



Graduation Thesis

To obtain the Industrial Engineering diploma

OPTION : DATA SCIENCE & ARTIFICIAL INTELLIGENCE

Unlocking Potential: Open Source LLMs in Rag Systems

Written by:

Ghribi Ouassim Abdelmalek

Jury:

President:

Prof. Philippe CUDRÉ-MAUROUX - National Polytechnic School of Algiers

Examiners:

Prof. Pierre SENELLART - National Polytechnic School of Algiers

Supervisors:

Mr. Oussama ARKI - National Polytechnic School of Algiers

Mr. Hachem BETROUNIE - BIGmama Technology

ENP 2024

10, Avenue des Frères Oudek, Hassen Badi, BP. 182, 16200 El Harrach, Alger,
Algérie.

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

In the Name of God, Most Gracious, Most Merciful

Acknowledgments

I extend my deepest gratitude to my thesis supervisors, Mr. Arki Oussama and Mr. Betrouni Hachem, for their indispensable guidance and unwavering support throughout this project. Their expertise and continuous encouragement have been pivotal in shaping its direction. I am sincerely thankful for their patience, insightful feedback, and generous investment of time, all of which have been crucial in refining this work.

On a personal level, I am profoundly grateful to my family and friends for their unwavering support and encouragement throughout this demanding yet rewarding journey. Their unwavering belief in my abilities has been a constant source of motivation, particularly during challenging times.

Moreover, I would like to express my appreciation to all individuals who directly or indirectly contributed to the completion of this thesis. Your support and encouragement have been invaluable and deeply cherished.

Special recognition goes to Mr. Hadj Khelil for his belief in me over the past two years. Progressing from a junior front-end developer to a team lead, overseeing an exceptional team of ambitious individuals, has been an enriching experience. Together, we have strived to develop software with the potential to make a positive impact on the world.

Abstract

Table of Contents

Acknowledgments	iv
Abstract	vi
List of Figures	vii
List of Tables	viii
List of Listings	ix
Acronyms	x
1 State of Play	1
1.1 Presenting BIGmama Technology	1
1.1.1 Mission	2
1.1.2 Vision	2
1.1.3 BIGmama Specificities	3
1.1.4 Products	4
1.1.5 Service: Asset based consulting (ABC)	4
1.1.6 Hyko	5
1.2 RAG: Enhancing LLMs with External Knowledge	6
1.3 Conclusion	8
2 State of the Art	9
2.1 Machine Learning	10
2.1.1 Definition	10
2.1.2 Relationships to other fields	10
2.1.3 Types of Machine Learning	11
2.1.4 Limitations of Machine Learning: Challenges and Considerations	12
2.2 Deep Learning	13
2.2.1 Definition	13
2.2.2 Deep Learning vs. Machine Learning	13
2.2.3 Deep Learning Applications	13

List of Figures

- 1.1 Hyko’s Landing page (<https://www.hyko.ai>) 6
- 1.2 Retrieval Augmented Generation. <https://www.promptingguide.ai/research/rag#introduction-to-rag> 7
- 2.1 Machine learning as subfield of AI, from [4]. 10
- 2.2 Types of machine Learning. 11

List of Tables

Listings

Glossary

Here are the main acronyms used in this document. The meaning of an acronym is usually indicated once, when it first appears in the text.

ABC	Asset Based Consulting
AI	Artificial Intelligence
ML	Machine Learning
DL	Deep Learning
NLP	Natural Language Processing
LLM	Large Language model
GPT	Generative Pretrained Transformer
BERT	Bidirectional Encoder Representations from Transformers
RAG	Retrieval-Augmented Generation

State of Play

“The true power of AI lies in its ability to augment human intelligence, not replace it.”

Andrew Ng, Co-founder of Coursera and former Chief Scientist at Baidu

In this chapter, we embark on a journey through the landscape of existing literature and research pertinent to the convergence of open source Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) systems, in collaboration with BIGmama Technology. Our objective is to establish a robust foundation for understanding the intricacies and potentials inherent in this symbiotic relationship. As we delve into the state of play, we aim to illuminate key theoretical underpinnings and empirical insights that underpin our study of open source LLMs and Rag systems within the context of BIGmama Technology. To initiate this exploration, we contextualize our inquiry within the broader discourse of natural language processing and information retrieval, elucidating the multifaceted challenges and opportunities presented by the integration of open source LLMs and Rag systems. Subsequently, we navigate through seminal concepts and methodologies, including the principles of LLM architecture, Rag framework components, and their applications in various domains. These foundational insights serve as pillars supporting our quest to explore the potential synergies and advancements offered by the convergence of open source LLMs and Rag systems within the framework of BIGmama Technology.

1.1 Presenting BIGmama Technology

BIGmama is an innovative startup founded in France with a subsidiary in Algeria, specializing in data science and artificial intelligence solutions. With over 9 years of experience as of 2024, the company has established itself as a leading provider of bespoke predictive applications tailored to meet the unique needs of its clients.

Guided by a distinguished board of directors comprising former CEOs of global conglomerates such as Danone and Safran, BIGmama boasts a team of highly skilled data scientists and software engineers. This multidisciplinary team, consisting of more than a dozen experts in their respective fields, brings a wealth of knowledge and expertise to the company’s endeavors.

BIGmama’s commitment to innovation and cutting-edge technologies has allowed them to deliver state-of-the-art solutions that leverage the power of data science and

artificial intelligence. With a strong focus on providing customized solutions, the company has consistently exceeded its clients' expectations, enabling them to gain valuable insights and make data-driven decisions.

Through its strategic partnerships and collaborations, BIGmama continues to push the boundaries of what is possible in the realms of data science and AI (Artificial Intelligence), positioning itself as a driving force in the ever-evolving technological landscape.

1.1.1 Mission

Beyond speeches, BIGmama's mission is to propose effective methodologies, action plans, and tools to:

- Solve problems with artificial intelligence tools.
- Put people at the heart of technology, i.e., the hybridization of AI.
- Make technology a common good, shareable, and accessible to the maximum number of people who can participate and contribute.

The company's commitment extends far beyond mere rhetoric; it is dedicated to developing practical solutions, methodologies, and actionable plans to leverage the power of artificial intelligence for problem-solving. BIGmama places a strong emphasis on ensuring that technology remains human-centric, fostering a harmonious integration and hybridization of AI with human intelligence.

Moreover, BIGmama recognizes the importance of democratizing access to technology, ensuring that it becomes a shared resource that can benefit society as a whole. The company aims to create an environment where the maximum number of individuals can actively participate and contribute to the advancement of technology.

Ultimately, BIGmama's mission is to position technology as a catalyst for freedom, empowering individuals and communities to unlock new possibilities and overcome barriers through innovative solutions.

1.1.2 Vision

"Standing on the Shoulders of Giants"

We find ourselves at a pivotal moment in human history, where the rapid advancements in artificial intelligence are poised to reshape our world in unprecedented ways. Groundbreaking technologies like ChatGPT, developed by industry giants, are at the forefront of this transformative wave, mobilizing vast resources to redefine the boundaries of what is possible.

However, BIGmama's approach is not to compete directly with these titans but rather to harness their innovations as a springboard for its own technological endeavors. By building upon the foundations laid by these industry leaders, BIGmama aims to leverage their advancements as a catalyst for its own visionary project.

Rather than reinventing the wheel, BIGmama recognizes the value in standing on the shoulders of giants, capitalizing on their groundbreaking work to propel its own unique solutions forward. This strategic approach not only allows for more efficient progress but also fosters an environment of collaboration and synergy within the broader technological ecosystem.

With a keen understanding of the rapidly evolving landscape, BIGmama remains agile and adaptable, poised to seize opportunities presented by the advancements of industry giants. By embracing a collaborative mindset and harnessing the collective wisdom of the technological community, BIGmama is well-positioned to contribute its own innovative solutions, shaping the future in alignment with its vision and mission.

1.1.3 BIGmama Specificities

One of the valuable heritages that BIGmama acquired during the 9 years of actively developing bespoke AI applications to its clients is its unique methodology of work. This methodology is centered around the following ideas :

- **Data science starts with problematization:**

At BIGmama, the approach to data science does not begin with data science itself but rather with reframing and problematization. The company's clients often arrive with a subject or topic rather than a clearly defined problem. BIGmama has developed a methodology to convert these topics into well-defined problems that data scientists can effectively tackle.

- **The data scientist is a "maverick":**

BIGmama recognizes that the data scientist is a "maverick" who cannot be restrained. The company understands that it cannot impose specific tools or methods on data scientists when it comes to solving problems.

- **Data science a tool to solve problems:**

BIGmama views data science not as an end in itself but as a means to solve problems. Often, the solution to their clients' problems lies outside the realm of data science tools.

- **Hybridization:**

BIGmama believes that the future of AI lies in what is commonly referred to as Hybrid AI (Hybrid Artificial Intelligence). This approach encompasses a set of

methodologies and techniques aimed at combining the potential of models with purely human knowledge. The company believes that this hybridization allows for putting humans at the heart of technological development and producing tools that are more efficient, easier to maintain, and less expensive.

1.1.4 Products

Historically, BIGmama is a startup specialized in the development of predictive applications for third parties. The company's approach is specific and based on the principles of hybridization between human and model capabilities. These **specificities** have led BIGmama to develop its own internal tools. These tools have evolved as the company's needs have changed, transitioning from Iko¹ to Hyko². BIGmama is now willing to make these tools available to the broader technology ecosystem.

As a startup with a rich history in predictive application development, BIGmama has cultivated a unique approach that emphasizes the seamless integration of human expertise and model capabilities. This distinctive methodology, rooted in the principles of hybridization, has driven the company to develop its own suite of proprietary internal tools.

Over time, as BIGmama's requirements have evolved, these tools have undergone continuous refinement and adaptation. The company has transitioned from its initial Iko platform to the more advanced Hyko solution, ensuring that its tools remain aligned with its ever-changing needs and the dynamic technological landscape.

In recognition of the value these tools hold for the broader tech community, BIGmama has made the strategic decision to make them available to the larger ecosystem. By sharing its internally developed tools, the company aims to foster collaboration, knowledge exchange, and the collective advancement of technological solutions.

1.1.5 Service: Asset based consulting (ABC)

The evolving business landscape, driven by rapid technological advancements and heightened competition, necessitates innovative solutions for optimizing operations and fostering sustainable growth. Traditional consulting approaches, such as report writing, are no longer sufficient to meet the demand for specialized solutions tailored to individual client needs. This shift has led to the emergence of asset-based consulting (ABC), a specialized field focused on creating, leveraging, and maximizing the value of a company's assets. In ABC, consultants collaborate closely with clients to identify and harness tangible and intangible assets, including technology, intellectual property, and human capital, to drive operational efficiency and gain a competitive edge. Unlike traditional consulting, ABC takes a holistic and internal perspective,

¹<https://big-mama.io/en/iko>

²<https://www.hyko.ai>

recognizing the unique assets and capabilities of each organization. Consultants draw on multidisciplinary expertise to assess, diagnose, and unlock hidden value through innovative asset management strategies. As industries become increasingly complex and competitive, the demand for asset-based consulting, particularly in leveraging AI assets, continues to grow across sectors. From optimizing supply chains to capitalizing on intellectual property, asset-based consulting offers a strategic framework and expertise to help organizations achieve their objectives in a dynamic business environment.

1.1.6 Hyko

Hyko is a platform that empowers you to become an AI toolmaker. It can be thought of as using "Lego" pieces from its vast toolbox, filled with AI models (over 130,000), APIs (such as Stability.ai, OpenAI, Cohere), web scraping tools, utility functions, and more.

Hyko is a powerful platform that puts the tools and resources necessary for building AI applications at your fingertips. By offering a comprehensive toolbox and streamlined workflow, Hyko empowers users to unleash their creativity and bring their AI-powered visions to life.

Many companies grapple with numerous manual tasks performed repetitively by employees, often numbering in the hundreds or thousands each month. Traditional automation tools struggled to automate these tasks fully due to their reliance on human reasoning. However, LLMs have revolutionized automation by exposing reasoning capabilities through an API.

In response to this advancement, Hyko emerged as an AI-first platform, designed not merely to streamline employee workflows but to entirely replace them from start to finish. The platform's automation builder offers unparalleled flexibility, enabling the automation of highly complex workflows with ease.

The future value of AI is anticipated to derive not solely from AI engineers or data scientists, but also from field experts who possess the capability to integrate their domain knowledge with AI technologies to address everyday challenges. This shift underscores the growing importance of domain-specific expertise in harnessing the potential of AI for practical applications.

This imperative is frequently encountered among Hyko users, who often find themselves in need of extracting vital insights from their diverse data repositories, which encompass web sources, PDF documents, and email archives. In response to this demand, Rag systems emerge as key facilitators, particularly when coupled with open-source LLMs such as Llama3, Phi3, and mixtral7B. Rag systems excel in the retrieval, aggregation, and generation of pertinent information, empowering users to distill valuable insights from their data reservoirs efficiently and effectively. This

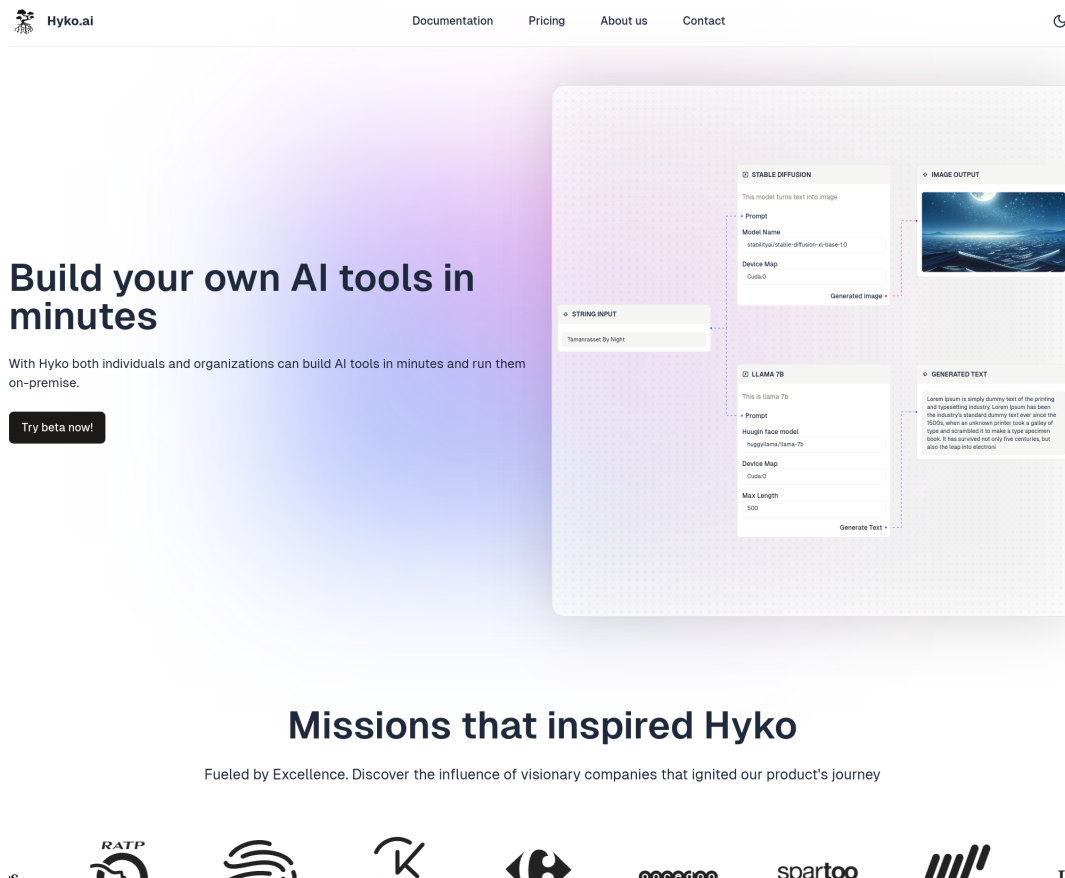


Figure 1.1: Hyko's Landing page (<https://www.hyko.ai>)

seamless integration enhances the capacity for nuanced understanding and actionable intelligence extraction, thereby augmenting the utility and versatility of Hyko across diverse information retrieval and analysis endeavors.

1.2 RAG: Enhancing LLMs with External Knowledge

LLMs showcase impressive capabilities but encounter challenges like hallucination [5], outdated knowledge, and non-transparent, untraceable reasoning processes. RAG has emerged as a promising solution by incorporating knowledge from external databases. This enhances the accuracy and credibility of the generation, particularly for knowledge-intensive tasks, and allows for continuous knowledge updates and integration of domain specific information. RAG synergistically merges LLMs' intrinsic knowledge with the vast, dynamic repositories of external databases [1].

General-purpose language models can be fine-tuned to achieve several common tasks such as sentiment analysis and named entity recognition. These tasks generally don't require additional background knowledge.

For more complex and knowledge-intensive tasks, it's possible to build a language model-based system that accesses external knowledge sources to complete tasks. This enables more factual consistency, improves reliability of the generated responses, and helps to mitigate the problem of "hallucination".

Meta AI researchers introduced a method called Retrieval Augmented Generation (RAG) [2] to address such knowledge-intensive tasks. RAG combines an information retrieval component with a text generator model. RAG can be fine-tuned and its internal knowledge can be modified in an efficient manner and without needing retraining of the entire model.

RAG takes an input and retrieves a set of relevant/supporting documents given a source (e.g., Wikipedia). The documents are concatenated as context with the original input prompt and fed to the text generator which produces the final output. This makes RAG adaptive for situations where facts could evolve over time. This is very useful as LLMs's parametric knowledge is static. RAG allows language models to bypass retraining, enabling access to the latest information for generating reliable outputs via retrieval-based generation.

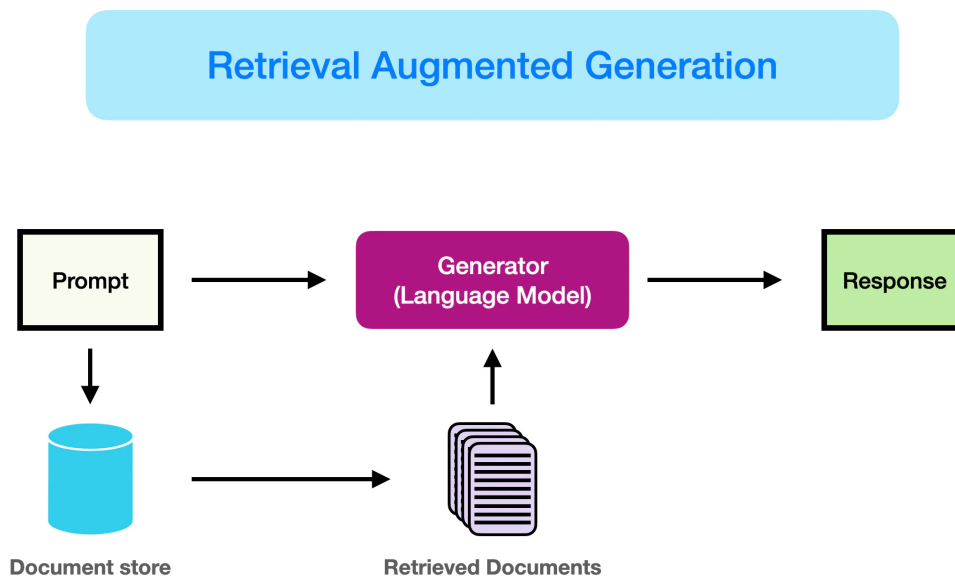


Figure 1.2: Retrieval Augmented Generation. <https://www.promptingguide.ai/research/rag#introduction-to-rag>

In short, the retrieved evidence obtained in RAG can serve as a way to enhance the accuracy, controllability, and relevancy of the LLM's response. This is why RAG can help reduce issues of hallucination or performance when addressing problems in a highly evolving environment.

1.3 Conclusion

In this chapter, we lay the foundation for the subsequent sections of this thesis. We introduce fundamental concepts in Retrieval-Augmented Generation (RAG) systems and Large Language Models (LLMs), along with a basic overview of the host company, BIGmama Technology.

State of the Art

In recent years, the remarkable advancements in Natural Language Processing (NLP) have been primarily driven by the development of LLMs. These LLMs, such as GPT (Generative Pretrained Transformer) and BERT (Bidirectional Encoder Representations from Transformers), have demonstrated remarkable capabilities in understanding and generating human-like text. However, despite their impressive performance, LLMs still face challenges in effectively retrieving and incorporating relevant context for generating accurate and coherent responses.

Enter RAG systems, a novel approach that seeks to overcome the limitations of traditional LLMs by integrating retrieval mechanisms with generation models. RAG systems combine the strengths of both retrieval and generation techniques to enhance the quality and relevance of generated text.

This chapter provides a comprehensive exploration of RAG systems, delving into their architecture, components, training processes, applications, advantages, and challenges. We begin by establishing a foundational understanding of LLMs and their evolution, laying the groundwork for understanding the need for RAG systems. We then proceed to dissect the intricacies of RAG systems, discussing the role of retrieval in providing context and the role of generation in producing fluent responses.

Through detailed examination and analysis, we uncover the inner workings of RAG systems, exploring how retrieval and generation components interact within the architecture. Real-world applications and use cases of RAG systems across various domains are elucidated, demonstrating their potential to revolutionize tasks such as question answering, dialogue generation, and content creation.

Furthermore, we evaluate the advantages and limitations of RAG systems compared to traditional LLMs and other approaches in NLP. By examining performance metrics, challenges, and future directions, we gain insights into the transformative impact of RAG systems on the field of natural language processing.

In summary, this chapter serves as a comprehensive guide to RAG systems, offering readers a deep dive into one of the most promising advancements in NLP. As we navigate through the complexities and potentials of RAG systems, we pave the way for understanding their role in shaping the future of human-computer interaction and language understanding.

2.1 Machine Learning

2.1.1 Definition

Machine learning (ML) is a specialized field within artificial intelligence (AI) dedicated to the development and analysis of statistical algorithms capable of learning from data and making predictions without explicit programming. In recent years, artificial neural networks (ANNs) have emerged as particularly powerful tools, showcasing superior performance compared to traditional methods. ML boasts a wide range of applications spanning various domains, including natural language processing, computer vision, speech recognition, email filtering, agriculture, and medicine. In business contexts, ML is commonly known as predictive analytics. While not all ML methods rely on statistical principles, computational statistics significantly influences the methodologies employed in the field [4].

2.1.2 Relationships to other fields

Artificial Intelligence (AI) encompasses the field of computer science dedicated to creating systems capable of emulating human-like intelligence, problem-solving, and decision-making. Machine Learning (ML) is a subset of AI focused on enabling computers to learn from data without being explicitly programmed. Within ML, Deep Learning stands out as a subfield that employs neural networks with multiple layers to learn complex representations of data, particularly effective in tasks like image and speech recognition. ML and Deep Learning are integral components of AI, providing the framework for developing intelligent systems capable of learning, reasoning, and adapting to new information, thereby advancing the capabilities of AI across various domains.

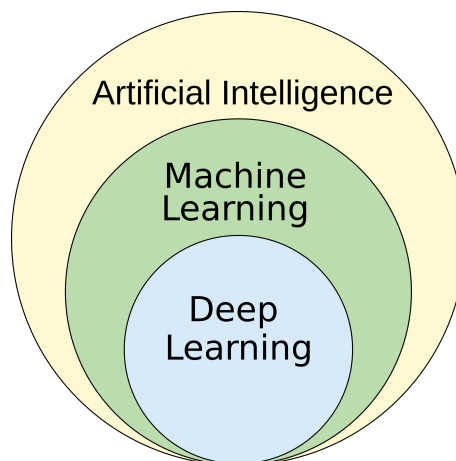


Figure 2.1: Machine learning as subfield of AI, from [4].

2.1.3 Types of Machine Learning

Machine Learning can be broadly categorized into several types based on the learning approach, the availability of labeled data, and the feedback mechanism. The main types of machine learning are:

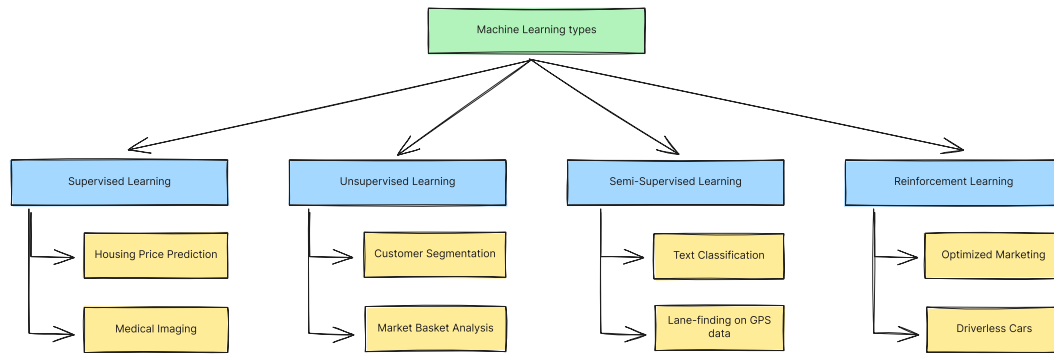


Figure 2.2: Types of machine Learning.

Supervised Learning

- In supervised learning, the algorithm learns from labeled data, where each input example is paired with the corresponding output label.
- The goal is to learn a mapping from inputs to outputs, allowing the algorithm to make predictions or classifications on unseen data.
- Common supervised learning tasks include regression (predicting continuous values) and classification (predicting discrete labels).

Unsupervised Learning:

- Unsupervised learning involves learning from unlabeled data, where the algorithm aims to find patterns, structures, or relationships in the data without explicit guidance.
- The algorithm discovers hidden structures or clusters within the data, facilitating tasks such as clustering, dimensionality reduction, and anomaly detection.

Semi-Supervised Learning:

- Semi-supervised learning lies between supervised and unsupervised learning, where the algorithm learns from a combination of labeled and unlabeled data.

- The presence of both labeled and unlabeled data allows the algorithm to leverage additional information and improve performance, particularly in scenarios where labeled data is scarce or expensive to obtain.

Reinforcement Learning:

- Reinforcement learning involves an agent learning to make decisions by interacting with an environment to achieve a specific goal.
- The agent receives feedback in the form of rewards or penalties based on its actions, enabling it to learn through trial and error.
- Reinforcement learning is commonly used in applications such as game playing, robotics, and autonomous systems.

2.1.4 Limitations of Machine Learning: Challenges and Considerations

Although machine learning is a powerful technique for extracting knowledge from data, it also has certain limitations that are important to consider:

- **Data dependency:** Machine learning heavily relies on the quality and quantity of training data. If the data is poorly labeled or unrepresentative, it can lead to errors in predictions.
- **Overfitting:** When the model is too complex compared to the training data, it can overfit and not generalize well to test data. This can also result in poor performance for new data.
- **Explainability:** Machine learning models can be very complex and difficult to understand. It can be challenging to understand how the model makes decisions and to explain these decisions to users.
- **Biased data:** Training data can be biased due to factors such as data selection, human biases, or measurement errors. This can lead to biased predictions for test data.
- **Lack of diversity:** Machine learning models may lack diversity in the types of data they can handle. For example, machine learning models may struggle to process unstructured data such as images, sounds, and texts.
- **Computational cost:** Machine learning algorithms may require high computational power and significant storage resources to process large amounts of data. This can be costly and time-consuming to train and implement models.

2.2 Deep Learning

2.2.1 Definition

Deep learning encompasses machine learning techniques that utilize artificial neural networks (ANNs) with representation learning. The term "deep" denotes the use of multiple layers within the network architecture. These methods can be supervised, semi-supervised, or unsupervised, depending on the learning approach employed.

Architectures such as deep neural networks, deep belief networks, recurrent neural networks, convolutional neural networks, and transformers are prominent examples of deep learning models. These models have found applications in diverse fields, including computer vision, speech recognition, natural language processing, machine translation, bioinformatics, drug design, medical image analysis, climate science, material inspection, and board game programs. In many cases, deep learning approaches have achieved results comparable to, and sometimes surpassing, human expert performance [3].

2.2.2 Deep Learning vs. Machine Learning

Deep learning stands apart from traditional machine learning in its approach to data and learning methods. Machine learning algorithms typically rely on structured, labeled data for predictions, where specific features are defined and organized into tables. While machine learning can handle unstructured data, it often requires pre-processing to structure it. In contrast, deep learning streamlines this process by directly processing unstructured data such as text and images. It automates feature extraction, reducing reliance on human experts. For instance, in categorizing pet photos, deep learning algorithms autonomously identify key features, like ears, crucial for distinguishing between animals. In contrast, machine learning requires manual feature hierarchy establishment by human experts.

2.2.3 Deep Learning Applications

Deep learning has a wide range of applications across various domains due to its ability to learn complex patterns from large volumes of data. Some of the different types of applications for deep learning include:

Image Recognition and Computer Vision:

- Deep learning is extensively used for tasks such as image classification, object detection, facial recognition, and image segmentation.
- Applications include self-driving cars, medical image analysis, surveillance systems, and augmented reality.

Natural Language Processing (NLP):

- Deep learning is employed for understanding and generating human language, enabling tasks such as sentiment analysis, machine translation, text summarization, and chatbots.
- Applications include virtual assistants, language translation services, social media sentiment analysis, and customer support chatbots.

Speech Recognition and Synthesis:

- Deep learning is utilized for speech recognition, speech-to-text conversion, and text-to-speech synthesis.
- Applications include voice assistants, speech-to-text transcription services, voice-controlled devices, and interactive voice response systems.

Recommendation Systems:

- Deep learning is used to build personalized recommendation systems that analyze user preferences and behavior to suggest relevant products, content, or services.
- Applications include movie or music recommendation platforms, personalized advertising, and e-commerce product recommendations.

Autonomous Vehicles:

- Deep learning plays a crucial role in enabling autonomous vehicles to perceive and navigate their environment.
- Applications include object detection, lane detection, traffic sign recognition, and path planning for self-driving cars.

These are just a few examples of the diverse applications of deep learning, demonstrating its versatility and impact across various industries and fields.

Bibliography

- [1] Yunfan Gao et al. *Retrieval-Augmented Generation for Large Language Models: A Survey*. 2024. arXiv: [2312.10997](#) [cs.CL].
- [2] Sebastian Riedel et al. *Retrieval Augmented Generation: Streamlining the creation of intelligent natural language processing models*. 2020. URL: [%5Curl%7Bhttps://ai.meta.com/blog/retrieval-augmented-generation-streamlining-the-creation-of-intelligent-natural-language-processing-models/%7D](#).
- [3] Wikipedia contributors. *Deep learning* — *Wikipedia, The Free Encyclopedia*. https://en.wikipedia.org/w/index.php?title=Deep_learning&oldid=1220437155. [Online; accessed 29-April-2024]. 2024.
- [4] Wikipedia contributors. *Machine learning* — *Wikipedia, The Free Encyclopedia*. https://en.wikipedia.org/w/index.php?title=Machine_learning&oldid=1220758567. [Online; accessed 28-April-2024]. 2024.
- [5] Yue Zhang et al. *Siren’s Song in the AI Ocean: A Survey on Hallucination in Large Language Models*. 2023. arXiv: [2309.01219](#) [cs.CL].