Building Intelligent Applications with Generative AI

Explore the potential of AI for next gen applications

Yattish Ramhorry

www.bpbonline.com

First Edition 2025

ISBN: 978-93-55519-139

LIMITS OF LIABILITY AND DISCLAIMER OF WARRANTY

The information contained in this book is true to correct and the best of author's and publisher's knowledge. The author has made every effort to ensure the accuracy of these publications, but publisher cannot be held responsible for any loss or damage arising from any information in this book.

All trademarks referred to in the book are acknowledged as properties of their respective owners but BPB Publications cannot guarantee the accuracy of this information.

Dedicated to

My beloved parents:

Mrs. Deomathee Ramhorry

Late. Parmanand Ramhorry

and

My sisters,

Nireshaa Lala *and* ***Nirvana Ramhorry***

About the Author

Yattish Ramhorry is the visionary CTO and founder of 4IR Technologies, a leading technology firm in Johannesburg, South Africa. With over 20 years of software development experience, he has spearheaded projects ranging from simple web apps to complex distributed systems.

A passionate mentor, Yattish supports aspiring software developers across Africa through the Google Developers Exchange program. He also organized the Johannesburg School of AI Meetups at Microsoft South Africa from 2018 to 2020.

As an advocate for Ethical AI, Yattish writes extensively on Software Development, Generative AI, and Responsible AI, with articles featured in prominent publications like Data Driven Investor and Analytics Vidhya. Yattish also serves as an AI Ethics Consultant at Ethical Intelligence, where he works on fostering ethical AI practices within organizations.

Building Intelligent Applications with Generative AI is not Yattish's first book. He has previously written and published *Building NFTs with Ethereum* in February 2023.

About the Reviewers

- **Dr Tariq Ahmad**, PhD CEng MIET has been working in the IT industry since 1994. He specializes in .NET and Python and has worked for KPMG and Sybase. He lives in England and currently works for a leading consulting company, helping clients understand and harness the power of NLP by combining his knowledge of NLP algorithms, techniques, and tools with his consulting skills to guide clients through the complex landscape of NLP. Prior to this, Tariq was a senior developer for a company that specialized in providing software solutions and services to the public sector.

- **Gaurav Deshmukh** is a Senior Software Engineer Tech lead with more than a decade of expertise in software engineering and technical management. His current focus is on utilizing his diverse skill set proficiency in Java, React, TypeScript, AWS, Docker, and Kubernetes to drive innovation and deliver high-impact No Code Low Code automation solutions and integration of Generative AI in existing software applications. Gaurav is passionate about mentorship and knowledge sharing and actively participates in professional organizations such as IEEE and ACM. He also serves as a mentor on platforms like ADPList and Criya. He is dedicated to continuous learning and staying updated with emerging technologies to drive digital transformation and create value in today's rapidly evolving landscape by contributing as a technical reviewer for various books on topics such as Engineering, DevOps, Kubernetes, Generative AI, and Cybersecurity.

Acknowledgement

I am deeply grateful to the individuals who have supported and guided me throughout the creation of this book, making it a labor of love.

I wish to express my gratitude to my mother Deomathee Ramhorry, and my sisters, Nireshaa Lala and Nirvana Ramhorry for their continued support and encouragement throughout the duration of this book, without which I could not have finished writing another book. I also want to honor the memory of my late father, Parmanand Ramhorry. His wisdom, strength, and love continue to inspire me every day. This book is a testament to the values and determination he instilled in me.

I am also grateful to Olivia Gambelin, founder of Ethical Intelligence for her valuable contributions and feedback that helped me shape **Chapter 13, Ethical Considerations of Generative AI.**

I am also grateful to BPB Publications for providing me with the opportunity to write another book for them. The guidance and expertise offered by their editors and technical reviewers were invaluable in bringing this book to life. The journey of writing and revising was long, but the collaboration with reviewers, technical experts, and editors made it a rewarding experience.

To all those who dream and believe, I encourage you to never give up on your hopes and aspirations, as they can indeed come true.

"No matter what people tell you, words and ideas can change the world"

— Robin Williams

I am thankful and grateful for this opportunity.

Preface

My journey into generative AI began in August 2020, when I was introduced to GPT-3. I was among a small group of 100 beta testers of this unique and innovative technology. At that time, I was unfamiliar with generative AI and the broader field of artificial intelligence, having only worked on a few small AI projects.

As I delved deeper into generative AI, I assembled a team to develop an application using GPT-3 to generate smart contracts for the Ethereum Blockchain. Although the idea was revolutionary at the time, the project did not progress as hoped. However, the experience was invaluable, leading me to explore various new avenues, including the creation of this book. This journey has opened up a world of opportunities and possibilities for further exploration and discovery.

Since then, there has been a significant progress in generative AI. AI and generative AI has the potential to disrupt every area of our lives, and if left unchecked, can cause harm in ways that we have not yet imagined. The time is apt for us to further our understanding of generative AI, and implement guardrails and safety mechanisms for the responsible use of generative AI.

Building Intelligent Applications with Generative AI aims to equip readers with the knowledge and essential tools to build intelligent applications using generative AI.

As much as this book is dedicated to understanding and building applications with generative AI, a portion of the book is also dedicated to the ethical and responsible use of generative AI.

It is important to raise awareness about the responsible use of AI among developers as early as possible. This book aims to instill this awareness by sharing valuable insights and strategies for building LLM powered applications responsibly.

The book is divided into practical exercises for developers to get familiar with the APIs and developer toolkits required to build LLM powered applications. The theoretical chapters are intended to provide developers with a glimpse into real-world applications of generative AI. We also explore open-source language models like LLama 2 and learn how they are a powerful way of building and deploying LLM applications cost effectively.

I hope this book provides you with a firm foundation for building creative and state-of-the-art generative AI applications!

Chapter 1: Exploring the World of Generative AI

In this introductory chapter, readers will be introduced to the exciting world of generative AI and its potential for building intelligent applications. Further, it provides a high level overview of what generative AI is and how it differs from traditional programming approaches.

This chapter also provides an introduction into **Large Language Models** (**LLMs**) and how they relate to Generative AI. The readers will also gain an understanding of how generative AI models like GPT-4 can generate realistic and contextually relevant content, opening up possibilities for automation and creativity.

Chapter 2: Use Cases for Generative AI Applications

In this chapter, readers explore a fascinating array of real-world use cases that showcase the transformative potential of generative AI applications. By examining diverse industries and domains, this chapter provides a comprehensive view of how generative AI is applied to solve complex problems and drive innovation.

This chapter also introduces some of the most popular generative AI tools like Copy.AI, DALL-E, Midjourney, Leonardo.AI, ElevenLabs, and Speechify used for text, image, video, and audio generation.

Chapter 3: Mastering the Art of Prompt Engineering

In this chapter, the readers delve into the art and science of prompt engineering, a crucial skill for effectively leveraging generative AI models. Crafting the right prompts are essential for obtaining desired outputs from these models.

This chapter equips the readers with the knowledge and skills necessary to design prompts that yield accurate, coherent and contextually relevant resources. The readers will also learn about the significance of prompt design in shaping generative AI outputs.

Chapter 4: Integrating Generative AI Models into Applications

This chapter serves as a foundational guide for developers who are new to generative AI and are interested in integrating it into their applications. The readers will gain a solid understanding of the fundamental concepts and techniques required to successfully build applications using generative AI models.

This chapter begins with an introduction to the key components in building applications with generative AI, including data preparation, model selection and integration strategies.

Chapter 5: Emerging Trends and the Future of Generative AI

In this chapter, developers who are new to generative AI will explore the exciting realm of emerging trends and the future possibilities of this rapidly evolving field. The readers will gain valuable insights into the latest advancements, research breakthroughs, and potential applications that lie ahead.

This chapter covers the current state of generative AI and its impact on various industries. It discusses recent trends and developments, including advancements in generative AI models, novel technologies for data generation, and the integration of generative AI with other technologies such as computer vision and others. The readers will also explore the cutting edge research areas in generative AI, including topics like few-shot learning, unsupervised learning, and multi-modal generation.

Chapter 6: Building Intelligent Applications with the ChatGPT API

The ChatGPT API offers developers the ability to integrate the power of conversational AI into their applications, enabling dynamic and interactive user experiences. We explore how to leverage the capabilities of ChatGPT to create Chatbots, virtual assistants, and other AI-driven applications.

Through hands-on examples, the readers will learn how to harness the power of ChatGPT API and unlock its full potential for building intelligent applications. Additionally, we will walk them through creating an end-to-end chat assistant system which provides young learners with information about dinosaurs. This chapter will also teach the readers how to submit various questions about dinosaurs, and how the chat assistant will respond with relevant answers.

Chapter 7: Retrieval Augmented Generation with Gemini Pro

In this practical, hands-on chapter, developers will get a glimpse into retrieval augmented generation using LangChain and Google's Gemini Pro large language model. For this project, developers will use a T4 GPU available on Google Colabs free tier.

Chapter 8: Generative AI Applications with Gradio

Gradio is an open-source Python library that facilitates the creation of interactive interfaces and web applications for machine learning and data science. It serves as a potent tool for building interactive demos and web applications seamlessly across various platforms. In this chapter, we build three demo applications, showcasing Gradio's potential to quickly create user interfaces with minimal coding.

We will also build a **Natural Language Processing** (**NLP**) application with text summarization, and an image captioning application to allow users to upload images, and generate unique captions. These NLP applications are built upon open-source language models like DistilBart and the Salesforce Blip image captioning model. Finally, we will build a chat interface using the `gr.ChatInterface` Gradio object and the OpenAI ChatGPT API.

Chapter 9: Visualize your Data with LangChain and Streamlit

In this practical hands-on chapter, the readers will learn how to use powerful tools like LangChain, Python and Streamlit to analyze CSV files. This chapter provides step-by-step guidance to create a functional and interactive assistant that can understand user queries and retrieve information from CSV datasets. By the end of this chapter, the readers will gain valuable experience in integrating generative AI into their applications and harnessing its capabilities for document-based interactions.

Chapter 10: Building LLM Applications with Llama 2

By the end of this chapter, the readers will learn how to use Llama 2, a powerful open-source language model from Meta (Facebook) AI. Through practical exercises, the readers will gain skills in setting up, and deploying Llama 2 for tasks including automated blog post generation and a pair programming assistant.

The readers will also discover the benefits of open-source development and learn how to evaluate the performance of LLM based applications using LangSmith. Additionally, we will use a local instance of Llama 2, downloaded from Hugging Face, and a serverless instance of Code Llama from Together.AI to build projects.

Chapter 11: Building an AI Document Chatbot with Flowise AI

For many developers, the fundamental question arises: how can they leverage their existing data effectively in generative AI applications? In this chapter, we delve into the utilization of Flowise AI, a user-friendly platform designed for building Chatbots that extract information from PDF documents and provide answers to questions.

What sets this approach apart is its “no-code” nature, eliminating the need for coding throughout the chapter. This feature makes Flowise AI an excellent choice for developers seeking to rapidly prototype and experiment with ideas and theories, enabling them to validate concepts swiftly and efficiently.

Chapter 12: Best Practices for Building Applications with Generative AI

In this chapter, the readers will learn the best practices and strategies for effectively building applications with GPT-4 and other generative AI models. They will learn essential strategies and techniques to optimize performance, reliability, and scalability of their applications while ensuring a seamless integration of generative AI.

Chapter 13: Ethical Considerations of Generative AI

In this final chapter of the book, we dive into the critical topic of ethical considerations surrounding generative AI. As generative AI becomes more prevalent in various applications, it is essential for developers to understand and address the ethical challenges and implications associated with these technologies. This chapter explores the potential risks, biases, privacy concerns, and societal impacts of generative AI, providing guidelines and frameworks for responsible development and deployment.

Code Bundle and Coloured Images

Please follow the link to download the
Code Bundle and the ***Coloured Images*** of the book:

https://rebrand.ly/o9dhcq8

The code bundle for the book is also hosted on GitHub at
https://github.com/bpbpublications/Building-Intelligent-Applications-with-Generative-AI.
In case there's an update to the code, it will be updated on the existing GitHub repository.

We have code bundles from our rich catalogue of books and videos available at **https://github.com/bpbpublications**. Check them out!

Errata

We take immense pride in our work at BPB Publications and follow best practices to ensure the accuracy of our content to provide with an indulging reading experience to our subscribers. Our readers are our mirrors, and we use their inputs to reflect and improve upon human errors, if any, that may have occurred during the publishing processes involved. To let us maintain the quality and help us reach out to any readers who might be having difficulties due to any unforeseen errors, please write to us at :

errata@bpbonline.com

Your support, suggestions and feedbacks are highly appreciated by the BPB Publications' Family.

Piracy

If you come across any illegal copies of our works in any form on the internet, we would be grateful if you would provide us with the location address or website name. Please contact us at **business@bpbonline.com** with a link to the material.

If you are interested in becoming an author

If there is a topic that you have expertise in, and you are interested in either writing or contributing to a book, please visit **www.bpbonline.com**. We have worked with thousands of developers and tech professionals, just like you, to help them share their insights with the global tech community. You can make a general application, apply for a specific hot topic that we are recruiting an author for, or submit your own idea.

Reviews

Please leave a review. Once you have read and used this book, why not leave a review on the site that you purchased it from? Potential readers can then see and use your unbiased opinion to make purchase decisions. We at BPB can understand what you think about our products, and our authors can see your feedback on their book. Thank you!

For more information about BPB, please visit **www.bpbonline.com**.

Join our book's Discord space

Join the book's Discord Workspace for Latest updates, Offers, Tech happenings around the world, New Release and Sessions with the Authors:

https://discord.bpbonline.com

Table of Contents

1. Exploring the World of Generative AI 1
Introduction 1
Structure 2
Objectives 2
An introduction to generative AI 3
The evolution of AI 3
Supervised learning *4*
Unsupervised learning *5*
Semi-supervised learning *6*
Generative AI and deep learning 7
Generative AI algorithms and techniques 11
Generative AI tools and resources 11
Understanding generative AI models 12
Types of generative models *12*
Understanding the differences between GANs vs VAEs *13*
Text-to-text *16*
Text-to-image *16*
Text-to-video and Text-to-3D *16*
Text-to-task *17*
Text-to-code *17*
Traditional programming versus Generative AI 18
Benefits of incorporating generative AI into applications 19
Conclusion 20
Points to remember 20
References 20

2. Use Cases for Generative AI Applications 21
Introduction 21
Structure 22
Objectives 22
Generative AI in content generation 22
Jasper.ai *23*

Copy.ai 24
Creative media: Music composition, art and design 25
DALL-E 2 25
Midjourney 27
Leonardo.Ai 30
Conversational agents powered by generative AI 32
Virtual reality and simulation using generative AI 34
Generative AI and metaverse: Re-imagining the internet 34
Generative AI and Roblox: Creating 3D worlds 34
Generative AI and video game development 35
Revolutionizing gaming with AI-powered conversations 35
Conclusion 36
Points to remember 37
References 37

3. Mastering the Art of Prompt Engineering 39
Introduction 39
Structure 40
Objectives 40
The role of prompt engineering 40
Shaping Generative AI outputs with prompts 40
Strategies for formulating effective prompts 42
Preparing the development environment 43
Adding delimiters to prompts 45
Structuring your output 46
Checking whether conditions are satisfied 47
Few-Shot prompting with base examples 49
Allowing the model time to think 49
Instructing the model to provide its own solution 51
Asking the model to adopt a persona 54
Prompt templates 54
Model limitations: Avoiding hallucinations 56
Prompt engineering for different application domains 57
Language and text domain 57
Image generation domain 58
Music composition domain 59

Refining prompts through iterative processes 60
Defining the goal 60
Writing the initial prompt 60
Testing the prompt 61
Analyzing the output 61
Refining the prompt 61
Iterating the prompt 62
Implementation 62
Case studies of successful prompt engineering in real-world applications 62
Conclusion 63
Points to remember 63
References 64

4. Integrating Generative AI Models into Applications 65
Introduction 65
Structure 66
Objectives 66
Generative AI models: Applications and use cases 66
ChatGPT 67
GPT 3 68
GPT-4 69
Uses of GPT-4 69
Duolingo 70
Be My Eyes 70
What else is GPT-4 capable of doing? 70
Whisper 70
Claude – Anthropic 71
Google Gemini 73
Llama 2 75
SeamlessM4T 75
DALL-E 2 76
Integrating generative AI models into applications 77
OpenAI 78
Chat 78
Embeddings 79
Fine-tuning 79

Claude, Anthropic *80*
Generative AI Studio, Google *81*
Understanding tokens in Large Language Models 83
Counting the cost of tokens *84*
Data preparation and preprocessing for generative AI 86
Handling large-scale data with generative AI 88
AI-driven coding with VS Code and GitHub Copilot 89
Conclusion 90
Points to remember 90
References 91

5. Emerging Trends and the Future of Generative AI 93
Introduction 93
Structure 93
Objectives 94
Overview of the current state of generative AI 94
Recent trends and developments in generative AI 95
Advancements in generative AI models 98
Multimodal language models *98*
Fine-tuning Large Language Models *99*
Vector embeddings *100*
Emerging research areas in generative AI 103
Few-shot learning *103*
Retrieval augmented generation *104*
Generative AI in natural language processing 105
Future applications and opportunities 106
Conclusion 109
Points to remember 109
References 109

6. Building Intelligent Applications with the ChatGPT API 111
Introduction 111
Structure 112
Objectives 112
Setting up the ChatGPT API 112
Language model entity types *113*

Implementing the ChatGPT API in applications 114
Moderating user inputs and system outputs 116
Preventing prompt injections 117
Using delimiters to prevent prompt injections 118
Chaining prompts 119
Enhancing the user experience 120
Fine-tuning with the ChatGPT API 122
Identifying when fine-tuning is needed 122
Pricing 123
Fine-tuning steps 123
OpenAI function calling 125
Conclusion 128
Points to remember 128
References 129

7. Retrieval Augmented Generation with Gemini Pro 131
Introduction 131
Structure 132
Objectives 132
Introduction to Gemini Pro 132
Understanding Retrieval Augmented Generation 132
Python pre-requisites and Gemini Pro access 134
Configuring LangChain with Gemini Pro 136
Creating vector embeddings 138
Building a vector database with Chroma 139
Creating the RAG RetrievalQA Component 140
Creating a prompt template with LangChain 140
Conclusion 142
Points to remember 142
References 143

8. Generative AI Applications with Gradio 145
Introduction 145
Structure 146
Objectives 146
Python frameworks for building web applications 146

Streamlit *146*
Gradio *146*
Plotly Dash *147*
Panel *147*
Anvil *147*
NLP tasks with a simple Gradio interface 147
Simple image captioning application with Gradio 151
Creating a chat application with Gradio 155
Conclusion 157
Points to remember 158

9. Visualize your Data with LangChain and Streamlit 159
Introduction 159
Structure 159
Objectives 160
Introduction to LangChain 160
Setting up the development environment 161
Setting up the agent 162
Creating agent.py *163*
Adding the query_agent function *164*
Setting up the StreamLit interface 165
Running the interface 169
Testing the chatbot's performance and accuracy 169
Conclusion 171
Points to remember 171

10. Building LLM Applications with Llama 2 173
Introduction 173
Structure 174
Objectives 174
Open-source Large Language Models 174
Security and data privacy concerns *175*
Accessibility and flexibility *175*
Cost-effective fine-tuning *175*
Community collaboration *175*
The future of open-source models 176

Introducing Llama 2 from Meta 176
Downloading the Llama 2 model 177
End-to-end blog generation platform 180
Installing dependencies *181*
Creating app.py for Python code *181*
Defining generate_blog function *181*
Setting up the Streamlit interface *183*
Testing the Streamlit application *185*
Pair programming with Llama 2 186
Setting up Serverless Code Llama *187*
Creating the Project files *188*
Writing technical documentation with Llama 2 ..... *189*
Using Llama 2 for code creation *190*
Debugging Code with Llama 2 ..... *193*
LangSmith tutorial: LLM evaluation 195
Creating a LangSmith account *195*
Creating a LangSmith API Key *196*
Creating a LangSmith project *197*
Integrating LangSmith into LangChain applications *198*
Viewing LangSmith tracing data *198*
Conclusion 200
Points to remember 201
References 202

11. Building an AI Document Chatbot with Flowise AI 203
Introduction 203
Structure 203
Objectives 204
Introduction to Flowise AI 204
Setting up Flowise AI 204
Understanding the document Chatbot use case 209
Building the Chatbot Workflow 209
Testing the Chatbot 211
Best practices and tips 212
Conclusion 213
Points to remember 213

12. Best Practices for Building Applications with Generative AI 215
Introduction 215
Structure 216
Objectives 216
Building applications with Generative AI 216
Model selection and evaluation 217
Open-source versus proprietary LLMs 218
Pros of a premium (paid) language model 218
Cons of a premium (paid) language model 219
Pros of using open-source language models 219
Cons of using open-source language models 220
Best practices for LLMOps 220
Difference between LLMOps and MLOps 222
Importance of LLMOps 223
Benefits of LLMOps 224
Understanding LLMOps components 224
Managing LLMOps with an LLMOps platform 225
Data privacy and security guidelines 225
Guidelines for developers 226
Guidelines for Generative AI Applications 226
Conclusion 226
Points to remember 227
References 227

13. Ethical Considerations of Generative AI 229
Introduction 229
Structure 230
Objectives 230
Addressing ethical practices in generative AI 230
Fairness issues in AI-generated content 231
Strategies for mitigating data biases 232
Synthetic data as a valuable tool to mitigate biases 232
Anonymizing or removing sensitive attributes 232
Addressing fairness in content generation and decision-making 233
Benefits of fairness in AI 233
Inclusive strategies for generative AI development 234

User data privacy concerns 234
Building trust and understanding with users 235
Strategies to build trust in LLM powered applications 236
Copyright issues in AI-generated content 237
Ethical use of AI in creative generation 238
Ethics in the development lifecycle 239
Environmental impact of generative AI 241
Conclusion 241
Points to remember 241
References 242

Index 243-248

Chapter 1
Exploring the World of Generative AI

Introduction

A fundamental shift is taking place in the world right now, and **large language models** (**LLMs**) are driving that shift. This disruptive new technology has changed how we think, play, work, and transact. Since the launch of ChatGPT to the public by OpenAI in November 2022, the world has witnessed firsthand, the immense potential of **artificial intelligence** (**AI**). AI, once accessible to only a select few, is now widely accessible by people of all ages, thanks to large language models like ChatGPT, Gemini, and Claude. With LLMs, your ability to create is magnified a thousand times.

Now, anyone can become an artist, musician, writer, poet, web developer, software engineer, and UI/UX designer. Applications like Midjourney, and DALL-E, allows anyone to create art that rivals the best artists in the art world, while applications like Udio, and Suno, allows anyone to perfectly create music that is unrecognizable as AI generated.

Another notable benefit of LLMs is its ability to comprehend and translate multiple languages, effectively breaking down language barriers. This means that content creators are no longer required to speak the language of the market they serve. For instance, with the help of LLMs, it is now possible to create content in French, German, and Spanish, thereby, reaching a broader audience than was previously possible.

However, this is not the only thing that LLMs can do. One of AI's key strengths is its ability to recognize patterns in data and make predictions based on those patterns. With LLMs, it

is now possible for anyone to forecast stock market trends, predict currency fluctuations, and discover new methods for analyzing enterprise data. Applications leveraging this capability can reveal trends and insights that human analysts might overlook.

Experts predict that from 2022 to 2026, the world will experience a transformation comparable to the changes that occurred from the 1900s to the 2000s. This means, nearly a century's worth of progress condensed into just four years. Some estimates even suggest that around 40% of jobs may be replaced by AI within the next three to five years. However, this shift is not something to be feared, as it also holds the potential to generate new opportunities and business models. Embracing this fundamental shift is crucial, as it presents more opportunities rather than threats.

This chapter introduces you to the exciting world of generative AI and its potential for building intelligent applications. It provides a high-level overview of what generative AI is and how it differs from traditional programming approaches. We also provide an introduction to LLMs and how they relate to Generative AI.

Structure

In this chapter, we will discuss the following topics:

- An introduction to generative AI
- Evolution of AI
- Generative AI and Deep Learning
- Generative AI algorithms and techniques
- Generative AI tools and resources
- Understanding generative AI models
- Traditional programming versus Generative AI
- Benefits of incorporating generative AI into applications

Objectives

By the end of this chapter, you will gain an understanding of how generative AI models like GPT-3 and GPT-4 can generate realistic and contextually relevant content, opening possibilities for automation and creativity.

An introduction to generative AI

Over the past few months, LLMs such as ChatGPT and Midjourney have taken the world by storm. Whether it is writing poetry, generating a creative image, or helping you plan a dinner for six, we are seeing a change in the performance of AI and its potential to drive enterprise value.

LLMs form a part of a different class of models known as **foundation models**. The term foundation models was first coined by a research team at *Stanford University* when they saw that the field of AI was converging into a new paradigm. In the past, AI applications were built by training a library of different AI models, where each AI model was trained on data to perform a *specific* task.

Stanford researchers predicted that we will witness a paradigm shift where we will have a foundational capability, or foundation model, that will drive all the same use cases and applications. So, the same applications that we were building in the past, with conventional AI, using the same model, could drive a number of additional applications.

The point is that the foundation model could be transferred to perform a number of different tasks. What makes foundation models powerful with its ability to perform multiple different tasks and functions, is that foundation models have been trained on large volumes of data, in an unsupervised manner, on unstructured data.

What this means, in the language domain, is that we feed terabytes of textual data to train the model. If we were to provide the model with a sentence like, *no use crying over* the model might respond with *spilled milk*.

It is the generative capability of the model, predicting and generating the next words in a sentence based on previous words it has previously seen, is what makes foundation models part of the field of AI referred to as *generative AI* since we are generating something new. In our sentence example, the next sequence of words in a sentence.

Foundation models have the potential to disrupt many industries, including health care, finance, retail, and customer service, among others. They can be used to detect fraud and provide personalized customer service.

The evolution of AI

In this section, we will explain artificial intelligence to those new to generative AI and machine learning. Since we are exploring generative AI, let us begin with some context. The two most commonly asked questions are: What is artificial intelligence, and *what is the difference between artificial intelligence and machine learning?*

Figure 1.1 graphically illustrates the relationship between AI and **machine learning** (**ML**):

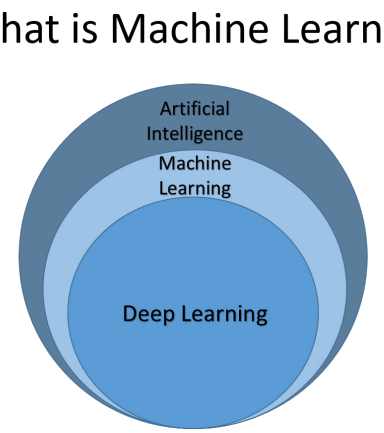

***Figure 1.1**: Illustrating the relationship between AI and ML*

AI is a subfield or branch of computer science that deals with systems that can reason, learn, and act autonomously. By *autonomously*, we mean software/agents that can act independently, without any human or other intervention.

In simpler terms, AI is the theory and development of computer systems and algorithms that can perform tasks that normally require human intelligence. So, we are essentially building machines that can act and think like humans.

One area of AI is machine learning, which involves training a model or models using input data. This subfield of AI enables the model to make accurate predictions from novel data that it has never encountered before, using the same process used to train the initial model.

Machine learning allows computers to learn without explicit programming instructions. Two of the most popular types of machine learning models are supervised and unsupervised.

Supervised learning

Supervised learning uses labeled data, which is labeled with a *tag*, like a name, type, or number. A simple example to illustrate supervised learning is teaching a computer how to recognize cats and dogs in pictures. You would show it many pictures of cats and dogs that are each labeled as either cat or dog. The computer would then learn the features that distinguish cats from dogs and use that knowledge to classify new images as cats or dogs.

Supervised learning is frequently utilized in finance and banking to detect credit card fraud. It is also utilized in text classification problems, where the objective is to forecast the class label of a specific text. Predicting the sentiment of a tweet or a product review is often another use case of text classification.

Some examples of supervised learning applications are spam detection, image and object recognition, and price prediction.

In addition, supervised learning can be used for anomaly detection. For instance, email spam detection is one of the widest anomaly detection algorithms today.

Supervised learning involves inputting testing data values (**x**) into the model and generating a prediction. This prediction is then compared to the data used to train the model.

Unsupervised learning

Unsupervised learning uses data that does not have a tag, that is, un-labeled data. Unsupervised learning involves looking at the raw data, and seeing if it falls into groups. Some examples of unsupervised learning applications are categorizing news articles, recognizing objects in images, predicting diseases using medical imaging, and segmenting customers or DNA patterns.

Google News categorizes articles on the same story from different online news outlets using unsupervised learning. This method involves discovering patterns in the raw data to see if they can be grouped together naturally.

The main difference between supervised and unsupervised models is that with supervised learning or supervised models, we use *labels*. Unsupervised learning, on the other hand, learns patterns from the data it is provided.

Understanding these basic concepts forms the basis of your understanding of generative AI. Let us dig a little deeper to show this graphically. *Figure 1.2* illustrates the differences between supervised and unsupervised learning:

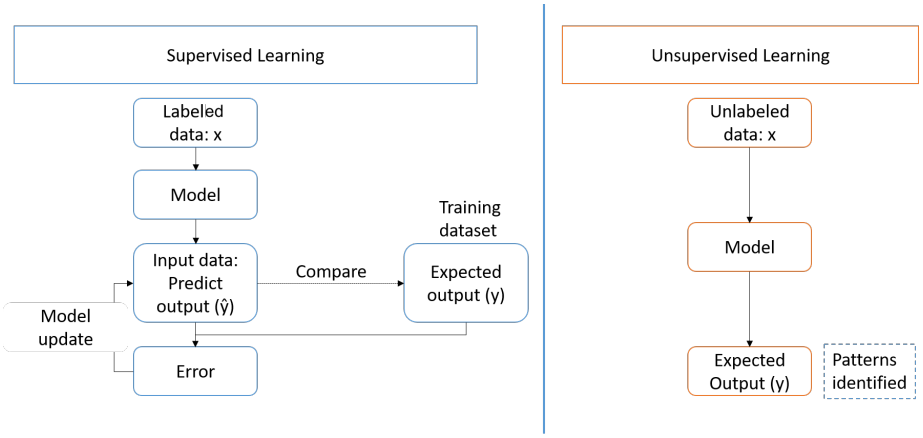

***Figure 1.2**: Supervised learning versus unsupervised learning*

When there is a significant difference between the predicted test data and actual training data, it is considered an error. To minimize this error, the model works towards reducing the gap between the predicted and actual values. This process is a common optimization problem.