionnaire in Appendix B: Survey Questionnaire as a guideline for data collection. To use your own data:

- Replace the placeholder survey data in the notebooks with your dataset. Ensure the data format aligns with the following structure:
 - A matrix or table where rows and columns represent factors.
 - Numerical values indicating relationships or dependencies between factors.
- Modify the data-loading section of the notebook to point to your dataset file (e.g., CSV or Excel format).

3. Running the Notebooks

1. Execute the notebook cells sequentially to process the input data.

2. For each notebook:

- `ISM.ipynb`: Generates a excel file with all required matrices and a structural model graph.
- `DEMATEL.ipynb`: Generates a excel file with all Influence matrices and produces a Causality vs. Centrality plot.

4. Reviewing and Exporting Results The generated tables and visualizations are saved automatically for review and inclusion in reports. Refer to Appendix A: Repository Access for detailed instructions on accessing the repository and retrieving output files.

4 LLM development

4.1 Training data

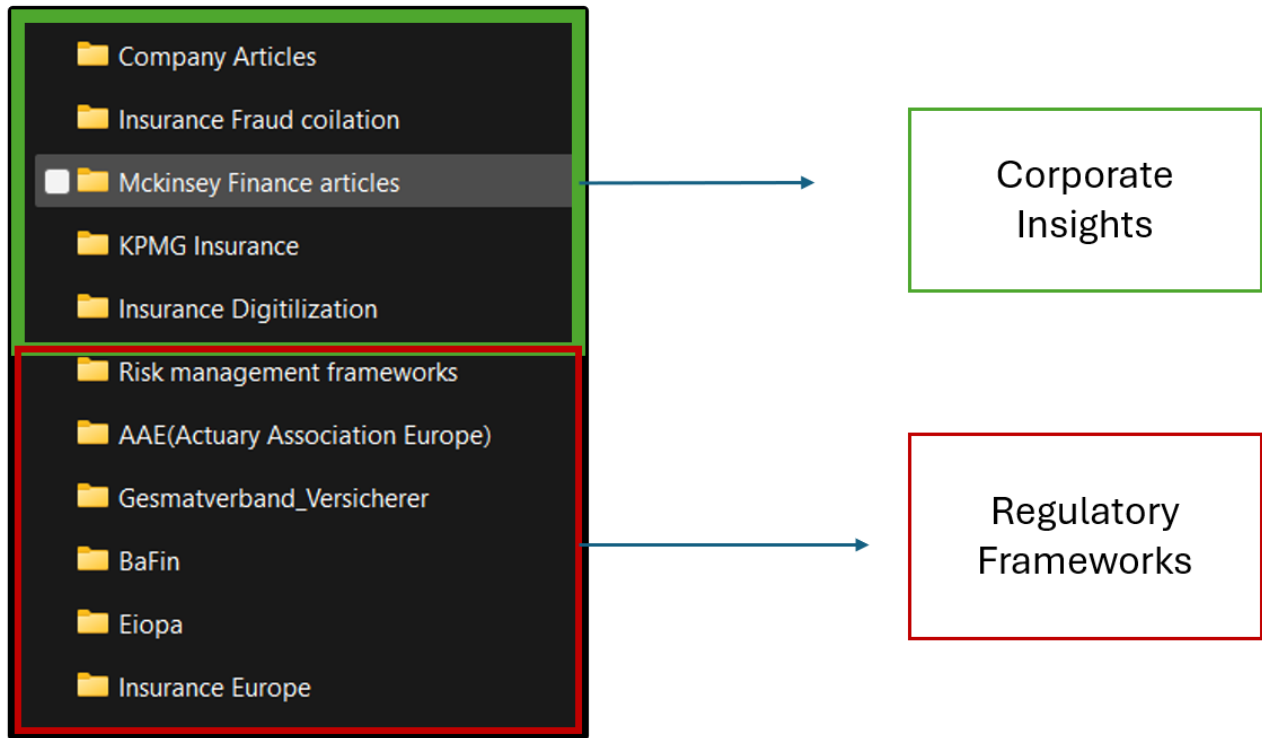


Figure 11: LLM Training data.

Corporate Insights : This category includes a collection of resources focused on company-specific practices, innovation, and strategies within the insurance and financial sectors. The training data comprises several subcategories. Company Articles consist of internal or public articles detailing corporate strategies, product launches, and operational methodologies. Insurance Fraud Collation provides comprehensive datasets addressing patterns, prevention techniques, and case studies on insurance fraud. McKinsey Finance Articles include industry insights provided by McKinsey, focusing on market trends, digital transformation, and financial strategies. KPMG Insurance Reports feature studies and whitepapers highlighting insurance practices, regulatory compliance, and emerging market opportunities. Finally, Insurance Digitalization emphasizes resources that highlight technological adoption and innovation in the insurance industry. These documents collectively enable the system to offer valuable insights and recommendations tailored to corporate contexts.

Regulatory Frameworks : The regulatory frameworks category includes authoritative sources that define policies, compliance standards, and risk management protocols. The training datasets utilized include Risk Management Frameworks, which outline strategies to identify, assess, and mitigate financial and operational risks. AAE (Actuarial Association Europe) provides reports from the actuarial community, addressing regulations and best practices in risk analysis. Gesamtverband Versicherer contains policies and guidelines from the German Insurance Association to ensure compliance with na-

tional insurance regulations. BaFin publications, from the Federal Financial Supervisory Authority of Germany, focus on financial stability and supervisory standards. Eiopa resources, from the European Insurance and Occupational Pensions Authority, address pan-European insurance regulations. Finally, Insurance Europe includes reports from the European insurance and reinsurance federation, detailing market trends and policy frameworks. These datasets empower the AI to provide contextually accurate responses regarding regulatory compliance and risk mitigation strategies.

Scientific Relevance : The dual focus on corporate insights and regulatory frameworks ensures that the AI system is equipped with both operational and compliance-related knowledge. By training on diverse sources of high-quality information, the system achieves several objectives. It offers contextual depth, enabling the AI to provide nuanced recommendations tailored to corporate strategies and regulatory compliance. Scalability is ensured, making the system applicable across multiple use cases in finance and insurance, including fraud detection, policy recommendations, and risk management. Additionally, the system achieves accuracy by leveraging domain-specific knowledge from authoritative sources. By integrating these datasets into the training pipeline, the AI system gains the capability to serve as a comprehensive tool for decision support in highly regulated and dynamic environments.

4.2 Overview of Chatbot Architecture

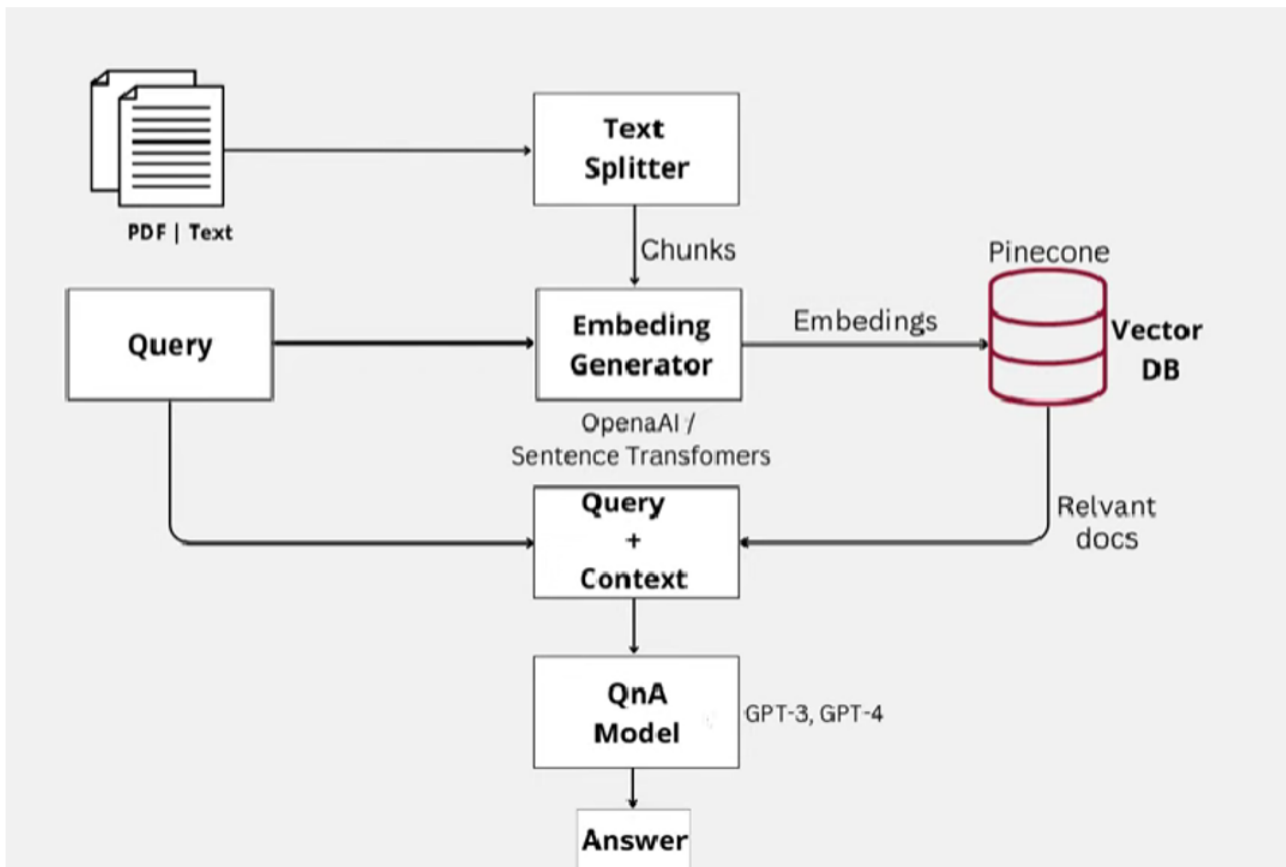


Figure 12: Chatbot Architecture.

Introduction : The Insurance Chatbot leverages state-of-the-art AI technologies to redefine user interaction within the insurance industry. It is designed to provide users with accurate, contextually

relevant responses to complex queries, showcasing its versatility across diverse applications such as policy information retrieval, digital transformation insights in Europe. Built using an advanced stack of technologies, including LangChain, Pinecone, and OpenAI's GPT-4o model which had higher context understanding compared to base models Rishi Bommasani et al. (2021), the chatbot seamlessly integrates semantic understanding and natural language processing to deliver a superior conversational experience.

The chatbot architecture incorporates critical components such as embedding generation, vector storage, contextual querying, and advanced Q&A modeling. Embedding generation ensures the system can process and understand the semantic and contextual relationships inherent in insurance-related data. The embeddings are stored and managed in Pinecone, a highly efficient and scalable vector database, facilitating rapid similarity searches. Contextual querying enables precise integration of user queries with relevant information retrieved from the vector database, while the Q&A model, powered by GPT-4o, ensures human-like response generation. Designed with scalability and robustness in mind, the system is capable of handling high query volumes and large datasets, making it a valuable tool for the insurance industry. By leveraging AI-driven advancements, Truong and Nguyen (2024) the chatbot addresses critical challenges, including fraud detection and regulatory compliance, and provides transformative solutions that drive innovation in digital insurance services.

Text Splitter : The Text Splitter is a fundamental component of the chatbot architecture, responsible for preprocessing large documents, such as PDFs, insurance policies, or datasets, by breaking them into smaller, manageable chunks. This preprocessing step is crucial to ensure efficient embedding generation and to preserve the semantic coherence of the data Lewis et al. (2020). By dividing extensive textual inputs into chunks, the Text Splitter reduces computational overhead and allows downstream components to handle data in a structured and optimized manner. This process is particularly significant in insurance-related applications, where documents are often lengthy, detailed, and complex.

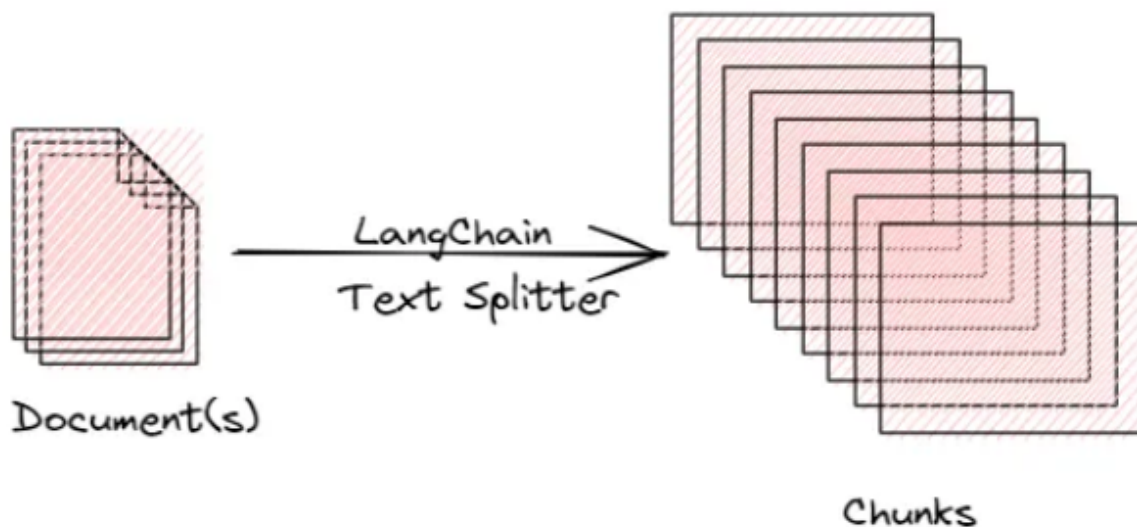


Figure 13: Text Splitting.

The splitting mechanism also ensures that the context within each chunk is retained, which is vital for the accurate representation of semantic relationships during embedding generation. This enables

the chatbot to process queries efficiently and respond with high precision, even when handling large datasets. For example, breaking down an insurance policy into smaller segments allows the system to retrieve and analyze specific clauses or sections relevant to the user's query without compromising the integrity of the data.

Vector Database and Embedding Generator : The chatbot employs **Pinecone** as its vector database to store and manage embeddings. Pinecone enables efficient similarity search and retrieval of high-dimensional data, ensuring relevant results are delivered quickly. It is designed to handle large datasets, such as comprehensive insurance regulations or customer support documentation, while maintaining scalability and reliability. Additionally, Pinecone provides real-time interaction with low-latency responses, making it an ideal choice for customer-facing applications where fast and accurate replies are critical which is extensively discussed in **subsection 4.3**, Together, the embedding generation and vector database components form the foundation for delivering precise and contextually relevant responses in real-time.

The core functionality of the chatbot lies in generating embeddings using **Sentence Transformers**. These embeddings are high-dimensional vectors that encode the semantic meaning and contextual nuances of the input text. By transforming textual data into numerical representations, the embeddings enable advanced querying capabilities. The chatbot leverages these embeddings to retrieve relevant documents from the database efficiently and match user queries with stored knowledge. This facilitates similarity-based search, which is crucial for handling complex insurance queries where semantic understanding is essential.

Query and Context Integration : The query and context integration stage is a critical component of the chatbot architecture. This process retrieves relevant embeddings from the vector database and integrates them with the user's query to provide accurate and contextually appropriate responses. By aligning retrieved data with the user's intent, the system ensures that the information delivered is both relevant and tailored to specific requirements. This integration process significantly enhances the overall conversational experience, particularly in high-stakes domains like insurance, where precision and contextual understanding are essential. The seamless combination of query and context allows the chatbot to address complex inquiries, ensuring user satisfaction and maintaining the system's credibility.

Q&A Model : The Q&A model, powered by **OpenAI's GPT-4o**, is the centerpiece of the chatbot's response generation mechanism. It is designed to generate human-like responses by leveraging its advanced context-aware reasoning capabilities. This enables the system to provide accurate and detailed answers, even for complex and nuanced queries. The model is highly adaptable, capable of addressing diverse topics such as policy recommendations, claims processes, and fraud detection. Its integration with Pinecone and the embedding generator ensures consistent performance across various tasks. By effectively processing both the user's query and the retrieved context, the Q&A model maintains a seamless conversational flow, delivering precise and meaningful responses that align with the user's needs.

4.3 Vector Databases in Chatbot Architecture

Vector databases are a specialized and increasingly indispensable component in managing and processing high-dimensional vector representations of data, particularly in advanced AI systems like chatbots. In the context of chatbot architectures, vector databases play a pivotal role in storing and retrieving embeddings, which are key to enabling efficient query resolution and context-aware interactions. This section provides a detailed exploration of the core aspects of vector databases, their functionalities, and their importance in leveraging Large Language Models (LLMs).

Feature	Traditional Database	Vector Database
Data Representation	Tables with rows and columns	Vectors
Querying	Exact matches, range queries, and joins	Similarity search
Scalability & Performance	Designed for general-purpose data storage and retrieval, with a focus on ACID	Optimized for high-throughput and low-latency retrieval of complex vector data
Indexing & Search Efficiency	KD-trees, R-trees, approximate nearest neighbor (ANN) search algorithms	B-trees, hash indexes, full-text search indices

Table 13: Comparison between Traditional Database and Vector Database.

Unlike traditional databases that store data in structured rows and columns, vector databases store data as numerical arrays in a high-dimensional space. These vectors encode the semantic meaning of various data types, including text, images, and audio. In chatbot systems, embeddings are derived from user queries and documents, enabling similarity-based searches that go beyond keyword matching. For instance, a user query is transformed into a vector representation, which is then compared to vectors stored in the database to retrieve the most contextually relevant information.

Key Characteristics of Vector Databases : Vector databases are a specialized class of databases designed to manage and query high-dimensional data efficiently. A distinguishing feature of vector databases is their ability to support **semantic search**, which retrieves information based on conceptual similarity rather than exact keyword matches. This capability enables systems to understand and process user intent more effectively, going beyond traditional exact-match retrieval methods. Additionally, vector databases are engineered for **scalability**, allowing them to efficiently handle complex, high-dimensional data representations across large-scale applications, including those involving billions of vectors.

One of the defining strengths of vector databases is their optimization for **real-time interaction**, making them particularly well-suited for high-throughput environments where low-latency responses are critical. These characteristics position vector databases as indispensable tools in applications such as recommendation systems, conversational AI, fraud detection, and personalized search engines.

Embeddings: The Foundation of Vector Databases: At the core of vector databases lies the concept of **embeddings**, which are dense, numerical representations of data that encapsulate its

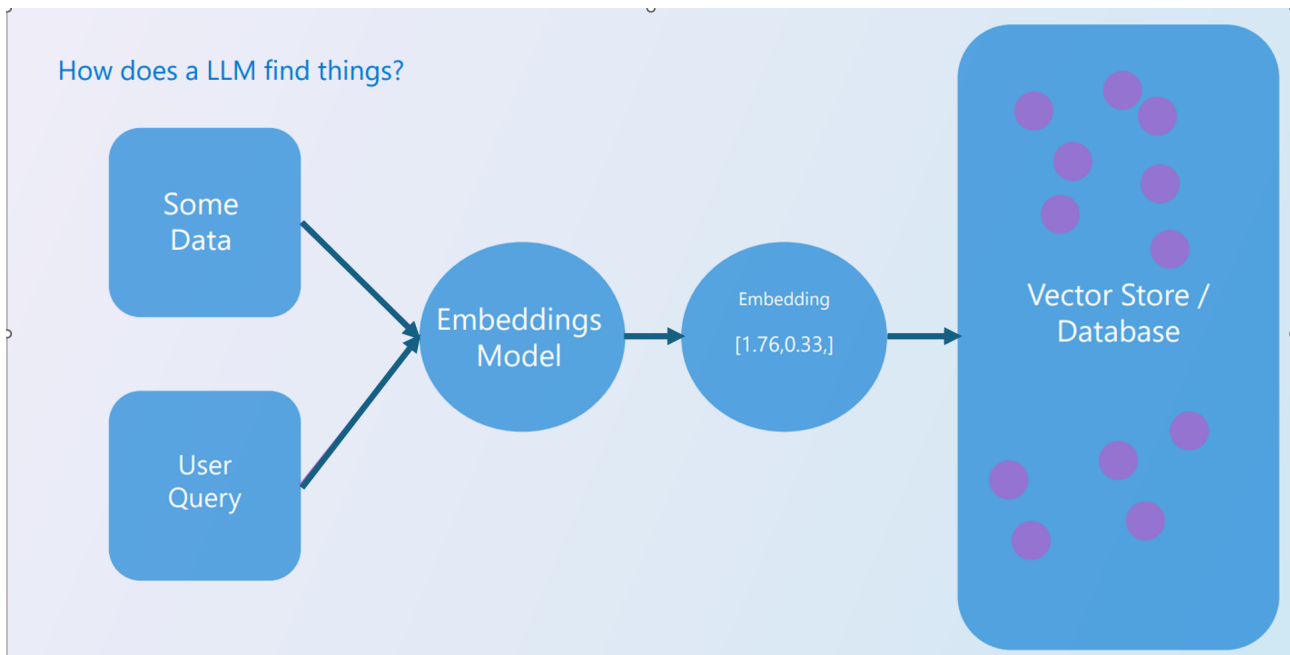


Figure 14: Overview of Embeddings in Vector Databases.

semantic and contextual relationships. These embeddings form the foundation for enabling advanced querying capabilities and efficient handling of unstructured data, such as text, images, and audio. Unlike traditional databases that rely on exact matches, embeddings empower vector databases to perform **semantic search**, facilitating information retrieval based on conceptual similarity rather than lexical overlap. In the domain of chatbot systems and conversational AI, embeddings are typically generated using state-of-the-art machine learning models, such as *OpenAI's Sentence Transformers*, *BERT*, or similar neural architectures. These models transform textual inputs, including user queries and documents, into high-dimensional vector representations. These vectors encode not only the *literal meaning* of the input but also its *contextual nuances*, making it possible to match queries with relevant content even in complex, nuanced interactions.

Once generated, embeddings are stored in vector databases, enabling **rapid and precise retrieval** Strubell et al. (2019) of information based on similarity metrics, such as cosine distance or Euclidean distance. For example, user queries can be matched against stored embeddings of documents, FAQs, or historical conversations, ensuring **contextually relevant responses** are delivered in real-time. This capability is particularly vital in high-performance systems, where response times measured in milliseconds are critical to maintaining a seamless user experience. Beyond chatbots, embeddings also power diverse applications such as **recommendation systems**, **fraud detection**, and **personalized search engines**, demonstrating their versatility in capturing and leveraging latent patterns in data. By bridging the gap between raw data and actionable insights, embeddings play a pivotal role in driving the functionality and scalability of modern AI-driven systems. The integration of embeddings into vector databases highlights the growing importance of **context-aware AI**, where systems not only respond accurately but also adapt dynamically to user intent and situational nuances. This synergy between embeddings and vector databases is shaping the future of information retrieval, pushing the boundaries of what is possible in fields ranging from conversational AI to large-scale recommendation systems, which revolutionize the information retrieval.

Similarity search is a foundational capability of vector databases, employing metrics such as cosine similarity, Euclidean distance, and Jaccard similarity. Cosine similarity measures the angle between

Metric	Description
Euclidean	Measures the straight-line distance between two points in a vector space. Sensitive to vector magnitudes. Often referred to as the L2 norm.
Manhattan	Computes the sum of absolute differences of corresponding coordinates. Known as L1 norm or taxicab geometry. Common in image retrieval and financial analysis.
Cosine	Measures the cosine of the angle between two vectors, emphasizing direction over magnitude. Useful for text similarity where term frequency matters less than contextual similarity.
Jaccard	Computes the overlap between two sets to gauge similarity. Commonly used for comparing customer purchase histories or document term overlaps.

Table 14: Summary of Distance Metrics for Vector Databases

vectors and is particularly effective for textual data, enabling the chatbot to identify semantically similar queries. Euclidean distance calculates the straight-line distance between points in the vector space, making it suitable for spatial data, while Jaccard similarity measures set overlap and is applicable in specialized use cases like recommendation systems. These metrics ensure robust matching capabilities, which are integral to the chatbot’s ability to respond accurately.

Indexing Strategy	Description
Flat Index	A simple structure that stores data points in a flat array or list without advanced preprocessing. Best suited for small datasets due to its simplicity and direct access.
Inverted File Index (IVF)	Uses a lookup table to map vectors to locations. Efficient for narrowing search spaces by associating vectors with metadata or dimensions.
Approximate Nearest Neighbors (ANN)	Constructs multiple search trees to partition the dataset and find approximate neighbors efficiently. Ideal for large datasets with high-dimensional data.
Product Quantization (PQ)	Compresses high-dimensional vectors into compact codes by dividing vectors into smaller sub-vectors and quantizing each segment. Optimized for storage efficiency.
Hierarchical Navigable Small World (HNSW)	Builds a graph structure that connects similar vectors, enabling fast and scalable nearest neighbor searches, even for massive datasets.

Table 15: Indexing Strategies for Vector Databases

To optimize performance, vector databases employ advanced indexing techniques. Hierarchical Navigable Small World (HNSW) organizes vectors in a graph structure, enabling efficient nearest-

neighbor searches. Product Quantization (PQ) compresses vectors to save storage while maintaining retrieval accuracy. Approximate Nearest Neighbors (ANN) techniques strike a balance between computational speed and search precision, making them ideal for large-scale applications. These techniques collectively enhance the operational efficiency of vector databases, allowing chatbots to process complex queries in real time. Querying in vector databases involves transforming user input into vector embeddings and retrieving the most relevant stored vectors. This enables chatbots to provide contextually enriched responses by leveraging pre-stored knowledge. Additionally, the ability to integrate vector databases seamlessly into LLM-powered architectures allows for a more dynamic and intelligent interaction experience. Chatbots leveraging such databases can deliver multi-turn conversations with contextual memory, ensuring a smooth and engaging user experience.

Feature	Vector	Vector Database
Definition	A mathematical representation of data as numerical arrays.	A specialized database designed to store, manage, and query vectors efficiently.
Data Focus	Represents individual data points in high-dimensional space.	Handles collections of vectors and facilitates similarity search across datasets.
Purpose	Captures semantic or contextual meaning of a single entity.	Enables rapid and efficient operations like similarity search, clustering, and retrieval.
Applications	Basis for embeddings in machine learning and AI.	Used in recommendation systems, fraud detection, and content moderation.
Scalability	Limited to representing specific data points.	Optimized for high-throughput environments with billions of vectors.

Table 16: Comparison Between Vector and Vector Databases

Vector databases significantly enhance chatbot capabilities. They enable personalized responses by leveraging historical user interactions, support efficient knowledge retrieval for complex queries, and handle real-time responses at scale. These features make vector databases a cornerstone of modern AI systems. For example, in e-commerce applications, chatbots powered by vector databases can provide highly tailored product recommendations based on user preferences and past behavior.

Despite their numerous benefits, vector databases face challenges. Managing high-dimensional data complexity requires robust algorithms and computational resources. Resource intensity remains a concern, as large-scale deployments demand substantial storage and processing power. Query optimization, which balances speed and accuracy, is another critical area for improvement. Addressing these challenges is vital for ensuring the scalability and efficiency of vector databases in real-world applications. Additionally, ensuring compliance with data privacy regulations, such as GDPR, presents another layer of complexity when storing and processing sensitive user data in vectorized formats.

The Future of Vector Databases: Future advancements in vector databases are poised to address these challenges. Cross-modal search capabilities, enabling similarity search across diverse data types

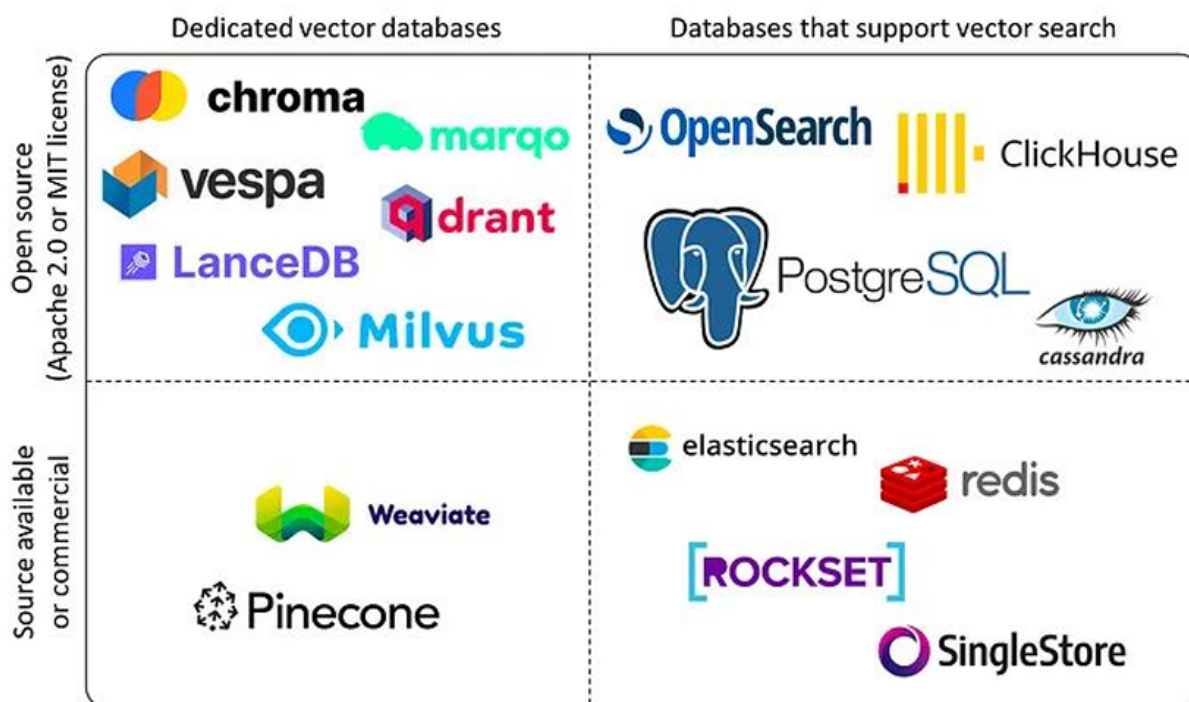


Figure 15: Market Overview.

such as text, images, and audio, will expand their applicability to multi-modal AI systems. Explainable AI features will enhance transparency by providing insights into the decision-making processes of similarity results. For instance, in scenarios where critical business decisions depend on AI-driven chatbot responses, explainability will ensure trustworthiness and regulatory compliance. Federated learning frameworks will support collaborative AI applications while maintaining data privacy, an essential requirement in regulated industries like healthcare and finance. Furthermore, the integration of vector databases with cutting-edge LLMs is likely to drive innovations in areas such as real-time multilingual translation Brynjolfsson et al. (2023), complex query answering, and adaptive learning systems. These advancements will enhance the capability of chatbots to understand and cater to diverse user needs more effectively. Additionally, the development of energy-efficient indexing and retrieval algorithms will mitigate resource intensity concerns, making vector databases more sustainable and cost-effective. Vector databases represent a transformative advancement in modern chatbot architectures, offering efficient semantic search, scalable query handling, and personalized user interactions. By integrating sophisticated indexing techniques and similarity metrics, they form a robust foundation for leveraging the power of LLMs. Future developments in transparency, cross-modal capabilities, and collaborative frameworks will further elevate their role in diverse domains, cementing their position as a critical technology for intelligent systems. As businesses continue to adopt AI-driven solutions, the role of vector databases will expand, supporting not only conversational AI but also a wide range of applications across industries such as healthcare, finance, and e-commerce. This highlights their versatility and underscores their importance in the broader AI ecosystem.

4.4 Post-Development Operations (LLMOPS)

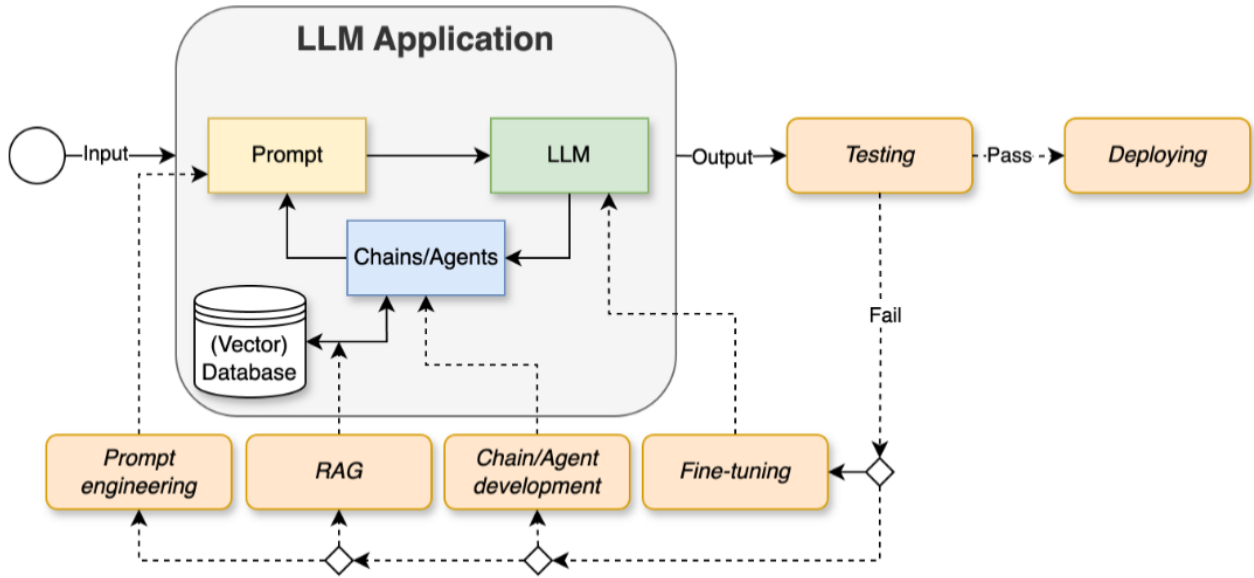


Figure 16: LLMOPS.

Deploying LLM Applications: A Streamlit and LLMOPS Perspective

The figure outlines a robust workflow for developing, validating, and deploying a Large Language Model (LLM) application Mahmoudabadi and Lalaei (2024) with a focus on operational scalability and reliability. Deploying an LLM-powered chatbot (see, e.g., Mahr et al. (2024)) on platforms like Streamlit combines the strengths of a user-friendly interface with a scalable backend, ensuring a seamless user experience. Leveraging Streamlit for deployment aligns with modern LLMOPS (LLM Operations) principles Sendas and Rajale (2024), enabling efficient management, monitoring, and optimization of LLM applications in production environments.

Streamlit provides a lightweight and interactive framework to deploy LLM-based applications Razak et al. (2024), offering several advantages. With minimal configuration, developers can convert Python scripts into fully functional web applications, making it an ideal choice for rapid prototyping and deployment. Combined with vector databases like Pinecone and orchestration tools such as LangChain, Streamlit ensures that the LLM application can handle multiple user queries concurrently without performance degradation. It supports seamless integration with backend components, including prompt engineering, retrieval-augmented generation (RAG), and vector search functionalities. In the chatbot code provided, Streamlit is used to build an intuitive interface where users can input their queries. The application processes these inputs, retrieves context from the Pinecone database, and generates responses using OpenAI's GPT models. The real-time capabilities of Streamlit allow users to experience the chatbot's performance interactively, making it particularly suitable for customer-facing insurance applications.

LLMOPS (Large Language Model Operations) is a critical framework that ensures the reliability, scalability, and continuous improvement of deployed LLM applications. In this workflow, data pipelines, managed using tools like Databricks or Airflow, handle unstructured data to create embeddings stored in vector databases. Embedding models, such as OpenAI or Hugging Face transformers,

generate embeddings that capture semantic relationships and ensure context-aware outputs. Validation and monitoring are integral components of LLMOPS Razak et al. (2024), ensuring rigorous evaluation of outputs and continuous monitoring post-deployment. Tools like Prometheus or Grafana can be integrated for performance tracking and anomaly detection. By integrating these principles, the chatbot leverages LLMOPS to address real-world challenges, including handling high query volumes, ensuring data security, and fine-tuning models based on user feedback. Streamlit complements this framework by providing a transparent and adaptable interface, making it easier to visualize LLM operations and their impact.

Deploying LLM applications with Streamlit and LLMOPS exemplifies a balance between theoretical advancements and practical implementation. LLMOPS Kalva (2024) ensures that the entire lifecycle of the chatbot—from development to deployment—is streamlined and optimized for performance and reliability. Meanwhile, Streamlit offers a scalable, user-centric interface that bridges the gap between cutting-edge AI capabilities and end-user accessibility. Together, these tools enable efficient deployment and management of LLM applications in high-stakes domains like insurance, where accuracy and responsiveness are paramount.

4.5 Setting Up and Deploying the Insurance Chatbot Application

1. **Install Prerequisites:** Ensure Python 3.7 or higher is installed. Then, install the required libraries using the following command:

```
pip install streamlit sentence-transformers pinecone-client openai
```

2. **Clone the Repository:** Download the application code from the repository and navigate to the directory:

```
git clone https://github.com/NagarjunaD024
cd LLM_Insurance_Chatbot
```

3. **Set Up API Keys:** Obtain API keys from Pinecone and OpenAI. Set these keys in your environment:

```
export PINECONE_API_KEY="your_pinecone_api_key"
export OPENAI_API_KEY="your_openai_api_key"
```

4. **Add Ngrok AuthToken:** If you are using Ngrok for local tunneling, you need to add the Ngrok AuthToken to authenticate your agent. Obtain the AuthToken from your Ngrok dashboard, then set it up using:

```
ngrok config add-authtoken <your_ngrok_authtoken>
```

To obtain your Ngrok AuthToken, visit the Ngrok AuthToken Dashboard.

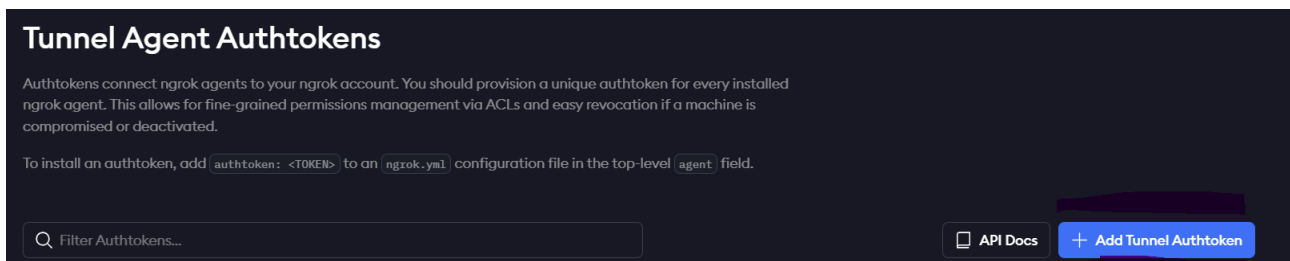


Figure 17: Ngrok Token.

5. **Run the Streamlit Application:** Start the chatbot application by running:

```
streamlit run app.py
```

6. **Interact with the Chatbot:** Open the application in your browser, input insurance-related queries, and receive contextually accurate responses.

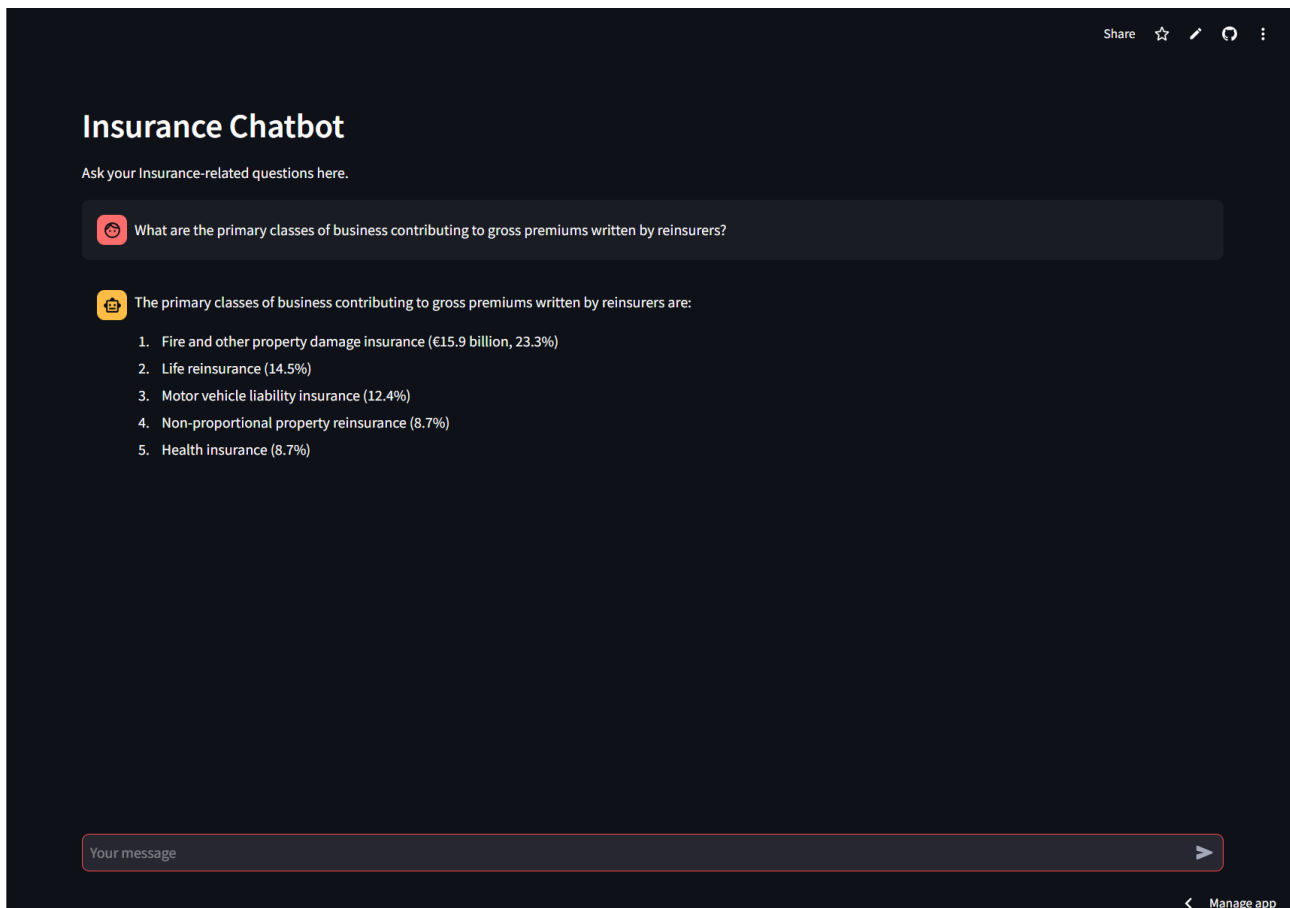


Figure 18: Streamlit UI.

4.6 Accessing the Deployed Insurance Chatbot on Streamlit

The Insurance Chatbot application has been deployed on Streamlit to provide users with an intuitive and interactive interface for querying insurance-related information. The deployment ensures seamless access to the chatbot's functionalities without requiring additional setup on the user's end.

To access the deployed chatbot, visit the following link:

[Insurance Chatbot on Streamlit](#)

Once on the page, users can input their queries in the provided interface, and the chatbot will process the input and return accurate, context-aware responses in real-time. This deployment leverages Streamlit's scalable framework to ensure reliability and low-latency interactions, catering to high-stakes insurance-related use cases.

5 Conclusion and outlook

The integration of Generative AI and LLMs into business workflows represents a transformative shift in organizational capabilities. However, the journey from development to deployment is fraught with technical, operational, and ethical challenges. By exploring frameworks for integration and evaluating their application in the insurance industry, this thesis contributes to the growing body of knowledge on how organizations can harness the power of LLMs responsibly Sam Bowman (2023) and effectively. As businesses navigate the evolving landscape of AI, the insights and methodologies presented here aim to serve as a guiding framework for realizing the full potential of Generative AI.

The findings suggest that embedding Generative AI in workflows enhances efficiency and decision-making processes while fostering transparency and trust. Moreover, the practical applications highlighted, such as claims automation, and personalized policy recommendations, demonstrate the scalability and adaptability of LLMs in tackling complex industry demands . As organizations continue to navigate the evolving landscape of AI (see, e.g.,Chen et al. (2023)), the methodologies and insights presented here serve as a foundational guide for integrating these transformative tools responsibly. This work advocates for a collaborative approach between technology providers, regulators, and industry stakeholders to ensure ethical, compliant, and impactful deployment of LLMs, ultimately setting a benchmark for the future of AI-driven business innovation.

5.1 Future outlook

Generative AI is at the cusp of revolutionizing industries through enhanced productivity, creative problem-solving, and knowledge democratization. However, this transformation necessitates a balance between innovation and ethical deployment

1. Economic and Workforce Transformation:

- Generative AI promises exponential economic growth by automating decision-making (see, e.g., Peng et al. (2023)), marketing, and software engineering tasks Michael Chui (2023).
- Knowledge workers must adapt to new roles focusing on Noy and Zhang (2023) creativity and oversight, while automation may challenge (see, e.g., Capraro et al. (2024)) traditional educational and skill hierarchies.

2. Socioeconomic Equity:

- Widespread adoption can reduce barriers in education and healthcare but risks deepening inequalities without robust policy interventions Noy and Zhang (2023).
- Ethical and transparent AI deployment is essential to address potential biases and accessibility concerns Capraro et al. (2024).

3. Technological Advancement:

- Innovations like retrieval-augmented generation and multimodal Si et al. (2024) capabilities expand the scope of AI applications, from precise fact-checking to novel idea generation (see, e.g., Si et al. (2024)).
- Models such as Llama 3 (see, e.g., Grattafiori et al.) demonstrate that open-source initiatives can drive global AI advancements while emphasizing scalability and safety.

4. Regulation and Governance:

- Frameworks like the EU AI Act set global benchmarks for safe and ethical AI usage Capraro et al. (2024).
- Striking a balance between stringent safety requirements and fostering innovation will define the future trajectory of AI adoption .

5. Sustainable Integration:

- Tools like ChatGPT and GitHub Copilot underline AI's role in improving productivity and democratizing access to advanced capabilities Noy and Zhang (2023).
- Future efforts should focus on enhancing usability for diverse user bases and fostering equitable AI integration across sectors Michael Chui (2023).

Generative AI's promise lies in its ability to complement human creativity while reshaping traditional workflows. By fostering inclusive policies, ensuring transparent (see, e.g., Yang et al. (2023)), and continually refining technological capabilities, AI can transition from a disruptive force to a universally empowering tool.

6 List of References

References

- Ai risk management framework | nist. URL <https://www.nist.gov/itl/ai-risk-management-framework>.
- Exploring large language models: A guide for insurance professionals. 2023. URL <https://www.milliman.com/en/insight/exploring-large-language-models-guide-insurance>.
- Large language models in underwriting and claims | munich re life us, 2024. URL <https://www.munichre.com/us-life/en/insights/future-of-risk/large-language-models-in-underwriting-and-claims.html>.
- Kartik Hosanagar and Ramayya Krishnan. Who profits the most from generative ai?, 2024. URL <https://sloanreview.mit.edu/article/who-profits-the-most-from-generative-ai/>.
- M. Abdekhoda, A. Dehnad, and J. Zarei. Factors influencing adoption of e-learning in healthcare: integration of utaut and ttf model. *BMC medical informatics and decision making*, 22(1):327, 2022. doi: 10.1186/s12911-022-02060-9.
- E. Brynjolfsson, D. Li, and L. Raymond. Generative ai at work, 2023. URL <http://arxiv.org/pdf/2304.11771>.
- V. Capraro, A. Lentsch, D. Acemoglu, S. Akgun, A. Akhmedova, E. Bilancini, J.-F. Bonnefon, P. Brañas-Garza, L. Butera, K. M. Douglas, J. A. C. Everett, G. Gigerenzer, C. Greenhow, D. A. Hashimoto, J. Holt-Lunstad, J. Jetten, S. Johnson, W. H. Kunz, C. Longoni, P. Lunn, S. Natale, S. Paluch, I. Rahwan, N. Selwyn, V. Singh, S. Suri, J. Sutcliffe, J. Tomlinson, S. van der Linden, P. A. M. van Lange, F. Wall, J. J. van Bavel, and R. Viale. The impact of generative artificial intelligence on socioeconomic inequalities and policy making. *PNAS Nexus*, 3(6):pgae191, 2024. doi: 10.1093/pnasnexus/pgae191. URL <https://academic.oup.com/pnasnexus/article/3/6/pgae191/7689236>.
- L. Chen, M. Zaharia, and J. Zou. How is chatgpt’s behavior changing over time?, 2023. URL <http://arxiv.org/pdf/2307.09009>.
- E. Falatoonitoosi, Z. Leman, S. Sorooshian, and M. Salimi. Decision-making trial and evaluation laboratory. *Research Journal of Applied Sciences, Engineering and Technology*, 5(13):3476–3480, 2013. ISSN 20407459. doi: 10.19026/rjaset.5.4475.
- Francisco Castro, Jian Gao, and Sébastien Martin. Does genai impose a creativity tax? *MIT Sloan Management Review*, 66(2), 2024. URL <https://sloanreview.mit.edu/article/does-genai-impose-a-creativity-tax/>.
- A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, A. Yang, A. Fan, A. Goyal, A. Hartshorn, A. Yang, A. Mitra, A. Sravankumar, A. Korenev, A. Hinsvark, A. Rao, A. Zhang, A. Rodriguez, A. Gregerson, A. Spataru, B. Roziere, B. Biron, B. Tang, B. Chern, C. Caucheteux, C. Nayak, and Bi. The llama 3 herd of models. URL <http://arxiv.org/pdf/2407.21783>.
- Q. Guo and W. Huang. Analyzing the diffusion of innovations theory. *Scientific and Social Research*, 6(12): 95–98, 2024. ISSN 2661-4332. doi: 10.26689/ssr.v6i12.8947.
- F. R. Janes. Interpretive structural modelling: a methodology for structuring complex issues. *Transactions of the Institute of Measurement and Control*, 10(3):145–154, 1988. doi: 10.1177/014233128801000306.
- R. Kalva. Optimizing e-commerce platforms with genai-driven devops and llmops a scalable framework for enhanced user experience. *Journal of Artificial Intelligence, Machine Learning and Data Science*, 2(4): 1782–1788, 2024. ISSN 2583-9888. doi: 10.51219/jaimld/rahul-kalva/396.
- F.-C. Kao, S.-C. Huang, and H.-W. Lo. A rough-fermtean dematel approach for sustainable development evaluation for the manufacturing industry. *International Journal of Fuzzy Systems*, 24(7):3244–3264, 2022. ISSN 2199-3211. doi: 10.1007/s40815-022-01334-8. URL <https://link.springer.com/article/10.1007/s40815-022-01334-8>.

- Le Nguyen. Generative ai and llms in insurance: Common risks and proven mitigation tactics. *Zelros*, 2023. URL <https://www.zelros.com/2023/10/12/generative-ai-and-llms-in-insurance-common-risks-and-proven-mitigation-tactics/>.
- P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2020. URL <http://arxiv.org/pdf/2005.11401>.
- H.-C. Lin, X. Han, T. Lyu, W.-H. Ho, Y. Xu, T.-C. Hsieh, L. Zhu, and L. Zhang. Task-technology fit analysis of social media use for marketing in the tourism and hospitality industry: a systematic literature review. *International Journal of Contemporary Hospitality Management*, 32(8):2677–2715, 2020. ISSN 0959-6119. doi: 10.1108/IJCHM-12-2019-1031.
- A. Mahmoudabadi and R. A. Lalaei. *Promoting Project Outcomes: A Development Approach to Generative AI and LLM-Based Software Applications’ Deployment*. 2024. doi: 10.2139/ssrn.4907319.
- F. Mahr, G. Angeli, T. Sindel, K. Schmidt, and J. Franke. A reference architecture for deploying large language model applications in industrial environments. In *2024 IEEE 30th International Symposium for Design and Technology in Electronic Packaging (SIITME)*, pages 19–23. IEEE, 2024. doi: 10.1109/siitme63973.2024.10814877.
- M. Mailizar, D. Burg, and S. Maulina. Examining university students’ behavioural intention to use e-learning during the covid-19 pandemic: An extended tam model. *Education and information technologies*, 26(6): 7057–7077, 2021. ISSN 1360-2357. doi: 10.1007/s10639-021-10557-5.
- E. H. Michael Chui. economic-of-generative-ai. 2023. URL https://threeoaksadvisory.com/staging1/wp-content/uploads/2024/03/the_economic-of-generative-ai.pdf.
- Michael Schrage and David Kiron. Intelligent choices reshape decision-making and productivity. *MIT Sloan Management Review*, 2024. URL <https://sloanreview.mit.edu/article/intelligent-choices-reshape-decision-making-and-productivity/>.
- E. Mollick. Reinventing the organization for genai and llms, 2024. URL <https://sloanreview.mit.edu/article/reinventing-the-organization-for-genai-and-llms/>.
- G. Muchenje and M. Seppänen. Unpacking task-technology fit to explore the business value of big data analytics. *International Journal of Information Management*, 69:102619, 2023. ISSN 0268-4012. doi: 10.1016/j.ijinfomgt.2022.102619. URL <https://researchportal.tuni.fi/en/publications/unpacking-task-technology-fit-to-explore-the-business-value-of-bi>.
- S. R. Natasia, Y. T. Wiranti, and A. Parastika. Acceptance analysis of nuadu as e-learning platform using the technology acceptance model (tam) approach. *Procedia Computer Science*, 197:512–520, 2022. ISSN 1877-0509. doi: 10.1016/j.procs.2021.12.168. URL <https://www.sciencedirect.com/science/article/pii/S1877050921023929>.
- Nikola Marangunic and A. Granić. Technology acceptance model: a literature review from 1986 to 2013. *Universal Access in the Information Society*, 2014. URL <https://www.semanticscholar.org/paper/Technology-acceptance-model%3A-a-literature-review-to-Marangunic-Grani%C4%87/e76223ae615bbcb44f3c2d3c99e316fe580e5777>.
- S. Noy and W. Zhang. *Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence*. 2023. doi: 10.2139/ssrn.4375283.
- Y. M. Omar, M. Minoufekr, and P. Plapper. Business analytics in manufacturing: Current trends, challenges and pathway to market leadership. *Operations Research Perspectives*, 6:100127, 2019. ISSN 2214-7160. doi: 10.1016/j.orp.2019.100127. URL <https://www.sciencedirect.com/science/article/pii/S2214716019300934>.
- S. Peng, E. Kalliamvakou, P. Cihon, and M. Demirer. The impact of ai on developer productivity: Evidence from github copilot, 2023. URL <http://arxiv.org/pdf/2302.06590>.
- H. Rafique, A. O. Almagrabi, A. Shamim, F. Anwar, and A. K. Bashir. Investigating the acceptance of mobile library applications with an extended technology acceptance model (tam). *Computers & Education*, 145:

- 103732, 2020. ISSN 0360-1315. doi: 10.1016/j.compedu.2019.103732. URL <https://www.sciencedirect.com/science/article/pii/S0360131519302854>.
- R. Ramakrishnan. A practical guide to gaining value from llms, 2024. URL <https://sloanreview.mit.edu/article/a-practical-guide-to-gaining-value-from-llms/>.
- A. Razak, A. Nazhan, K. Adha, W. A. F. Adzlan, M. A. Ahmad, and A. Azman. *Adapting Safe-for-Work Classifier for Malaysian Language Text: Enhancing Alignment in LLM-Ops Framework*. 2024.
- Renée Richardson Gosline, Yunhao Zhang, Haiwen Li, Paul Daugherty, Arnab D. Chakraborty, Philippe Roussiere, and Patrick Connolly. Nudge users to catch generative ai errors. *MIT Sloan Management Review*, 65(4), 2024. URL <https://sloanreview.mit.edu/article/nudge-users-to-catch-generative-ai-errors/>.
- Rishi Bommasani, Drew A. Hudson, E. Adeli, R. Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, E. Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, Chris Donahue, M. Doumbouya, Esin Durmus, and Stefano Ermon. On the opportunities and risks of foundation models. *arXiv.org*, 2021. URL <https://www.semanticscholar.org/paper/On-the-Opportunities-and-Risks-of-Foundation-Models-Bommasani-Hudson/76e9e2ec3de437ffb30d8b7b629f7fe3e61de5c2>.
- Sam Bowman. Eight things to know about large language models. *arXiv.org*, 2023. URL <https://www.semanticscholar.org/paper/Eight-Things-to-Know-about-Large-Language-Models-Bowman/23a183676b28269e7a427c41da7329b6326a9f17>.
- G. Sehgal and S. Jain. Nutrition education in college students using diffusion of innovation theory: An interventional study. *Preventive Medicine: Research & Reviews*, 1(1):47–51, 2024. ISSN 2950-5836. doi: 10.4103/PMRR.PMRR{\textunderscore}37{\textunderscore}23.
- N. Sendas and D. Rajale. Future trends in mlps. In *The Definitive Guide to Machine Learning Operations in AWS*, pages 371–423. Apress, Berkeley, CA, 2024. ISBN 979-8-8688-1076-3. doi: 10.1007/979-8-8688-1076-3{\textunderscore}10. URL https://link.springer.com/chapter/10.1007/979-8-8688-1076-3_10.
- Y.-Y. Shih and C.-Y. Chen. The study of behavioral intention for mobile commerce: via integrated model of tam and ttf. *Quality & Quantity*, 47(2):1009–1020, 2013. ISSN 0033-5177. doi: 10.1007/s11135-011-9579-x.
- C. Si, D. Yang, and T. Hashimoto. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers, 2024. URL <http://arxiv.org/pdf/2409.04109>.
- E. Strubell, A. Ganesh, and A. McCallum. Energy and policy considerations for deep learning in nlp, 2019. URL <http://arxiv.org/pdf/1906.02243>.
- Sushil. Interpreting the interpretive structural model. *Global Journal of Flexible Systems Management*, 13(2): 87–106, 2012. ISSN 0974-0198. doi: 10.1007/s40171-012-0008-3. URL <https://link.springer.com/article/10.1007/s40171-012-0008-3>.
- Tiago Oliveira and Maria Fraga Martins. Literature review of information technology adoption models at firm level. *Electronic Journal of Information Systems Evaluation*, 14(1):pp110–121–pp110–121, 2011. ISSN 1566-6379. URL <https://academic-publishing.org/index.php/ejise/article/view/389>.
- H.-L. Truong and N. N. T. Nguyen. Culao - constructing utilities of large language models in resource-constrained environments. In *Proceedings of the 2024 International Conference on Information Technology for Social Good*, pages 100–104, New York, NY, USA, 2024. ACM. doi: 10.1145/3677525.3678648.
- Tucker J. Marion and Frank Piller. When generative ai meets product development, 2024. URL <https://sloanreview.mit.edu/article/when-generative-ai-meets-product-development/>.
- L. G. Wallace and S. D. Sheetz. The adoption of software measures: A technology acceptance model (tam) perspective. *Information & Management*, 51(2):249–259, 2014. ISSN 0378-7206. doi: 10.1016/j.im.2013.12.003. URL <https://www.sciencedirect.com/science/article/pii/S0378720614000032>.

J. Yang, H. Jin, R. Tang, X. Han, Q. Feng, H. Jiang, B. Yin, and X. Hu. Harnessing the power of llms in practice: A survey on chatgpt and beyond, 2023. URL <http://arxiv.org/pdf/2304.13712>.

Statutory Declaration

I hereby declare that the submitted thesis is my own original work and was written independently without unauthorized assistance. All sources and references used in the creation of this thesis have been appropriately cited and acknowledged. This thesis has neither been previously submitted for examination nor published elsewhere.

Gi. Nogueira

A Repository Access

The Jupyter notebooks and resources for this project are available in the following GitHub repository:

```
https://github.com/NagarjunaD024
```

To access the notebooks:

1. Clone the repository using:

```
git clone https://github.com/NagarjunaD024/Framework_Methodology
```

2. Navigate to the directory and open the desired notebook (ISM.ipynb or DEMATEL.ipynb) in Jupyter.

B Survey Questionnaire

The survey questionnaire used to collect data for this project is provided below:

- **Question 1:** Please rate the importance of the following factors on a scale of 1 to 5.
- **Question 2:** How frequently do you experience dependencies between these factors?
- **Question 3:** Please specify any additional factors influencing the system.