AN INDUSTRY INTERNSHIP REPORT ON

# ANALYTICS ENGINEERING

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

## BATCHELOR OF TECHNOLOGY IN

**COMPUTER ENGINEERING**

BY

**JIYAN AFASARPASHA PATIL (GR. No.12220055)**

UNDER THE MENTORSHIP OF

Dr. Sonali Antad



**DEPARTMENT OF COMPUTER ENGINEERING**

BANSILAL RAMNATH AGARWAL CHARITABLE TRUST’S

## VISHWAKARMA INSTITUTE OF TECHNOLOGY,

**PUNE – 411037.**

(An Autonomous Institute affiliated to Savitribai Phule Pune University)

**2024 – 2025**



**DEPARTMENT OF COMPUTER ENGINEERING**

### BANSILAL RAMNATH AGARWAL CHARITABLE TRUST’S

**VISHWAKARMA INSTITUTE OF TECHNOLOGY,**

**PUNE – 411037.**

(An Autonomous Institute affiliated to Savitribai Phule Pune University)

**CERTIFICATE**

This is to certify that the industry internship report entitled **ANALYTICS ENGINEERING** submitted by **JIYAN AFASARPASHA PATIL (GR. No.12220055)** is approved for partial fulfillment of the requirements for the award of degree of Batchelor of Technology in Computer Engineering of Vishwakarma Institute of Technology, Savitribai Phule Pune University. This report is a record of bonafide work carried out as a part of his/her internship in **MSCI** during the academic year 2024**–**25, Semester**–**8.

Industry Mentor Internal College Mentor

External Examiner Head of Department

Date:

<Scanned copy of duly signed Completion Certificate issued by Internship offering

Industry>

**ACKNOWLEDGEMENT**

I would like to express my sincere gratitude to Prof. Dr. Rajesh Jalnekar, Director, Vishwakarma Institute of Technology, Pune, for providing the infrastructure, resources, and an enriching academic environment that made this internship possible. His constant encouragement and vision have been truly motivating.

I extend my heartfelt thanks to Prof. Dr. Sandip Shinde, Head of the Department, Computer Engineering, for his continuous support and guidance throughout this journey. His leadership and commitment to academic excellence have been invaluable.

I am deeply thankful to my industry manager and mentor for their unwavering support, mentorship, and technical guidance during the course of the internship. I also wish to acknowledge the entire team for their collaborative spirit and knowledge sharing, which helped me bridge the gap between academic learning and industry practices. Their insights and feedback have greatly contributed to my personal and professional growth.

I would also like to acknowledge the invaluable support of my academic guide, Dr. Sonali Antad for her mentorship, encouragement, and timely advice throughout the internship. Her dedication and availability to resolve queries played a crucial role in the successful completion of this work.

Finally, I express my sincere appreciation to colleagues and teammates for their constant encouragement, motivation, and support throughout this internship.  
  
  
  
Jiyan Afasarpasha Patil

**CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Sr. No.** | **Topic** | **Page No.** |
| **1** | **Abstract** | 6 |
| **2** | **About the Company** | 6 |
| **3** | **Introduction** | 8 |
| **5** | **Problem Statement** | 13 |
| **6** | **Scope and Objectives** | 15 |
| **7** | **Proposed Methodology** | 16 |
| **8** | **Outcome** | 25 |
| **9** | **Conclusion and Future Scope** | 27 |
| **10** | **References** | 27 |

#### I. Abstract

MSCI, a globally recognized provider of investment tools and analytics, has been serving the global investment community for over 50 years. Their expertise in research, data, and technology helps clients understand and analyze the factors influencing risk and return, enabling them to build more effective portfolios with confidence. Clients rely on MSCI's solutions to enhance their investment processes. As a software developer intern at MSCI, I had the opportunity to work with advanced technologies, collaborate with industry experts, and contribute to significant projects. This internship report offers an overview of my experiences, projects, and lessons learned during my tenure, which lasted approximately 12 months, from June 12, 2023, to May 31, 2024. During this time, I undertook various tasks and responsibilities within the software development team.

#### II. About the company

MSCI Inc. stands at the forefront of the investment industry, providing invaluable decision support tools and services for the global investment community. The company's extensive portfolio includes indexes, portfolio analytics, and ESG (Environmental, Social, and Governance) ratings, all catering to a diverse clientele comprising asset managers, pension funds, sovereign wealth funds, and other institutional investors. MSCI's offerings have global reach and impact, with its indexes, in particular, being instrumental in tracking the performance of stock markets and various asset classes across the world. The MSCI World Index, a widely recognized benchmark, monitors the performance of large- and mid-cap stocks in 23 developed countries, showcasing MSCI's significance in the financial landscape.

The company also offers a range of other indexes, from country-specific and sectoral indexes  
to thematic ones. Investors can leverage MSCI's portfolio analytics tools to construct, manage, and optimize their investment portfolios, incorporating risk and return calculations. Additionally, MSCI's ESG ratings empower investors to evaluate the ESG performance of companies, enabling more informed and socially responsible investment decisions. As a global enterprise, MSCI operates in over 20 countries, with its headquarters situated in New York City. The company's commitment to innovation and a deep understanding of client needs positions it for sustained growth. MSCI is well- aligned with the burgeoning trend of sustainable investing, catering to the evolving preferences of investors committed to ESG responsibility.

**Key Offeriings at MSCI :**

* Indexes: MSCI's indexes serve as reliable tools for tracking global stock markets and various asset classes. These indexes underpin trillions of dollars in assets worldwide, exemplifying their impact and credibility.
* Portfolio Analytics: MSCI's portfolio analytics tools provide investors with essential  
  resources for constructing, managing, and benchmarking their portfolios. These tools offer  
  the ability to calculate risk and return, optimizing portfolios for the best outcomes.
* Sustainability Ratings: MSCI’s sustainability ratings provide investors with a valuable metric to assess the environmental, social, and governance-related performance of companies. In an era of increasing emphasis on responsible and impact-driven investing, these ratings enable investors to make informed decisions that align with long-term sustainability goals and values.

**Noteworthy Industry Trends:**

* Rise of Sustainable Investing: MSCI's ESG ratings are pivotal in helping investors identify and invest in companies that are committed to ESG responsibility, responding to the growing demand for socially responsible investments.
* Growth of Passive Investing: Passive investing, typified by tracking market indexes like  
  MSCI's, is on the ascent due to its cost-efficiency and effectiveness.
* Globalization of the Investment Industry: Investors increasingly seek opportunities in global markets, and MSCI's global indexes play a significant role in enabling investors to gain exposure to these markets, ensuring their portfolios remain diversified and well-balanced.

**Technology and Data**

Investment professionals around the world are accelerating advances in technology and data science to generate alpha, serve clients and expand the frontiers of innovative investing. MSCI’s solutions are designed to help investors harness and leverage new technology and data to build better portfolios.

That includes collecting and processing enormous datasets that span investment disciplines, as well as marrying quantitative analysis, research and modeling with technology that helps investors navigate a rapidly changing landscape. Investors turn to MSCI’s solutions to help them:

* Analyze sources of risk and return
* Create multi-asset-class portfolios and investment products that capture market opportunities.
* Harness investment benchmarks and select securities to advance strategies.
* Understand the portfolio implications of climate change and the resilience of companies to
* long-term climate risks.
* Deepen insight into factor investing.
* Elevate transparency and improve reporting.

Technology is the engine that delivers research, models, data, and analytical capabilities. For over  
a decade, MSCI has been using artificial intelligence (“AI”) and natural-language processing (“NLP”) to enhance MSCI’s content and data collection process, transform the client experience, and drive operational efficiencies. Given the years of experience, MSCI is uniquely positioned to connect relevant data to deliver meaningful and easy-to-consume insights, empowering its clients with quality data to create differentiated investment solutions. To further advance MSCI’s capabilities, it has leveraged partnerships with Google Cloud and Microsoft and MSCI has several AI projects in development toimprove and accelerate data production quality, model development, and client service engagement.

#### III. Introduction

MSCI aims to enhance transparency in financial markets, ultimately empowering the investment community to make more informed decisions. By providing comprehensive insights and tools, MSCI endeavors to facilitate better decision-making processes among investors. This commitment to transparency not only fosters trust within the financial sector but also promotes greater efficiency and effectiveness in investment strategies. Through these efforts, MSCI seeks to equip investors with the knowledge and resources necessary to navigate the complexities of financial markets confidently.

MSCI’s primary goal is to empower their clients to make more informed investment choiceswhich lies at the core of their mission. They strive to achieve this through their innovation engine, which is focused on several key objectives:

1. Improving their client services to adapt to the changing dynamics of markets and the diverse needs of investors.
2. Pioneering new research and methodologies to enhance their ability to evaluate investment risks effectively.
3. Assisting their clients in better understanding the global investment landscape, including emerging markets, as well as public and private real estate opportunities.
4. Developing robust frameworks to measure and analyze environmental, social, and governance factors, along with other risk and performance metrics.
5. Providing comprehensive coverage and analytical insights into global data to deliver maximum value to their clients.

MSCI’s competitive advantage is anchored on several key pillars:

1. **Solutions**: They offer multi-asset class models that enable clients to construct and assess portfolios consistently. From product development to risk management and performance attribution, they provide comprehensive support throughout the investment process across all major asset classes. Their suite of products, data, analytics, and decision-support tools collectively empower clients to make betterinformed investment decisions.

1. **Service**: Their dedicated team of specialists ensures the optimized implementation of products and services. Global client service professionals deliver round-theclock support and issue resolution in eight languages. They utilize quantitative and qualitative metrics, including client feedback and service effectiveness assessments, to continuously enhance the client experience.

1. **Research**: With a team of over 200 multidisciplinary research professionals possessing deep expertise in capital markets, they produce thought leadership regularly, offering timely and relevant insights. Leveraging technology and data science, they drive product development and scalable solutions, ensuring ongoing innovation and relevance in their offerings.

At the forefront of innovation, MSCI provides decision-support tools alongside thorough research, transparent methodologies, and a global outlook, catering to the needs of the world's leading investment decision-makers. While each client's requirements are distinct, asset managers, banks, wealth managers, and other financial intermediaries leverage their thought leadership and actionable insights to make more informed investment decisions. MSCI’s offerings include valuable decision-support tools tailored to clients aiming to enhance portfolio performance, refine product development strategies, mitigate unintended risks, comply with regulatory reporting requirements, implement sustainable investing strategies efficiently, and drive growth, scalability, and innovation while operating costeffectively.

In powering better investment decisions, MSCI employs several key strategies:

1. **Relentless Innovation**: Since the inception of our global equity indexes over four decades ago, we have continuously leveraged our extensive data repositories to enhance our models, analytics, and indexes. This ongoing effort has resulted in the development of frameworks for factor classification, detailed analytical investing, market data and analysis for private real estate, and benchmarks for financial instruments.

1. **Flexible Technology**: Our open-architecture analytics platform seamlessly integrates with our clients' data, systems, and workflows. This integration provides our clients with access to high-quality, comprehensive views across their enterprise, enabling strategic risk analysis and addressing investor inquiries effectively.

1. **Differentiated Content**: Our data-rich content, supported by deep capital markets expertise and scientific methods leveraging artificial intelligence and other technological advancements, encompasses a wide range of liquid and illiquid global asset classes.

1. **Actionable Solutions**: MSCI supports various aspects of the investment process, offering clarity in complex global financial markets by transforming vast data points into actionable insights and solutions.

In evolving portfolio solutions, MSCI offers tailored services across various areas:

1. **Investment Policy & Benchmarks**: MSCI provides pension fund managers and asset owners with solutions to define policy benchmarks for their portfolios. Asset owners and managers utilize MSCI's indexes and analytical tools to assess risk and return drivers. Additionally, MSCI's analytics tools assist in confirming that risk exposures align with governance controls, while enriched data enables institutional clients to compare and report sector exposures against benchmarks.

1. **Portfolio Construction & Asset Allocation**: Wealth managers leverage MSCI's factors, modeling, and portfolio construction tools to meet client needs, while asset owners and managers employ factors and ESG criteria to evaluate assets. Asset managers utilize MSCI's factor and index tools to identify risk sources and enhance competitiveness. Real estate and ESG data, analytics, and research support allocation strategies for asset and wealth managers.

1. **Security Selection**: Institutional investors utilize MSCI's ESG ratings for in-depth analysis supporting due diligence activities, while hedge fund managers rely on MSCI's research-derived data. Original research and thought leadership in real estate and impact investing aid clients in making informed decisions. MSCI's Global Industry Classification Standard (GICS), developed with S&P Dow Jones, provides consistent industry and sector identification for industry participants.

1. **Risk Management, Monitoring & Reporting**: Central banks and sovereign wealth funds use MSCI's research, data, and analytics tools for stress-testing and assessing investment risk and returns. Chief investment officers utilize MSCI's high-quality data for developing and testing investment strategies. MSCI's products and tools assist banks in measuring liquidity risk and complying with regulatory mandates, with risk and performance models aligned with asset managers' decisionmaking processes.

In addressing practical investment challenges, MSCI's primary areas of focus encompass Analytics,Environmental, Social, Governance, Real Estate, Factors, and Indexes:

* 1. **Analytics**: MSCI provides institutional investors with a comprehensive perspective on risk and return through its Analytics division. The integrated view incorporates research-enhanced content and tools across major asset classes and timeframes via the MSCI BeonTM platform.
  2. **Sustainability & Climate Engineering:** MSCI’s Sustainability and Climate division focuses on evaluating and modeling long-term environmental and climate-related risks. It provides institutional investors with critical insights to understand how sustainability factors may impact investment strategies and asset performance over time. The advanced research team conducts in-depth analysis of thousands of companies globally, supporting the development of specialized indexes and tools that promote more informed and responsible investment decisions.
  3. **Real Estate**: Within Real Estate, MSCI delivers research and business intelligence services spanning public and private assets. The flexible products and services assist institutional investors, real estate owners, managers, and brokers globally throughout the investment lifecycle, including asset allocation, strategy development, execution, and performance reporting.

* 1. **Factors**: MSCI's Factor solutions provide a framework for assessing, implementing, and reporting factor exposures across equities, fixed income, and multi-asset classes. This aids investors in understanding portfolio risk and return drivers to optimize and build factor-based solutions.

* 1. **Indexes**: MSCI Indexes facilitate portfolio construction and monitoring with a consistent approach across all markets, including the core Modern Index Strategy. Additionally, MSCI offers index licensing for exchange-traded funds and exchangelisted futures and options, ensuring seamless integration into investment strategies.

### V. Project Statement

This internship report provides an overview of my experience at MSCI, where I was engaged in the development of several features for an internal project named **Ask Services (Changed Name for Security Concerns)**. This project focuses on building a **chatbot capable of answering complex internal queries** within MSCI, aimed at enhancing internal support and operational efficiency.

My primary focus during the internship was on learning and applying advanced technologies in modern software development. I worked extensively with **Python**, particularly in backend development and logic orchestration, contributing to enhancements that made the existing system architecture more dynamic, modular, and scalable.

A significant portion of my work centered on **Langflow**, a low-code platform built on top of **LangChain**, which facilitates the rapid development of **LLM-powered applications** and **agentic workflows**. LangChain provides a flexible framework for chaining together language model operations, memory modules, tools, and external data sources, enabling more complex and context-aware interactions. By leveraging LangChain’s core abstractions—such as **Chains**, **Agents**, **Tools**, and **Vector Stores**—I was able to construct advanced workflows that enabled seamless information retrieval and conversational logic.

Within Langflow, I developed multiple **custom components** tailored to specific enterprise needs. These components extended LangChain’s capabilities and were integrated into pipelines involving **ChromaDB**, **embedding models**, and **REST APIs**, allowing the chatbot to query internal systems in real-time and provide more accurate, contextual responses to user queries. I also worked on creating agentic flows that dynamically invoked components multiple times based on evolving input, showcasing the power of **LangChain Agents** when combined with Langflow’s visual interface.

Beyond these tools, I also worked with **Azure** for cloud services and **Terraform** for infrastructure-as-code to manage cloud resources reliably. The internship also involved aspects of **Spring** for backend microservices, and we plan to incorporate **React** in the future to improve the user interface and overall usability of the chatbot system.  
In addition to the technical aspects, I gained valuable experience in **task management using JIRA**, and further developed my **soft skills**, which are essential for effective communication and collaboration. The internship emphasized teamwork and agile practices, enabling us to deliver results that aligned with the broader goals of the organization.  
  
The report begins by introducing MSCI and its position in the financial services industry, highlighting the growing importance of intelligent systems to manage and utilize internal data effectively. It then delves into the foundational concepts of software development, cloud infrastructure, and AI integration, all applied within the context of my contributions to the Ask Services.

All tasks during the internship were assigned and managed using **Atlassian Jira**, a widely adopted project management and issue tracking tool developed by Atlassian. Jira serves as a centralized platform for planning, tracking, and managing work throughout the software development lifecycle, enabling teams to collaborate efficiently and deliver high-quality solutions on schedule.

One of the standout features of Jira is its **support for Agile methodologies**, including **Scrum**, making it well-suited for dynamic and iterative development environments. During the internship, we followed the **Agile Scrum** framework, organizing our work into **sprints**, which are time-boxed iterations that allow for continuous delivery and iterative improvements. Each task or feature was defined as a **user story** and assigned a corresponding **story point** to estimate effort and complexity, facilitating more accurate sprint planning and velocity tracking.

Jira’s flexibility and extensive customization options enabled us to tailor workflows, issue types, and fields to align with our team's specific development process. This ensured better alignment with team objectives and streamlined task management.

In addition to its core capabilities, Jira provides a range of **integrated features and add-ons** that enhance team productivity and visibility. These include **built-in reporting**, **burndown charts**, and **sprint dashboards** for monitoring progress, as well as seamless integrations with version control systems, CI/CD pipelines, and communication tools. The Jira Marketplace further extends functionality with thousands of third-party apps and plugins, supporting a wide variety of use cases.

With its robust features, user-friendly interface, and deep support for Agile practices, Jira played a critical role in managing project tasks and ensuring effective collaboration throughout the internship.

A significant portion of my work involved **confidential use cases** tailored to internal MSCI workflows, so details remain undisclosed. However, the report covers the challenges encountered, such as integrating disparate systems and optimizing dynamic workflows, and summarizes the key insights gained. These insights underscore the **potential benefits of integrating AI-powered support systems** like Ask Services **Ask Services (Changed Name for Security Concerns)** into enterprise environments.

### VI. Scope and Objectives

During the early stages of my internship, my primary objective was to familiarize myself with the existing codebase and bridge any technological knowledge gaps by gaining proficiency in the technologies used within the project. One of the key tasks was to incorporate and understand the existing Spring Boot codebases, ensuring a solid foundation for contributing effectively to the ongoing development efforts.

During my internship, I immersed myself in the world of Python development, aiming not only to acquire proficiency in the language but also to cultivate a deep understanding of its application in real-world scenarios. I actively participated in coding tasks and projects, sharpening my coding skills, and grasping the nuances of designing, implementing, and maintaining Python-based web applications. Through collaborative efforts with fellow team members, I gained valuable insights into best practices and industry standards in Python development.

Beyond Python, I delved into the realm of cloud computing with Azure. As cloud technology continues to revolutionize modern software development, I familiarized myself with Azure cloud services and their practical applications. Engaging in the deployment and management of applications on Azure, I gained hands-on experience and insight into the role of cloud computing in driving innovation and scalability in software development.

During my internship, I worked extensively with Langflow, an intuitive low-code/no-code platform for building LLM-powered applications. My focus was on developing custom components within Langflow to extend its functionality and integrate it with various data sources and APIs. This involved understanding the underlying Python logic and how it interacts with Langflow’s visual interface. Langflow’s UI allows users to design and manage LLM workflows through a drag-and-drop canvas, making it easy to build, test, and iterate on conversational agents, data pipelines, or AI tools without deep coding expertise. Through this experience, I not only strengthened my Python development skills but also gained hands-on exposure to building modular, reusable components tailored to specific business needs.

In addition to developing custom components in Langflow, I also worked on enhancing the integration with ChromaDB, a vector database used for storing and retrieving high-dimensional embeddings. My contributions focused on improving the vectorization process by optimizing how textual data is embedded using models like OpenAI’s text-embedding-ada-002. I gained a strong understanding of how embeddings are generated, stored, and queried to enable semantic search and context-aware responses in LLM workflows. This work allowed me to explore the practical applications of embeddings in AI systems, such as improving retrieval accuracy and building more intelligent, data-driven components.

### VII. Proposed Methodology

The development approach for the **Ask Services (Changed Name for Security Concerns)** project was grounded in modern software engineering principles and focused on transforming a rule-based chatbot into a dynamic, intelligent, and scalable system. The initial phase involved a thorough assessment of the existing architecture, which relied on static backend documents to generate responses. This setup lacked flexibility, semantic understanding, and the ability to fetch or update content dynamically. To overcome these limitations, the project adopted a more advanced technology stack, incorporating Langflow for visual workflow orchestration, ChromaDB for vector-based information retrieval, and OpenAI’s embedding models to semantically process textual content.

The core solution involved designing and implementing a custom Langflow component that integrated multiple functionalities into a single, seamless workflow. This component was responsible for making API calls to internal systems, such as Confluence, retrieving relevant documentation, processing and cleaning the content, generating embeddings using the text-embedding-ada-002 model from OpenAI, and storing these vectors in ChromaDB. With this new vectorized data layer, the chatbot was now able to respond to user queries by performing semantic similarity searches rather than relying on rigid, predefined answers. This significantly improved both the accuracy and relevance of chatbot responses.

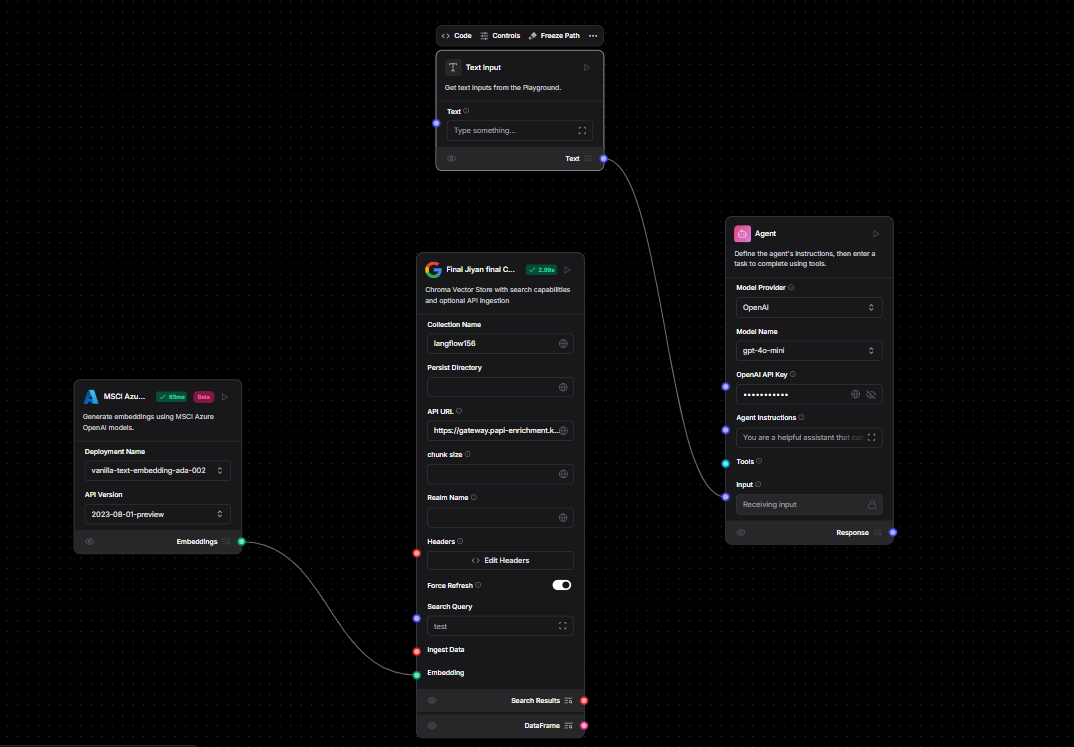


Fig.1.1

Langflow’s intuitive drag-and-drop interface was utilized to visually design and test the workflows. This environment allowed for easier debugging, faster iteration cycles, and clearer visualization of how data flowed between components. The architecture was further enhanced by replacing the static document dependency with a dynamic Confluence component, enabling the system to fetch the latest internal content in real time. Unlike the previous system, which lacked a vector database, the new setup used ChromaDB to store high-dimensional embeddings, enabling smarter and more context-aware interactions.

Throughout the process, collaborative development practices were followed, with regular peer reviews, testing cycles, and refinements based on feedback. The methodology not only ensured a robust technical implementation but also aligned the project goals with practical business needs, resulting in a powerful internal tool capable of enhancing operational efficiency through intelligent automation.

The tools and technologies used in all these projects are mentioned below:

#### 1. Python

Python played a pivotal role in all three projects I contributed to during my internship. It served as the primary programming language for implementing various functionalities and automating processes. Python is a versatile and powerful programming language renowned for its simplicity and readability. Developed in the late 1980s by Guido van Rossum, Python was designed with an emphasis on code clarity and ease of use. Its syntax is straightforward and easy to understand, making it an ideal choice for beginners and experienced programmers alike. With a rich standard library and a vast ecosystem of thirdparty packages, Python can be used for a wide range of applications, from web development and data analysis to artificial intelligence and scientific computing.

One of Python's key strengths lies in its flexibility and scalability. It supports multiple programming paradigms, including procedural, object-oriented, and functional programming, allowing developers to choose the approach that best fits their project requirements. Python's dynamic typing and automatic memory management contribute to its flexibility, enabling rapid development and iteration of code. Additionally, Python's extensive standard library provides built-in support for various tasks, such as file I/O, networking, and regular expressions, further enhancing its versatility and scalability.

Moreover, Python's popularity continues to grow steadily, driven by its vibrant community and widespread adoption in both industry and academia. The availability of comprehensive documentation, tutorials, and online resources makes it easy for developers to learn and master Python. As a result, Python has become a go-to language for software development across diverse domains, empowering developers to build robust, efficient, and scalable solutions for a wide range of applications.

#### 2. Microsoft Azure

The cloud computing platform used by all three projects currently is Microsoft Azure. All the databricks notebooks, storage containers, logic apps, web apps, cosmos containers etc, are hosted here. Azure, commonly referred to as Microsoft Azure, is a cloud computing platform and set of services provided by Microsoft. It offers a wide range of cloud-based services, including computing, storage, networking, databases, machine learning, analytics, and more, enabling organizations to build, deploy, and manage applications and services through Microsoft's global network of data centers.

Azure provides Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS) offerings, allowing users to choose the level of abstraction and control they need for their applications and workloads. With Azure, users can easily scale resources up or down based on demand, pay only for what they use, and access a wide array of tools and technologies to accelerate innovation and digital transformation.

Key features of Azure include robust security and compliance capabilities, hybrid cloud support for seamless integration between on-premises and cloud environments, built-in artificial intelligence (AI) and machine learning services, and a vibrant ecosystem of thirdparty integrations and marketplace solutions.

Overall, Azure empowers businesses of all sizes and industries to leverage the power of cloud computing to drive agility, scalability, and efficiency, enabling them to innovate faster, reduce costs, and deliver greater value to their customers.

#### 3. Chroma DB

ChromaDB is an open-source vector database designed to efficiently store and retrieve high-dimensional embeddings used in modern AI applications, particularly those involving large language models (LLMs). In Langflow, ChromaDB serves as a core component in Retrieval-Augmented Generation (RAG) workflows, enabling the system to perform semantic search and provide context-aware, intelligent responses based on vector similarity.

ChromaDB is optimized for use cases where speed, scalability, and simplicity are essential. It allows embeddings—numerical representations of text or structured data generated by models like OpenAI’s text-embedding-ada-002—to be stored and indexed for fast retrieval. These embeddings represent the semantic meaning of data, allowing ChromaDB to match user queries not just based on keywords but on conceptual relevance.

In the Langflow platform, ChromaDB is seamlessly integrated through custom components and the visual interface. It enables users to create workflows where documents, API responses, or structured content are ingested, vectorized using embedding models, and stored in ChromaDB collections. When a user query is issued, the query is first transformed into an embedding, and ChromaDB retrieves the most semantically relevant documents based on vector similarity. These results are then passed to the LLM for generating a contextually informed response—a process known as Retrieval-Augmented Generation (RAG).

ChromaDB supports real-time embedding updates, configurable similarity thresholds, and fine-tuned retrieval parameters. It is designed to be lightweight and developer-friendly, making it suitable for both experimentation and production deployment within Langflow-powered applications.

**Key capabilities of ChromaDB include:**

* Efficient Vector Storage and Retrieval: Enables rapid similarity search across large collections of embeddings using approximate nearest neighbor algorithms.
* Seamless Embedding Integration: Compatible with a wide range of embedding models including OpenAI, Cohere, Hugging Face, and more.
* Scalable Architecture: Suitable for local development as well as scalable cloud deployment when embedded into larger pipelines.
* Tight Langflow Integration: Supports drag-and-drop configuration in Langflow’s UI, allowing non-technical users to build RAG pipelines with minimal code.
* Support for Modular Pipelines: ChromaDB works alongside other Langflow components like API fetchers, document loaders, and response generators to power dynamic AI systems.
* Role in RAG Systems: Acts as the retrieval engine that feeds relevant content into language models for accurate, grounded, and explainable outputs.

By leveraging ChromaDB in Langflow, developers and data scientists can move beyond static prompt-based AI responses and build intelligent agents that understand context, search dynamically, and deliver precise answers grounded in curated enterprise knowledge.

#### 4. Agentic Langflow

Agentic flows in Langflow represent a powerful paradigm in building intelligent, iterative, and autonomous workflows where AI agents can make decisions, interact with components dynamically, and query tools multiple times based on context. Designed around the concept of agents with reasoning abilities, Langflow enables developers to create workflows where an agent can process a user query, determine missing information, and decide how to retrieve and integrate relevant data in multiple steps—simulating human-like thought processes.

One of the most impactful use cases is when the agent is combined with custom components that perform data retrieval, embedding generation, and vector store queries (e.g., using ChromaDB). This setup is ideal for building RAG (Retrieval-Augmented Generation) pipelines, where:

The agent receives a natural language query.

* It invokes a custom component (e.g., an API + embedding generator) to fetch and embed fresh content.
* These embeddings are stored in ChromaDB for semantic search.
* The agent queries ChromaDB as needed (even multiple times) to retrieve contextually relevant chunks.
* The agent integrates the retrieved context into its response, ensuring accurate, informed, and grounded answers.

A screenshot of a chat

AI-generated content may be incorrect.

Fig. 1.2

#### 5. AZURE And Terrafrom

Terraform is an open-source Infrastructure as Code (IaC) tool developed by HashiCorp, designed to automate the provisioning, configuration, and management of infrastructure across various cloud platforms. It allows users to define both simple and complex infrastructure in human-readable configuration files using the **HashiCorp Configuration Language (HCL)**. These files describe the desired state of infrastructure, which Terraform then uses to deploy and manage resources in a consistent, repeatable manner.

One of Terraform's key strengths is its **declarative approach**, meaning users describe *what* they want their infrastructure to look like rather than *how* to create it. Terraform then takes care of creating, updating, or destroying resources to match the declared configuration. This model ensures consistency and eliminates many of the manual steps that often lead to configuration drift or deployment errors.

Terraform supports multiple cloud providers such as **Azure, AWS, GCP, and many others** through its provider plugins. In the context of Azure, Terraform can be used to provision a wide range of resources including Azure VMs, storage accounts, networking components, and cognitive services like **Azure OpenAI**. For instance, in the **Ask Services (Changed Name for Security Concerns)** project, Terraform was used to automate the deployment of **Azure OpenAI resources**, enabling a scalable and reproducible setup of the underlying AI infrastructure.

Another powerful feature of Terraform is its use of **plans** and **state files**. Before applying any changes, users can generate a **Terraform plan**, which provides a preview of the infrastructure changes that will be made. The **Terraform state** file tracks the current state of the deployed infrastructure, allowing Terraform to detect drift and apply only necessary changes during subsequent runs.

Terraform also supports modularization, enabling infrastructure to be broken down into reusable components (modules), which promotes scalability and maintainability. With built-in support for variables, outputs, dependencies, and lifecycle management, Terraform makes it easier to manage complex cloud environments across different stages of development, testing, and production.

A diagram of a company

AI-generated content may be incorrect.

Fig.1.3

Security and collaboration are enhanced through integrations with Terraform Cloud or other CI/CD platforms, allowing teams to review, approve, and deploy infrastructure changes via version-controlled pipelines.

#### 6. OpenAI Embedding Models and Prompt Engineering

OpenAI provides state-of-the-art language models and embedding models that enable developers to build intelligent applications capable of understanding, generating, and analyzing natural language. Among these, **embedding models** like text-embedding-ada-002 are specifically designed to convert text into high-dimensional vector representations that capture semantic meaning. These vector embeddings are essential for tasks such as semantic search, document clustering, similarity detection, recommendation systems, and Retrieval-Augmented Generation (RAG) pipelines.

In the context of the **Ask Services** project, OpenAI embeddings were used to vectorize internal documents and API responses, which were then stored in **ChromaDB**, a vector database optimized for fast similarity searches. This allowed the chatbot to retrieve the most relevant pieces of information based on a user's query by performing vector similarity search, significantly enhancing the accuracy and contextual relevance of the responses.

The embeddings are generated by passing input text through the OpenAI API, which returns a fixed-length vector representation. These vectors capture not only syntactic features but also deeper semantic relationships between words and phrases. Embeddings are fundamental to building **RAG-based architectures**, where they enable the system to retrieve and ground responses in actual data, ensuring both relevance and reliability in outputs generated by large language models (LLMs).

In parallel, **Prompt Engineering** played a vital role in crafting effective interactions with the LLM. Prompt engineering involves designing and optimizing the input text (prompt) provided to the LLM to elicit desired behavior or responses. Through prompt tuning, chain-of-thought reasoning, and zero-shot or few-shot examples, the LLM's behavior can be guided to produce more accurate, context-aware, and consistent outputs.

In this project, prompt engineering techniques were applied to improve the chatbot's conversational flow and to structure queries to the LLM based on retrieved documents or user intents. This helped minimize hallucination, improve answer grounding, and ensure that the system aligned closely with internal MSCI knowledge sources.

The combined use of OpenAI embeddings and prompt engineering empowered the Ask Services chatbot to go beyond simple keyword matching. It enabled deep semantic understanding, context-aware search, and robust LLM responses — forming the core of an intelligent, scalable internal support tool.

#### 

#### 7. Atlassian Jira

All tasks are assigned to users using Atlassian Jira. Jira is a widely used project management and issue tracking tool developed by Atlassian. It provides teams with a centralized platform to plan, track, and manage their work throughout the software development lifecycle. With Jira, teams can create, prioritize, and assign tasks, track progress, and collaborate effectively to deliver high-quality software products on time.

One of the key features of Jira is its flexibility and customization options, allowing teams to adapt the tool to their specific workflows and processes. Jira supports various project management methodologies, including Agile, Scrum, Kanban, and more, making it suitable for teams of all sizes and across different industries. Users can create custom workflows, issue types, and fields to reflect their unique requirements and preferences, ensuring that Jira aligns with their way of working.

In addition to its core project management capabilities, Jira offers a range of integrated features and add-ons to enhance productivity and collaboration. These include built-in reporting and analytics tools, integrations with other popular development tools and services, and a marketplace with thousands of third-party apps and extensions. With its user-friendly interface, powerful features, and extensive customization options, Jira has become a go-to solution for teams looking to streamline their project management processes and deliver better results.

### VIII. Outcome

**Building an intelligent internal support chatbot using Langflow and ChromaDB**

• We also created a  **prototype** for an internal **Formula Builder chatbot** that leverages **Langflow** and **OpenAI** to assist **non-technical users**—such as project managers, consultants, and accountants—in understanding and working with complex internal formulas used across various business systems.

This chatbot was designed not just for explanation, but also for **interactive formula generation and guidance**. Users could input natural language questions, the system would then retrieve relevant formula logic from **context documents** stored, human-readable explanations or even generate similar formulas tailored to user intent.

MSCI sought to improve internal team productivity by automating responses to common service-related queries using AI. Employees often needed to reference internal APIs, configuration files, and documentation scattered across various systems, resulting in inefficiencies and delays.

• Solution: As part of my internship, I contributed to the design and development of a Langflow-based conversational agent that integrates ChromaDB for vector search, OpenAI’s embedding models for semantic understanding, and custom Langflow components to dynamically retrieve API responses. The solution included a RAG (Retrieval-Augmented Generation) architecture that allowed the agent to fetch context-relevant documents before generating answers.

• Outcome: The deployed chatbot provided accurate, real-time support for internal teams by intelligently querying the underlying system multiple times through agentic loops, reducing turnaround time for service requests and promoting self-service. The successful implementation demonstrated scalable architecture and became a foundation for further LLM-based internal tools.

**Creating reusable agentic flows for dynamic, multi-step decision making**

• The team needed a mechanism to orchestrate complex interactions across internal APIs where the response from one service could dynamically inform the next query—something traditional bots could not handle effectively.

• Solution: I implemented agentic Langflow flows that allowed an agent to intelligently call custom components iteratively until a complete answer was derived. This included decision loops and memory components that enabled context-aware decision trees.

• Outcome: These agentic flows significantly increased the system's ability to handle ambiguous or multi-layered questions by continuously querying internal components and data sources until a satisfactory resolution was reached. This modular, agent-based approach improved both the flexibility and depth of the chatbot’s capabilities.

**Enabling vectorization and embedding of internal knowledge for semantic search**

• Internal documents, such as API references and service metadata, were not easily searchable or understandable via simple keyword-based search systems.

• Solution: I developed a Langflow component pipeline to process internal XML and JSON documents, convert them to readable formats, and vectorize them using OpenAI’s text-embedding-ada-002 model. These embeddings were stored in ChromaDB, enabling semantic search across diverse document types.

• Outcome: The chatbot could retrieve contextually relevant content based on user queries with high accuracy, even when the phrasing varied. This transition from keyword to semantic search greatly improved the discoverability and usability of internal documentation.

**Management of Various Tasks:**

Managing multiple tasks and priorities within the project taught me the importance of effective time management and organization. Prioritizing tasks based on urgency and impact helped me stay focused and productive, ensuring timely delivery of project milestones.

**Soft Skills Development:**

Collaborating with managers, and developers, honed my interpersonal and teamwork skills. Learning  
to communicate effectively, both verbally and in writing, was crucial for gathering requirements,  
providing updates, and soliciting feedback throughout the project lifecycle.

**Communication with Team:**

Regular communication with team members, including daily stand-ups and sprint reviews, fostered a collaborative and transparent work environment. Clear and concise communication helped ensure alignment on project goals, tasks, and timelines, facilitating smooth coordination and progress tracking.

**Adaptability and Problem-Solving:**

Encountering challenges and unforeseen obstacles during the project allowed me to develop  
adaptability and problem-solving skills. Leveraging critical thinking and creativity, I was able to  
overcome technical hurdles and find innovative solutions to meet project requirements.

### IX. Conclusion and Future Scope

This internship served as a pivotal foundation for transitioning from academic concepts to real-world enterprise AI solution engineering. It provided practical exposure to cutting-edge technologies like Langflow for agentic orchestration, OpenAI embeddings for semantic search, and ChromaDB for scalable vector storage. While infrastructure automation with Terraform introduced real DevOps practices and IaC (Infrastructure as Code) discipline.

Working on agentic flows that involved iterative querying of custom components offered deep insights into building intelligent, modular systems capable of dynamic reasoning. Moreover, applying prompt engineering techniques in a production-like environment improved understanding of LLM behavior and optimization. This internship has significantly enhanced both technical competency and architectural thinking within the domain of AI-driven enterprise applications.

Looking ahead, this experience can be expanded by exploring deeper integrations of LangChain or Langflow with cloud-native services, scaling agentic systems and adopting advanced retrieval frameworks. Diving further into areas like secure multi-agent collaboration, LLM fine-tuning, and hybrid cloud deployments would provide added value in a career focused on intelligent, secure, and scalable AI infrastructure.

### X. References

1. <https://www.msci.com/documents/1296102/34522823/MSCI_Company_Brochure.pdf>
2. <https://www.msci.com/>
3. <https://en.wikipedia.org/wiki/MSCI>
4. [https://www.investopedia.com/terms/m/msci.asp#:~:text=Investopedia%20%2F%20Jake](https://www.investopedia.com/terms/m/msci.asp#:~:text=Investopedia%20%2F%20Jake%20Shi-,What%20Is%20MSCI%3F,institutional%20investors%20and%20hedge%20funds)

[%20Shi-](https://www.investopedia.com/terms/m/msci.asp#:~:text=Investopedia%20%2F%20Jake%20Shi-,What%20Is%20MSCI%3F,institutional%20investors%20and%20hedge%20funds)

[,What%20Is%20MSCI%3F,institutional%20investors%20and%20hedge%20funds.](https://www.investopedia.com/terms/m/msci.asp#:~:text=Investopedia%20%2F%20Jake%20Shi-,What%20Is%20MSCI%3F,institutional%20investors%20and%20hedge%20funds)

1. [https://www.computer.org/publications/tech-news/trends/sustainability-digitization-intech](https://www.computer.org/publications/tech-news/trends/sustainability-digitization-in-tech)
2. [Full Year and Fourth Quarter 2023 - MSCI Inc.](https://www.bing.com/ck/a?!&&p=d60bfdf7cc31ea8d09c905a89797dec489c0b54c09a027ea0b1e7c49ed69d0ebJmltdHM9MTc0NjA1NzYwMA&ptn=3&ver=2&hsh=4&fclid=37f0b99d-9c18-68f5-258b-acea9daa69ae&psq=reaserch+paper+where+msci+company+is+metioned+and+its+achievements&u=a1aHR0cHM6Ly9pci5tc2NpLmNvbS9zdGF0aWMtZmlsZXMvYjY1ODEzNWQtYmFkOC00OWQ1LTk1MjItM2JiOWJjMjg4MGNm&ntb=1) October 2023