



ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA (ICAI)

Master in Big Data

Final Master's Thesis

**Business processes automation with artificial
intelligence**

Author

Tomás Alcántara Carrasco

Directed by

Jorge Gómez Berenguer

Madrid

January 2025

Tomás Alcántara Carrasco, declara bajo su responsabilidad, que el Proyecto con título **Business processes automation with artificial intelligence** presentado en la ETS de Ingeniería (ICAI) de la Universidad Pontificia Comillas en el curso académico 2024/25 es de su autoría, original e inédito y no ha sido presentado con anterioridad a otros efectos. El Proyecto no es plagio de otro, ni total ni parcialmente y la información que ha sido tomada de otros documentos está debidamente referenciada.

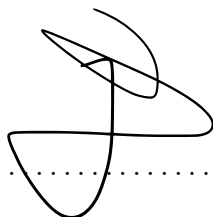
Fdo.:

Fecha: ..06.. / ..01.. / ...2025..

Autoriza la entrega:

EL DIRECTOR DEL PROYECTO

Jorge Gómez Berenguer



Fdo.:

Fecha: ..06.. / ...01.. / ...2025..

V. B. DEL COORDINADOR DE PROYECTOS

Carlos Morrás Ruiz-Falcó

Fdo.:

Fecha: / /

AUTORIZACIÓN PARA LA DIGITALIZACIÓN, DEPÓSITO Y DIVULGACIÓN EN RED DE PROYECTOS FIN DE GRADO, FIN DE MÁSTER, TESIS O MEMORIAS DE BACHILLERATO

1º. Declaración de la autoría y acreditación de la misma.

El autor D. Tomás Alcántara Carrasco **DECLARA** ser el titular de los derechos de propiedad intelectual de la obra: Business processes automation with Artificial Intelligence, que ésta es una obra original, y que ostenta la condición de autor en el sentido que otorga la Ley de Propiedad Intelectual.

2º. Objeto y fines de la cesión.

Con el fin de dar la máxima difusión a la obra citada a través del Repositorio institucional de la Universidad, el autor **CEDE** a la Universidad Pontificia Comillas, los derechos de digitalización, de archivo, de reproducción, de distribución y de forma gratuita y no exclusiva, por el máximo plazo legal y con ámbito universal, de comunicación pública, incluido el derecho de puesta a disposición electrónica, tal y como se describen en la Ley de Propiedad Intelectual. El derecho de transformación se cede a los únicos efectos de lo dispuesto en la letra a) del apartado siguiente.

3º. Condiciones de la cesión y acceso

Sin perjuicio de la titularidad de la obra, que sigue correspondiendo a su autor, la cesión de derechos contemplada en esta licencia habilita para:

- (a) Transformarla con el fin de adaptarla a cualquier tecnología que permita incorporarla a internet y hacerla accesible; incorporar metadatos para realizar el registro de la obra e incorporar “marcas de agua” o cualquier otro sistema de seguridad o de protección.
- (b) Reproducir la en un soporte digital para su incorporación a una base de datos electrónica, incluyendo el derecho de reproducir y almacenar la obra en servidores, a los efectos de garantizar su seguridad, conservación y preservar el formato.
- (c) Comunicarla, por defecto, a través de un archivo institucional abierto, accesible de modo libre y gratuito a través de internet.
- (d) Cualquier otra forma de acceso (restringido, embargado, cerrado) deberá solicitarse expresamente y obedecer a causas justificadas.
- (e) Asignar por defecto a estos trabajos una licencia Creative Commons.
- (f) Asignar por defecto a estos trabajos un HANDLE (URL *persistente*).

4º. Derechos del autor.

El autor, en tanto que titular de una obra tiene derecho a:

- (a) Que la Universidad identifique claramente su nombre como autor de la misma

- (b) Comunicar y dar publicidad a la obra en la versión que ceda y en otras posteriores a través de cualquier medio.
- (c) Solicitar la retirada de la obra del repositorio por causa justificada.
- (d) Recibir notificación fehaciente de cualquier reclamación que puedan formular terceras personas en relación con la obra y, en particular, de reclamaciones relativas a los derechos de propiedad intelectual sobre ella.

5º. Deberes del autor.

- (a) El autor se compromete a:
- (b) Garantizar que el compromiso que adquiere mediante el presente escrito no infringe ningún derecho de terceros, ya sean de propiedad industrial, intelectual o cualquier otro.
- (c) Garantizar que el contenido de las obras no atenta contra los derechos al honor, a la intimidad y a la imagen de terceros.
- (d) Asumir toda reclamación o responsabilidad, incluyendo las indemnizaciones por daños, que pudieran ejercitarse contra la Universidad por terceros que vieran infringidos sus derechos e intereses a causa de la cesión.
- (e) Asumir la responsabilidad en el caso de que las instituciones fueran condenadas por infracción de derechos derivada de las obras objeto de la cesión.

6º. Fines y funcionamiento del Repositorio Institucional.

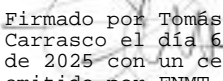
La obra se pondrá a disposición de los usuarios para que hagan de ella un uso justo y respetuoso con los derechos del autor, según lo permitido por la legislación aplicable, y con fines de estudio, investigación, o cualquier otro fin lícito. Con dicha finalidad, la Universidad asume los siguientes deberes y se reserva las siguientes facultades:

- La Universidad informará a los usuarios del archivo sobre los usos permitidos, y no garantiza ni asume responsabilidad alguna por otras formas en que los usuarios hagan un uso posterior de las obras no conforme con la legislación vigente. El uso posterior, más allá de la copia privada, requerirá que se cite la fuente y se reconozca la autoría, que no se obtenga beneficio comercial, y que no se realicen obras derivadas.
- La Universidad no revisará el contenido de las obras, que en todo caso permanecerá bajo la responsabilidad exclusiva del autor y no estará obligada a ejercitar acciones legales en nombre del autor en el supuesto de infracciones a derechos de propiedad intelectual derivados del depósito y archivo de las obras. El autor renuncia a cualquier reclamación frente a la Universidad por las formas no ajustadas a la legislación vigente en que los usuarios hagan uso de las obras.
- La Universidad adoptará las medidas necesarias para la preservación de la obra en un futuro.

- La Universidad se reserva la facultad de retirar la obra, previa notificación al autor, en supuestos suficientemente justificados, o en caso de reclamaciones de terceros.

Madrid, a ...06...deenero.....de .2025.

ACEPTA

Fdo.:  Firmado por Tomás Alcántara Carrasco el día 6 de enero de 2025 con un certificado emitido por FNMT.

Motivos para solicitar el acceso restringido, cerrado o embargado del trabajo en el Repositorio Institucional:



ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA (ICAI)

Master in Big Data

Final Master's Thesis

**Business processes automation with artificial
intelligence**

Author

Tomás Alcántara Carrasco

Directed by

Jorge Gómez Berenguer

Madrid

January 2025

Resumen

En el ámbito empresarial actual, los consultores de venture capital, dedican una cantidad considerable de tiempo a la elaboración de informes de mercado. Esta tarea puede ser muy tediosa y demandante. Bayesian_x, una empresa matriz, cuenta con Research_x, una empresa asociada que se especializa en la creación de estos informes para asesorar a fondos de capital de riesgo. Research_x emplea técnicas avanzadas de análisis de datos e inteligencia artificial para proporcionar información valiosa a sus clientes, ayudándoles a tomar decisiones informadas y estratégicas.

La automatización de procesos empresariales mediante inteligencia artificial representa un avance significativo en la optimización de recursos y en la mejora de la precisión y rapidez en la generación de informes de mercado.

El objetivo principal de este Trabajo Fin de Máster es desarrollar un motor de automatización tecnológica que utilice inteligencia artificial para automatizar en su totalidad los procesos involucrados en la creación de estos informes de mercado. Esta automatización permitirá reducir significativamente el tiempo necesario para elaborar los informes, lo que no solo mejorará la eficiencia operativa, sino que también incrementará la ventaja competitiva de nuestra empresa.

Palabras clave: inteligencia artificial, automatización de procesos empresariales, informes de mercado, eficiencia operativa.

Abstract

In today's business environment, venture capital consultants spend a considerable amount of time preparing market reports. This task can be very tedious and demanding. Bayesian_x, a parent company, has an associated company, Research_x, which specializes in the creation of these reports to advise venture capital funds. Research_x employs advanced data analysis techniques and artificial intelligence to provide valuable information to its clients, helping them make informed and strategic decisions.

The automation of business processes through artificial intelligence represents a significant advance in resource optimization and in improving the accuracy and speed of market report generation.

The main objective of this Master's Thesis is to develop a technological automation engine that uses artificial intelligence to fully automate the processes involved in the creation of these market reports. This automation will significantly reduce the time required to prepare the reports, which will not only improve operational efficiency but also increase the competitive advantage of our company.

Keywords: artificial intelligence, business process automation, market reports, operational efficiency.

“Dei Fortioribus Adsunt.”

Agradecimientos

Queridos profesores, familiares y amigos:

Al llegar al final de este camino académico, quiero expresar mi más profundo agradecimiento a todos los que han sido esenciales para la realización de mi Trabajo de Fin de Máster.

A los profesores, gracias por su dedicación y compromiso. Su experiencia y pasión por la enseñanza han sido una fuente de inspiración. Cada consejo y palabra de ánimo han sido invaluable para mi crecimiento académico y personal.

A mi familia, mi pilar fundamental, no tengo palabras suficientes para expresar mi gratitud. Desde el primer día han estado junto a mí, apoyándome en cada paso. Han sido mi mayor fuente de motivación y fortaleza. Gracias por su amor incondicional y paciencia infinita. Este logro no habría sido posible sin su constante apoyo y sacrificio.

A mis amigos, gracias por su compañía, a las risas compartidas y a los momentos de desconexión que me ayudaron a recargar energías. Sus palabras de aliento y amistad sincera han sido un bálsamo en los momentos de mayor estrés y me han recordado la importancia de disfrutar el camino. Este año en Madrid será inolvidable, sin importar lo que el futuro depare.

A Marta, por ser mi apoyo constante y estar siempre a mi lado. Tu amor y comprensión han sido esenciales para mantenerme enfocado y motivado. Gracias por ser mi luz en los días más difíciles.

Y finalmente, a mí mismo, gracias por la valentía de enfrentar este desafío, por superar mis propias dudas y perseverar incluso cuando parecía imposible. Gracias por no dejar de creer en mí y recordarme que puedo lograrlo. Este logro es una prueba tangible de mi fuerza y determinación, y estoy verdaderamente orgulloso de lo que he conseguido.

Este TFM no es solo el resultado de mi esfuerzo individual, sino del apoyo de muchas personas que han dejado una huella imborrable en mi vida. Gracias por ser parte de mi historia y por hacer de este momento un recuerdo inolvidable. Espero que estos agradecimientos reflejen mi profundo aprecio y gratitud hacia cada uno de ustedes.

Con eterno agradecimiento,

Tomás.

Contents

1 Introduction	1
1.1 Problem description and motivation	1
1.2 Objectives	2
1.3 Document structure	2
1.4 Bayesian _x and Research _x	3
2 Theoretical basis and state of the art	5
2.1 Introduction to artificial intelligence and machine learning	5
2.1.1 Definitions and basic concepts	5
2.1.2 History and evolution	9
2.2 Advances in deep learning and generative AI	13
2.2.1 Deep neural networks	13
2.2.2 Generative text AI and LLMs	16
2.3 Applications of AI in market studies	18
2.3.1 Automated data collection	18
2.3.2 Market analysis	19
2.3.3 Automated report generation	21
2.4 State of the art in business process automation	22
2.4.1 Current and future trends	22
2.4.2 Case studies and real applications	22
2.4.3 State of the art competitors	23
3 Design of the tool	25
3.1 Introduction	25
3.2 Overview of the tool	26
3.2.1 Layer 0: Raw data inputs	26
3.2.2 Layer 1: Processed data inputs	27
3.2.3 Layer 2: Processing	28
3.2.4 Layer 3: Outputs	28
3.2.5 Integration of technologies	29
3.3 Architecture diagram of the tool	30
3.4 Main components	30
3.4.1 Layer 0: Raw data inputs	30
3.4.2 Layer 1: Processed data inputs	31
3.4.3 Layer 2: Processing	45
3.4.4 Layer 3: Outputs	46

4 Results and discussion	47
4.1 Layer 1: Processed data inputs	47
4.1.1 Results	47
4.1.2 Discussion	58
4.2 Layer 2: Processing	59
4.2.1 Results	59
4.2.2 Discussion	61
4.3 Layer 3: Outputs	62
4.3.1 Results	62
4.3.2 Discussion	63
5 Metrics and evaluation of results	65
5.1 Introduction	65
5.2 Metrics and evaluation for generative AI models	65
5.2.1 Logprobs and Perplexity	65
5.2.2 BLEU and ROUGE metrics	68
5.2.3 Semantic Accuracy	72
5.2.4 Ground Truth	74
5.2.5 Global KPIs for generative AI	74
5.3 Metrics and evaluation for XGBoost models	74
5.3.1 Introduction to XGBoost metrics	74
5.3.2 Key metrics for regression models	75
5.3.3 Ground Truth	76
5.3.4 Global KPIs	76
6 Future lines and conclusions	77
6.1 Future lines	77
6.2 Conclusions	77
Appendix	80
A Additional information	81
A.1 First ideas and theories in the field of artificial intelligence	81
A.2 Vector databases and embeddings: a primer on Pinecone	83
A.2.1 What are embeddings?	83
A.2.2 How Pinecone utilizes embeddings?	84
A.2.3 Visual representation	84
B Gantt diagram	85
Bibliografia	87

List of Figures

2.1	Overview of areas of artificial intelligence. Adapted from reference [44].	5
2.2	Emergence of LLMs as an intersection between NLP and DL. Adapted from reference [25].	8
2.3	Diagram illustrating the relationship between Text Generation, Deep Learning, Generative AI, LLMs, and ChatGPT.	9
2.4	Timeline of key events in the history of AI.	11
2.5	Comparison of images generated by MidJourney in April 2022 (up) and in April 2023 (down). Adapted from reference [1].	11
2.6	First example of an image of the Midjourney model. Adapted from reference [2].	12
2.7	Second example of an image of the Midjourney model. Adapted from reference [3].	12
2.8	Architecture of a CNN. Adapted from reference [4].	14
2.9	Architecture of a RNN. Adapted from reference [5].	14
2.10	Architecture of an Autoencoder. Adapted from reference [6].	15
2.11	Architecture of a GAN. Adapted from reference [7].	15
2.12	Venn diagram illustrating the market estimation Bottom-Up approach with identifying products, services and finance metrics.	20
3.1	Overview of the tool's architecture.	26
3.2	Detailed diagram with tool parts.	30
3.3	Venn diagram illustrating that the union of quantitative inputs and quality inputs form the core of the content of market reports	32
3.4	Actual diagram of the first workflow.	38
3.5	Actual diagram of the second workflow.	40
3.6	Actual diagram of the third workflow.	42
3.7	Actual diagram of the fourth workflow.	45
3.8	Actual diagram of the chaining workflow.	46
4.1	Eurocebollas website with unstructured information.	52
4.2	Legalitas website with unstructured information.	55
4.3	Universal Music website with unstructured information.	57
4.4	One of the slides of the Legálitas market report	62
4.5	Slides of Eurocebollas market reports.	62
4.6	Slides of Universal Music market reports.	63
5.1	Logprobs and Perplexity representation.	67
5.2	Illustration of cosine similarity in 2D. The angle θ determines the similarity between the vectors \vec{A} and \vec{B} .	73

A.1	From left to right: a diagram suggesting how the eyes might transmit a unified picture of reality to the brain, a Purkinje neuron in the human cerebellum, and a diagram showing the flow of information through the hippocampus. Diagram by S. Ramón y Cajal. Adapted from reference ^[8]	. . . 81
A.2	Comparison between a biological neuron and an artificial neuron. Adapted from reference ^[9] 82
A.3	How embedding works. Adapted from reference ^[10] 83
A.4	Visual representation of an embedding. Adapted from reference ^[11] 84
B.1	Gantt chart for TFM project timeline. 85

List of Tables

4.1	Company data: employee numbers and revenue. Red cells with question marks indicate missing values.	48
4.2	Company data: employee numbers and revenue. Green cells indicate previously missing values now updated.	49
4.3	Company data: financial metrics. Red cells with question marks indicate missing values.	50
4.4	Company data: financial metrics. Green cells indicate previously missing values now updated.	51

Listings

3.1	Example of detailed computer products in Amazon	34
4.1	Product details for Cooked Onion.	53
4.2	Service details for Legal Support Services.	56
4.3	Service details for Music and Entertainment Services.	57

Acronyms

<i>ICAI</i>	Instituto Católico de Artes e Industrias
<i>FMT</i>	Final Master's Thesis
<i>AI</i>	Artificial Intelligence
<i>LLM</i>	Large Language Model
<i>MECE</i>	Mutually Exclusive Collectively Exhaustive
<i>ML</i>	Machine Learning
<i>DL</i>	Deep Learning
<i>NLP</i>	Natural Language Processing
<i>CNN</i>	Convolutional Neural Network
<i>GPT</i>	Generative Pre-trained Transformer
<i>NER</i>	Named Entity Recognition
<i>NMT</i>	Neural Machine Translation
<i>RL</i>	Reinforcement Learning
<i>RNN</i>	Recurrent Neural Network
<i>SMT</i>	Statistical Machine Translation
<i>SVM</i>	Support Vector Machine
<i>DNN</i>	Deep Neural Network
<i>CKS</i>	Central Knowledge System
<i>BPA</i>	Bussiness Process Automation
<i>IoT</i>	Internet of Things
<i>RPA</i>	Robotic Process Automation
<i>IPA</i>	Intelligent Process Automation
<i>REST</i>	Representational State Transfer
<i>IPA</i>	Application Programming Interface
<i>USPs</i>	Unique Selling Propositions

Symbols

z	Weighted sum (2.1)
w_i	Weights (2.1)
x_i	Wnputs value (2.1)
b	Bias term (2.1)
L	Loss function (2.2)
W	Weights (2.2)
a	Activation (2.2)
W_{new}	Updated weights (2.2)
W_{old}	Current weights (2.2)
η	Learning rate (2.2)

Chapter 1

Introduction

1.1 Problem description and motivation

Currently, one of the main challenges faced by venture capital consultants is the considerable amount of time they must invest in conducting market studies. This process, crucial for making informed decisions, can be extremely labor-intensive and consume valuable resources. In response to this problem, Bayesian_x has emerged as an innovative company dedicated to producing market reports in an automated manner, enabled by artificial intelligence (AI). Bayesian_x aims not only to accelerate the process of data collection and analysis but also to improve the precision and depth of the insights obtained.

Artificial intelligence has gained significant prominence recently, primarily due to substantial advances in the field of deep learning. These advances, driven by factors such as having increased data processing capabilities and the availability of large volumes of information, have enabled the development of more sophisticated and efficient algorithms. In particular, generative AI and natural language processing (NLP) have shown considerable potential in various applications, from content creation to understanding and generating human language in an increasingly natural and precise manner.

This technological context has opened up a new window of business opportunities. Numerous companies are exploring and leveraging these emerging technologies to automate processes that previously required intensive human intervention. Automation through AI is redefining industries by enabling the execution of complex tasks more quickly and at a lower cost, leading to a significant increase in operational efficiency. These improvements not only reduce expenses and the time invested in key processes but also provide a competitive advantage by allowing companies to focus on strategic and higher-value activities.

This study addresses the automation of business processes oriented towards market analysis, market studies, and the preparation of market reports using artificial intelligence, with the objective of reducing time and operational costs. By implementing these technologies, businesses can enhance their operational efficiency, thereby gaining a competitive edge. The specific processes to be automated will be detailed in chapter [3](#), providing a clear and structured overview of the application areas and expected benefits.

1.2 Objectives

- **Main objective:**

- Implement technologies/software that allow the automation of market report creation, consequently reducing the time and resources needed to produce these reports.

- **Specific objectives:**

- Implement automated collection engines to capture characteristics of **products** and **services** associated with companies.
- Integrate and analyze detailed financial information (revenue, net profit, cash flow, total assets, and shareholder's equity, ...), building a database with more than 200 financial variables from others and performing interpolation and prediction models with XGBoost and algebra.
- Data processing through the development of an efficient workflow, called *chain-ing*.
- Generate automated reports efficiently (*text-to-slide*), reducing the time and resources needed for their production. Slides are automatically populated with the information received from the processing.

1.3 Document structure

- **Chapter 1: Introduction**

This chapter provides the context of the project, describing the problem, objectives, document structure, and the theoretical and methodological basis that underpin the research.

- **Chapter 2: Theoretical basis and state of the art**

This chapter offers a comprehensive review of the existing literature, covering basic concepts, recent advancements, and practical applications in artificial intelligence and machine learning.

- **Chapter 3: Design of the tool**

Here, the design of the developed tool is detailed, including its architecture, main components, and how various technologies are integrated.

- **Chapter 4: Results and discussion**

This chapter presents the results obtained after implementing the tool, followed by a discussion on their relevance, implications, and comparison with the state of the art.

- **Chapter 5: Future work and conclusions**

The final chapter reflects on the work done, summarizing the conclusions reached and proposing possible future research directions to continue and expand the project.

1.4 Bayesian_x and Research_x

Bayesian_x is an innovative company specializing in the automation of business processes through artificial intelligence (AI). Their mission is to transform operational efficiency and deliver significant value across various industries. Combining management consulting, machine learning engineering, and data architecture, Bayesian_x develops solutions to optimize workflows and service models.

Research_x, an associated company, acts as an on-demand research and analytics back-office, supporting private equity and corporate clients with actionable market intelligence for strategic decision-making.

My role in the company

As an AI Engineer at Bayesian_x, my work involves implementing advanced technologies such as deep learning and natural language processing (NLP). These technologies, enabled by increased data processing capabilities and large data volumes, help automate complex tasks, reduce costs and time, and allow the company to focus on higher-value strategic activities.

For more information, visit: [Bayesian_x](#).

Chapter 2

Theoretical basis and state of the art

2.1 Introduction to artificial intelligence and machine learning

The project is situated within the field of Natural Language Processing (NLP), a branch of Artificial Intelligence (AI) that specializes in the processing of unstructured text data. While a comprehensive review of AI fundamentals has been conducted, the primary focus consistently remains on NLP and deep learning, particularly the emergence and development of large language models (LLMs).

2.1.1 Definitions and basic concepts

Artificial Intelligence (AI) is the branch of computer science dedicated to creating systems capable of performing tasks that typically require human intelligence. These tasks include learning, reasoning, problem-solving, perception, and language understanding. AI aims to replicate or simulate human cognitive functions, enabling machines to perform complex functions autonomously.

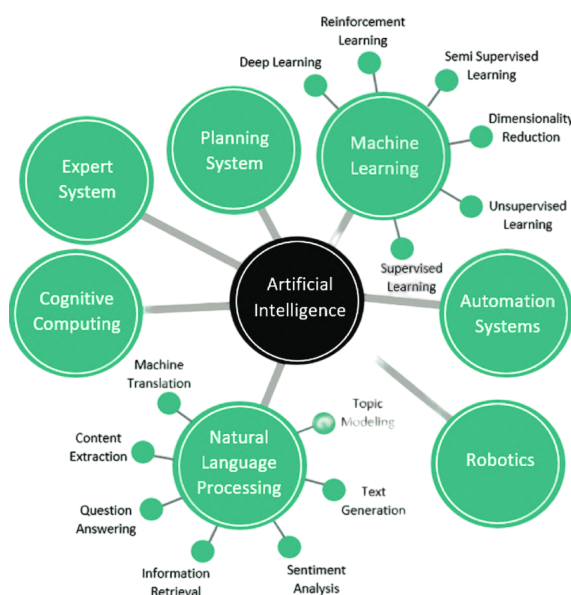


Figure 2.1: Overview of areas of artificial intelligence. Adapted from reference [44].

This branch of computer science encompasses various areas, as shown in Figure 2.1. Below, we will focus on some of the most important areas, from which generative AI will emerge. First, we provide a general description of these areas. Finally, we will delve into generative AI, which is the primary technology we use for this project. These areas are divided into three main categories (Machine Learning, Deep Learning, and Natural Language Processing). Generative AI and the large language models (LLMs) subfield¹ emerge as an intersection of these categories.

Machine Learning (ML) is a subset of AI that focuses on developing systems that can learn from and make decisions based on data. ML algorithms build models from sample data, known as training data, to make predictions or decisions without being explicitly programmed to perform the task. We will explain some of the most important areas below:

- **Supervised Learning:** in supervised learning, the model is trained on a labeled dataset, meaning each training example is paired with an output label. The model learns a mapping from inputs to outputs, which can be used to predict the labels of new, unseen data. Common applications of supervised learning include classification and regression tasks. Classification involves predicting discrete labels, such as identifying spam emails, while regression involves predicting continuous values, such as forecasting stock prices. [11].
- **Unsupervised Learning:** in unsupervised learning, the model is trained on data that has no labels. The objective is to identify hidden patterns or intrinsic structures in the input data. Techniques include clustering and association. Clustering algorithms, such as K-means and hierarchical clustering, group data points with similar characteristics, which is useful in customer segmentation and image compression. Association algorithms, such as the Apriori algorithm, find rules that describe large portions of the data, often used in market basket analysis. [27].
- **Reinforcement Learning:** reinforcement learning (RL) is a type of machine learning where an agent learns to make decisions by performing actions in an environment to maximize cumulative reward. The agent learns from the consequences of its actions rather than from being told what to do by a teacher. This approach is particularly effective in scenarios where the optimal solution is not immediately apparent, such as in game playing, robotics, and autonomous driving. [37].
- **Deep Learning (DL)** is also a subset of machine learning that uses neural networks with many layers (hence "deep") to model complex patterns in data. It has revolutionized fields such as computer vision and natural language processing (NLP). For example, deep learning models like convolutional neural networks (CNNs) excel at image recognition tasks, while recurrent neural networks (RNNs) and transformers are highly effective for sequential data such as text. One of the most significant differences between machine learning and deep learning is that while machine learning requires more human intervention to select and design relevant features, deep learning has less need for feature engineering, as it can learn representations directly from raw data.

¹<https://explodingtopics.com/blog/llms-vs-generative-ai>

Natural Language Processing (NLP) is a subfield of AI and computational linguistics that focuses on the interaction between computers and humans through natural language. The goal of NLP is to enable computers to understand, interpret, and generate human language in a way that is both meaningful and useful. NLP encompasses a range of tasks, each with specific applications and techniques, including:

- **Text Classification:** Text classification involves assigning predefined categories to text data. This is a fundamental task in NLP with applications such as spam detection in emails, sentiment analysis in social media, and topic labeling in news articles. Traditional approaches to text classification include methods like Naive Bayes and Support Vector Machines (SVM). However, more advanced techniques, such as deep learning models using recurrent neural networks (RNNs) and transformers, have significantly improved performance. These models can capture complex patterns and dependencies in text, making them highly effective for classification tasks. [33].
- **Named Entity Recognition (NER):** NER is the process of identifying and classifying entities within a text into predefined categories, such as names of people, organizations, locations, dates, and more. This task is essential for information extraction, enabling applications like search engines, recommendation systems, and automated customer service to function more effectively. NER systems typically use a combination of rule-based approaches and machine learning techniques to achieve high accuracy. [41].
- **Machine Translation:** Machine translation involves automatically converting text from one language to another. Early approaches relied heavily on rule-based systems, which were limited in their ability to handle the complexities of human language. Modern systems, however, use statistical machine translation (SMT) and neural machine translation (NMT). NMT models, particularly those based on the Transformer architecture, such as Google's Transformer, have revolutionized the field by providing more accurate and fluent translations. [31].
- **Speech Recognition:** Speech recognition technology converts spoken language into text. This technology is used in various applications, including virtual assistants (e.g., Siri, Alexa), transcription services, and hands-free control systems. Deep learning models, especially convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have significantly enhanced the accuracy and robustness of speech recognition systems. These models can effectively handle the variability and nuances of human speech. [17].
- **Question Answering:** Question answering systems aim to automatically answer questions posed by humans in natural language. These systems are integral to the functionality of chatbots, virtual assistants, and customer support platforms. Advanced models such as BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer) have achieved state-of-the-art performance in this area. They can understand and generate contextually relevant responses, making interactions more natural and efficient. [21].
- **Text Generation:** Text generation involves creating coherent and contextually relevant text based on a given input. This task has applications in content creation, text summarization, and dialogue generation. Generative models like GPT-3 have

demonstrated remarkable capabilities in producing human-like text across various domains. These models can generate news articles, creative writing, and even code snippets, showcasing the versatility of NLP technologies. [12].

Generative AI involves models that can generate new data samples similar to the training data. This includes techniques such as Generative Adversarial Networks (GANs) and variational autoencoders. These models have applications in creating realistic images, videos, and even synthetic data for training other AI models.

Generative text AI and Large Language Models (LLMs) arise from the intersection of Natural Language Processing (NLP) and Deep Learning. Generative text AI leverages NLP techniques to understand and generate text, while utilizing deep learning through the use of deep neural networks to learn and generate complex text data. Generative AI, in a broader sense, uses deep learning to generate complex data, including images and audio.

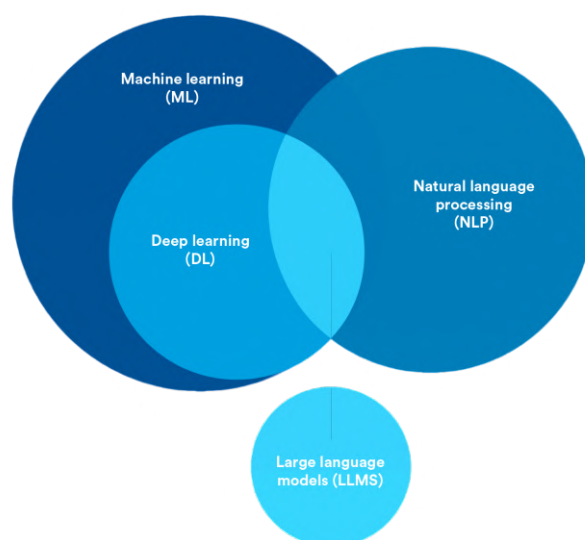


Figure 2.2: Emergence of LLMs as an intersection between NLP and DL. Adapted from reference [25].

Within the realm of Generative text AI, LLMs, like GPT, represent a specific application focused on language-related tasks, such as translation, summarization, and text generation. These models employ deep learning architectures, such as transformers, to process vast amounts of textual data and learn complex patterns. LLMs are pre-trained on extensive corpora of text data to understand the nuances of human language, enabling them to generate coherent and contextually relevant text [12, 22].

It is important to emphasise that, while Generative AI encompasses a broad range of data generation capabilities, including text, images, and audio, LLMs are specifically designed for language-related tasks. They combine NLP and deep learning techniques to manage and interpret large datasets of textual information, showcasing advanced capabilities in text generation and understanding.

Finally to stress the difference between GPT and ChatGPT. GPT (Generative Pre-trained Transformer) is a general-purpose large language model pre-trained on a vast

corpus of text [12]. ChatGPT, on the other hand, is a fine-tuned version of GPT, specifically adapted for conversational contexts. This fine-tuning process is known as transfer learning.

Transfer learning involves taking a pre-trained model and adapting it to a specific task by training it further on a smaller, task-specific dataset. This approach leverages the general knowledge learned during the initial pre-training phase and refines it to improve performance on specialized tasks [45]. In the case of ChatGPT, transfer learning allows the model to generate more coherent and relevant responses in a conversational setting by fine-tuning GPT on dialogue-specific data.

Finally, the final and more specific Venn diagram with the last mentioned is attached in figure 2.3.

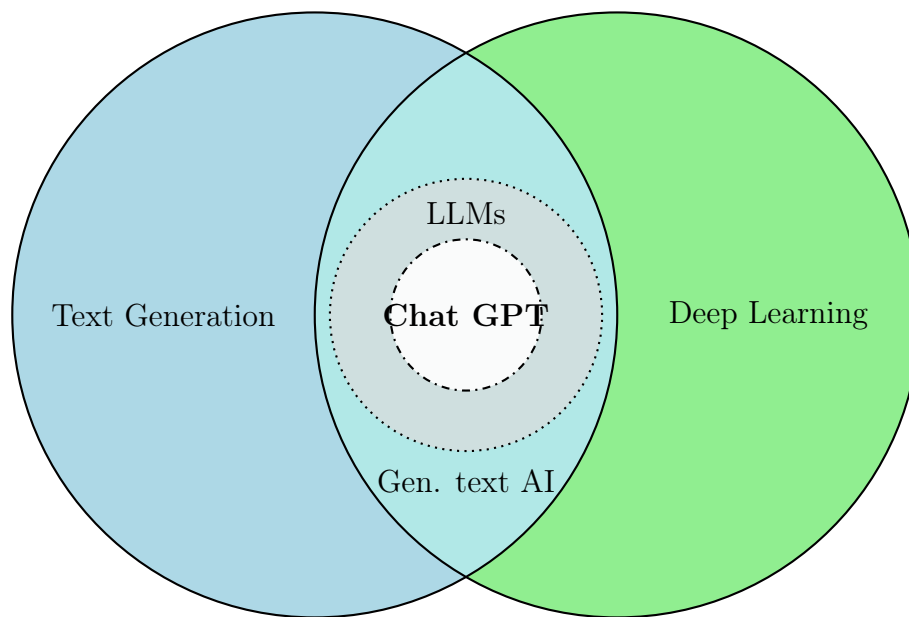


Figure 2.3: Diagram illustrating the relationship between Text Generation, Deep Learning, Generative AI, LLMs, and ChatGPT.

2.1.2 History and evolution

The origins of AI can be traced back to ancient times, with myths and legends about automatons and artificial beings created by gods. Philosophers and inventors throughout history have pondered the idea of creating machines that could mimic human intelligence. However, the formal development of AI as a scientific discipline began in the mid-20th century.

In the 1940s and 1950s, the groundwork for AI was laid with the development of digital computers, which were capable of performing calculations much faster than human beings. During this period, Alan Turing, a British mathematician and logician, made significant contributions to the field of computer science. In his seminal 1950 paper "Computing Machinery and Intelligence", Turing proposed the idea of a machine that could think and introduced the famous "Turing Test" as a criterion for machine intelligence. The Turing

Test involves a human evaluator who interacts with both a machine and a human without knowing which is which. If the evaluator cannot reliably distinguish the machine from the human based on their responses, the machine is said to have demonstrated intelligent behavior [49].

The term "artificial intelligence" was coined in 1956 during the Dartmouth Conference, organized by John McCarthy, Marvin Minsky, Nathaniel Rochester, and Claude Shannon. This conference is considered the official birth of AI as a field of study. The attendees of the conference proposed that "every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it". This ambitious statement set the stage for future research and development in AI [34].

Following the Dartmouth Conference, the field of AI experienced rapid growth. Researchers developed early AI programs such as the Logic Theorist, which was designed to mimic the problem-solving skills of a human, and the General Problem Solver, which aimed to solve a wide range of problems using heuristic search techniques. During this period, the concept of artificial neural networks also emerged, with Frank Rosenblatt's development of the perceptron in 1958. The perceptron was an early attempt to create a machine that could learn from experience in a manner analogous to human learning.

Despite these early successes, AI research faced several challenges in the subsequent decades. The limitations of early neural networks, coupled with a lack of computational power and data, led to periods of reduced funding and interest, often referred to as "AI winters." However, advances in computer hardware, particularly the development of powerful graphics processing units (GPUs), and new algorithms revitalized the field in the late 20th and early 21st centuries.

In recent years, AI has advanced rapidly, particularly in the field of natural language processing. One of the most notable developments is the creation of Large Language Models (LLMs) like OpenAI's GPT (Generative Pre-trained Transformer) series. The latest iteration, GPT-3, and its fine-tuned counterpart, ChatGPT, have demonstrated unprecedented capabilities in understanding and generating human-like text. This progress is largely due to significant improvements in computational power, the availability of massive datasets for training, and advancements in deep learning architectures, such as transformers [22].

ChatGPT, in particular, represents a breakthrough in conversational AI. By leveraging transfer learning, where a pre-trained model is fine-tuned on a smaller, task-specific dataset, ChatGPT can generate coherent and contextually relevant responses in real-time conversations. This ability has made it a valuable tool in various applications, from customer service to personal assistants and beyond [12, 43].

The advancements in AI over the past few years highlight the transformative potential of combining powerful algorithms, extensive training data, and sophisticated hardware. As AI continues to evolve, it holds the promise of further enhancing human capabilities and revolutionizing industries across the globe [45].

Timeline of key events in AI history

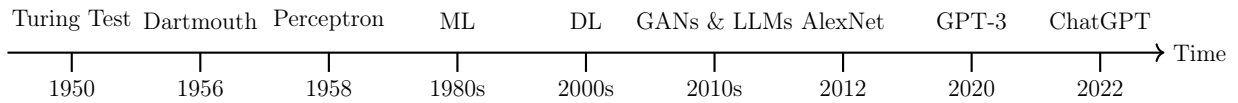


Figure 2.4: Timeline of key events in the history of AI.

Generative AI has also shown remarkable progress over the past few years, particularly in the quality and realism of images it can produce. MidJourney, an AI-based image generation platform, is a prime example of this evolution. Comparing images generated by MidJourney a few years ago with those generated recently highlights the advancements in generative AI technologies.



Figure 2.5: Comparison of images generated by MidJourney in April 2022 (up) and in April 2023 (down). Adapted from reference ²

These images illustrate the significant improvements in texture, detail, and overall realism, showcasing the potential of generative AI to create high-quality visual content.

Below are some examples of pictures of the current Midjourney model with its prompts.

²<https://hipertextual.com/2023/04/asi-avanzado-inteligencia-artificial-generadora-imagenes-primer-ano-vida>

Prompt

“A realistic photograph of a Dutch classroom with 4-year-old children playing with blocks. At the front of the class, there’s a girl with brown hair, brown eyes, light skin, and a black t-shirt. The scene should be colorful with beautiful lighting, and the classroom should not look too crowded. The children are wearing different clothes, but the focus is not on them; instead, it’s on capturing the natural, lively atmosphere of the classroom as they play with blocks.”



Figure 2.6: First example of an image of the Midjourney model. Adapted from reference ³

Prompt

“Sci-fi surrealism pancakes paper cutout, Alice in Wonderland themed, cutout using pancakes instead of paper.”



Figure 2.7: Second example of an image of the Midjourney model. Adapted from reference ⁴

³<https://www.midjourney.com/jobs/2028d4f1-5bb0-430c-b148-41367b7aef59?index=0>

⁴<https://www.midjourney.com/jobs/73968b5f-d735-4631-85a2-8e992330cc02?index=0>

2.2 Advances in deep learning and generative AI

2.2.1 Deep neural networks

Deep Neural Networks (DNNs) are a class of machine learning models composed of multiple layers of neurons, each layer transforming the input data in increasingly complex ways. These networks are designed to automatically learn features and patterns from large amounts of data, making them exceptionally powerful for tasks such as image and speech recognition, natural language processing, and more.

Structure and functioning

A typical deep neural network consists of an input layer, multiple hidden layers, and an output layer. Each layer contains nodes (or neurons) that are connected by edges with associated weights. The primary operations within a neural network include:

- **Weighted sums:** each neuron calculates a weighted sum of its inputs.

$$z = \sum_{i=1}^n w_i x_i + b \quad (2.1)$$

where z is the weighted sum, w_i are the weights, x_i are the input values, and b is the bias term.

- **Activation functions:** the weighted sum is passed through an activation function (such as ReLU, sigmoid, or tanh) to introduce non-linearity into the model, enabling it to learn more complex patterns.

- **ReLU (Rectified Linear Unit):** $f(z) = \max(0, z)$

- **Sigmoid:** $f(z) = \frac{1}{1+e^{-z}}$

- **Tanh:** $f(z) = \tanh(z)$

- **Backpropagation:** the network adjusts its weights based on the error of its predictions through a process called backpropagation, which involves computing gradients of the loss function with respect to each weight.

$$\frac{\partial L}{\partial W} = \frac{\partial L}{\partial a} \cdot \frac{\partial a}{\partial z} \cdot \frac{\partial z}{\partial W} \quad (2.2)$$

where L is the loss function, W are the weights, a is the activation, and z is the weighted sum.

The weights are then updated in the opposite direction of the gradient:

$$W_{\text{new}} = W_{\text{old}} - \eta \frac{\partial L}{\partial W} \quad (2.3)$$

where W_{new} are the updated weights, W_{old} are the current weights, η is the learning rate, and $\frac{\partial L}{\partial W}$ is the gradient of the loss function with respect to the weights.

Types of deep neural networks

These are the main types of deep neural networks tailored for different types of data and tasks:

- **Convolutional Neural Networks (CNNs):** designed for processing grid-like data such as images. CNNs use convolutional layers to automatically and adaptively learn spatial hierarchies of features. To see how they work it is recommended to visit this website, which provides you with interactive plots.

$$\text{CNN} \begin{cases} \text{2D version: } \text{Interactive plot for 2D CNN} \\ \text{3D version: } \text{Interactive plot for 3D CNN} \end{cases}$$

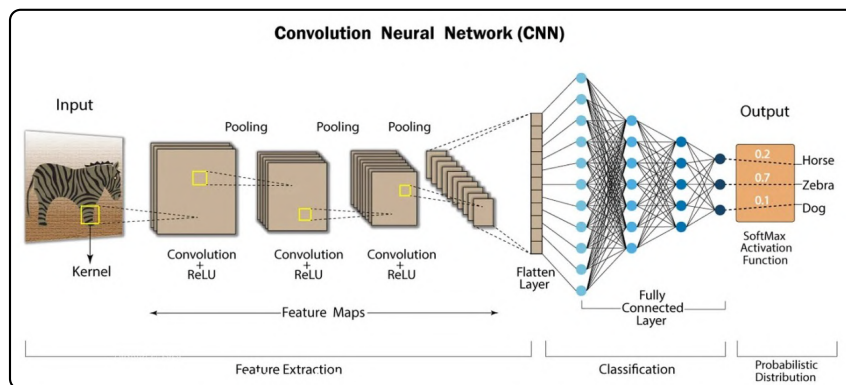


Figure 2.8: Architecture of a CNN. Adapted from reference ⁵

- **Recurrent Neural Networks (RNNs):** ideal for sequential data, such as time series or text. RNNs have connections that form directed cycles, allowing them to maintain a state that can capture information about previous elements in the sequence.

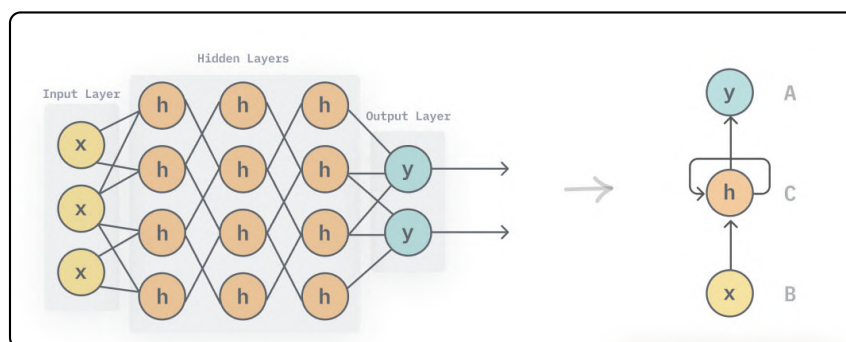


Figure 2.9: Architecture of a RNN. Adapted from reference ⁶

- **Autoencoders:** used for tasks such as dimensionality reduction and feature learning. They learn to encode the input data into a lower-dimensional representation and then decode it back to the original form.

⁵<https://www.linkedin.com/pulse/what-convolutional-neural-network-cnn-deep-learning-nafiz-shahriar/>

⁶<https://www.v7labs.com/blog/recurrent-neural-networks-guide>

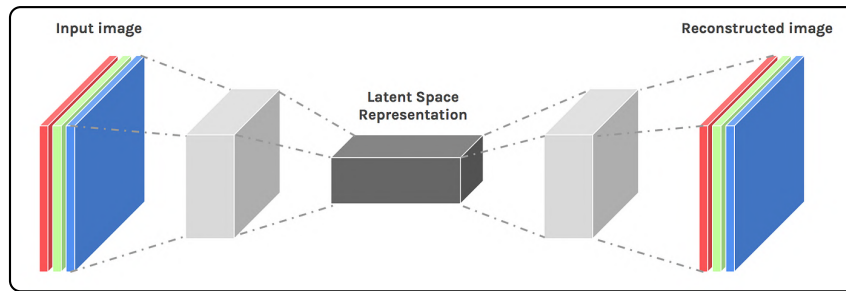


Figure 2.10: Architecture of an Autoencoder. Adapted from reference ⁷

- **Generative Adversarial Networks (GANs):** consist of two networks (a generator and a discriminator) that compete against each other to create realistic data samples. GANs are widely used for generating images, videos, and other types of data.

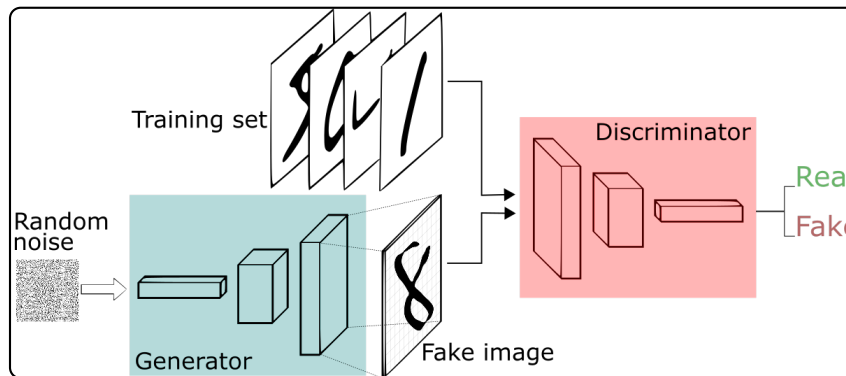


Figure 2.11: Architecture of a GAN. Adapted from reference ⁸

Applications of deep neural networks

Deep neural networks have been successfully applied in various domains due to their ability to model complex relationships in data. Some of the key applications include:

- **Computer vision:** DNNs, especially Convolutional Neural Networks (CNNs), have achieved state-of-the-art performance in tasks such as image classification, object detection, and image segmentation. Applications include facial recognition, autonomous driving, and medical image analysis ³.
- **Speech recognition:** DNNs have significantly improved the accuracy of speech recognition systems. These systems convert spoken language into text and are used in virtual assistants (e.g., Siri, Alexa), transcription services, and voice-controlled applications ¹⁷.

⁷<https://towardsdatascience.com/autoencoders-bits-and-bytes-of-deep-learning-eaba376f23ad>

⁸<https://sthalles.github.io/intro-to-gans/>

- **Natural Language Processing:** Recurrent Neural Networks (RNNs) and transformers are types of DNNs used extensively in NLP tasks such as language translation, sentiment analysis, and text generation. Models like BERT and GPT-3 have set new benchmarks in language understanding and generation [14,22]. It is within this framework that we will move.

- **Healthcare:** DNNs are used for predictive analytics, diagnosis, and personalized treatment recommendations. For example, they can analyze electronic health records to predict disease outbreaks or patient outcomes [13].
- **Finance:** in finance, DNNs are applied in algorithmic trading, credit scoring, fraud detection, and risk management. They analyze vast amounts of financial data to identify patterns and make predictions [19].
- **Gaming and Reinforcement Learning:** DNNs are used in reinforcement learning to create agents that can learn to play games at a superhuman level. Notable examples include AlphaGo and AlphaStar, which have defeated human champions in Go and StarCraft II, respectively [16,20].

Key models and algorithms

Some of the key models and algorithms in the realm of deep neural networks include:

- **AlexNet:** pioneered the use of deep convolutional networks for image classification, demonstrating the potential of deep learning on large-scale datasets.
- **VGGNet:** showcased the importance of network depth by using very deep networks with small convolutional filters.
- **ResNet:** introduced residual connections to solve the problem of vanishing gradients, allowing for the training of even deeper networks.

- **Transformers:** transformers have revolutionized natural language processing by employing self-attention mechanisms to effectively handle long-range dependencies in text data. Unlike traditional recurrent neural networks (RNNs), transformers process all elements of the input sequence simultaneously, allowing for greater parallelization and efficiency. The self-attention mechanism enables the model to weigh the importance of different words in a sentence, regardless of their position, leading to improved performance in tasks such as machine translation, text summarization, and language modeling. This architecture has been fundamental, for example, in GPT, BERT, T5 and XLNet models.

2.2.2 Generative text AI and LLMs

Generative models in NLP have progressed significantly, thanks to the advent of advanced architectures like transformers, as mentioned in [2.2.1]. Below are some of the most influential generative models:

GPT-4, developed by OpenAI, is one of the most advanced language models to date. It uses the transformer architecture to generate coherent and contextually relevant text. GPT-4 is pre-trained on a diverse range of internet text and can perform tasks such as translation, question-answering, and text completion with minimal fine-tuning [22].

BERT, developed by Google AI, revolutionized NLP by introducing bidirectional training of transformers, which looks at both the left and right context in all layers. This bidirectional approach enables BERT to understand the context of a word based on its surroundings, leading to significant improvements in tasks like text classification and question answering [18].

T5 (Text-to-Text Transfer Transformer), also developed by Google Research, treats every NLP problem as a text-to-text problem. This unified approach simplifies the task structure and allows T5 to achieve state-of-the-art results across a wide range of NLP tasks by framing them all as converting one text to another. This model has demonstrated impressive performance in translation, summarization, and question-answering tasks [15].

Advances in generative text AI techniques

The development of generative models has been driven by several key techniques and innovations:

Self-attention mechanisms introduced by transformers, allow models to weigh the importance of different words in a sequence, enabling them to capture long-range dependencies more effectively than traditional RNNs [14].

Transfer learning involves pre-training a model on a large dataset and then fine-tuning it on a smaller, task-specific dataset. This approach has proven to be highly effective in NLP, as demonstrated by models like GPT and BERT [15].

Benefits of transfer learning:

- **Efficiency:** reduces the need for large task-specific datasets.
- **Performance:** enhances model performance by leveraging knowledge from the pre-training phase.
- **Versatility:** enables models to be adapted to a wide range of tasks with minimal additional training.

Fine-Tuning and domain adaptation Fine-tuning involves adjusting a pre-trained model to perform specific tasks by training it on a task-specific dataset. Domain adaptation goes a step further by tailoring models to perform well in particular domains (e.g., medical text, legal documents) [15].

Future directions

The future of generative AI in NLP holds exciting possibilities. Areas of ongoing research include improving model efficiency, enhancing the interpretability of AI systems, and addressing ethical concerns related to AI-generated content [22].

- **Model efficiency:** researchers are exploring ways to make generative models more efficient, reducing the computational resources required for training and inference [15].
- **Interpretability:** understanding how generative models make decisions is crucial for building trust and ensuring the reliability of AI systems [22].
- **Ethics and bias:** addressing biases in AI-generated content and ensuring ethical use of generative models are key priorities for the field [18].

2.3 Applications of AI in market studies

Tools and techniques squared in black are the ones we have used in the project and will be explained in detail in the section 3.

2.3.1 Automated data collection

Automated data collection refers to the use of technology to gather information without manual intervention. This process is crucial for efficiently obtaining large volumes of data from various sources, enabling timely and accurate analysis. In the context of market studies, automated data collection helps in aggregating data from different segments, such as sales records, customer feedback, financial reports, and competitor analysis. Here, we discuss the general paradigms and tools used for automated data collection [7, 10, 23, 30, 36, 38].

General paradigms

Web scraping: this involves extracting data from websites. Tools like BeautifulSoup, Scrapy, and Selenium are commonly used for this purpose. Web scraping scripts can automatically navigate websites, extract relevant information, and store it in a structured format [8, 9, 30]. This is the one of the ways that we have used to collect data.

APIs (Application Programming Interfaces): many websites and services provide APIs that allow automated access to their data. APIs are particularly useful for accessing real-time data. Examples include financial market APIs, social media APIs, and e-commerce APIs. This is the other way that we have used to collect data.

Sensor data collection: in industries such as manufacturing and logistics, sensors are used to collect real-time data on various parameters. IoT (Internet of Things) devices can automatically gather and transmit data, which can then be processed and analyzed.

Automated surveys and feedback forms: platforms like Google Forms, SurveyMonkey, and Typeform allow for the automated collection of survey data. These tools can send surveys to target audiences, collect responses, and compile the data for analysis.

Tools and techniques

Web scraping tools

- **BeautifulSoup:** a Python library for parsing HTML and XML documents. It creates parse trees that are helpful for extracting data from HTML [\[30\]](#).
- **Scrapy:** an open-source and collaborative web crawling framework for Python. It is used to extract data from websites and process them as per the needs [\[8\]](#).

- **Selenium:** a portable framework for testing web applications. It is also used for web scraping when interaction with JavaScript-loaded content is required [\[9\]](#).

API integration

- **REST APIs:** representational State Transfer APIs are commonly used for web services that allow for interaction with RESTful web services.
- **GraphQL APIs:** these provide a more flexible alternative to REST, allowing clients to request exactly the data they need.

Internally developed tools

- **Researcher_x:** is a tool designed to address the limitation of Large Language Models (LLMs) in providing transparent data sources. It performs semantic searches on Google using keywords to retrieve valuable information from the internet, including links to relevant PDF and HTML documents. This information is then parsed using content analysis technology, making Researcher_x essential for projects that require efficient data collection and processing from large online datasets.
- **Central Knowledge System (CKS):** is an advanced RAG (Retrieval-Augmented Generation) system integrated with Pinecone, where financial information is stored. This information is gathered through various workflows and the researcher_x tool. This setup allows for efficient retrieval and analysis of large datasets, facilitating comprehensive financial insights and decision-making.
- **Workflows for products and services:** different workflows with the n8n tool for retrieving information from websites.

2.3.2 Market analysis

The application of AI algorithms in financial analysis has revolutionized the way market studies and financial insights are conducted. Two primary approaches used in market analysis are top-down and bottom-up methods, each providing unique perspectives and advantages.

Top-Down approach

Starts with the broader macroeconomic environment and works down to the individual companies. This method is often demand-oriented, focusing on the perspective of consumers and overall market demand. Analysts begin by examining global economic trends, industry performance, and sector health before drilling down to individual companies. This approach helps in understanding the larger market dynamics and identifying the sectors with growth potential [18].

Bottom-Up approach

This is the approach we will take. Is more robust for conducting detailed market research as it starts from the micro level, focusing on individual companies and their offerings. This method is supply-oriented, looking at the perspective of companies providing products and services. The primary steps involved include:

- Identifying market constituents: determine the companies that form the market by identifying those producing or distributing products and services of types A, B, and C.
- Product and service logic: develop a logic to understand which products and services each company uses. Cross-reference this information with financial data to estimate the market size accurately.
- Calculating market size: define the market by summing up the total revenues of all companies within the country that produce, distribute, or manufacture the identified products and services, as well as other financial metrics.

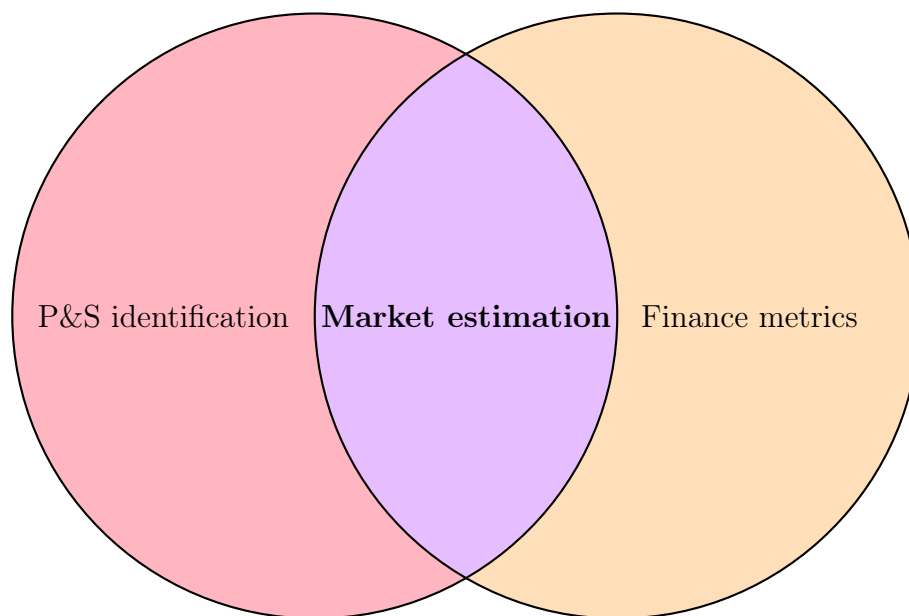


Figure 2.12: Venn diagram illustrating the market estimation Bottom-Up approach with identifying products, services and finance metrics.

Using this approach, analysts can answer key questions about the market, such as:

- Which companies are growing the fastest?
- Which companies are the most profitable?
- Which companies have the highest revenues (market dominance)?
- Is the market highly consolidated or fragmented?
- Is there a risk of economic disruption in the market?
- Are new products and services emerging, and what is the value proposition driving company success?
- Is there a relationship between product/service characteristics and higher sales?
- How does the location of these companies correlate with higher sales?

This comprehensive analysis provides valuable insights for market reports, addressing a wide range of questions critical for strategic planning and decision-making. This approach will be the one we will use in the project and the inputs will be taken from those mentioned in [2.3.1](#)

2.3.3 Automated report generation

Automated report generation leverages AI and machine learning technologies to create comprehensive, accurate, and timely reports without human intervention. This process enhances efficiency, reduces errors, and allows for the analysis of vast amounts of data in real-time.

Technologies used in automated report generation include:

- **Natural language processing (NLP):** NLP systems transform structured data into readable text. These systems are capable of producing narratives that explain data insights, trends, and anomalies [\[24\]](#).
- **Data visualization tools:** tools such as Tableau and Power BI automatically generate visualizations like charts and graphs that complement textual reports, providing a more comprehensive view of the data [\[39\]](#).
- **AI-Powered analytics platforms:** platforms like IBM Watson Analytics and Google Cloud AutoML use AI to analyze data and generate reports, offering predictive insights and recommendations based on the data [\[5\]](#).
- **Workflow automation tools:** tools such as **n8n** and Apache NiFi automate data workflows, ensuring that data from various sources is collected, processed, and integrated seamlessly [\[23, 40\]](#).

- **Python libraries:** libraries like Pandas and Matplotlib are extensively used in automated report generation for data manipulation and visualization [28,35].

2.4 State of the art in business process automation

2.4.1 Current and future trends

Business process automation (BPA) is continuously evolving, driven by advancements in artificial intelligence, machine learning, and other emerging technologies. The current and future trends in BPA include:

Artificial intelligence and machine learning: AI and ML are being increasingly integrated into BPA to enhance decision-making processes, predict outcomes, and automate complex tasks. These technologies enable systems to learn from data, adapt to new inputs, and perform tasks that traditionally required human intelligence [2].

Robotic process automation (RPA): RPA involves the use of software robots or "bots" to automate routine, repetitive tasks. RPA tools are becoming more sophisticated, with capabilities to handle unstructured data and integrate with AI for more complex processes [1].

Intelligent process automation (IPA): IPA combines RPA with AI technologies such as natural language processing (NLP) and computer vision to automate end-to-end business processes. This trend is expected to transform how businesses operate, providing more agility and efficiency [51].

Cloud computing: the adoption of cloud-based BPA solutions allows for greater scalability, flexibility, and cost savings. Cloud platforms provide the infrastructure needed to support large-scale automation projects and facilitate remote work [26].

Internet of things (IoT): IoT devices generate vast amounts of data that can be used to automate and optimize business processes. The integration of IoT with BPA enables real-time monitoring, predictive maintenance, and enhanced operational efficiency [32].

Hyperautomation: Hyperautomation is the expansion of automation capabilities across the organization by combining multiple technologies such as RPA, AI, ML, and advanced analytics. It aims to automate as many business processes as possible, leading to significant improvements in efficiency and productivity [29].

2.4.2 Case studies and real applications

The implementation of business process automation in various industries has demonstrated its potential to streamline operations, reduce costs, and enhance productivity. Here are some practical examples:

- **Finance:** in the financial sector, BPA is used to automate tasks such as invoice processing, fraud detection, and customer service. For instance, JPMorgan Chase implemented an AI-driven system called *COiN* that automates the review of legal documents, significantly reducing the time and effort required for this process [4].

- **Healthcare:** BPA in healthcare includes the automation of administrative tasks such as patient scheduling, billing, and claims processing. The Mayo Clinic has utilized RPA to automate the retrieval and processing of patient data, improving accuracy and freeing up staff for more critical tasks [42].
- **Manufacturing:** in the manufacturing industry, BPA is applied to optimize supply chain management, quality control, and production scheduling. Siemens has implemented IoT and AI technologies to create a "smart factory" that enhances production efficiency and reduces downtime [46].
- **Retail:** retailers use BPA to manage inventory, process orders, and personalize customer experiences. Walmart has employed AI and ML to optimize its supply chain operations, resulting in faster restocking and reduced operational costs [47].
- **Human resources:** BPA tools are used in HR for automating recruitment, onboarding, and employee performance management. Unilever has leveraged AI to streamline its hiring process, using AI-driven assessments to evaluate candidates, thus reducing hiring time and improving the quality of hires [50].

2.4.3 State of the art competitors

In the realm of competitive intelligence and market report automation, several advanced tools have emerged, leveraging artificial intelligence to enhance data gathering and analysis. Among the notable competitors in this field are Crayon and Speak. These platforms exemplify the cutting-edge capabilities currently available, providing businesses with critical insights and strategic advantages.

Crayon is a comprehensive competitive intelligence platform designed to help businesses stay ahead of their competition by tracking and analyzing data from multiple sources in real-time. One of Crayon's standout features is its intelligent filtering system, which eliminates 99% of noisy data, allowing users to focus on the most critical 1% relevant to their competitive strategy. This significantly reduces distractions and enhances efficiency. Crayon also employs AI-powered summarization to create ready-to-share summaries of news articles, blog posts, and press releases, streamlining the process of information dissemination. Additionally, the platform's automatic tagging feature categorizes new intelligence into over 80 subcategories, minimizing the need for manual organization [6].

Furthermore, Crayon utilizes natural language processing (NLP) for sentiment analysis, providing a graded scale of sentiment for insights, which helps users understand market sentiment more accurately. The importance scoring system, powered by machine learning models, dynamically evaluates and prioritizes intelligence based on its market impact, ensuring that users can quickly identify and respond to critical insights. Crayon

also features anomaly detection, tracking trends across competitors' digital footprints and alerting users to significant anomalies, enabling timely market responses. Its website tracking capability has monitored millions of web page changes with automatic annotations, offering a detailed view of competitor activities and updates.

Speak, on the other hand, focuses on automating competitive intelligence and market reporting. It excels in automated data collection, gathering information from various sources including web pages, social media, and market reports. This extensive data collection is followed by AI-driven insight generation, where the platform processes and analyzes the collected data to produce actionable insights. Speak also offers customizable reports, allowing businesses to tailor the generated market reports to their specific needs and goals. The platform's real-time update feature ensures that users have the most current information on competitor activities, providing a significant edge in strategic planning and market positioning [48].

Both Crayon and Speak demonstrate how AI can automate and enhance the creation of market reports and competitive intelligence. These tools reduce the manual effort required and increase the accuracy and relevance of the information provided. However, while these platforms offer substantial capabilities, they do not fully replicate the comprehensive, multi-layered approach of the tool developed in this project. The unique advantage of our tool lies in its specialized design for market report automation, incorporating a sophisticated architecture that seamlessly integrates data collection, processing, and presentation. This ensures a higher degree of customization, accuracy, and efficiency, tailored specifically for private equity firms and their strategic needs.

Chapter 3

Design of the tool

3.1 Introduction

In today's fast-paced business environment, the ability to quickly and accurately gather, analyze, and report market data is crucial for maintaining a competitive edge. The traditional methods of conducting market studies, which often involve extensive manual effort, are increasingly being supplanted by automated solutions leveraging advanced technologies. As it has been reported in section [1](#), Bayesian_x has developed an innovative tool aimed at revolutionizing this process by integrating artificial intelligence (AI) and machine learning (ML) techniques to automate market studies and report generation.

This chapter delves into the design and implementation of this cutting-edge tool, highlighting its architecture, key components, functionalities, and technical underpinnings. The tool's primary objective is to significantly reduce the time and resources required to produce comprehensive and precise market reports, thereby enhancing operational efficiency and providing valuable insights for strategic decision-making.

The sections that follow will provide a detailed overview of the tool, starting with its modular architecture, which ensures seamless interaction between various components involved in data collection, processing, and reporting. We will explore each major component, including:

- Layer 0 (raw data inputs): Researcher_x [\(2.3.1\)](#), REST APIs, webscrapping.
- Layer 1 (processed data inputs): product and service workflows and financial information database.
- Layer 2 (processing): Chaining.
- Layer 3 (outputs): Market reports.

The necessary *glues* to unify all parts into a single functional tool: n8n, python codes, MongoDB and Pinecone (CKS) has been explained in [\(2.3.1\)](#)

Technical implementation details will be covered, showcasing examples of code and explaining how various technologies and programming languages have been utilized. Additionally, we will delve into the concept of the context window in large language models

(LLMs) and explain how Pinecone helps manage this context for effective data queries and retrievals.

Comment again that the boxes in black in the previous sections will form the framework in which the project moves.

And lastly, the challenges encountered during the development and implementation of the tool, along with the solutions adopted, will be discussed. In the next section will also highlight lessons learned and offer recommendations for future implementations.

3.2 Overview of the tool

The tool developed by Bayesian_x is designed to automate and enhance the process of market study and report generation through a sophisticated multi-layered architecture. This architecture is visualized in the accompanying diagram (3.1), which highlights the interaction and flow between various components of the system.

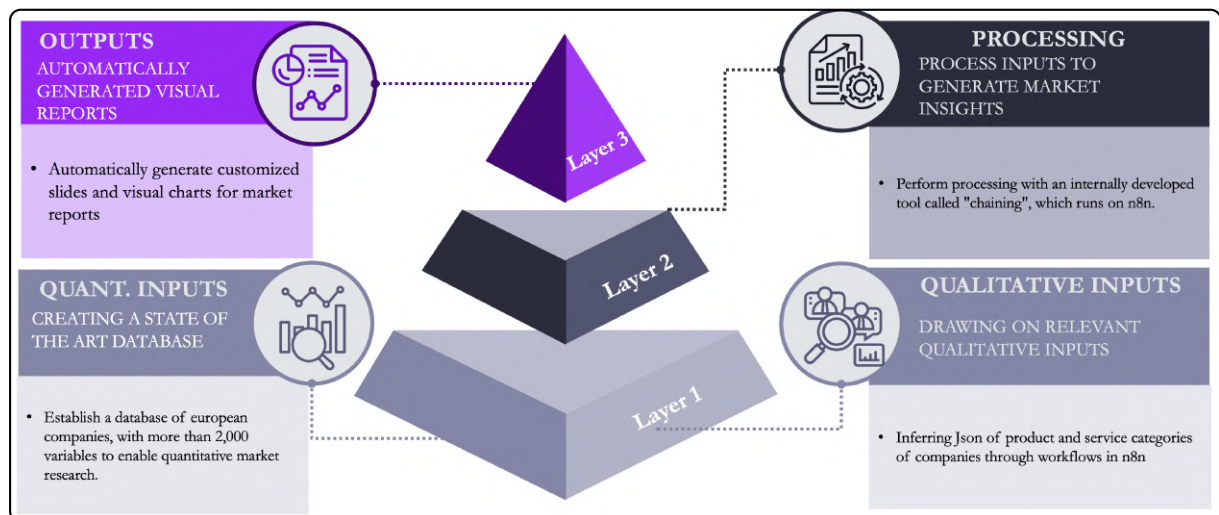


Figure 3.1: Overview of the tool's architecture.

It is important to note that this is an overview of the tool. The details and detailed architecture are described in section 3.3 below.

3.2.1 Layer 0: Raw data inputs

At the foundational level, the tool begins with the collection of raw data inputs. These inputs are sourced from several avenues including Researcher_x, REST APIs, and web scraping techniques. This layer ensures that a wide array of unprocessed data is available for further analysis and transformation.

Researcher_x is a custom-developed tool designed to facilitate the collection of raw market data. It aggregates data from various sources, including public databases, industry reports, and proprietary datasets. This tool automates the initial phase of data gathering,

ensuring a comprehensive and up-to-date collection of relevant market information.

Web scraping techniques are employed to extract data from websites and online platforms. Using tools like BeautifulSoup and Scrapy, the tool can systematically gather information from multiple web sources, including company websites, online news articles, and market analysis blogs. This method is essential for capturing real-time data and trends that are not readily available through other means.

The tool also leverages REST APIs to access structured data from external systems and services. APIs provide a standardized way to retrieve data from various platforms, such as financial databases, social media channels, and market research firms. This approach allows the tool to integrate diverse data sources efficiently, ensuring that the collected data is both relevant and comprehensive.

By combining these methods, the tool ensures a robust and diverse collection of raw data inputs, setting a strong foundation for subsequent processing and analysis.

In collecting raw data inputs, we encountered several challenges. Ensuring data quality and consistency across different sources required rigorous validation and cleaning processes. We tackled this by leveraging Researcher_x to aggregate data from reliable sources and using targeted web queries with site-specific searches (e.g., site:https...) to ensure relevance. Web scraping faced difficulties such as handling dynamic content and anti-scraping measures, mitigated with advanced techniques and rotating proxies.

3.2.2 Layer 1: Processed data inputs

The raw data collected in Layer 0 is then processed to generate more structured and usable data sets. This involves two primary components: organizing quantitative financial information into a comprehensive database and compiling qualitative data from product and service workflows.

Firstly, the financial information database includes over 200 variables that provide a detailed quantitative analysis of market data. The missing information within this dataset has been interpolated and extrapolated using the XGBoost algorithm, ensuring completeness and accuracy. This step is crucial for generating a reliable and actionable data foundation that supports in-depth market analysis.

Secondly, the qualitative data from product and service workflows is organized into structured JSON formats. This information captures the intricate details of various products and services, allowing for a nuanced understanding of market dynamics. The qualitative insights are essential for complementing the quantitative data, providing a holistic view of the market.

Once the information is processed and retrieved, it is stored in MongoDB. Pinecone, on the other hand, is used for retrieval-augmented generation (RAG) during the information retrieval in the workflows. This ensures that the data can be efficiently accessed and utilized in generating insights and reports.

By effectively processing and organizing the raw data, Layer 1 establishes a robust and comprehensive data foundation that is essential for generating accurate and insightful market reports.

For quantitative data, the challenge was incomplete datasets, resolved using XGBoost for interpolation and extrapolation. In qualitative data, the inconsistency in LLM outputs was addressed by implementing structured prompts with three distinct roles and refining them using the OpenAI Playground. Ensuring the effectiveness of RAG with Pinecone required precise prompt crafting. Lastly, setting up MongoDB with CapRover was challenging but overcame with persistent troubleshooting. These processed data sets are stored in MongoDB and Pinecone for efficient access, establishing a robust foundation for accurate market reports.

The major challenge was scaling the solution for many companies. While workflows worked well for 10-20 companies, our cloud n8n service ran out of memory for more extensive use. This was resolved by deploying n8n on our DigitalOcean servers via CapRover, though the installation was challenging due to configuration and credential issues.

3.2.3 Layer 2: Processing

The core processing of the tool is handled by an internally developed system called *Chaining*, which operates on the n8n platform. This layer is responsible for the sophisticated analysis and synthesis of the processed data inputs, transforming them into valuable market insights. The Chaining process is designed to be highly efficient and scalable, ensuring that the tool can handle large volumes of data seamlessly.

The Chaining system was initially developed as a comprehensive processing framework. I have made several modifications and adaptations to tailor it specifically to our use case. These enhancements ensure that the tool meets our unique requirements and optimizes its performance for our specific data and analysis needs. The Chaining process returns the processed text, which is then included in the final slides of the market reports.

It's important to note that, similar to the Chaining system, the Researcher_x tool was also pre-developed. My contributions involved customizing and refining these tools to align them with the objectives and constraints of our project.

By leveraging and adapting these pre-existing tools, we have been able to streamline the development process, ensuring that the core functionalities are robust and well-integrated to support our market study and report generation needs.

3.2.4 Layer 3: Outputs

The final layer involves generating the outputs, which are automatically produced visual reports. These reports include customized slides and visual charts tailored to specific market studies. The automated generation of these outputs significantly reduces the manual effort traditionally required, providing timely and precise insights for strategic decision-making.

A key component that connects the Chaining process to the final slides of the market report is a tool developed by another team member called “text-to-slide.” This tool takes the processed text generated by Chaining and converts it into visually appealing slides. The integration of text-to-slide ensures that the final reports are both informative and visually engaging.

By incorporating this automated text-to-slide functionality, the tool enhances the efficiency and accuracy of report generation, allowing for rapid production of high-quality market insights.

3.2.5 Integration of technologies

To unify all these components into a single, functional tool, several key technologies are employed. These technologies act as a ‘nexus’, ensuring that all parts of the tool work together seamlessly. They include the n8n automation platform, Python for scripting and data manipulation, MongoDB for database management, Pinecone for managing the context window in large language models (LLMs) and retrieval-augmented generation (RAG) for CKS, and the text-to-slide tool for generating the final visual reports.

n8n serves as the backbone of the automation process, orchestrating the various workflows and ensuring smooth data flow between different components. Python is utilized for scripting and data manipulation tasks, providing the flexibility and power needed to handle complex data processing operations. MongoDB is used for storing and managing processed data, offering a robust and scalable solution for database management. Pinecone plays a crucial role in managing the context window in LLMs and implementing RAG for CKS, enhancing the tool’s ability to retrieve and utilize relevant information efficiently. Finally, the text-to-slide tool bridges the gap between the processed text from Chaining and the final visual reports, ensuring that the output is both informative and visually engaging.

These transversal tools are critical for the functionality and effectiveness of the entire system. The following section will delve into each of these technologies in greater detail, explaining their specific roles and how they contribute to the overall workflow.

3.3 Architecture diagram of the tool

This section breaks down the detailed architecture of the tool into an easy to visualise diagram. In the following subsection, a comprehensive analysis of each main component and the sub-modules needed to link all the parts together will be carried out.

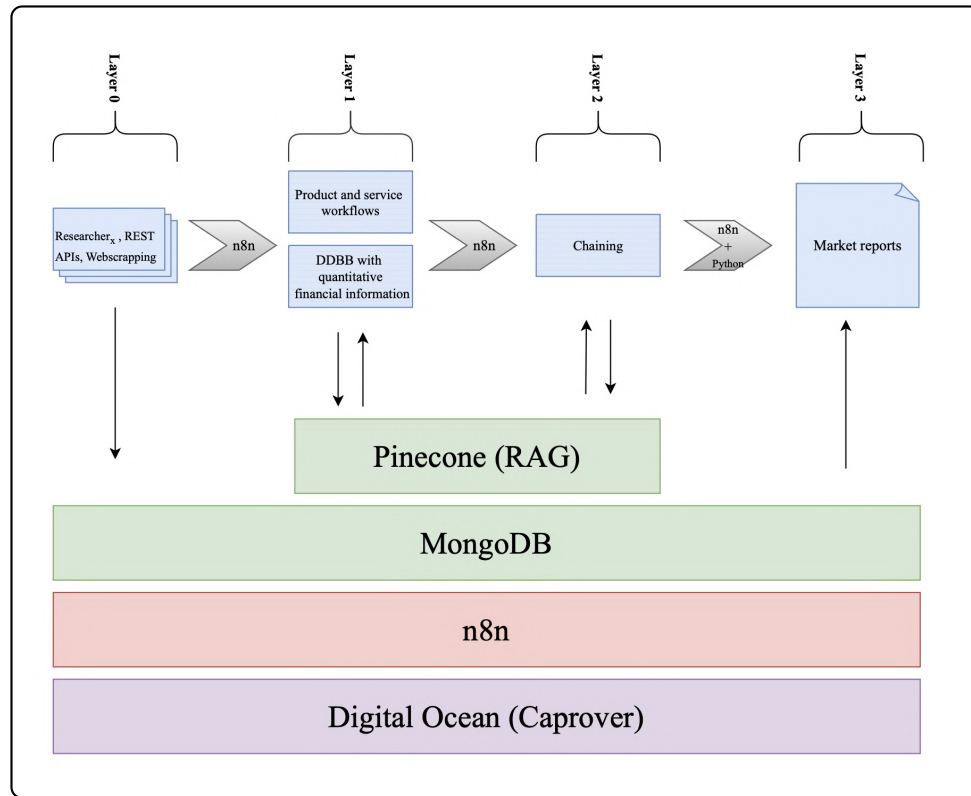


Figure 3.2: Detailed diagram with tool parts.

3.4 Main components

Following the overview in section [3.2](#), the parts and tools involved in the process are detailed below. The tool comprises 4 key components:

3.4.1 Layer 0: Raw data inputs

Layer 0 is the foundational layer of the tool, responsible for the collection of raw data inputs. This layer incorporates three main methods: web scraping, the use of Researcher_x, and REST APIs.

Web scraping

Web scraping is a crucial method for gathering data from various online sources. In our tool, we utilize a service called Serper, which facilitates the initial process of identifying and collecting web pages that contain relevant information. Serper is an API that interfaces with search engines, providing us with URLs of pages that match our specified queries. This API allows for efficient and targeted data collection by narrowing down the

vast amount of information available on the web to only the most pertinent sources.

Once we have the URLs from Serper, we employ a custom-developed scraper created by Bayesian_x. This scraper is designed to extract specific data points from the identified web pages. The scraper runs on DigitalOcean, a cloud infrastructure provider, and is managed by Caprover. This setup ensures that the web scraping process is scalable and can handle large volumes of data efficiently.

Researcher_x

Researcher_x is another integral component of Layer 0. This tool was developed by a team member at Bayesian_x to streamline the collection of market data from various sources. Researcher_x aggregates data from public databases, industry reports, and proprietary datasets, providing a comprehensive and up-to-date collection of relevant market information. It automates the initial phase of data gathering, reducing the time and effort required to compile extensive datasets manually. This tool is essential for ensuring that our data inputs are both diverse and current, enabling more accurate and relevant market analyses and, above all, a trace of the web pages that are accessed.

REST APIs

The third method employed in Layer 0 involves the use of REST APIs to access structured data from external systems and services. We utilize several APIs to gather detailed information about companies and market trends. For instance, we use LinkedIn's API to extract data about company profiles, employee counts, and industry classifications. Additionally, we leverage APIs from other business information providers, such as Crunchbase and Glassdoor, to obtain insights into company financials, market positions, and employee reviews.

These APIs offer a standardized way to retrieve data, ensuring that the information is consistent and reliable. By integrating multiple APIs, we can cross-reference data from different sources, enhancing the accuracy and depth of our datasets. This approach allows us to compile a comprehensive view of the market landscape, which is critical for generating valuable insights.

In summary, Layer 0 of our tool is designed to collect a wide array of raw data inputs through web scraping with Serper and a custom scraper, the Researcher_x tool, and various REST APIs. This foundational layer ensures that we have a robust and diverse dataset for further processing and analysis in the subsequent layers.

3.4.2 Layer 1: Processed data inputs

Layer 1 is the most crucial layer in the architecture, as it processes raw data inputs into structured and usable datasets. As shown in [3.1](#), this layer handles two main types of inputs: quantitative and qualitative.

By integrating both quantitative and qualitative data inputs, Layer 1 ensures that the necessary information is available for comprehensive market analysis. The quantitative data provides a numerical basis for evaluating company performance, while the qualitative data enriches the reports with detailed descriptions of products and services. These two inputs, once processed and transformed through Layer 2 (Chaining) and Layer 3 (Output), populate the slides of the market reports. Thus, the inputs from Layer 1 can be considered the nucleus of the market reports.

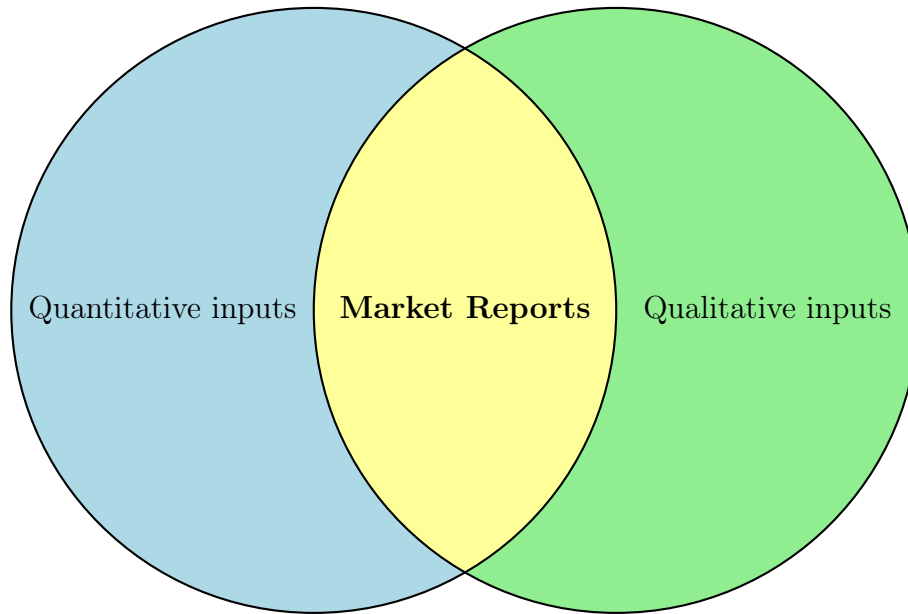


Figure 3.3: Venn diagram illustrating that the union of quantitative inputs and quality inputs form the core of the content of market reports

Quantitative inputs

The quantitative inputs in Layer 1 are derived from extensive web scraping of important company databases. These databases include a wide array of financial metrics that are essential for comprehensive market analysis. By aggregating data from various reliable sources, we ensure that our database encompasses critical financial variables that provide a detailed snapshot of each company's financial health and performance. Some of the most significant variables collected include:

- Employee Costs (Year 9)
- Total Assets (LYA, Year 3, Year 9)
- Shareholder Funds (LYA, Year 3, Year 9)
- Non-current Liabilities (LYA, Year 3, Year 9)
- Fixed Assets (LYA, Year 3, Year 9)
- Revenue (LYA, Years 1-4, Year 9)
- Ordinary Results (LYA, Years 1-4, Year 9)
- Cost of Goods Sold (COGS) (LYA, Year 3, Year 9)

- R&D Spend (LYA, Year 3, Year 9)
- Cyber Risk Rating
- Cash Flow (LYA, Year 1, Year 9)
- Year-over-Year (YoY) Growth Rate for Employees (Years 1-4)
- Compound Annual Growth Rate (CAGR) for Employees and Financial Metrics over various periods

In total, we track 202 distinct financial variables. However, due to the nature of the data sources, some data points may be missing. To address this, we use XGBoost, a powerful machine learning algorithm, to interpolate and extrapolate missing values. XGBoost is particularly effective for this task because it handles large datasets efficiently and can model complex relationships between variables.

For instance, if we have the Total Assets for Years 3 and 9 but are missing the data for Year 5, XGBoost can predict this intermediate value based on the available data and the trends observed in other variables. Additionally, some financial variables can be derived algebraically from others. For example, Total Liabilities can be calculated by subtracting Shareholder Funds from Total Assets. Similarly, the Equity Ratio can be computed as Shareholder Funds divided by Total Assets.

The implementation of these calculations and machine learning models is done using Python. Python's extensive libraries and frameworks, such as pandas for data manipulation, scikit-learn for machine learning, and XGBoost for predictive modeling, provide robust tools for handling and processing large datasets.

Example of calculations and interpolations

1. **Interpolating Missing Values:** if Total Assets data for Year 5 is missing, XGBoost uses the data from surrounding years and other related variables to predict the missing value. This interpolation ensures that the dataset remains continuous and reliable, which is crucial for trend analysis and forecasting.

2. **Algebraic derivations:**

- **Total liabilities:** this can be calculated as:

$$\text{Total Liabilities} = \text{Total Assets} - \text{Shareholder Funds} \quad (3.1)$$

By using available data for Total Assets and Shareholder Funds, we can accurately derive the Total Liabilities.

- **Equity Ratio:** This is computed as:

$$\text{Equity Ratio} = \frac{\text{Shareholder Funds}}{\text{Total Assets}} \quad (3.2)$$

This ratio provides insights into the financial stability of a company, indicating the proportion of a company's assets that are financed by shareholders' equity.

3. Year-over-Year (YoY) growth and CAGR calculations:

- **YoY growth for employees:**

$$\text{YoY Growth Rate} = \frac{\text{Employees in Year } n - \text{Employees in Year } (n-1)}{\text{Employees in Year } (n-1)} \times 100 \quad (3.3)$$

This metric helps in understanding the annual growth trends in the workforce.

- **CAGR for financial metrics:**

$$\text{CAGR} = \left(\frac{\text{Ending Value}}{\text{Beginning Value}} \right)^{\frac{1}{n}} - 1 \quad (3.4)$$

Where n is the number of years. CAGR provides a smoothed annual growth rate over a specified period, which is useful for long-term growth analysis.

Storage and scalability

All processed data, both interpolated and derived, is stored in MongoDB. MongoDB offers the flexibility and scalability needed to handle large volumes of diverse data. Its document-oriented database structure allows for the efficient storage of complex and variable data types, making it ideal for our extensive and varied dataset.

Our goal is to build a comprehensive database encompassing 1,000,000 companies, making it one of the largest and most detailed business datasets in Europe. This extensive database will provide invaluable insights for market analysis and strategic decision-making. The scalability of MongoDB ensures that as our dataset grows, the performance and accessibility of the data remain robust and efficient.

By meticulously processing and organizing these quantitative inputs, Layer 1 establishes a robust foundation for the subsequent analysis and reporting stages. This thorough and detailed approach ensures that the tool delivers accurate, comprehensive, and actionable market insights.

Qualitative inputs

The qualitative inputs consist of a series of JSON files that contain detailed information about categories of products and services for specific companies. These JSON files include various attributes such as product names, target audiences, key features, unique selling propositions (USPs), pricing details, performance metrics, and more. Below is an example of the JSON structure for a company like Amazon in the computers category:

```
1 {  
2   "Category Details": {  
3     "Category Name": "Computers",  
4     "Product Names": [  
5       "HP 17 Business Laptop",  
6       "Acer Aspire 3 A315-24P-R7VH Slim Laptop",  
7       "HP 2023 Newest Chromebook Laptop",  
8       "Dell Optiplex 7050 SFF Desktop PC (Renewed)",  
9       "HP 2022 Newest All-in-One Desktop",  
    ]  
  }  
}
```

```

10     "HP Elite Desktop Computer PC"
11 ],
12 "Target Audience": "General consumers, students, professionals,
    businesses",
13 "Key Features": [
14     "Various display sizes (14 inch, 15.6 inch, 17.3 inch, 21.5
        inch)",
15     "Different processors (Intel Core, AMD Ryzen, Intel Celeron)",
16     "RAM options ranging from 4GB to 32GB",
17     "Storage options (SSD, HDD, eMMC)",
18     "Operating systems (Windows, ChromeOS)",
19     "Wi-Fi and Bluetooth connectivity",
20     "Special features like backlit keyboard, fingerprint reader,
        anti-glare coating"
21 ],
22 "Unique Selling Proposition (USP)": "A wide range of computers for
    different needs, combining performance with specific features
    (e.g., backlit keyboard, fingerprint reader), and participation
    in the Climate Pledge Friendly program for sustainability.",
23 "Price": {
24     "Retail Price": {
25         "HP 17 Business Laptop": "$497.99",
26         "Acer Aspire 3 A315-24P-R7VH Slim Laptop": "$299.99",
27         "HP 2023 Newest Chromebook Laptop": "$184.70",
28         "HP 2022 Newest All-in-One Desktop": "$395.00",
29         "HP Elite Desktop Computer PC": "$189.99"
30     },
31     "Wholesale Price": "N/A"
32 },
33 "Performance": {
34     "Efficiency": "Various processing power based on different CPUs
        ",
35     "Reliability": "Reliability assured by brands like HP, Dell,
        and Acer",
36     "Compatibility": "Compatible with various peripherals and
        software due to diverse hardware interfaces and operating
        systems",
37     "Scalability": "Scalability depends on the specific product
        with options for upgrades in RAM, storage",
38     "Battery Life": [
39         "Up to 4 Hours",
40         "5 to 7 Hours",
41         "8 to 10 Hours",
42         "11 Hours & Up"
43     ],
44     "Speed/Response Time": "Dependent on CPU processor speed (
        ranging from 1 GHz to 4 GHz and above)",
45     "Capacity/Volume": "Storage options from 64GB eMMC to 1TB SSD
        or HDD"
46 }
47 }
48 }

```

Listing 3.1: Example of detailed computer products in Amazon

To obtain these detailed JSON files, a structured process involving multiple workflows has been developed, all revolving around a critical tool: the Central Knowledge System (CKS), which is a Retrieval-Augmented Generation (RAG) system based on Pinecone.

The advantages of using a RAG instead of using simply calls to OpenAI API are detailed in chapter [6.2](#)

Process for obtaining qualitative inputs

First workflow: storing information about product and service categories

The first workflow is designed to initiate the extraction and storage of information related to categories of products and services for a given company. This workflow begins with three key inputs: the company name, the category to search for (products or services), and the country where the research will be conducted. These inputs set the foundation for a series of processes that ultimately lead to storing structured information in Pinecone.

1. Initial inputs:

- **Company name:** the official name of the company as known internationally.
- **Category to search:** specifies whether the focus is on ‘products’ or ‘services.’
- **Country:** the country where the research is being conducted to ensure regional relevance.

2. **Retrieving official company name:** the workflow starts with a call to the OpenAI API ChatGPT 4o, which is used to retrieve the official name of the company in the specified country. This ensures that the research is conducted using the correct and recognized name of the company within the given regional context.

3. **Generating site-specific search query:** a JavaScript program is then executed to append the official company website to the search query. This is done by constructing a search query in the format:

site: https://officialcompanywebsite.com category

For example, for Telefónica services in Spain, the query would be:

site: https://telefonica.es services

The use of ‘site:’ ensures that the retrieved information is confined to the official company website, thereby minimizing noise and irrelevant data from other sources.

4. **Fetching webpage links using Serper:** Serper is an API service that interfaces with search engines to provide links to relevant web pages. This service takes the site-specific search query generated in the previous step and returns a list of URLs that match the search criteria. Serper helps in narrowing down the search to the most pertinent pages that contain the needed information about the company’s products or services.

5. **Processing retrieved links:** the links obtained from Serper are then processed through a series of transformations using custom JavaScript code. These transformations prepare the URLs for the subsequent HTTP requests and ensure that the data pipeline

is streamlined.

6. Scraping web pages: an HTTP request is made to a codebase that has been API-fied and is running on DigitalOcean, managed by CapRover. This codebase performs web scraping on the URLs provided by Serper. Web scraping involves extracting specific information from the web pages, such as product or service details, descriptions, and other relevant attributes.

7. Storing data in Google Drive: the scraped information is processed and then stored in a Google Drive document. This is facilitated by using Google OAuth2 for secure access and integration. The data is formatted and saved as a PDF file in Google Drive, ensuring that it is easily accessible and can be referenced or shared as needed.

8. Ingesting data into Pinecone: the final step in this workflow is to ingest the processed information into Pinecone. Pinecone is a vector database designed for high-dimensional data. The data from the PDF is embedded and stored in Pinecone with the following configurations:

- **Metadata:** the company name is used as metadata for indexing and retrieval purposes.
- **Embedding algorithm:** the algorithm used for embedding is 'Cohere-embed-multilingual-v3.0,' which supports multiple languages and provides high-quality embeddings.
- **Metric:** the cosine similarity metric is used to measure the similarity between vectors.
- **Number of dimensions:** the embeddings are 1024-dimensional, providing a rich representation of the data.
- **Cloud provider:** AWS is used as the cloud provider for hosting the Pinecone instance.
- **Chunk size:** the data is ingested in chunks of 1000, ensuring efficient processing and storage.

This completes the first workflow, laying the groundwork for the subsequent workflows that will further refine and expand the dataset. The initial setup and storage in Pinecone provide a robust foundation for detailed research and data extraction in the following steps.

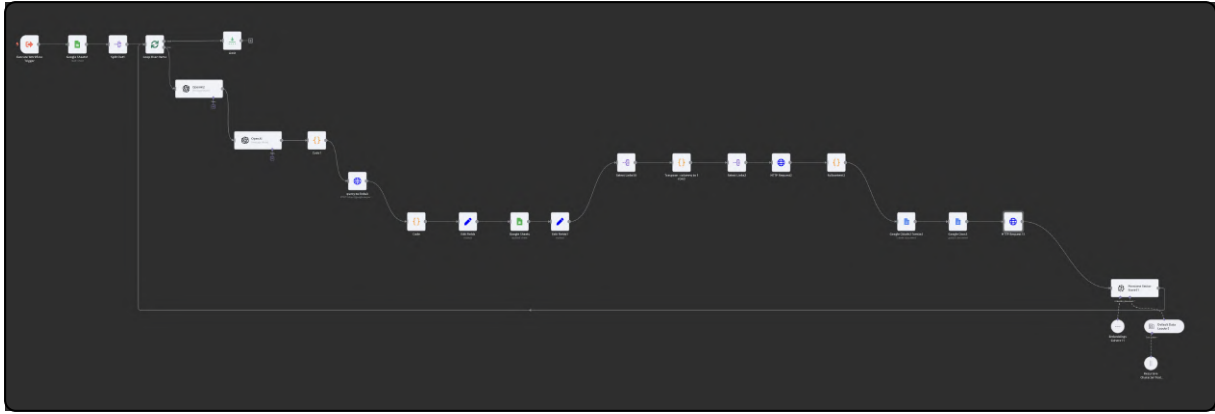


Figure 3.4: Actual diagram of the first workflow.

Second workflow: creating a list of product and service categories

The second workflow is designed to iterate over each company and retrieve relevant information from Pinecone to create a comprehensive list of product and service categories. This workflow leverages the data stored in Pinecone from the first workflow and refines it to produce structured lists that are stored in MongoDB. Below is a detailed explanation of each step involved in this workflow.

1. **Iteration over companies:** once the information is stored in Pinecone, the workflow initiates a loop that iterates over each company. This loop can handle multiple companies, allowing for scalable and extensive data processing.

2. **Retrieving information from Pinecone:** for each company, the workflow retrieves relevant information from Pinecone using the following prompt:

“categories of products/services of the company”

This process involves vectorizing the prompt and comparing it with the stored vectors in Pinecone. Pinecone uses an embedding algorithm to perform this comparison, ensuring that the retrieved information is the most relevant to the query.

3. **Chunking process in Pinecone:** Pinecone handles the chunking process by breaking down the data into manageable pieces. The prompt is vectorized and compared with internal vectors stored in Pinecone. This comparison uses the same embedding algorithm (‘Cohere-embed-multilingual-v3.0’) and cosine similarity metric as in the first workflow. The chunking process ensures efficient retrieval of relevant data segments.

4. **Configuration settings:** the configuration settings for this retrieval process are as follows:

- **Elector store retriever:** 50
- **Embedding algorithm:** Cohere-embed-multilingual-v3.0
- **Cosine similarity metric:** used for comparing vectors
- **OpenAI Model:** used to transform the retrieved responses into a more human-readable format

5. **Merging retrieved information:** the workflow then proceeds with a node that merges all the retrieved information into a single item. This step consolidates the data to ensure that the subsequent processing stages can handle it efficiently.

6. **Creating a list of categories:** a call to the OpenAI API is made to create a list of all product or service categories. The prompts used in this step are crafted with three specific roles to ensure clarity and accuracy in the generated list.

Please, note that the reformulated prompts are attached below. The actual prompts are not attached for privacy reasons.

System prompt

The prompt describes a task where an intelligent assistant needs to extract and categorize data from a complex text containing information from various sources. The goal is to compile a comprehensive list of unique categories mentioned in the text, rather than specific items. Even though the text may include repeated mentions and omissions, the assistant should identify the broader category each item belongs to and list these unique categories accurately, even if mentioned only once. The list should be concise, without repetitions, and clearly formatted.

Assistant prompt

At a general level, the prompt describes a task where an intelligent assistant processes a complex text to achieve the following:

Data extraction: The assistant needs to read and extract information from a complex text containing various items.

Category identification: Identify the broad categories to which these items belong.

Duplication elimination: Ensure each category appears only once in the final list, removing any repetitions.

Handling omissions: Recognize and handle situations where some mentions might be omitted in the text.

Clear formatting: Present the list of categories as a bullet-point list, formatted simply and clearly.

The expected output is a list of unique categories, formatted as:

["Category1", "Category2", "Category3", ...]

User Prompt

The prompt provides a text containing information about a company's products. The task is to extract and compile a comprehensive list of all unique product categories mentioned in the text. The goal is to identify the broader category for each product and list these categories without repetitions. The list should be concise and clearly formatted.

7. **Decomposing categories and storing in MongoDB:** once the comprehensive list of categories is obtained, a custom code is used to decompose these categories into individual items. These items are then stored in MongoDB. MongoDB's flexible document-oriented structure allows for efficient storage and retrieval of the decomposed

product and service categories, ensuring that they can be accessed and utilized in subsequent workflows.

This completes the second workflow, which effectively creates a structured and comprehensive list of product and service categories for each company. This refined data is crucial for the detailed research and data extraction processes that follow in the subsequent workflows.

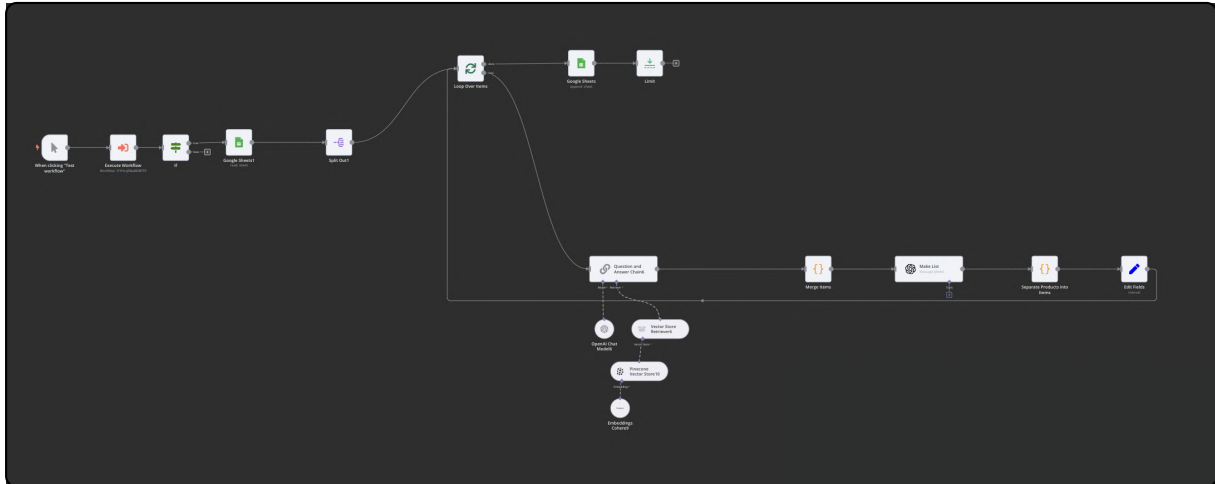


Figure 3.5: Actual diagram of the second workflow.

Third workflow: storing detailed information for each product or service

The third workflow is similar to the first workflow but focuses on extracting and storing detailed information for each specific product or service offered by the company. This workflow builds on the comprehensive list of categories generated in the second workflow and dives deeper into the specifics of each product or service. Below is a detailed explanation of each step involved in this workflow.

1. **Retrieving categories from MongoDB:** the workflow begins by retrieving the list of product and service categories from MongoDB. This list was created in the second workflow and serves as the starting point for detailed research.

2. **Splitting categories:** using a node split operation, the workflow separates each product or service category along with its associated company. This step ensures that each product or service is processed individually in the subsequent loop.

3. **Loop for each product or service:** the workflow then enters a loop that iterates over each product or service category for the company. For each iteration, the following process is executed:

4. **Generating site-specific search query:** a JavaScript program (detailed in Section A) is executed to append the official company website to the specific product or service. This is done by constructing a search query in the format:

site: <https://officialcompanywebsite.com> specific product or service

For example, for Amazon laptops, the query would be:

site: https://amazon.es laptops

This ensures that the retrieved information is confined to the official company website, focusing specifically on the product or service in question.

5. Fetching webpage links using Serper: Serper is again utilized to fetch links to relevant web pages. This time, the search query is more specific, targeting individual products or services. Serper returns a list of URLs that contain detailed information about the specified product or service.

6. Processing retrieved links: the links obtained from Serper are processed through a series of transformations using custom JavaScript code. These transformations prepare the URLs for the subsequent HTTP requests and ensure that the data pipeline is streamlined.

7. Scraping web pages: an HTTP request is made to a codebase that has been API-fied and is running on DigitalOcean, managed by CapRover. This codebase performs web scraping on the URLs provided by Serper. The scraping process extracts detailed information about the product or service, including descriptions, specifications, and other relevant attributes.

8. Storing data in Google Drive: the scraped information is processed and then stored in a Google Drive document. This is facilitated by using Google OAuth2 for secure access and integration. The data is formatted and saved as a PDF file in Google Drive, ensuring that it is easily accessible and can be referenced or shared as needed.

9. Ingesting data into Pinecone: the final step in this workflow is to ingest the processed information into Pinecone. Pinecone is a vector database designed for high-dimensional data. The data from the PDF is embedded and stored in Pinecone with the following configurations:

- **Metadata:** the specific product or service name is used as metadata for indexing and retrieval purposes and the company.
- **Embedding algorithm:** the algorithm used for embedding is 'Cohere-embed-multilingual-v3.0,' which supports multiple languages and provides high-quality embeddings.
- **Metric:** the cosine similarity metric is used to measure the similarity between vectors.
- **Number of dimensions:** the embeddings are 1024-dimensional, providing a rich representation of the data.
- **Cloud provider:** AWS is used as the cloud provider for hosting the Pinecone instance.
- **Chunk size:** the data is ingested in chunks of 1000, ensuring efficient processing and storage.

This completes the third workflow, which focuses on extracting and storing detailed information for each specific product or service offered by the company. The detailed data stored in Pinecone provides a robust foundation for further analysis and reporting.

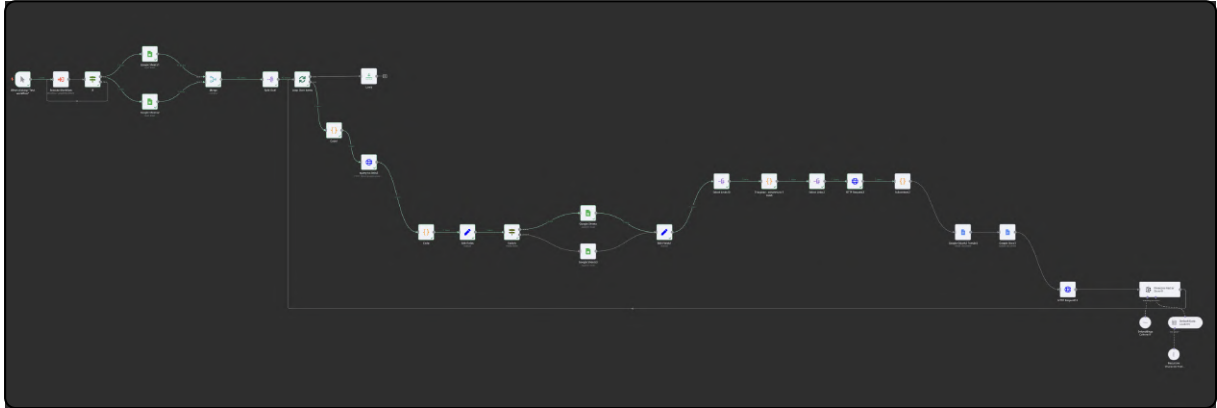


Figure 3.6: Actual diagram of the third workflow.

Fourth workflow: inferring detailed information for JSON fields

The fourth workflow is designed to infer detailed information for JSON fields of products and services, building on the data retrieved and processed in the previous workflows. This workflow follows a similar structure to the second workflow but focuses on filling in the specific fields of the JSON templates for products and services. Below is a detailed explanation of each step involved in this workflow.

1. **Retrieving categories from MongoDB:** the workflow begins by retrieving the list of product and service categories from MongoDB, along with their associated web pages and countries. This data serves as the starting point for detailed information extraction.

2. **Splitting categories:** using a node split operation, the workflow separates each product or service category along with its associated company, web page, and country. This step ensures that each product or service is processed individually in the subsequent loop.

3. **Loop for each product or service:** the workflow then enters a loop that iterates over each product or service category for the company. For each iteration, the following process is executed:

4. **Switch node for product or service:** a switch node is used to direct the workflow based on whether the current item is a product or a service. This ensures that the appropriate processing steps are applied to each type of item.

5. **Generating retrieval prompts:** for products, the following prompt is used to retrieve information from Pinecone:

“send me back all the information you have about the products category:
product category (e.g., laptops)”

For services, the prompt is:

“send me back all the information you have about the services category: service category (e.g., customer service)”

These prompts are used to retrieve the most relevant information stored in Pinecone for each specific product or service category.

6. Retrieving information from pinecone: Pinecone handles the retrieval process by comparing the vectorized prompts with the stored vectors. This comparison uses the same embedding algorithm ('cohere-embed-multilingual-v3.0') and cosine similarity metric as in the previous workflows, ensuring that the most relevant data is retrieved.

7. Merging retrieved information: the workflow proceeds with a JavaScript code node that merges all the retrieved information into a single item. This step consolidates the data to ensure that the subsequent processing stages can handle it efficiently.

8. Inferring JSON fields using OpenAI: an openai node is used to infer the fields of the json templates for products and services. The prompts used in this step are crafted with three specific roles to ensure clarity and accuracy in the generated json:

For products:

System prompt

This prompt instructs the assistant to analyze provided information, categorize it accurately, and fill in a JSON structure. If any data is missing or unknown, the assistant should indicate this with "null" and avoid fabricating information. The focus is on precise and thorough data organization.

Assistant prompt

This prompt instructs the assistant to examine the text provided, categorize the information accurately, and populate the fields in a JSON structure. The assistant should ensure precise organization and follow any given instructions or priorities.

User prompt

The prompt describes a task where a text containing characteristics of product categories is provided. The task is to complete a specific JSON structure with information extracted from that text. The goal is to fill in key details such as the category name, product names, target audience, key features, unique selling proposition (USP), weight, shelf life, prices (retail and wholesale), and various aspects of performance (efficiency, reliability, compatibility, scalability, battery life, response time, capacity/volume, energy consumption, and operating conditions like temperature and humidity).

For services:

System prompt

This prompt instructs the assistant to collect and organize detailed information about different service characteristics into a structured JSON format. If any information is missing or unknown, the assistant should indicate this with "null" without fabricating any data.

Assistant prompt

This prompt instructs the assistant to analyze the provided text and organize the extracted information into a structured JSON format. The assistant should accurately categorize the data and follow any specific instructions or priorities given.

User prompt

The prompt describes a task where a text containing characteristics of categories of services is provided. The task is to complete a specific JSON structure with information extracted from that text. The goal is to fill in key details such as the category name, service names, target audience, key features, unique selling proposition (USP), typical service duration, delivery methods, and degree of customization (including level of adaptability, level of interaction, and degree of standardization). Additionally, it requires details about performance and agreements (such as support channels and typical success criteria) and pricing information (including typical pricing model and typical price range).

The provided JSON structure should be filled with the relevant details extracted from the text.

9. Storing data in MongoDB: finally, the inferred json data is stored and persisted in mongodb. mongodb's flexible document-oriented structure allows for efficient storage and retrieval of the detailed product and service information, ensuring that it can be accessed and utilized in subsequent analysis and reporting.

This completes the fourth workflow, which focuses on inferring detailed information for json fields of products and services and storing this structured data in mongodb. The detailed data provides a robust foundation for further analysis and decision-making.

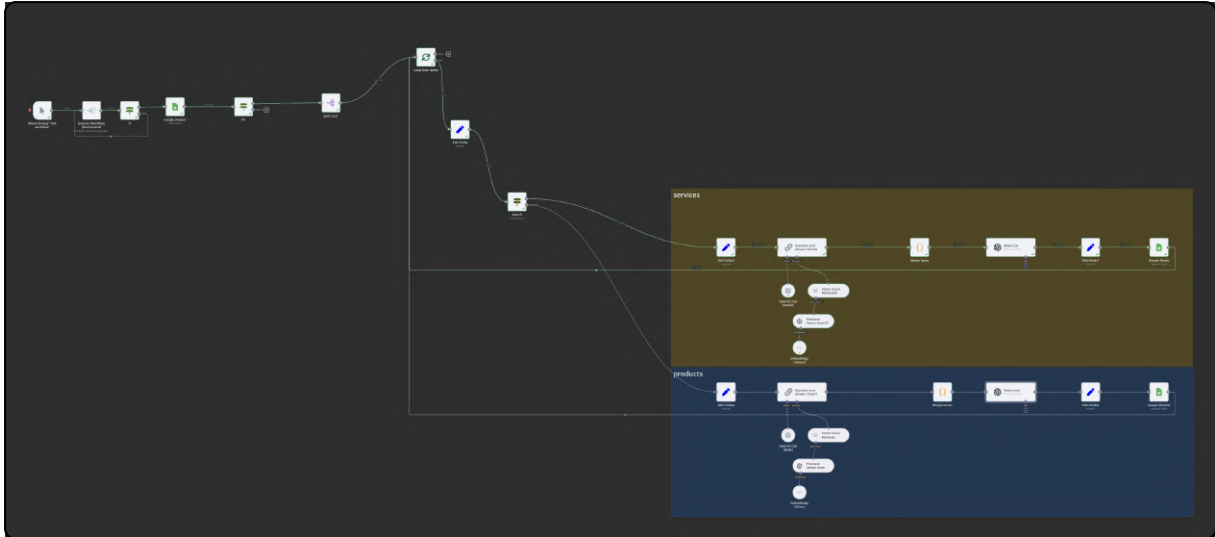


Figure 3.7: Actual diagram of the fourth workflow.

3.4.3 Layer 2: Processing

The tool developed for this layer was originally created by another team member and has been adapted for this specific use case. Its primary function is to act as the processing component of the tool, transforming the retrieved data into processed and structured information that will later be used to populate the slides of the reports.

Initially, the tool retrieves various inputs from MongoDB, which include JSON files and information outputs from Layer 1. This data forms the basis for the processing activities. Next, the titles of the slides to be populated are selected, ensuring these titles are relevant and cover the key aspects of the report.

Once the inputs and slide titles are defined, the tool performs a series of processes using Large Language Models (LLMs). This processing involves several stages:

First, a function is used to obtain general information about the companies, ensuring a clear understanding of what each company does. Subsequently, the tool conducts more in-depth research on the market, products and services, differentiated value proposition, revenues, profits, and growth of each company. This step is crucial for collecting detailed and specific data necessary for analysis.

In an advanced stage of processing, a language model such as GPT-4 is employed to act as a management consultant. This model summarizes the content about each company into specific dimensions such as company overview, products and services, value proposition, and key figures. The model ensures that only relevant figures and numbers are provided, omitting any irrelevant or non-existent data.

Finally, the tool generates the outputs in text format, which are used to populate the slides of the reports. These outputs include detailed and structured summaries covering the most important aspects of each company and its relationship with the investment fund. The generated information is clear, concise, and suitable for presentation in professional reports, facilitating the creation of high-quality and relevant content. This comprehensive

process ensures that a complete and accurate view of the investment portfolio is obtained, based on data processed and analyzed through advanced natural language processing techniques.

The workflow for this process is attached below.

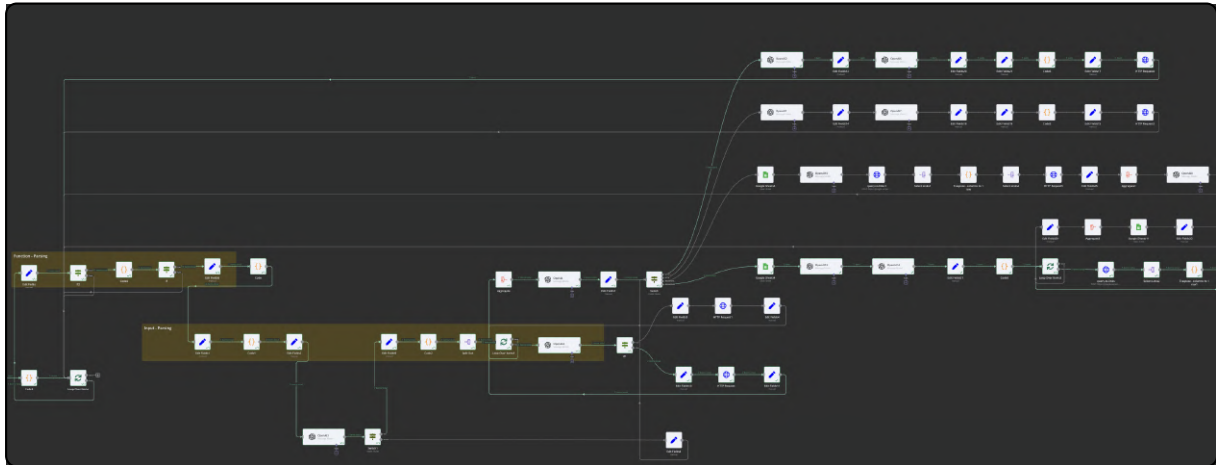


Figure 3.8: Actual diagram of the chaining workflow.

3.4.4 Layer 3: Outputs

The connection between the chaining (processing) and the outputs (market reports) is established through a Python code. These market reports are primarily used to advise private equity firms. The reports are designed to provide comprehensive insights and analysis on various companies and their market positions, financial performance, and growth potential.

Private equity firms rely on these reports to make informed investment decisions, strategize market entries, and manage portfolio companies effectively. The reports cover various aspects such as company overviews, product and service details, value propositions, key financial figures, and market trends.

A crucial point to emphasize is that **the entire process described ensures that information from unstructured web pages is automatically transformed into structured information to slides**. This 100% automated process leverages advanced data processing and natural language processing techniques to extract, analyze, and present relevant information in a structured and visually appealing format. This automation significantly enhances efficiency, accuracy, and the ability to handle large volumes of data, making it an invaluable tool for private equity advisory and decision-making.

Chapter 4

Results and discussion

In this chapter, we present and discuss the results obtained from the implementation of the three layers of the tool: processed data inputs (Layer 1), processing (Layer 2), and outputs (Layer 3). Each section will detail the outcomes, analyze the effectiveness of the methods employed, and address any challenges encountered during the development and execution phases.

4.1 Layer 1: Processed data inputs

4.1.1 Results

Quantitative data:

- Leverage existing financial data for several companies.
- Interpolated and extrapolated missing data points using XGBoost.
- Derived additional financial variables through algebraic operations.
- Stored the processed data in MongoDB, creating a robust and scalable database.

Before showing the financial information, some variables are briefly explained:

CAGR (LYA-Y3): Compound Annual Growth Rate over the last three years.

EBITDA (LYA): Earnings Before Interest, Taxes, Depreciation, and Amortization for the last year available.

Cost per employee (LYA): Average cost per employee for the last year available.

Cost per employee (Y3): Average cost per employee three years ago.

Finally, some of the more than 200 total variables are shown, indicating that missing values have been updated with the XGBoost algorithm.

Company Name	Employees (LYA)	Employees (Y1)	Employees (Y3)	Revenue (LYA) (\$)
SHELL	123,124	93,000	87,000	286,534,000
VOLKSWAGEN AG	?	643,297	662,575	259,859,000
UNIPER GLOBAL COMMODITIES SE	879	849	825	?
GLENCORE PLC	140,000	135,000	?	241,976,000
TOTALENERGIES SE	147,095	?	105,476	201,468,000
APPLE OPERATIONS INTERNATIONAL LIMITED	56,639	52,563	47,337	227,605,000
BP PLC	217,646	67,500	65,845	?
MERCEDES-BENZ GROUP AG	168,797	172,425	288,481	152,569,000

Table 4.1: Company data: employee numbers and revenue. Red cells with question marks indicate missing values.

Company Name	Employees (LYA)	Employees (Y1)	Employees (Y3)	Revenue (LYA) (\$)
SHELL	123,124	93,000	87,000	286,534,000
VOLKSWAGEN AG	646,837	643,297	662,575	259,859,000
UNIPER GLOBAL COMMODITIES SE	879	849	825	245,372,000
GLENCORE PLC	140,000	135,000	145,000	241,976,000
TOTALENERGIES SE	147,095	101,279	105,476	201,468,000
APPLE OPERATIONS INTERNATIONAL LIMITED	56,639	52,563	47,337	227,605,000
BP PLC	217,646	67,500	65,845	190,163,000
MERCEDES-BENZ GROUP AG	168,797	172,425	288,481	152,569,000

Table 4.2: Company data: employee numbers and revenue. Green cells indicate previously missing values now updated.

Company Name	CAGR (LYA-Y3) (%)	EBITDA (LYA) (\$)	Cost/Employee (LYA) (\$)	Cost/Employee (Y3) (\$)
SHELL	3.22	47,852	?	74.32
VOLKSWAGEN AG	?	38,654	72.50	68.40
UNIPER GLOBAL COMMODITIES SE	5.48	12,807	56.23	?
GLENCORE PLC	?	31,002	64.89	60.78
TOTALENERGIES SE	1.13	22,143	69.34	65.23
APPLE OPERATIONS INTERNATIONAL LIMITED	2.49	?	84.56	80.45
BP PLC	2.27	29,361	?	65.12
MERCEDES-BENZ GROUP AG	2.00	23,945	79.15	75.14

Table 4.3: Company data: financial metrics. Red cells with question marks indicate missing values.

Company Name	CAGR (LYA-Y3) (%)	EBITDA (LYA) (\$)	Cost/Employee (LYA) (\$)	Cost/Employee (Y3) (\$)
SHELL	3.22	47,852	78.12	74.32
VOLKSWAGEN AG	1.66	38,654	72.50	68.40
UNIPER GLOBAL COMMODITIES SE	5.48	12,807	56.23	52.11
GLENCORE PLC	4.44	31,002	64.89	60.78
TOTALENERGIES SE	1.13	22,143	69.34	65.23
APPLE OPERATIONS INTERNATIONAL LIMITED	2.49	41,206	84.56	80.45
BP PLC	2.27	29,361	69.23	65.12
MERCEDES-BENZ GROUP AG	2.00	23,945	79.15	75.14

Table 4.4: Company data: financial metrics. Green cells indicate previously missing values now updated.

Note that some of the values presented in the tables are fictitious in order to preserve privacy.

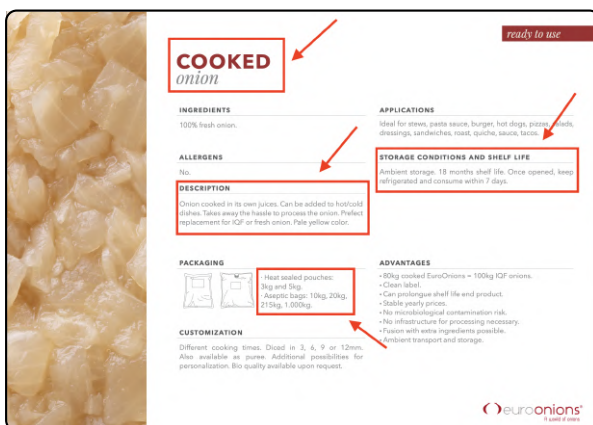
Qualitative data:

- Extracted detailed product and service information from various web sources.
- Categorized data into structured JSON files.
- Stored qualitative data in Pinecone for efficient retrieval.

Here are some examples of product and service categories from various companies.

Eurocebollas - Cooked onion

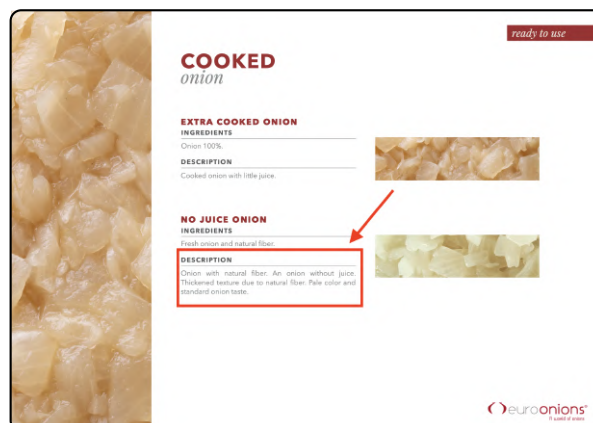
Below is a screenshot of the eurocebollas website. It can be seen how the information is unstructured and after passing through layer 1 (product and service workflows) it is structured in a json, ready to be saved in MongoDB for later use.



(a) Eurocebollas website number 1.



(b) Eurocebollas website number 2.



(c) Eurocebollas website number 3.

Figure 4.1: Eurocebollas website with unstructured information.

Here is the structured json:

```
1 {
2   "Product Basics": {
3     "Product Name": "Cooked Onion",
4     "Product Category": "Food Ingredients",
5     "Product Type": "Prepared Ingredients",
6     "SKU": "67890",
7     "Barcode/EAN/UPC": "0987654321098",
8     "Product Lifecycle Stage": "Mature",
9     "Target Audience": "Food Industry and Foodservice",
10    "Key Features": "Onion cooked in its own juices, can be added to
    hot/cold dishes, eliminates hassle of processing onion, perfect
    replacement for IQF or fresh onion",
11    "Unique Selling Proposition (USP)": "Simplifies production and
    supply processes, saves on personnel and energy costs,
    homogeneous ingredient, reduces hassle of processing fresh
    onions",
12    "Variants": {
13      "Sizes": [3, 5, 10, 20, 215, 1000],
14      "Colors": ["Pale yellow"],
15      "Flavors": []
16    }
17  },
18  "Physical Characteristics": {
19    "Dimensions": {
20      "Height": "Variable",
21      "Width": "Variable",
22      "Depth": "Variable"
23    },
24    "Weight": "Variable",
25    "Materials/Ingredients": "100% fresh onions",
26    "Texture": "Melt-in-your-mouth",
27    "Color Options": ["Pale Yellow"],
28    "Durability": "Long-lasting when stored properly",
29    "Assembly Requirements": "None",
30    "Maintenance Requirements": "Store in a cool, dry place",
31    "Safety Features": "None",
32    "Warranty/Guarantee Period": "6 months"
33  },
34  "Performance": {
35    "Efficiency": "High, reduces preparation time",
36    "Reliability": "Consistent quality",
37    "Compatibility": "Compatible with various dishes and cuisines",
38    "Scalability": "Easily scalable for large production needs",
39    "Battery Life": "N/A",
40    "Speed/Response Time": "Instant use",
41    "Capacity/Volume": "N/A",
42    "Energy Consumption": "N/A",
43    "Operating Conditions": {
44      "Temperature": "Room temperature",
45      "Humidity": "N/A"
46    },
47    "Shelf Life": "1 year"
48  },
49  "Brand and Manufacturer": {
50    "Brand Name": "EuroOnionsis",
```

```
51     "Manufacturer": "Eurocebollas",
52     "Country of Origin": "Spain",
53     "Brand Reputation": "Worldwide leader in the production of Ready-
        to-Use Ingredients"
54 },
55 "Pricing": {
56     "Retail Price": "N/A",
57     "Wholesale Price": "N/A",
58     "Discount Policies": "N/A",
59     "Financing Options": "N/A",
60     "Value for Money": "High, due to labor and cost savings"
61 }
62 }
```

Listing 4.1: Product details for Cooked Onion.

Legalitas - Legal support services

Servicios empresa

Favorito

Contigo Empresa

La tranquilidad de tenerlo todo controlado

- Resolvemos todas tus dudas legales. Sin límite. 24h si tu caso es urgente.
- Revisamos tus contratos y redactamos los documentos que necesitas.
- Cancelamos por ti los servicios y seguros que quieras dar de baja.
- Gestionamos tus reclamaciones con clientes y proveedores, negociando por ti cuando sea necesario.
- Te ayudamos a recuperar el dinero que te deben tus clientes.
- Tu gestor personal se ocupa de llevar tu contabilidad al día, asesorarte y gestionar tus obligaciones fiscales.
- Gestionamos tus alertas y notificaciones con Hacienda.
- Plataforma online donde elaborar presupuestos, enviar facturas, ver tus KPI, hablar con tu gestor y llevar tu agenda.
- Elaboramos y presentamos tus impuestos, libros oficiales y cuentas anuales.
- Realizamos tus solicitudes de certificados o modificación de datos en Hacienda.
- Y además, para que también estés protegido en tu vida personal, tienes incluido Contigo Premium.

112,00€ / mes + IVA. Servicio anual

Ahora 22% descuento. ¡Por 85€ al mes + IVA!

Descubre Contigo Empresa

Servicios empresa

Asesoría Legal Online

Cuenta con un abogado siempre que lo necesitas

- Resolvemos todas tus dudas legales. Sin límite. 24h si tu caso es urgente.
- Revisamos tus contratos y redactamos los documentos que necesitas.
- Cancelamos por ti los servicios y seguros que quieras dar de baja.
- Gestionamos tus reclamaciones con clientes y proveedores, negociando por ti cuando sea necesario.
- Te ayudamos a recuperar el dinero que te deben tus clientes.
- Tu gestor personal se ocupa de llevar tu contabilidad al día, asesorarte y gestionar tus obligaciones fiscales.
- Gestionamos tus alertas y notificaciones con Hacienda.
- Plataforma online donde elaborar presupuestos, enviar facturas, ver tus KPI, hablar con tu gestor y llevar tu agenda.
- Elaboramos y presentamos tus impuestos, libros oficiales y cuentas anuales.
- Realizamos tus solicitudes de certificados o modificación de datos en Hacienda.
- Y además, para que también estés protegido en tu vida personal, tienes incluido Contigo Premium.

45,00€ / mes + IVA. Servicio anual

Descubre Asesoría Legal Online

Servicios empresa

Gestoría Online Fiscal y Contable

Expertos en gestionar tu negocio

- Resolvemos todas tus dudas legales. Sin límite. 24h si tu caso es urgente.
- Revisamos tus contratos y redactamos los documentos que necesitas.
- Cancelamos por ti los servicios y seguros que quieras dar de baja.
- Gestionamos tus reclamaciones con clientes y proveedores, negociando por ti cuando sea necesario.
- Te ayudamos a recuperar el dinero que te deben tus clientes.
- Tu gestor personal se ocupa de llevar tu contabilidad al día, asesorarte y gestionar tus obligaciones fiscales.
- Gestionamos tus alertas y notificaciones con Hacienda.
- Plataforma online donde elaborar presupuestos, enviar facturas, ver tus KPI, hablar con tu gestor y llevar tu agenda.
- Elaboramos y presentamos tus impuestos, libros oficiales y cuentas anuales.
- Realizamos tus solicitudes de certificados o modificación de datos en Hacienda.
- Y además, para que también estés protegido en tu vida personal, tienes incluido Contigo Premium.

100,00€ / mes + IVA. Servicio anual

Descubre Gestoría Online

Asesoría legal y gestoría online para el día a día de tu empresa

¿Eres emprendedor? ¿Tienes una empresa o eres autónomo? Sea cual sea tu negocio, te ayudamos.

Para empresas

Te ayudamos a tomar las mejores decisiones.

Para autónomos

Te ayudamos a tenerlo todo en tu día a día. Ágil y sencillo.

Descubre cómo podemos ayudarte para que tú puedas dedicarte enteramente a tu negocio

Dinos cuál es tu situación.

Quiero montar mi propia empresa

Ya tengo un negocio

¿Qué subvenciones puedo pedir?

Quiero pedir el pago único del paro

Para empresas

¿Estás pensando en montar tu propia empresa?

Enhorabuena. Es una gran decisión. Seguro que llevas semanas preparando tu modelo de negocio, analizando a tus clientes potenciales, estudiando financiación... Y en Legalitas podemos ayudarte porque ahora, más que nunca, es muy importante que estés asesorado por verdaderos expertos.

¿Sabes qué tipo de empresa o sociedad quieres crear? ¿Qué pasos seguir y qué documentación necesitas? ¿Y cómo gestionar los trámites de constitución o tus obligaciones con la agencia tributaria? No te preocupes, con Legalitas crear tu empresa nunca había sido tan fácil. Te ayudamos.

(a) Legalitas website number 1.

(b) Legalitas website number 2.

documentos que necesitas.

- Cancelamos por ti los servicios y seguros que quieras dar de baja.
- Gestionamos tus reclamaciones con clientes y proveedores, negociando por ti cuando sea necesario.
- Tu gestor personal se ocupa de llevar tu contabilidad al día, asesorarte y gestionar tus obligaciones fiscales.
- Gestionamos tus alertas y notificaciones con Hacienda.
- Plataforma online donde elaborar presupuestos, enviar facturas, ver tus KPI, hablar con tu gestor y llevar tu agenda.
- Elaboramos y presentamos tus impuestos, libros oficiales y cuentas anuales.
- Realizamos tus solicitudes de certificados o modificación de datos en Hacienda.
- Y además, para que también estés protegido en tu vida personal, tienes incluido Contigo Plus.

75,00€ / mes + IVA. Servicio anual

Ahora 21% descuento. ¡Por 59€ al mes + IVA!

Descubre el Contigo Autónomo

documentos que necesitas.

- Cancelamos por ti los servicios y seguros que quieras dar de baja.
- Gestionamos tus reclamaciones con clientes y proveedores, negociando por ti cuando sea necesario.
- Tu gestor personal se ocupa de llevar tu contabilidad al día, asesorarte y gestionar tus obligaciones fiscales.
- Gestionamos tus alertas y notificaciones con Hacienda.
- Plataforma online donde elaborar presupuestos, enviar facturas, ver tus KPI, hablar con tu gestor y llevar tu agenda.
- Elaboramos y presentamos tus impuestos, libros oficiales y cuentas anuales.
- Realizamos tus solicitudes de certificados o modificación de datos en Hacienda.
- Y además, para que también estés protegido en tu vida personal, tienes incluido Contigo Plus.

35,00€ / mes + IVA. Servicio anual

Descubre el Asesoría Legal Online

documentos que necesitas.

- Cancelamos por ti los servicios y seguros que quieras dar de baja.
- Gestionamos tus reclamaciones con clientes y proveedores, negociando por ti cuando sea necesario.
- Tu gestor personal se ocupa de llevar tu contabilidad al día, asesorarte y gestionar tus obligaciones fiscales.
- Gestionamos tus alertas y notificaciones con Hacienda.
- Plataforma online donde elaborar presupuestos, enviar facturas, ver tus KPI, hablar con tu gestor y llevar tu agenda.
- Elaboramos y presentamos tus impuestos, libros oficiales y cuentas anuales.
- Realizamos tus solicitudes de certificados o modificación de datos en Hacienda.
- Y además, para que también estés protegido en tu vida personal, tienes incluido Contigo Plus.

60,00€ / mes + IVA. Servicio anual

Descubre el Gestoría Online

+ 20

20 años de experiencia

20 años de experiencia, capital 100% español. Red Nacional de 340 despachos. Legaltech española líder en asesoramiento jurídico para familias autónomas y pymes.

+1M

1 millón de consultas al año

Más de 800 abogados en toda España, 300.000 clientes individuales y 10M o los que prestamos servicio a través de importantes compañías.

24/365

24h al día los 365 días del año

Atención telefónica de 09:00 a 21:00 horas de lunes a viernes y un servicio de urgencia para aquellos asuntos legales que no puedan esperar.

(c) Legalitas website number 3.

Figure 4.2: Legalitas website with unstructured information.

Here is the structured json:

```
1 {
2   "Category Details": {
3     "Category Name": "Legal Support Services",
4     "Service Names": [
5       "Legal Advisory for Companies",
6       "Online Tax and Accounting Management",
7       "Online Freelancer Registration",
8       "With You Company",
9       "With You Freelancer"
10    ],
11    "Target Audience": "Companies, freelancers, and entrepreneurs",
12    "Key Features": "Comprehensive legal and management advisory,
13                    company and freelancer registration management, contract review
14                    , debt recovery, online platform for accounting and tax
15                    management.",
16    "Unique Selling Proposition (USP)": "Global legal and management
17                    advisory service, 24/7 support, discounts on annual services.",
18    "Typical Service Duration": "Annual",
19    "Delivery Methods": "In-person and online through a digital
20                    platform, phone, and WhatsApp.",
21    "Degree of Customization": {
22      "Level of Adaptability": "High, with personalized programs
23      based on client needs.",
24      "Level of Interaction": "High, including personalized attention
25      and urgent services.",
26      "Degree of Standardization": "Moderate, with standard services
27      that can be customized."
28    },
29    "Performance and Agreements": {
30      "Support Channels": "Personal managers, 24/7 customer service
31      via phone and WhatsApp.",
32      "Typical Success Criteria": "Efficient business management,
33      cost reduction, and compliance with legal standards."
34    },
35    "Pricing": {
36      "Typical Pricing Model": "Subscription-based or one-time
37      payment per service.",
38      "Typical Price Range": "Variable, depending on the complexity
39      and volume of services required."
40    }
41  }
42 }
```

Listing 4.2: Service details for Legal Support Services.

Universal Music - Music and entertainment services.



Figure 4.3: Universal Music website with unstructured information.

Here is the structured json:

```
{
  "Category Details": {
    "Category Name": "Music and Entertainment Services",
    "Service Names": [
      "Music Production",
      "Music Distribution",
      "Music Publishing",
      "Merchandise",
      "Artist Management",
      "Concerts and Live Events",
      "Music Licensing"
    ],
    "Target Audience": "Music Lovers, Artists, Record Labels, Media Outlets",
    "Key Features": "Comprehensive music production, global distribution, publishing rights management, artist merchandising, extensive artist management, live event organization, music licensing for media and advertising."
  }
}
```

```

15     "Unique Selling Proposition (USP)": "Global leader in music-based
      entertainment, diverse portfolio of artists, innovative digital
16         services, and extensive global reach.",
17     "Typical Service Duration": "Varies by service (e.g., ongoing for
      management, per event for concerts)",
18     "Delivery Methods": "In-person for live events, digital platforms
      for music distribution and licensing, personalized for artist
19         management.",
20     "Degree of Customization": {
21         "Level of Adaptability": "High, with personalized programs for
      artists and tailored solutions for media outlets.",
22         "Level of Interaction": "High, including personalized attention
      for artists and regular interaction with media and
23         advertising partners.",
24         "Degree of Standardization": "Moderate, with standard music
      distribution and licensing processes that can be customized.
25     },
26     "Performance and Agreements": {
27         "Support Channels": "Dedicated artist managers, customer
      service for music consumers, and business support for media
28         partners.",
29         "Typical Success Criteria": "Successful artist careers, high-
      quality music production, extensive global distribution, and
30         strong media partnerships."
31     },
32     "Pricing": {
33         "Typical Pricing Model": "Revenue share for music sales and
      streaming, fixed fees for live events and merchandise,
34         licensing fees for media usage.",
35         "Typical Price Range": "Variable, depending on the service type
      and artist popularity."
36     }
37 }

```

Listing 4.3: Service details for Music and Entertainment Services.

4.1.2 Discussion

Layer 1 is the most crucial layer in the architecture, as it forms the core foundation of the market reports. By integrating both quantitative and qualitative data inputs, Layer 1 ensures that the necessary information is available for comprehensive market analysis. The quantitative data provides a numerical basis for evaluating company performance, while the qualitative data enriches the reports with detailed descriptions of products and services. These two inputs, once processed and transformed through Layer 2 (Chaining) and Layer 3 (output), populate the slides of the market reports. Thus, the inputs from Layer 1 can be considered the nucleus of the market reports. The challenges and solutions that we have encountered during the process have been detailed in [3.2](#)

4.2 Layer 2: Processing

4.2.1 Results

Layer 2 focuses on processing structured data inputs to generate meaningful insights. The chaining process, implemented on the n8n platform, is crucial in transforming raw data into valuable market intelligence. This section provides an overview of the intermediate steps and methodologies used, serving primarily as a nexus in the overall process. The detailed results here are limited and consist mainly of text that will later be used to populate the slides. Comprehensive results and true insights derived from this process will be showcased in section [4.3](#).

Eurocebollas:

Title slide: Customers metrics - Eurocebollas

Measure:

Market Share in Targeted Segments Percent of Unprofitable Customers Number of Referrals to New Customers

Rationalize:

Cost per New Customer Acquired Percent of Customer Queries Not Satisfied by Initial Response Revenue from Cross-Market Activities

Improve:

Time to Resolve Customer Concerns or Complaints Quality Ratings from Premium Customers

Automate:

Automate CRM workflows to drive Profit Contribution by Segment Automate customer service queries resolution to drive Time to Resolve Customer Concerns or Complaints

Automate marketing analytics to drive Customer Response Rate to Campaigns

Legálitas:

Title slide: Operations metrics - Legálitas

Measure:

Matter Workflow Efficiency: Track and analyze the average time spent at each stage of case processing (e.g., client intake, document review, case preparations).

Tech Utilization Rate: Measure the percentage utilization of AI-driven tools and software in daily legal tasks (e.g., document review, research).

Error Rate in Automated Processes: Track the frequency of errors in processes where automation is employed (e.g., conflict-of-interest checks, compliance tracking).

Rationalize:

Operational Efficiency: Analyze and understand the full scope of time and resource utilization across different departments and processes.

Supplier Innovation Alignment: Assess the contribution of supplier innovation to Legálitas's service capabilities and competitive edge.

Documentation Quality Management: Evaluate the processes involved in document management, focusing on version control, error rates, and access rights.

Improve:

Improve Process Efficiency through Workflow Automation

Boost Transparency through Client Communication Portals

Enhance Expertise Utilization through Advanced Legal Research Tools

Improve Document Handling through Smart Document Management

Automate:

Automate Client Conflict-of-Interest Checks

Automate Document Drafting and Review

Automate Compliance Tracking

Automate Time Tracking and Billing

Universal music:

Title slide: Universal Music Group's competitive advantage

Strong Human Capital:

- Led by industry veteran Sir Lucian Grainge
- Workforce marked by versatile roles across business functions
- Commitment to driving equality
- Enhances their human capital strength

Comprehensive Product Portfolio:

- UMG boasts a diversified product range
- Includes music production, distribution, promotion services

Effective Market Strategy:

- UMG's market strategy is powered by analytics
- Predictive sales models
- Proprietary media and data platform
- Ensures effective market reach and high digital engagement

Customer Service Excellence:

- UMG's direct-to-fan tools provide fan analytics
- Customized experiences
- Caters to customers' music consumption preferences

Operational Efficiency:

- Successful strategic alliances and acquisitions
- Broad global operational base
- Effective resource management

4.2.2 Discussion

The Chaining process proved to be highly effective in automating data processing and generating insights. By leveraging the n8n platform, we were able to create scalable and flexible workflows that handle complex data transformations. The modifications made to the existing Chaining system ensured its applicability to our specific use cases. Despite the success, some challenges were encountered, such as optimizing workflow performance and handling edge cases in data processing.

4.3 Layer 3: Outputs

4.3.1 Results

Layer 3 is focused on generating visual and textual outputs from the processed data. As the final phase, emphasise that what is important and worthwhile in the whole process is to collect **dispersed information from the internet and directing it through a complex process to produce market analysis slides**. The automated generation of market reports significantly reduces the manual effort traditionally required and provides timely insights for strategic decision-making.

Here are some of the slides attached:

Legálitas

transformx 4 - PORTFOLIO VALUE CREATION DEEP DIVE

Operations Metrics - Legálitas

Legálitas Selected Interventions - Operations

	MEASURE	RATIONALIZE	IMPROVE	AUTOMATE
OPERATIONS	<ul style="list-style-type: none">Matter Workflow Efficiency: Track and analyze the average time spent at each stage of case processing (e.g., client intake, document review, case preparations).Tech Utilization Rate: Measure the percentage utilization of AI-driven tools and software in daily legal tasks (e.g., document review, research).Error Rate in Automated Processes: Track the frequency of errors in processes where automation is employed (e.g., conflict-of-interest checks, compliance tracking).	<ul style="list-style-type: none">Operational Efficiency: Analyze and understand the full scope of time and resource utilization across different departments and processes.Supplier Innovation Alignment: Assess the contribution of supplier innovation to Legálitas's service capabilities and competitive edge.Documentation Quality Management: Evaluate the processes involved in document management, focusing on version control, error rates, and access rights	<ul style="list-style-type: none">Improve Process Efficiency through Workflow Automation:Boost Transparency through Client Communication PortalsEnhance Expertise Utilization through Advanced Legal Research Tools:Improve Document Handling through Smart Document Management:	<ul style="list-style-type: none">Automate Client Conflict-of-Interest Checks.Automate Document Drafting and Review.Automate Compliance TrackingAutomate Time Tracking and Billing

Source: transformx, APRIL 2024 1

Figure 4.4: One of the slides of the Legálitas market report

Eurocebollas

transformx 4 - PORTFOLIO VALUE CREATION DEEP DIVE

Customers Metrics - Eurocebollas

Eurocebollas Selected Interventions - Customers

	MEASURE	RATIONALIZE	IMPROVE	AUTOMATE
CUSTOMERS	<ul style="list-style-type: none">Market Share in Targeted SegmentsPercent of Unprofitable CustomersNumber of Referrals to New Customers	<ul style="list-style-type: none">Cost per New Customer AcquiredPercent of Customer Queries Not Satisfied by Initial ResponseRevenue from Cross-Market Activities	<ul style="list-style-type: none">Time to Resolve Customer Concerns or ComplaintsQuality Ratings from Premium Customers	<ul style="list-style-type: none">Automate CRM workflows to drive Profit Contribution by SegmentAutomate customer service queries resolution to drive Time to Resolve Customer Concerns or ComplaintsAutomate marketing analytics to drive Customer Response Rate to Campaigns

Source: transformx, APRIL 2024 1

(a) One of the slides of the Eurocebollas market report.

transformx 4 - PORTFOLIO VALUE CREATION DEEP DIVE

Innovation Metrics - Eurocebollas

Eurocebollas Selected Interventions - Innovation

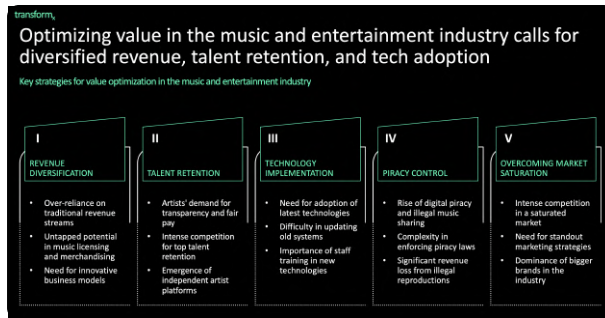
	MEASURE	RATIONALIZE	IMPROVE	AUTOMATE
INNOVATION	<ul style="list-style-type: none">Number of New Projects Launched Based on Client InputNumber of Life-Cycle Extension ProjectsActual Versus Desired Mix of Projects	<ul style="list-style-type: none">Number of New Projects or Concepts Presented for DevelopmentActual Versus Budgeted Spending on Projects at Each Development Stage	<ul style="list-style-type: none">Average Time Spent by Projects at the Development, Test, and Launch StagesNumber of Projects Delivered on TimeTotal Time from Concept to Market	<ul style="list-style-type: none">Automate project tracking systems to drive Number of Projects Delivered on TimeAutomate customer feedback collection and analysis to drive Number of New Projects Launched Based on Client Input

Source: transformx, APRIL 2024 1

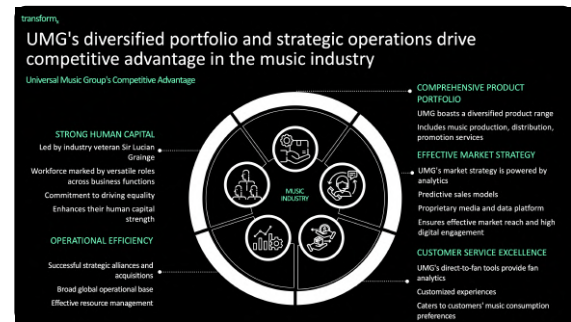
(b) Other of the slides of the Eurocebollas market report.

Figure 4.5: Slides of Eurocebollas market reports.

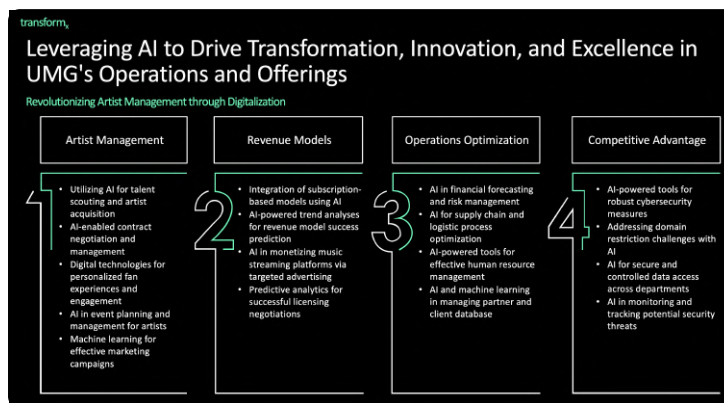
Universal Music



(a) One of the slides of the Universal Music market report.



(b) Other of the slides of the Universal Music market report.



(c) Other of the slides of the Universal Music market report.

Figure 4.6: Slides of Universal Music market reports.

4.3.2 Discussion

The ability to generate automated market reports is a significant advancement in improving efficiency and reducing manual effort. The text-to-slide component played a crucial role in seamlessly integrating textual insights into visual presentations. This automation not only saves time but also ensures consistency and accuracy in reporting. Challenges included ensuring the relevance and clarity of the generated slides, which were addressed through iterative improvements and user feedback.

Chapter 5

Metrics and evaluation of results

5.1 Introduction

Measuring and evaluating the performance of the models implemented in this project is crucial for ensuring the quality and reliability of the results. This chapter presents the metrics, key performance indicators (KPIs), and evaluation methodologies for both the generative AI models and the XGBoost model. The importance of the ground truth as a benchmark for validation is also highlighted.

It is important to note that, although the real data is not accessible for this project due to a change of company, the metrics and evaluation methodologies discussed here are grounded in the theoretical framework and best practices in artificial intelligence and machine learning that should have been followed during the development of this project.

5.2 Metrics and evaluation for generative AI models

5.2.1 Logprobs and Perplexity

Logprobs

Logprobs, short for "logarithmic probabilities," represent the logarithm (usually on a natural basis) of the probabilities assigned by a language model to each token it generates. When a large language model (LLM) predicts the next token in a sequence, it assigns a probability to each possible token based on its training data and learned patterns. The log probability of a token is defined as:

$$\text{logprob} = \ln(p(w_i)) \tag{5.1}$$

where w_i is the i -th token, and $p(w_i)$ is the probability assigned to that token by the model.

For example, consider a model generating the phrase "Processes automation through workflow". Using [5.1](#), the model might assign the following probabilities to each token:

- **Token 1 ("Processes"):** $P(\text{"Processes"}) = 0.3 \Rightarrow \text{logprob} = \ln(0.3) \approx -1.20$
- **Token 2 ("automation"):** $P(\text{"automation"}) = 0.4 \Rightarrow \text{logprob} = \ln(0.4) \approx -0.92$

- **Token 3 ("through"):** $P(\text{"through"}) = 0.2 \Rightarrow \text{logprob} = \ln(0.2) \approx -1.61$
- **Token 4 ("workflow"):** $P(\text{"workflow"}) = 0.1 \Rightarrow \text{logprob} = \ln(0.1) \approx -2.30$

The logprob values are particularly useful in evaluating the confidence of the model in its predictions. Lower logprob values (closer to zero) indicate higher confidence.

Currently, "logprobs" parameter is accessible via the OpenAI API with an additional layer provided by Azure, for example. However, OpenAI has announced plans to integrate this feature directly into their public API in the future, making it more widely available.

Perplexity

Perplexity is a metric used to evaluate the quality of a language model by measuring how well it predicts a sequence of tokens. It is defined as the inverse probability of the test set normalized by the number of tokens, or equivalently, as the exponent of the negative average log probability:

$$\text{Perplexity} = e^{-\frac{1}{N} \sum_{i=1}^N \ln(p(w_i))} \quad (5.2)$$

where:

- N : number of tokens in the sequence.
- $p(w_i)$: probability assigned by the model to the i -th token.
- $\ln(p(w_i))$: the natural logarithm of the probability assigned by the model to the i -th token (*logprob*)

The perplexity value can be interpreted as the effective "branching factor" of the model. A lower perplexity indicates that the model is more confident and accurate in its predictions. For example, if a model generates the sequence "Processes automation through workflow" with the probabilities specified earlier, the perplexity would be calculated, following [5.2](#), as:

$$\begin{aligned} \text{Perplexity} &= e^{-\frac{1}{4} [\ln(0.3) + \ln(0.4) + \ln(0.2) + \ln(0.1)]} \Rightarrow \\ &\Rightarrow e^{-\frac{1}{4} [-1.20 - 0.92 - 1.61 - 2.30]} \Rightarrow \\ &\Rightarrow e^{-\frac{1}{4} \times -6.03} = e^{1.5075} \approx 4.51 \end{aligned}$$

This value of perplexity reflects the effective "branching factor" of the model, which can be interpreted as the number of reasonable options the model considers at each step of token prediction. A branching factor of 4.51 means that, on average, the model is effectively choosing between approximately 4.51 plausible tokens for each position in the sequence. Lower perplexity indicates greater certainty, as the model focuses on fewer high-probability options, whereas higher perplexity implies greater uncertainty or a wider distribution of probabilities across many tokens.

Domains and thresholds

Although perplexity is not currently available directly via the OpenAI API, it can be computed using the formula above based on the log probabilities of tokens. Implementing perplexity thresholds is highly useful for monitoring the performance of agents, swarms of agents, and modules like the Keeper in a system architecture. For example:

- Define a perplexity threshold, such as $\text{Perplexity} < 10$, to ensure the generated responses remain coherent and meaningful.
- Trigger corrective actions if perplexity exceeds a specified limit, such as refining prompts, reinitializing agents, or incorporating additional context into the system.

The domains of these metrics are:

- **Logprobs:** defined for probabilities in the range $(0, 1]$, with log probabilities $(-\infty, 0]$.
- **Perplexity:** typically, $\text{Perplexity} > 1$.

This is the the graphic with both functions ($g=\text{perplexity}$, $f=\text{logprob}$):

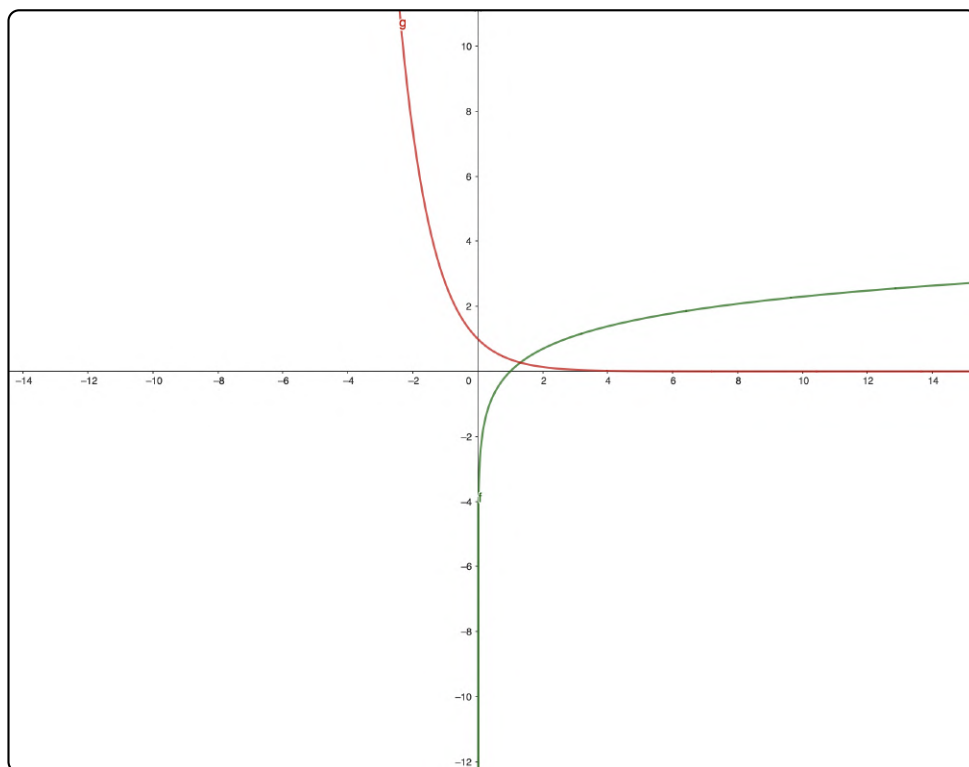


Figure 5.1: Logprobs and Perplexity representation.

Setting thresholds for these metrics can help detect and manage anomalies in real-time, enhancing the robustness and reliability of agent-based systems.

5.2.2 BLEU and ROUGE metrics

The evaluation of generated text against a reference or ground truth is a critical aspect of assessing the quality of generative models. Two widely used metrics in natural language processing for this purpose are BLEU and ROUGE. These metrics evaluate different aspects of the generated text, such as its precision, recall, and overall fidelity to the reference text.

BLEU (Bilingual Evaluation Understudy)

BLEU is a widely used metric for evaluating the quality of text generated by a machine learning model, especially in machine translation tasks. It compares the generated text (candidate) with a reference text based on overlapping n-grams, with a focus on precision.

BLEU measures how similar the generated text is to the reference by comparing n-grams (sequences of words of length n). The more n-grams overlap, the higher the BLEU score. To ensure that the candidate text is not excessively short, a brevity penalty (BP) is applied.

The BLEU score is calculated using the following formula:

$$\text{BLEU} = \text{BP} \cdot e^{\sum_{n=1}^N w_n \ln p_n} \quad (5.3)$$

Where:

- BP (Brevity Penalty): penalizes candidates shorter than the reference.
- w_n : weight assigned to each n-gram precision (commonly, all w_n are equal, e.g., $w_n = \frac{1}{N}$).
- p_n : n-gram precision, calculated as:

$$p_n = \frac{\text{Number of overlapping n-grams}}{\text{Total number of n-grams in the candidate}} \quad (5.4)$$

Brevity Penalty: the brevity penalty ensures that overly short candidates are penalized. It is defined as:

$$\text{BP} = \begin{cases} 1, & \text{if } c > r \\ e^{1-\frac{r}{c}}, & \text{if } c \leq r \end{cases} \quad (5.5)$$

Where:

- c : length of the candidate (generated) text.
- r : length of the reference text.

Example (with unigrams and bigrams):

1. **Reference text:** "Improve process efficiency through workflow automation."

2. **Candidate text:** "Improve process efficiency with workflow automation."

Step 1: calculate unigram precision:

- Unigrams in reference: {"Improve", "process", "efficiency", "through", "workflow", "automation"}
- Unigrams in candidate: {"Improve", "process", "efficiency", "with", "workflow", "automation"}
- Overlapping unigrams: {"Improve", "process", "efficiency", "workflow", "automation"}
- Precision (p_1): $\frac{5}{6} \approx 0.833$

Step 2: calculate bigram precision:

- Bigrams in reference: {"Improve process", "process efficiency", "efficiency through", "through workflow", "workflow automation"}
- Bigrams in candidate: {"Improve process", "process efficiency", "efficiency with", "with workflow", "workflow automation"}
- Overlapping bigrams: {"Improve process", "process efficiency", "workflow automation"}
- Precision (p_2): $\frac{3}{5} = 0.6$

Step 3: calculate brevity penalty:

- Candidate length (c): 6 tokens.
- Reference length (r): 6 tokens.

Since $c = r$, we have:

$$\text{BP} = 1$$

Step 4: combine results:

Using equal weights ($w_1 = w_2 = 0.5$) and following [5.3](#):

$$\begin{aligned} \text{BLEU} &= \text{BP} \cdot e^{\frac{1}{2} \ln(0.833) + \frac{1}{2} \ln(0.6)} \Rightarrow \\ &\Rightarrow \text{BLEU} = e^{0.5 \cdot \ln(0.833) + 0.5 \cdot \ln(0.6)} \\ \ln(0.833) &\approx -0.183, \quad \ln(0.6) \approx -0.511 \end{aligned}$$

Therefore:

$$\begin{aligned} \text{BLEU} &= e^{0.5 \cdot (-0.183) + 0.5 \cdot (-0.511)} \Rightarrow \\ &\Rightarrow \text{BLEU} = e^{-0.347} \approx 0.707 \end{aligned}$$

Limitations:

- BLEU does not account for synonyms or paraphrasing; it requires exact matches of n-grams.
- It penalizes valid but creative phrasing that differs from the reference text.
- It is a purely lexical metric and does not consider semantic meaning or context beyond n-grams.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

ROUGE is a widely used metric for evaluating the quality of text generation tasks, especially in summarization. Unlike BLEU, which focuses on **precision**, ROUGE emphasizes **recall**. It measures how much of the content in the reference text is captured in the generated text by comparing overlapping units such as n-grams, sequences, or even sentence-level structures

In BLEU, the denominator is the **total number of n-grams generated** by the model (candidate), focusing on precision. In ROUGE, the denominator is the **total number of n-grams in the reference text**, emphasizing recall.

ROUGE evaluates the quality of generated text by measuring the overlap between the generated text (candidate) and the reference text. The focus is on recall, assessing how much of the reference's content is reproduced in the candidate.

There are several variants of ROUGE, each tailored to different tasks:

- **ROUGE-N**: measures recall of overlapping n-grams between the candidate and reference.
- **ROUGE-L**: measures the longest common subsequence (LCS) between the candidate and reference. It captures sentence-level structure and coherence.
- **ROUGE-S**: measures the recall of skip bigrams, allowing for non-consecutive matches to evaluate broader semantic relationships.

For simplicity, we focus on ROUGE-1 (unigrams) and ROUGE-2 (bigrams). The formula for ROUGE-N is:

$$\text{ROUGE-N} = \frac{\text{Number of overlapping n-grams}}{\text{Total number of n-grams in the reference}} \quad (5.6)$$

Where:

- Number of overlapping n-grams : count of n-grams that appear in both the candidate and the reference.
- Total number of n-grams in the reference : total number of n-grams in the reference text.

Example (ROUGE-1 and ROUGE-2):

1. **Reference text**: "Improve process efficiency through workflow automation."
2. **Candidate text**: "Improve process efficiency with workflow automation."

Step 1: Calculate ROUGE-1 (unigrams):

- Unigrams in reference: {"Improve", "process", "efficiency", "through", "workflow", "automation"}
- Unigrams in candidate: {"Improve", "process", "efficiency", "with", "workflow", "automation"}

- Overlapping unigrams: {"Improve", "process", "efficiency", "workflow", "automation"}
- ROUGE-1:

$$\text{ROUGE-1} = \frac{\text{Number of overlapping unigrams}}{\text{Total number of unigrams in the reference}} = \frac{5}{6} \approx 0.833$$

Step 2: Calculate ROUGE-2 (bigrams):

- Bigrams in reference: {"Improve process", "process efficiency", "efficiency through", "through workflow", "workflow automation"}
- Bigrams in candidate: {"Improve process", "process efficiency", "efficiency with", "with workflow", "workflow automation"}
- Overlapping bigrams: {"Improve process", "process efficiency", "workflow automation"}
- ROUGE-2:

$$\text{ROUGE-2} = \frac{\text{Number of overlapping bigrams}}{\text{Total number of bigrams in the reference}} = \frac{3}{5} = 0.6$$

In this example:

- ROUGE-1 = 0.833 (83.3% of the unigrams in the reference are present in the candidate).
- ROUGE-2 = 0.6 (60% of the bigrams in the reference are present in the candidate).

Limitations:

- ROUGE, like BLEU, does not handle synonyms or paraphrasing well, as it relies on exact matches.
- It is biased towards longer texts since longer candidates have more chances of matching n-grams in the reference.
- ROUGE does not consider semantic meaning or grammatical correctness.

ROUGE is particularly effective for tasks like summarization, where recall is critical, and the goal is to capture as much of the reference content as possible. However, it should be complemented with other metrics, such as semantic similarity, to evaluate meaning and context.

5.2.3 Semantic Accuracy

Semantic accuracy evaluates how well the generated text aligns with the intended meaning or semantics of the reference text. Unlike BLEU and ROUGE, which focus on lexical overlap (exact matches of n-grams), semantic accuracy measures the conceptual similarity between texts. This is particularly important for tasks where flexibility in wording is allowed, such as open-ended generative tasks or dialogue systems.

To compute semantic accuracy, embeddings are used to represent the generated text and the reference text in a high-dimensional vector space. These embeddings encode the semantic meaning of the text, rather than its lexical content, enabling comparisons beyond exact word matches.

Commonly used embedding models include:

- **Word2Vec** and **GloVe**: capture word-level semantics.
- **Sentence Transformers** (e.g., BERT, SBERT): encode sentence-level meaning.

The similarity between the generated and reference texts is then computed using a metric like cosine similarity, which is defined as:

$$\text{Cosine Similarity} = \frac{\vec{A} \cdot \vec{B}}{\|\vec{A}\| \|\vec{B}\|} \quad (5.7)$$

Where:

- \vec{A} : Embedding vector of the generated text.
- \vec{B} : Embedding vector of the reference text.
- $\vec{A} \cdot \vec{B}$: Dot product of the two vectors.
- $\|\vec{A}\|, \|\vec{B}\|$: Magnitudes (norms) of the vectors.

Consider the following texts:

- **Reference text**: "Improve process efficiency through workflow automation."
- **Generated text**: "Improve process efficiency with workflow automation."

Step 1: generate embeddings:

- Using a sentence embedding model, compute the vector representations of both texts:
 - \vec{A} (reference): [0.75, 0.21, 0.33, ...].
 - \vec{B} (generated): [0.72, 0.20, 0.34, ...].

Step 2: compute cosine similarity: Using [5.7](#):

$$\text{Cosine Similarity} = \frac{\vec{A} \cdot \vec{B}}{\|\vec{A}\| \|\vec{B}\|}$$

$$\vec{A} \cdot \vec{B} = 0.75 \cdot 0.72 + 0.21 \cdot 0.20 + 0.33 \cdot 0.34 + \dots = 0.987$$

$$\|\vec{A}\| = \sqrt{0.75^2 + 0.21^2 + 0.33^2 + \dots}, \quad \|\vec{B}\| = \sqrt{0.72^2 + 0.20^2 + 0.34^2 + \dots}$$

Assuming $\|\vec{A}\| = 1.05$ and $\|\vec{B}\| = 1.02$:

$$\cos(\theta) = \frac{0.987}{1.05 \cdot 1.02} \approx 0.95$$

A cosine similarity of 0.95 indicates a very high semantic alignment between the reference and generated texts, despite differences in word choice or phrasing.

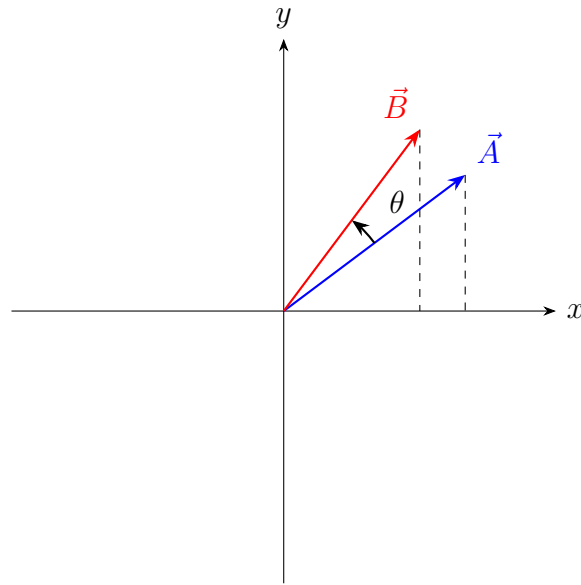


Figure 5.2: Illustration of cosine similarity in 2D. The angle θ determines the similarity between the vectors \vec{A} and \vec{B} .

Advantages:

- Captures semantic meaning beyond exact word matches, making it more robust to paraphrasing and synonym usage.
- Useful for evaluating open-ended or creative tasks where lexical overlap is less important.

Limitations:

- Heavily dependent on the quality of the embedding model; poor embeddings can lead to inaccurate similarity scores.
- Computationally more expensive compared to BLEU and ROUGE.
- Does not explicitly measure grammatical correctness or coherence.

Semantic accuracy provides a complementary perspective to BLEU and ROUGE by focusing on meaning rather than form. It is particularly suited for tasks like dialogue generation, summarization, and other applications where flexibility in expression is allowed, but the semantic intent must remain consistent.

5.2.4 Ground Truth

In the context of generative AI, the ground truth is defined as the reference data or expected content used to validate the quality and accuracy of AI-generated market reports. This ground truth should have been derived from:

- **Historical reports:** existing market studies and venture capital consulting documents served as benchmarks for evaluating the structure, tone, and content of the AI-generated reports.
- **Validated external data:** data from trusted financial and industry databases was used to cross-check the factual accuracy of generated insights, such as revenue or market trends.
- **Expert reviews:** domain experts reviewed a subset of generated reports, providing feedback on their accuracy, relevance, and adherence to consulting standards. In this case, Jorge Gómez was in charge of reviewing the information.

By comparing the AI-generated outputs to this ground truth, the system's performance was rigorously evaluated, ensuring reliability and relevance for venture capital applications.

5.2.5 Global KPIs for generative AI

To assess the overall impact of using generative AI for automating market reports, the following key performance indicators (KPIs) were defined:

- **Efficiency improvement:** the automation implemented reduced report execution time from 60 to 3 FTE days, decreasing this time on a 95%. This exceeded the predefined KPI of reducing report execution time by at least 90%, highlighting the success of the implemented solution.
- **Error rate in factual content:** monitored to ensure the accuracy of financial metrics and industry data generated by the AI system, with a target error rate below 5%. Evaluated using cosine similarity between the embeddings of AI-generated content and reference reports to ensure semantic alignment with ground truth data.

These KPIs provided a comprehensive evaluation framework, demonstrating the system's capability to deliver high-quality, relevant, and efficient market analyses for venture capital consultants.

5.3 Metrics and evaluation for XGBoost models

5.3.1 Introduction to XGBoost metrics

In this project, XGBoost was used to interpolate missing data in a dataset of companies that included financial and operational metrics obtained through web scraping. The dataset contained columns such as the number of employees, revenue, and total assets, but some values were missing due to inconsistencies in the original sources. XGBoost was

utilized to fill these missing values by identifying patterns in the available metrics.

The evaluation of XGBoost results should have been carried out using a set of metrics tailored for regression tasks to ensure that the predictions were consistent and accurate in the business context.

5.3.2 Key metrics for regression models

Mean Absolute Error (MAE)

MAE measures the average absolute errors between the real values (when available) and the values predicted by XGBoost:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (5.8)$$

For example, if the real number of employees in three companies was [100.000, 50.000, 80.000] and XGBoost predicted [95.000, 52.000, 85.000], the MAE is calculated as:

$$\text{MAE} = \frac{|100.000 - 95.000| + |50.000 - 52.000| + |80.000 - 85.000|}{3} = 4.000$$

Mean Squared Error (MSE) and Root Mean Squared Error (RMSE)

These metrics emphasize larger errors, penalizing them more severely than MAE, which is useful when avoiding significant deviations in predictions.

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (5.9)$$

$$\text{RMSE} = \sqrt{\text{MSE}} \quad (5.10)$$

Following the same example:

$$\text{MSE} = \frac{(100.000 - 95.000)^2 + (50.000 - 52.000)^2 + (80.000 - 85.000)^2}{3} = 18.000.000$$

$$\text{RMSE} = \sqrt{18.000.000} \approx 4.244$$

R-Squared (R^2)

R^2 measures the proportion of variability in the data that can be explained by the model. It is useful for assessing how well the model uses the available metrics to predict the missing values:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (5.11)$$

For example, if \bar{y} (mean of the real values) is 76.667, and the real and predicted values are the same as before, R^2 would be:

$$R^2 = 1 - \frac{(5.000)^2 + (-2.000)^2 + (5.000)^2}{(100.000 - 76.667)^2 + (50.000 - 76.667)^2 + (80.000 - 76.667)^2}$$

5.3.3 Ground Truth

In this project, the ground truth is defined as the actual or expected values used to validate the interpolations made by the XGBoost model. Specifically, the ground truth can be derived from the following sources:

- **Available data in other columns or rows:** for instance, if employee data for Company A in 2020 is missing, but data from 2019 and 2021 is available, the latter can serve as a reference.
- **External data comparisons:** if external sources, such as publicly available financial databases, provide values for the same companies and metrics, these can act as ground truth for validation purposes.

5.3.4 Global KPIs

To assess the overall impact of using XGBoost for missing data interpolation, the following key performance indicators (KPIs) were defined:

- **Percentage of completed cells:** 95% of the dataset was successfully completed.
- **Acceptable MAE:** for example, a MAE below 5.000 employees or \$1M in revenue was defined as a success criterion.
- **Execution time:** the complete interpolation of the dataset took less than 10 minutes, ensuring efficiency for large datasets.

Chapter 6

Future lines and conclusions

6.1 Future lines

Based on the results and discussions presented in the previous chapters, several areas for improvement and expansion have been identified for the future development of Bayesian_x's tool. Key future lines include:

- **Database expansion:** increase the number of companies in the MongoDB database to include one million companies, enhancing the breadth and depth of market analysis.
- **Workflow optimization:** refine and optimize existing workflows in n8n to increase efficiency and reduce processing time, including improving the performance of the Chaining process and integration with other tools and platforms.
- **NLP improvements:** continue enhancing the natural language processing models used for generating market reports, training more advanced and industry-specific models, and optimizing the context window to handle large data volumes effectively.
- **Automated data updates:** implement automated systems for periodic data updates to ensure the information used in market analyses is always current and relevant.

6.2 Conclusions

In this project, an innovative tool was developed and tested to automate the process of market studies and report generation using advanced artificial intelligence and machine learning technologies. Bayesian_x's tool has proven effective in collecting, processing, and analyzing large volumes of data, providing valuable insights for strategic decision-making.

The modular design of the tool, with its distinct layers for data input, processing, and output generation, has enabled seamless and efficient integration of various technologies and methods. From data collection through web scraping and APIs to advanced natural language processing models and automated report generation, each component has played a crucial role in the project's success.

A key aspect of the tool is the use of the Retrieval-Augmented Generation (RAG) system with Pinecone, compared to simply making calls to the OpenAI API. The RAG system with Pinecone offers several significant advantages:

- **Improved accuracy:** by retrieving relevant information from a structured knowledge base, RAG ensures that the generated outputs are accurate and based on verified data, reducing the likelihood of irrelevant or incorrect content.
- **Scalability:** Pinecone's efficient indexing and retrieval capabilities handle large volumes of data, ensuring the system remains scalable and robust. As the data volume grows, the retrieval process remains fast and efficient.
- **Context window limitation handling:** Large Language Models (LLMs) have a limited context window. RAG mitigates this limitation by retrieving only the most relevant documents to inform the generation process, ensuring detailed and contextually accurate responses without being hindered by the context window limitations of LLMs.

Throughout the development and testing of the tool, several challenges were faced and overcome, allowing for the identification of areas for future improvement and expansion. The proposed future lines ensure that the tool continues to evolve and improve, providing increasingly precise, efficient, and valuable market analyses.

Regarding our most direct competitors, tools like Crayon and Speak offer advanced solutions for competitive intelligence and market report generation, but my tool presents significant advantages over them. Crayon focuses on intelligent filtering and summarizing competitive intelligence data in real time, minimizing information overload by filtering out 99% of irrelevant data. However, my tool goes further by integrating both quantitative and qualitative data, providing a more comprehensive market view. Additionally, while Crayon automatically classifies and performs sentiment analysis on data, my tool organizes this data into detailed JSON structures, allowing for deeper customization in the reports. Furthermore, although Crayon offers anomaly detection and website tracking, my tool employs advanced techniques like Retrieval-Augmented Generation (RAG) with Pinecone, enhancing the accuracy and relevance of the retrieved data.

On the other hand, Speak automates data collection and report generation similarly to my tool. However, my system uses tools like n8n and Python for a more robust and scalable integration of workflows, ensuring more efficient and detailed data processing. Speak allows for customizable reports, but my tool has the additional advantage of incorporating processed financial and qualitative data using advanced algorithms like XGBoost to interpolate and extrapolate missing data. Moreover, while Speak provides real-time updates, my tool benefits from handling large data volumes and utilizing advanced Natural Language Processing (NLP) and RAG with Pinecone, improving the precision and relevance of the generated reports.

However, my tool also presents some disadvantages. The integration of multiple technologies and platforms, such as n8n, MongoDB, and Pinecone, can increase the complexity of the initial setup and ongoing maintenance. Additionally, the need for proprietary servers on DigitalOcean and the use of CapRover for management can impose additional

burdens in terms of resources and required technical skills. Although scalability issues have been addressed, handling a large number of companies still poses challenges in terms of performance and efficient memory usage.

In conclusion, the project has laid the foundation for a robust and scalable platform that can transform how market studies are conducted, providing significant value to private equity firms and other entities requiring detailed and reliable market analysis. Automating these processes not only reduces manual effort but also enhances the accuracy and relevance of generated insights, positioning Bayesian_x as a leader in automated market intelligence.

Appendix A

Additional information

This chapter introduces concepts that are interesting but a bit off topic and therefore not introduced in the main document.

A.1 First ideas and theories in the field of artificial intelligence

Santiago Ramón y Cajal and neurons: Santiago Ramón y Cajal was a pioneering neuroscientist whose work laid the foundation for our understanding of neural networks. In the late 19th and early 20th centuries, Ramón y Cajal used a staining technique developed by Camillo Golgi, known as the Golgi stain, to visualize neurons. This technique allowed him to observe the complex structures and connections of neurons in unprecedented detail. Through his meticulous drawings and detailed studies, Ramón y Cajal demonstrated that the nervous system is made up of individual cells, which he called neurons. His work revealed that neurons communicate with each other through specialized connections called synapses.

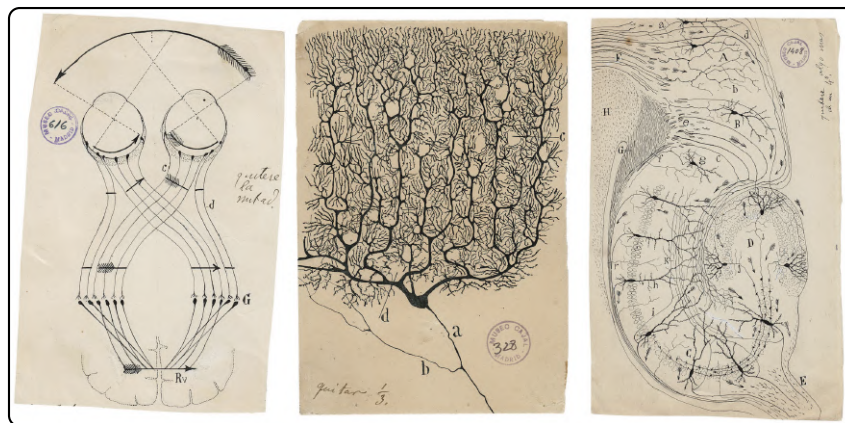


Figure A.1: From left to right: a diagram suggesting how the eyes might transmit a unified picture of reality to the brain, a Purkinje neuron in the human cerebellum, and a diagram showing the flow of information through the hippocampus. Diagram by S. Ramón y Cajal. Adapted from reference [\[1\]](https://www.nytimes.com/es/2017/02/21/espanol/cultura/santiago-ramon-y-cajal-el-hombre-que-dibujó-los-secretos-del-cerebro.html)

¹<https://www.nytimes.com/es/2017/02/21/espanol/cultura/santiago-ramon-y-cajal-el-hombre-que-dibujó-los-secretos-del-cerebro.html>

Ramón y Cajal's discoveries were groundbreaking and challenged the prevailing theory of his time, which posited that the nervous system was a continuous network of fibers. Instead, he showed that neurons are discrete entities that interact through contact points, a concept that became known as the "neuron doctrine". This discovery was crucial for the development of neuroscience and later influenced the field of artificial intelligence. In particular, the concept of neurons as individual processing units inspired the design of artificial neural networks, which are used in modern AI systems to process information in a way that mimics the human brain.

Ramón y Cajal's work highlighted the importance of understanding how neurons communicate and process information. This understanding is fundamental to both biological and artificial neural networks. By mapping the intricate networks of neurons, Ramón y Cajal provided a blueprint for researchers in AI to develop models that emulate the brain's ability to learn and adapt. His contributions laid the groundwork for future research in both neuroscience and AI, bridging the gap between biology and technology

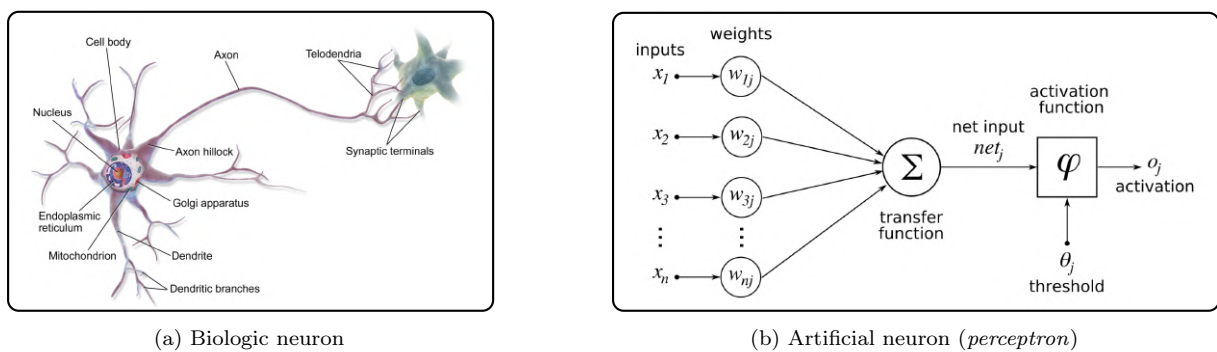


Figure A.2: Comparison between a biological neuron and an artificial neuron. Adapted from reference ²

Alan Turing and the Turing test: Alan Turing, a British mathematician and logician, made significant contributions to the field of computer science and artificial intelligence. In 1950, Turing published a seminal paper titled "Computing Machinery and Intelligence", in which he proposed the idea of a machine that could exhibit intelligent behavior indistinguishable from that of a human. To test this hypothesis, Turing introduced the famous "Turing Test," which has become a fundamental concept in AI.

In the Turing Test, a human evaluator interacts with both a machine and a human through a computer interface, without knowing which is which. The evaluator engages in a natural language conversation with both participants, asking questions and receiving responses. If the evaluator cannot reliably distinguish the machine from the human based on their responses, the machine is said to have demonstrated intelligent behavior. The Turing Test was groundbreaking because it shifted the focus of AI from the question of whether machines could think to whether machines could convincingly imitate human intelligence.

Turing's work laid the foundation for the development of modern AI by emphasizing the importance of natural language processing and human-computer interaction. The Turing Test remains a benchmark for assessing a machine's ability to exhibit human-like

²https://afit-r.github.io/ann_fundamentals

intelligence. Although no machine has yet passed the Turing Test in a manner that satisfies all critics, the test continues to inspire research and debate in the field of AI.

Turing's contributions extend beyond the Turing Test. He also developed the concept of the "universal Turing machine," a theoretical device that can simulate the logic of any computer algorithm. This concept is fundamental to computer science and underpins the design of modern computers. Turing's insights into the nature of computation and intelligence have had a lasting impact on the development of AI, and his ideas continue to influence researchers and engineers as they strive to create machines that can think and learn like humans.

A.2 Vector databases and embeddings: a primer on Pinecone

In the realm of advanced data management and retrieval, vector databases like Pinecone have emerged as crucial tools. These databases leverage the power of embeddings to enhance the efficiency and accuracy of data retrieval processes. Here, we explore the fundamental concepts and functionalities of Pinecone and its role in handling embeddings.

A.2.1 What are embeddings?

Embeddings are dense vector representations of data. Unlike traditional data representations, embeddings capture the semantic meaning and context of the data, making them particularly useful in natural language processing (NLP) and machine learning tasks. For instance, in text analysis, words or sentences are converted into fixed-length vectors that encode their semantic similarities.

To generate embeddings, various techniques can be employed, including neural network-based models like Word2Vec, GloVe, and BERT. These models train on large corpora of text and learn to represent words or phrases in a high-dimensional space where semantically similar terms are located closer together.

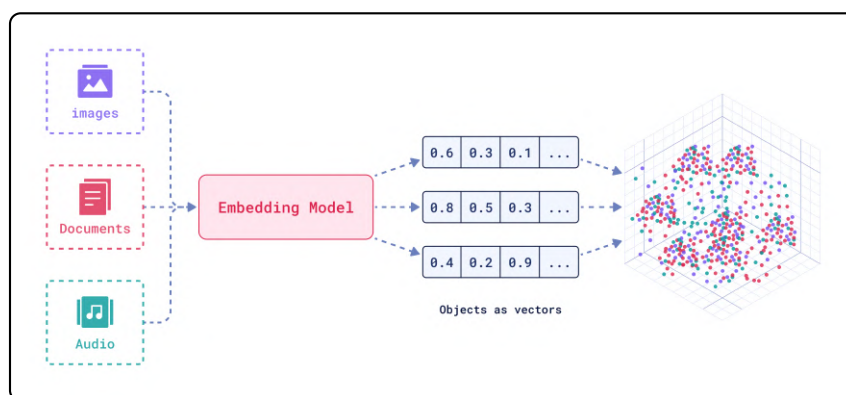


Figure A.3: How embedding works. Adapted from reference ³

³<https://qdrant.tech/articles/what-are-embeddings/>

A.2.2 How Pinecone utilizes embeddings?

Pinecone is a specialized vector database designed to handle these high-dimensional embeddings efficiently. Here's how it works:

Data ingestion: Pinecone allows for the ingestion of various types of data, which are then converted into embeddings using appropriate models. These embeddings are stored in Pinecone's vector database.

Indexing and storage: the embeddings are indexed in a way that supports fast similarity searches. Pinecone uses advanced indexing techniques to ensure that the search operations are highly efficient, even with large datasets.

Similarity search: one of the core functionalities of Pinecone is its ability to perform similarity searches. When a query embedding is provided, Pinecone quickly retrieves the most similar embeddings from its database. This is particularly useful for applications like recommendation systems, anomaly detection, and semantic search.

A.2.3 Visual representation

Below is a diagram to better understand the concept:

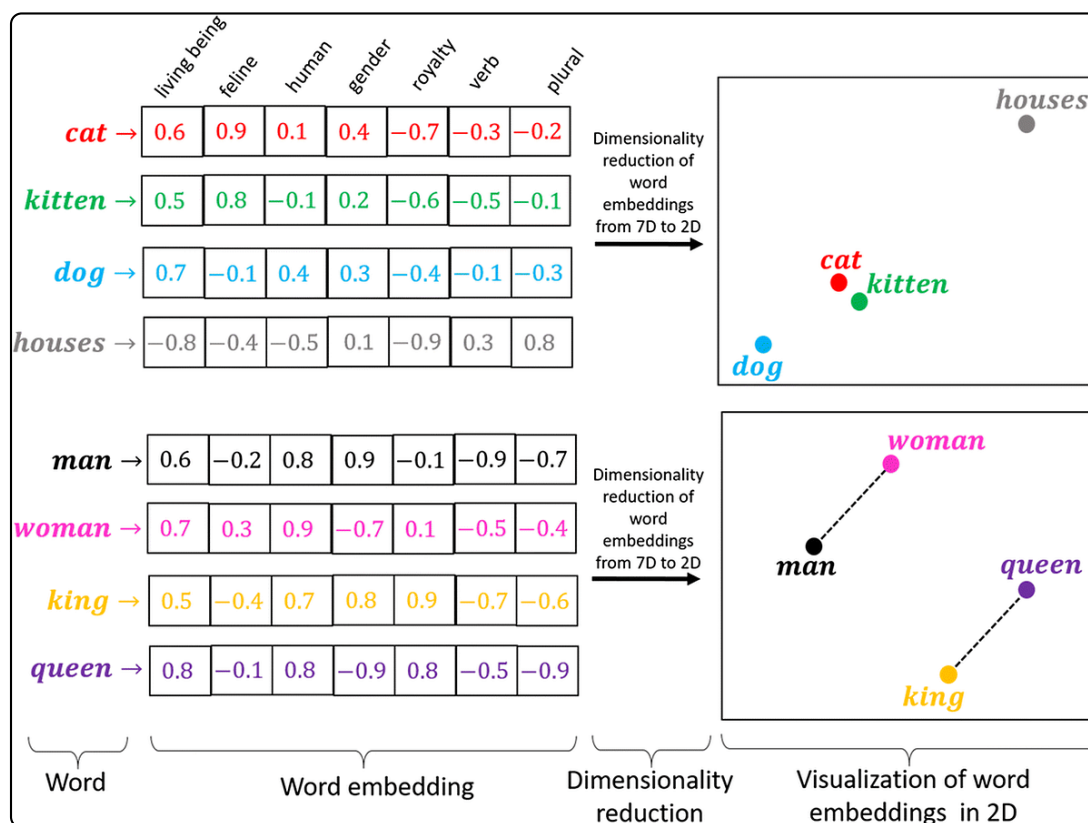


Figure A.4: Visual representation of an embedding. Adapted from reference ⁴

⁴<https://medium.com/@eugene-s/the-rise-of-embedding-technology-and-vector-databases-in-ai-4a8db58eb332>

Appendix B

Gantt diagram

Gantt chart for TFM project

Tasks		Date (Weeks) : 01/03/2024 - 27/06/2024																	
Nº	Description	W01	W02	W03	W04	W05	W06	W07	W08	W09	W10	W11	W12	W13	W14	W15	W16	W17	W18
T01	Data collection																		
T02	Quantitative data processing																		
T03	Qualitative data processing																		
T04	Chaining system development																		
T05	Automatic report generation																		

Figure B.1: Gantt chart for TFM project timeline.

Bibliography

- [1] R. Van Der Aalst. Impact of robotic process automation on business process management. *IEEE Computer Society*, 2018. <https://ieeexplore.ieee.org/document/8429527>.
- [2] Charu C. Aggarwal. Machine learning in business: An overview. *Springer*, 2020. <https://link.springer.com/book/10.1007/978-3-030-15729-6>.
- [3] Geoffrey E. Hinton Alex Krizhevsky, Ilya Sutskever. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 2012. <https://dl.acm.org/doi/10.1145/3065386>.
- [4] Michael Chui. Four fundamentals of workplace automation. *McKinsey Quarterly*, 2016. <https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/four-fundamentals-of-workplace-automation>.
- [5] John Chung. Google cloud automl: Making ai accessible to every business, 2018. <https://cloud.google.com/blog/products/ai-machine-learning/automl-making-ai-accessible-to-every-business>.
- [6] Crayon. Ai for competitive intelligence | crayon, 2023. <https://www.crayon.co/product/crayon-ai>.
- [7] Pinecone Developers. Pinecone documentation, 2023. <https://docs.pinecone.io/>.
- [8] Scrapy Developers. Scrapy documentation, 2023. <https://docs.scrapy.org/en/latest/>.
- [9] Selenium Developers. Selenium documentation, 2023. <https://www.selenium.dev/documentation/en/>.
- [10] Talend Developers. Talend documentation, 2023. <https://help.talend.com/>.
- [11] Pedro Domingos. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87, 2012. <https://dl.acm.org/doi/10.1145/2347736.2347755>.
- [12] Alec Radford et al. Language models are unsupervised multitask learners. *OpenAI Blog*, 2019. <https://openai.com/research/language-models-are-unsupervised-multitask-learners>.
- [13] Andre Esteva et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542:115–118, 2017. <https://www.nature.com/articles/nature21056>.

- [14] Ashish Vaswani et al. Attention is all you need. *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017. <https://arxiv.org/abs/1706.03762>.
- [15] Colin Raffel et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 2020. <https://arxiv.org/abs/1910.10683>.
- [16] David Silver et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529:484–489, 2016. <https://www.nature.com/articles/nature1961>.
- [17] Geoffrey Hinton et al. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012. <https://ieeexplore.ieee.org/document/6296526>.
- [18] Jacob Devlin et al. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2019. <https://arxiv.org/abs/1810.04805>.
- [19] James Heaton et al. Deep learning for finance: Deep portfolios. *arXiv preprint arXiv:1608.07257*, 2017. <https://arxiv.org/abs/1608.07257>.
- [20] Oriol Vinyals et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575:350–354, 2019. <https://www.nature.com/articles/s41586-019-1724-z>.
- [21] Pranav Rajpurkar et al. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, 2016. <https://aclanthology.org/D16-1264/>.
- [22] Tom Brown et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. <https://arxiv.org/abs/2005.14165>.
- [23] Apache Software Foundation. Apache nifi documentation, 2023. <https://nifi.apache.org/docs.html>.
- [24] Albert Gatt and Emiel Krahmer. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170, 2018. <https://www.jair.org/index.php/jair/article/view/11173>.
- [25] Genetec. The implications of large language models in physical security, 2023. <https://www.genetec.com/blog/cybersecurity/the-implications-of-large-language-models-in-physical-security>.
- [26] I. Abaker Targio Hashem. The rise of cloud computing in business process automation. *Elsevier*, 2015. <https://www.sciencedirect.com/science/article/abs/pii/S0167739X15001215>.
- [27] Geoffrey E Hinton, David E Rumelhart, and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986. <https://www.nature.com/articles/323533a0>.

- [28] John D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. <https://ieeexplore.ieee.org/document/4160265>.
- [29] Deloitte Insights. Hyperautomation: A catalyst for digital transformation. *Deloitte*, 2020. <https://www2.deloitte.com/global/en/insights/focus/industry-4-0/hyperautomation.html>.
- [30] Richard Jones. Beautiful soup documentation, 2023. <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>.
- [31] Philipp Koehn. *Statistical Machine Translation*. Cambridge University Press, 2009. <https://www.cambridge.org/core/books/statistical-machine-translation/1F1B61A3E9E34054E8D8E2D2C5061D6F>.
- [32] Giacomo Morabito Luigi Atzori, Antonio Iera. The internet of things: A survey. *Elsevier*, 2010. <https://www.sciencedirect.com/science/article/abs/pii/S1389128609001568>.
- [33] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999. <https://mitpress.mit.edu/9780262133609/foundations-of-statistical-natural-language-processing/>.
- [34] John McCarthy, Marvin Minsky, Nathaniel Rochester, and Claude Shannon. A proposal for the dartmouth summer research project on artificial intelligence. In *Dartmouth Conference*, 1956. <http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html>.
- [35] Wes McKinney. Data structures for statistical computing in python. *Proceedings of the 9th Python in Science Conference*, pages 51–56, 2010. <https://conference.scipy.org/proceedings/scipy2010/pdfs/mckinney.pdf>.
- [36] Microsoft. Sql server integration services (ssis), 2023. <https://docs.microsoft.com/en-us/sql/integration-services/sql-server-integration-services?view=sql-server-ver15>.
- [37] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013. <https://arxiv.org/abs/1312.5602>.
- [38] Inc. MongoDB. Mongodb documentation, 2023. <https://docs.mongodb.com/>.
- [39] Daniel G. Murray. *Tableau Your Data!: Fast and Easy Visual Analysis with Tableau Software*. John Wiley & Sons, 2013. <https://www.wiley.com/en-us/Tableau+Your+Data%21%3A+Fast+and+Easy+Visual+Analysis+with+Tableau+Software-p-9781118612040>.
- [40] n8n. n8n documentation, 2023. <https://docs.n8n.io/>.
- [41] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1):3–26, 2007. <https://www.jbe-platform.com/content/journals/10.1075/li.30.1.03nad>.

- [42] Healthcare IT News. Mayo clinic’s rpa journey: Automating administrative tasks to improve patient care. *Healthcare IT News*, 2020. <https://www.healthcareitnews.com/news/mayo-clinics-rpa-journey-automating-administrative-tasks-improve-patient-care>.
- [43] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *OpenAI*, 2018. <https://openai.com/research/language-understanding>.
- [44] ResearchGate. Classification of artificial intelligence (ai) systems: Open ai chatgpt like other large language models, 2023. https://www.researchgate.net/figure/Classification-of-artificial-intelligence-AI-systems-Open-AI-ChatGPT-like-other-large_fig1_373938258.
- [45] Sebastian Ruder. *Neural Transfer Learning for Natural Language Processing*. PhD thesis, National University of Ireland, Galway, 2019. https://ruder.io/thesis/neural_transfer_learning_for_nlp.pdf.
- [46] Siemens. Creating a smart factory with iot and ai. *Siemens*, 2021. <https://new.siemens.com/global/en/company/stories/industry/smart-factory.html>.
- [47] R. Smith. How walmart is using ai to revolutionize supply chain management. *Forbes*, 2020. <https://www.forbes.com/sites/rsmith/2020/02/14/how-walmart-is-using-ai-to-revolutionize-supply-chain-management/>.
- [48] Speak. Track your competitors with speak: The 1 competitive intelligence platform, 2023. <https://aidude.info/services/Speak>.
- [49] Alan M. Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950. <https://doi.org/10.1093/mind/LIX.236.433>.
- [50] Unilever. Ai-powered recruitment at unilever: Enhancing efficiency and quality of hires. *Unilever*, 2019. <https://www.unilever.com/news/news-search/2019/ai-powered-recruitment.html>.
- [51] Leslie Willcocks. Intelligent process automation: The future of business process management. *Cognition*, 2020. <https://www.cognition.com/intelligent-process-automation>.