ROYAUME DU MAROC
UNIVERSITE MOHAMMED PREMIER
Ecole Nationale des Sciences Appliquées (ENSA)
Oujda - Maroc

Mémoire de Fin d'Etudes

Filière : Ingénierie du Data Science et Cloud Computing

Designing, Developing, and Deploying Innovative Solutions using Artificial Intelligence and Natural Language Processing in a Multidisciplinary Context

Réalisé par :

MEHDI Ibrahim

Jury d'examen

Pr. KOULALI Mohmed Amine
Pr. AZGHIOU Kamal
Pr. NACIRI Rachid

Encadrants

Entreprise	PERFETTO Anna
	MIRON Jean-Raphaël
Ensa	KOULALI Mohamed Amine

Année universitaire
2022/23

Dedication

For my beloved Family

Acknowledgments

As I stand at the crossroads of completing my internship report, I find myself reflecting on the transformative experiences that have shaped my professional journey. Throughout this endeavor, I have been fortunate to receive support from numerous individuals and I would like to express my sincere appreciation.

First and foremost, I extend my heartfelt gratitude to my supervisors for their guidance and mentorship throughout the internship. Their expertise and encouragement provided me with the necessary tools to navigate the challenges and make meaningful contributions to the team.

I am also indebted to the entire team for their warm welcome and collaborative spirit. Working alongside such dedicated professionals has been inspiring, and I am grateful for the knowledge and insights they shared during our interactions.

Furthermore, I would like to acknowledge the organization's commitment to nurturing talent. The opportunity to be part of this team has been invaluable in honing my skills and gaining a deeper understanding of the industry.

In addition, I want to thank my family and friends for their unwavering support and understanding during this period. Their encouragement gave me the strength to persevere when faced with obstacles, and I am truly grateful for their presence in my life.

Lastly, I would like to express my appreciation to all those whose contributions may not be directly evident but have played a crucial role in making this internship a success.

As I move forward in my professional journey, I carry with me the knowledge and experiences gained during this internship. Each step has been instrumental in shaping my growth, and I am eager to apply the lessons learned in future endeavors.

MEHDI Ibrahim

Abstract

This report presents the findings and outcomes of three projects that utilized artificial intelligence (AI) to enhance operations in the chemical regulations industry.

The first project aimed to develop an AI-powered solution for predicting the outcomes of Challenge Tests, which evaluate the effectiveness of cosmetic products in inhibiting the growth of microorganisms. By analyzing the formulation of cosmetic products using machine learning algorithms, the solution provided accurate predictions of Challenge Test outcomes. This optimization allowed formulators to improve their formulations and increase the likelihood of passing the tests, thereby reducing the time and resources needed for product development. This projects was conducted in collaboration with the L'Occitane Group and integrated into the Cosmetic Factory software, a Product Lifecycle Management (PLM) solution.

The second project focused on developing an AI solution for reading Safety Data Sheets (SDS). The existing solution used a static code to search for keywords in PDFs and match them with a database. However, this approach had limitations in handling different languages and variations in key-value pairs. The new solution proposed the use of large language models and context injection to improve extraction accuracy and efficiency. This enhanced the ability to extract key information from SDS, improving safety and compliance in the chemical regulations industry.

The third project aimed to develop a chatbot powered by a large language model to provide support and answer frequently asked questions. However, the cost of fine-tuning the model with the company's database was prohibitive. As an alternative, the Langchain framework was proposed, which utilized context injection to provide relevant answers based on the user's question and available context documents. This chatbot solution improved customer support and efficiency in addressing client inquiries.

Résumé

Ce rapport présente les résultats et les retombées de trois projets qui ont utilisé l'intelligence artificielle (IA) pour améliorer les opérations dans l'industrie des réglementations chimiques.

Le premier projet visait à développer une solution alimentée par l'IA pour prédire les résultats des Challenge Tests, qui évaluent l'efficacité des produits cosmétiques pour inhiber la croissance des microorganismes. En analysant la formulation des produits cosmétiques à l'aide d'algorithmes d'apprentissage automatique, la solution fournissait des prédictions précises des résultats des tests. Cette optimisation a permis aux formulateurs d'améliorer leurs formulations et d'augmenter les chances de réussite des tests, réduisant ainsi le temps et les ressources nécessaires pour le développement de produits. Ce projet a été mené en collaboration avec le groupe L'Occitane et intégré dans le logiciel Cosmetic Factory, une solution de gestion du cycle de vie des produits (PLM).

Le deuxième projet était axé sur le développement d'une solution d'IA pour la lecture des fiches de données de sécurité (FDS). La solution existante utilisait un code statique pour rechercher des mots clés dans des fichiers PDF et les faire correspondre avec une base de données. Cependant, cette approche présentait des limites pour traiter différentes langues et variations dans les paires clé-valeur. La nouvelle solution proposait l'utilisation de grands modèles de langage et l'injection de contexte pour améliorer la précision et l'efficacité de l'extraction. Cela a renforcé la capacité à extraire des informations clés des FDS, améliorant la sécurité et la conformité dans l'industrie des réglementations chimiques.

Le troisième projet visait à développer un chatbot alimenté par un grand modèle de langage pour fournir un support et répondre aux questions fréquemment posées. Cependant, le coût de l'adaptation fine du modèle avec la base de données de l'entreprise était prohibitif. En guise d'alternative, le cadre Langchain a été proposé, utilisant l'injection de contexte pour fournir des réponses pertinentes en fonction de la question de l'utilisateur et des documents de contexte disponibles. Cette solution de chatbot a amélioré le support client et l'efficacité pour répondre aux demandes des clients.

Contents

List of Figures

Acronyms

Acronym	Definition
AI	Artificial Intelligence
NLP	Natural Language Processing
PDF	Portable Document Format
SDS/FDS	Safety Data Sheet/ Fiche de Données de Sécurité
PLM	Product Life Management
CT	Challenge Test
REACH	Registration, Evaluation, Authorisation and Restriction of Chemicals
CLP	Classification, Labelling and Packaging

Chapter 1

Introduction

In the ever-evolving landscape of the chemical regulations industry, advancements in technology have been instrumental in streamlining operations, improving safety measures, and fostering sustainable practices. As part of an enriching internship experience, this report sheds light on the transformative role of artificial intelligence (AI) in revolutionizing various facets of the chemical regulations sector. Throughout the internship tenure, I had the privilege of contributing to three distinct projects that harnessed AI's potential to address critical challenges and enhance industry practices.

This internship afforded the opportunity to explore the intersection of cutting-edge technologies and the intricate world of chemical regulations. Each project served as a stepping stone towards creating more efficient, reliable, and responsive solutions, ultimately contributing to the overall growth and advancement of the industry.

As a primary focus, the projects emphasized the strategic integration of AI technologies to tackle industry-specific issues effectively. By leveraging AI-powered machine learning algorithms and large language models, these projects sought to optimize various processes and improve the overall decision-making within the sector.

The objectives of the projects were carefully designed to align with the industry's needs, emphasizing the potential to transform how businesses operate, comply with safety standards, and drive product development. By overcoming the traditional limitations and exploring new paradigms, the projects showcased the immense promise of AI in reshaping the chemical regulations landscape.

Throughout this report, I will delve into the methodologies, challenges, and outcomes of each project, while highlighting the significance of their contributions to the industry. Moreover, the report will also cover the invaluable insights gained during the internship, including a deeper understanding of the industry's regulatory framework and how AI can be applied strategically to enhance its efficiency and efficacy.

As AI continues to gain prominence as a powerful tool in numerous industries, its potential in the chemical regulations sector remains unparalleled. This internship report aims to underscore the invaluable role of AI-driven innovations in fostering a safer, more sustainable, and technologically advanced chemical reg-

ulations industry. By understanding the implications of these projects, we embark on a journey towards a future where AI and chemical regulations coalesce to bring about transformative change.

1.1 Presentation of the host company

In this chapter, I am providing a comprehensive overview of the host company, outlining the objectives of my internship and setting the context for this report.

EcoMundo [3] specialises in chemical substances, their impact on human health and the environment, and the European and international regulations governing chemical risk (REACH, CLP, Cosmetics, Biocides, Medical devices, etc.).

They provide expert services and software to support the marketing of industrial products, enabling companies to manage the risks associated with chemical substances.

EcoMundo's strength lies in the combination of three complementary fields of expertise:

- **Chemistry/Toxicology:** In the domain of Chemistry/Toxicology, EcoMundo possesses a deep understanding of chemical properties and conducts thorough toxicological assessments. They analyze the behavior and potential hazards of chemical substances, assessing their impact on human health and the environment. This includes evaluating exposure routes, potential long-term effects, and safe levels of exposure. By staying up-to-date with the latest scientific research and regulatory developments, EcoMundo provides valuable insights into chemical risks and strategies for risk mitigation.
- **Regulations** EcoMundo's expertise in Regulations enables them to guide companies through the complex landscape of international regulations governing chemical substances. They help clients understand and comply with various regulatory frameworks, ensuring the safe and legal use of chemical substances. EcoMundo assists companies in navigating registration and notification processes, preparing regulatory dossiers, and complying with labeling and packaging requirements. They continuously monitor regulatory updates, keeping clients informed of any changes that may impact their products or processes.
- **Software development** Additionally, EcoMundo excels in Software Development, providing specialized software tools to streamline the management of chemical risks and regulatory compliance. Their software solutions enable companies to efficiently assess the compliance of their products with relevant

regulations. EcoMundo's software assists in generating safety data sheets (SDS), conducting hazard assessments, and providing product labeling information. The software modules they offer cover various aspects, including substance registration, data management, and risk assessment. Designed to integrate seamlessly with existing company systems, EcoMundo's software provides a comprehensive solution for effective chemical risk management.

1.1.1 History

— **2001**

Pierre Garçon participates in the European projects EDIT and ECODIS on the traceability of hazardous substances and environmental data.

— **2007**

Entry into force of the REACH Regulation: Europe establishes means to ensure a high level of protection against risks related to substances. Pierre Garçon and Jean-Raphaël Miron join forces to found the company EcoMundo. The initial core activity: compliance with REACH.

— **2010**

After 2 years of development, launch of SaaS software solutions dedicated to industrialists for mastering compliance with REACH.

— **2011**

EcoMundo's team grows from 2 to 27 employees.

— **2012**

Opening of a new office in Vancouver, Canada, to provide regulatory expertise to international companies.

— **2013**

EcoMundo diversifies its areas of regulatory expertise and meets the industry's new needs related to the European regulations on Cosmetics and Biocides.

— **2014**

EcoMundo increases its capacities and becomes one of the main European actors concerning REACH authorization dossiers.

— **2016**

Launch of the COSMETIC Factory software solution that revolutionizes cosmetic regulatory management. It notably allows for the automation of DIP creation.

— **2017**

Following a fundraising round, EcoMundo's capital is raised to 1 million euros. The number of employees continues to increase!

— **2018**

Opening of a new branch in Seoul, South Korea, mainly dedicated to cosmetics. The Vancouver office is transferred to Montreal, Canada.

— **2019**

Opening of a new branch in Barcelona, Spain.

— **2020**

Opening of a branch in London, United Kingdom. The teams now consist of 43 employees in Paris, 3 in Montreal, 4 in Seoul, and 2 in Barcelona. Finally, the offices in Paris are renovated to provide employees with an environment in line with our values.

1.2 Organizational framework

This section provides a consice overview of Ecomundo's internal organization, delineating the functional departments and collaborative networks that constitute the company's framework. During my internship at Ecomundo, I was assigned as an AI Engineer Intern inside the IT-R&D department, under the supervision of both Chief Technical director Jean-Raphaël MIRON and AI Engineer Anna PERFETTO.

1.2.1 Human Resources

The Human Resources department plays a pivotal role in managing and developing the organization's workforce. It is responsible for various activities, including recruitment, talent acquisition, training and development, employee relations, performance evaluation, and compensation management. By ensuring a skilled and motivated workforce, the Human Resources department contributes to the overall success and growth of Ecomundo. The director of this department is the Chief Financial Officer Simon PICCA.

1.2.2 Corporate Communication

Effective communication is crucial for any organization's success, and Ecomundo recognizes this importance by maintaining a dedicated Corporate Communication department. This department is responsible for managing both internal and external communication activities. It encompasses public relations, media relations, branding, and corporate messaging. Through strategic communication initiatives, the Corporate Communication department promotes Ecomundo's brand image, enhances stakeholder relationships, and facilitates the dissemination of information to employees and external audiences. The director of this department is the Marketing and Communication Director Laure SCHMITT.

1.2.3 Scientific Expertise

As a company specializing in environmental sciences, Ecomundo places great emphasis on scientific expertise. The Scientific Expertise department consists of subject matter experts who possess in-depth knowledge and experience in various scientific domains. These experts provide valuable insights, technical guidance, and scientific support across different projects and initiatives

undertaken by Ecomundo. Their contributions ensure that the organization's activities align with the latest scientific advancements and industry best practices. The director of this department is the Chief Scientist Officer Benoît SOTTON.

1.2.4 Regulatory Affairs

Compliance with regulations and standards is of paramount importance for Ecomundo. The Regulatory Affairs department is responsible for monitoring and ensuring adherence to applicable regulations and standards. This includes staying updated on regulatory changes, preparing and submitting regulatory documentation, conducting regulatory assessments, and liaising with regulatory bodies. By actively managing regulatory affairs, Ecomundo demonstrates its commitment to maintaining compliance and upholding the highest ethical and legal standards. The director of this department is the Legal Affairs Director and Head of Authorisation Béatrice ZAREMBA.

1.2.5 Sales

The Sales department serves as the key driver of revenue generation for Ecomundo. It is responsible for identifying and acquiring new customers, nurturing existing client relationships, and promoting the organization's products and services. The Sales team collaborates closely with other departments to understand customer needs, develop tailored solutions, and provide exceptional customer experiences. By effectively positioning Ecomundo's offerings in the market, the Sales department contributes to the organization's growth and market competitiveness. The director of this department is the Chief Operation Officer (COO) Fangcun ZHOU.

1.2.6 Innovation & Software

To stay at the forefront of technological advancements, Ecomundo maintains an Innovation & Software department. This department fosters a culture of innovation within the organization by exploring emerging technologies, conducting research, and developing software solutions. The Innovation & Software team collaborates with other departments to identify opportunities for process improvement, streamline operations, and enhance productivity. Through its innovative approach, this department enables Ecomundo to adapt to changing market dynamics and deliver cutting-edge solutions to its clients. The director of this department is the Chief Operation Officer (COO) Fangcun ZHOU.

1.2.7 IT

The IT department plays a critical role in managing Ecomundo's information technology infrastructure. It ensures the smooth operation of computer

systems, networks, and software applications throughout the organization. The IT department is responsible for system administration, network management, cybersecurity, technical support, and data management. By maintaining a robust and secure IT infrastructure, this department enables efficient and secure access to information, facilitates effective communication, and supports the organization's overall business operations. The director of this department is the Co-Founder and the Chief Technical Director Jean-Raphaël MIRON .

1.2.7.1 DevOps

The DevOps sub-department combines software development and operations to enable efficient and reliable software deployment, infrastructure management, and continuous integration and delivery. DevOps professionals collaborate with software developers, system administrators, and quality assurance teams to streamline development cycles, automate processes, and improve overall software reliability and performance.

1.2.7.2 R&D (Research & Development)

The R&D sub-department within the IT department focuses on exploring new technologies, conducting research, and developing innovative solutions. R&D professionals work closely with other departments to identify opportunities for technological advancements, prototype new features, and enhance existing software solutions. Their research-driven approach ensures that Ecomundo remains at the forefront of technological innovation within the environmental sciences domain.

1.2.7.3 Quality

The Quality sub-department is responsible for ensuring that software and systems developed within Ecomundo meet established quality standards. Quality professionals conduct comprehensive testing, implement quality assurance processes, and adhere to best practices throughout the software development lifecycle. Their efforts contribute to the delivery of reliable, user-friendly, and high-quality software solutions to Ecomundo's clients.

1.2.7.4 Data

The Data sub-department handles data management within the IT department. Data professionals are responsible for data analysis, database administration, data integration, and data security. They collaborate with other teams to ensure data integrity, facilitate data-driven decision-making processes, and maintain the confidentiality and privacy of sensitive information. Through effective data management, this sub-department provides valuable insights and supports evidence-based decision making within Ecomundo.

1.2.7.5 Dev

The Dev sub-department focuses on software development within the IT department. Dev professionals are responsible for writing and maintaining code, implementing new features or enhancements, and ensuring software solutions align with project requirements and specifications. They collaborate with other teams to develop robust and scalable software applications that meet the needs of Ecomundo's clients.

1.2.7.6 LAB

The LAB sub-department is involved in testing and quality control activities within the IT department. LAB professionals conduct experiments, analyze results, and ensure compliance with regulatory standards and best practices. Their expertise in quality control and testing procedures ensures that software solutions developed within Ecomundo are reliable, accurate, and in compliance with industry standards.

1.2.7.7 Product

The Product sub-department works closely with other teams within the IT department to oversee the development and management of software products. Product professionals collaborate with stakeholders to define product requirements, prioritize feature development, and ensure alignment with customer needs and market trends. Their role encompasses product strategy, roadmap planning, and product lifecycle management.

1.3 Internship objectives

The main objectives of this internship are to design, develop, test and document innovative solutions that adress business and R&D challenges. The area of focus includes supporting and improving rule-based expert systems that leverage knowledge and logic to provide answers to complex tasks that are traditionally done by a human. Additionally, developping AI solutions in different areas using Natural language processing and Machine/Deep Learning. Once these AI solutions are designed and developped, the next crucial step is to deploy them effectively. This envolves ensuring their quality and integrating them into the existing technical infrastructure of the organization. Rigourous testing procedures, such as unit testing, integration testing and performance testing are employed to validate the reliability and efficiency of the solutions. And to keep the projects well defined for future references, we worked on writing well documenting our work in order to provide clear intructions and guidelines for the code, deployment steps, maintenance and error handling.

1.4 Existing solutions

Ecomundo has many software in its arsenal, but in the context of this report, I worked with two applications: Cosmetic factory and SDS factory:

1.4.1 Cosmetic Factory

Cosmetic factory is an AI-Powered PLM Software that has more than 12 key features that supports clients at each stage of a cosmetic product lifecycle from marketing brief to the worldwide market launch, while ensuring the best level of expertise. It digitalizes the entire development process of the products:

— Choice of raw materials
— Formulation
— Regulatory and Environmental Compliance
— Marketing constraints

This application is powered an expert database containing more than 300.000 Chemical substances, +28.000 INCIs, +380 international regulations, +3900 toxicological profiles, An embedded database of all classifications (GHS, CLP, HAZCOM ...)

Figure 1.1 – Cosmetic Factory

From the figure 1.1, two modules are in the context of this report:

1. Formulation
   (a) **Smart Theoretical Formulation:** CF helps saving time from the beginning of the formulation process thanks too the assisted theoretical formulation:
      — Prepare the theoretical formulas according to the requirements of the associated product brief (target markets, naturalness, biodegradability, COSMOS standards...)
      — Reduce low value-added exchanges between the different stakeholders thanks to the embedded expertise
   (b) **Bench Formulation:** Adjust the theoretical formulas on the bench. Create a test library and record all associated analytical results (physicochemical and organoleptic properties, etc.), as well as other parameters. Re-validate the formula's compliance with the product brief with the AI FORMULA Check.
   (c) **Pilot Batch management:** Prepare the industrialization of the product with the management of all the parameters related to the formula and the necessary equipment.
   (d) **Test Management:** Plan, manage and follow the progress of all the tests related to the products (regulatory, tolerance, efficacy...) in COSMETIC Factory.
   (e) **Powerful Tools to accelerate formulation:**
      — **AI Formula Check:** Calculate global compliance in one click!
      — **Raw material Comparison Engine:** Compare the PMs with a multi-sourcing filter system
      — **AI Challenge test prediction:** Calculate the percentage of success of the challenge tests.
      — **Formula Comparison Engine:** Compare the formulas and tests with an advanced multi-criteria filter system:
         — comparison of private labels (naturalness, COSMOS, biodegradability)
         — concentrations in raw materials
         — concentrations in ingredients
         — go or no go results according to FORMULA Check
         — Analytical data
2. AI - Formula Check
   (a) **Global Compliance Calculation:**FORMULA Check instantly calculates the compliance of a raw material, a formula, a finished product or a packaging. The tool makes it possible to know instantly if a product, and even before that, if a formula under development is in compliance with the targeted markets, without waiting for an analysis by regulatory and toxicological services.

(b) **Oniscient Expert, Always Available:**FORMULA Check is like having an omniscient regulatory expert available at all times, with perfect and exhaustive knowledge of regulations (rules, constraints on use, product positioning) and ingredients.

(c) **Smart And Proactive Formulation:**FORMULA Check automatically provides expert advice on compliance and allows formulators to make the most informed decisions before proceeding further, modifying formulas, or initiating testing (only when necessary).

1.4.2 SDS Factory

This application comes in the context of this report as the after-effect of the new developped PDF Reader. PDF Reader allows the clients to create products from Safety data Sheets (see section 3.2.1.1) to be able to use them in other applications. As shown in figure 1.2, SDS Factory has many features that allows to manage SDSs in general from creation, translation to distribution.

Figure 1.2 – SDS Factory

Chapter 2

Methods and tools

2.1 Transformers

Transformers [11] have been introduced in 2017 to replace the Recurrent Neural Networks architectures using self-attention mechanisms and the encoder-decoder architecture that hugely improved the performance of such models.

2.1.1 Word Embedding

Word embedding is a fundamental concept in natural language processing that facilitates computers in comprehending words by transforming them into dense numerical vectors. These vectors exist within a multi-dimensional space, where words with similar semantic meanings exhibit proximate or analogous representations. As an example, the classic Word2Vec is a prominent and widely adopted algorithm for creating word embeddings [8].

The underlying principle behind word embeddings is to capture the contextual and semantic relationships between words. Rather than relying on sparse and discrete one-hot encoding representations, where each word is assigned a unique index in a high-dimensional vector space, word embeddings attempt to project words into a continuous and compact vector space. This process allows for a more efficient representation, as similar words or words used in comparable contexts end up with vectors that are closer together, enabling the model to leverage this spatial proximity to infer semantic associations.

By capturing and representing the underlying semantic relationships between words in a dense and continuous vector space, word embeddings have opened up new possibilities for machines to interpret and generate human language effectively. As research continues to advance, the field of embeddings continues to evolve, leading to even more sophisticated and contextually rich representations of language.

2.1.2 Attention

The attention mechanism allows the model to effectively capture relevancy between words in a sentence by calculation attention scores. The scores are calculated using the dot product between each word and the target word of interest. Softmax function is then used to transform the scores into 'meaningful' weights that help the model emphasis important words while de-emphasising less significant ones. This mechanism gives the model the superpower of understanding the contextual nuances and relationships between words, which is the ideal goal for Natual Language processing tasks.

2.1.3 Encoder-Decoder Architecture

2.1.3.1 Encoder

In the encoder, the input words are first embedded into a high-dimentional vector space. then we add to these embeddings to capture the position of the token in the sequence. Each token is transformed into three vectors: Query (Q), Key (K) and Value (V). This transformation is performed using learnable weight matrices:

$$Q = X \cdot W_Q$$
$$K = X \cdot W_K$$
$$V = X \cdot W_V$$

Self attention scores are then calculated as the dot product. For query vector q_i and a key vector k_j, the attention is:

$$Attention(q_i, k_j) = q_i.k_j/\sqrt{d}$$

With d the dimention of the key vectors.
Attention weights are simply the softmax of the attention weights:

$$\alpha_{ij} = softmax(Attention(q_i, k_j)$$

The final output of the attention mechanism is obtained by taking the weighted sum of the value vectors V.

$$AttentionOutput(q_i) = \sum_j \alpha_{ij} V_j$$

2.1.3.2 Decoder

The decoder is responsible for generating the output sequence step by step. It uses both self-attention mechanisms and an additional attention mechanism that allows it to focus on the relevant parts of the encoder's output. This attention mechanism is often referred to as the encoder-decoder attention.
Let's assume that the encoder produces the final hidden states $H = \{h_1, h_2, ..., h_n\}$, where n is the number of tokens in the input sequence. The decoder takes

these hidden states as input and processes them step by step to generate the output sequence.

At each decoding step t, the decoder receives the previously generated token and its corresponding embedding y_{t-1} and also an attention context vector c_t, which is a weighted sum of the encoder's hidden states based on attention scores. The process of obtaining the context vector is similar to the one used in the encoder's self-attention mechanism.

The decoder's self-attention is also calculated based on query (Q), key (K), and value (V) transformations, similar to the encoder. However, to prevent the decoder from attending to future tokens, a masking mechanism is applied to the attention scores.

The decoder's self-attention scores are calculated as follows:

$$Q_t = y_{t-1} \cdot W'_Q$$
$$K_t = y_{t-1} \cdot W'_K$$
$$V_t = y_{t-1} \cdot W'_V$$
$$\text{Attention}_t(q_i, k_j) = \begin{cases} \frac{Q_t \cdot K_j^T}{\sqrt{d}} & \text{if token } j \leq t \\ -\infty & \text{otherwise} \end{cases}$$

Here, d represents the dimension of the key vectors.

The attention weights for the decoder's self-attention are computed as follows:

$$\alpha_{ij} = \text{softmax}(\text{Attention}_t(q_i, k_j))$$

The context vector c_t is then calculated as the weighted sum of the encoder's hidden states using the attention weights:

$$c_t = \sum_{j=1}^{n} \alpha_{ij} V_j$$

After obtaining the context vector c_t, the decoder combines it with the input embedding y_{t-1} and feeds it through a decoding transformer layer to produce the final hidden state for the current decoding step:

$$h_t = \text{DecoderLayer}(y_{t-1} + c_t, h_{t-1})$$

Finally, the output at each step is generated using a linear transformation and a softmax activation:

$$\hat{y}_t = \text{softmax}(h_t \cdot W_o)$$

Where W_o is a learnable weight matrix, and $\hat{y}_t$ represents the probability distribution over the target vocabulary for the next token in the output sequence.

The decoding process continues until an end-of-sequence token is generated or a predefined maximum length for the output sequence is reached. The trained model can be used for tasks such as machine translation, text summarization, and more, by feeding the input sequence to the encoder and iteratively decoding the output sequence using the decoder.

2.1.4 HuggingFace

Huggingface [5] is an opensource known for their contributions for the transformers library, one of the most used libraries for NLP and AI tasks in general. Transfomers library provides a wide range of pre-trained models for multiple purposes such as text classification, NER, text generation, translation, etc... It is build on top of the PyTorch and TensorFlow frameworks. Huggingface has played an important role in democratising NLP and making it more accessible, they provide really powerful language models such as BERT, GPT and RoBERTa that can either be used directly, or for transfer learning. They also provide a large range of datasets that can be used for model fine-tuning for various specific tasks.

2.1.5 OpenAI & GPT

2.1.5.1 Generative Pretrained Transformers

GPT models are Language models that have been trained on huge amounts of data such as books, web data and human conversations for the purpose of developping a general understanding of the language.

GPT-1 It all started with GPT the first release by OpenAI [9] in 2018. It had 117 million parameters. It was mainly trained to be able to correctly predict the next word in a sentence. It suffered from the lack of understanding when given a longer context which resulted in incoherent outputs.

GPT-2 In 2019, OpenAI released GPT-2 with 1.5 billion parameter, about 10 times larger than its predecessor. It outperformed the first gen model in the ability of generating coherent text. Due to concerns for misuse, the use of this model was initially very limited. Which will lead OpenAI to perform different types of trainings for their future models.

GPT-3 In their paper "Language Models are Few Shot learners" [1], it was demonstrated how a large model like GPT 3 with a 175 billion can easily learn and adapt to new tasks only using few examples (few-shot learning). Obviously, it also demonstrated better language understanding and impresive performance on multiple NLP tasks such as summarization, question answering and translation

2.2 LangChain

With large language models emerging as a revolutionary technology, Langchain [6], an opensource framework, allows developpers to build applications that were near impossible to make before. Some good examples of those applications are Document Question Answering, Chatbots, Agents.

2.3 Ensemble Learning

Ensemble learning is a Machine Learning technique that combines multiple base models in order to produce a more powerful model that can achieve higher accuracies. By combining the predictions of each base model, we can capture the diverse knowledge and expertise of these models which leads to an improved overall performance. In order to achieve the most optimal results, many aspects need to be considered:

2.3.1 Diversity of Base models

When we talk about ensemble learning, we usually gather multiple models that are considered weak learners. These models have, individually, slightly better performance than random prediction. In order to squeeze the maximum out of this method, the models should be diverse. Diversity ensures that each model brings unique information and makes different errors, reducing the chances of the resulting model being biased or the results being fluctuated by a single model's shortcomings. We can push this further by feeding each model a different subset of our dataset.

2.3.2 Aggregation

Each model in Ensemble learning should predict his own output, then the final result is the aggregation of the multiple outputs. It can be a voting-based method which involves casting votes from each model. The voting can be either HARD, which basically means the majoritarian result is the final result. Or it can be SOFT, which is a weighted vote based on Confidence intervals. The other way to do it is by averaging the outputs. In this method we take the weighted average of the predictions from all the underlying models to get the final prediction.

2.4 Gradient Boosting

Gradient boosting is a type of ensemble learning. The key idea behind it is to iteratively build an ensemble of weak prediction models, where each sub model corrects the errors made by the other models. This iterative process follows the gradient of the loss functions. A very known model for Gradient Boosting is XGBoost:

2.4.1 XGBoost

A Known Gradient boosting machine learning model for his performance and versatility in solving multiple supervised learning tasks. It works as follows:

1. XGBoost starts by initializing a set of weak base models, usually decision trees called a regression tree. Each tree is trying to correct the mistake made by the previous tree. These trees are build sequentially.
2. To define whether the model is performing well, and to guide it in the right direction, a metric or Objective function needs to be defined and optimized during the training process. It is composed of a loss function and a regularization term.
3. The model then computes the gradients of the loss function. It also computes the Hessian matrix to get further infromation about the curvature of the loss function.
4. The gradients and hessian values are used to determine the best splits when constructing the trees. Each branch is added based on the evalution of these splits.
5. Regularization penalties are added in order to prevent overfitting and control the complexity of the model to keep it simple. Prunning is also added at the end to further improve the performance and generalization of the model.
6. XGBoost uses early stopping by monitoring the performance of the model on the validation set and stops the learning process when the performance starts to degrade.

2.5 Summary

Chapter 2 of this report delves into the tools and methodologies employed in the projects discussed in Chapter 3. The projects involved the utilization of various AI technologies to enhance operations in the chemical regulations industry. These technologies encompassed a wide array of methods, including Transformers, word embeddings, attention mechanisms, and encoder-decoder architectures. Additionally, the use of well-known libraries and frameworks like Hugging Face, OpenAI's GPT, and Langchain played a significant role in the successful implementation of these projects. Furthermore, the adoption of ensemble learning and gradient boosting techniques for Challenge Test predictions showcased the versatility of AI methodologies in addressing different tasks. Chapter 3 will further elaborate on the practical applications of these methodologies in the specific projects carried out, illustrating their impact on optimizing cosmetic product formulations, improving Safety Data Sheets extraction, and enhancing customer support through the implementation of an AI-powered chatbot.

Chapter 3

Projects

3.1 Challenge Test Predictions

3.1.1 Presentation of the project

All creams, foundations, and shampoos that become tomorrow's BEST SELLER are the result of a long Research & Development (R&D) process. Before being launched into the market, they undergo microbiology tests, which can make or break their success. If they pass the microbiology tests, it will be a great achievement. If they fail, it will be a nightmare for the formulators who must adjust their formulations to turn their aspirations into reality.
Help has arrived from Cosmetic Factory, a PLM cosmetic software that assists R&D laboratories in speeding up the launch of new cosmetic products. In partnership with the Occitane Group, Ecomundo has developed an Artificial Intelligence (AI)-powered solution to anticipate the outcomes of the Challenge Test (CT). It serves as an intelligent formulation aid for cosmeticians, and was unveiled during the Cosmetic 360 Conference in October 2022.
What is a Challenge Test? The Challenge Test is a mandatory pre-market test required by regulations, based on the ISO 11930 method. It involves inoculating a particular concentration of bacteria, such as Escherichia coli, Candida albicans, Staphylococcus aureus, etc. As show in figure 3.1, in order for a product to pass the test, a specific log reduction in the number of microorganisms recovered must be achieved at designated intervals, typically after 2, 7, 14, and 28 days.

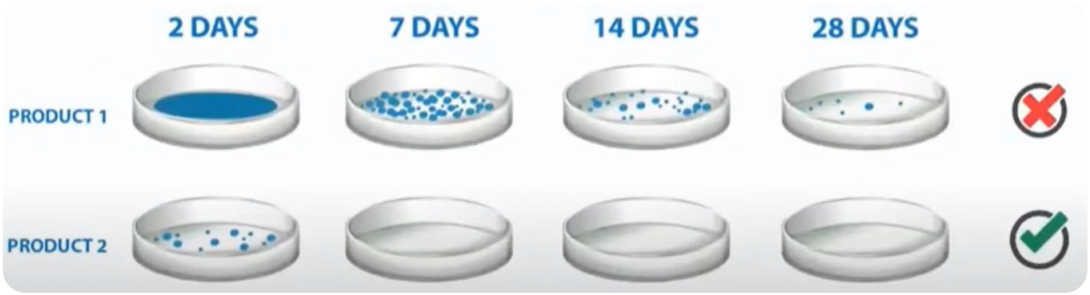


Figure 3.1 – Challenge Test

If the specified criteria are not met, as was the case with PRODUCT 1, cosmeticians are required to reformulate and test again until they develop a product that passes the CT, like PRODUCT 2. As a result, one can imagine the laborious nature of the process, as well as the waste of time and resources involved. This is where artificial intelligence comes to the rescue, with its predictive capabilities.

3.1.2 How it is currently done

EcoMundo has developed a hybrid AI solution, which combines an expert system (ES) and machine learning (ML) that predicts the result of the CT as shown in figure 3.2, in other words the efficiency of the conservation system. At the moment the solution has been implemented in Cosmetic Factory in the Formulas module, in the Prediction wizard (kyoto server).

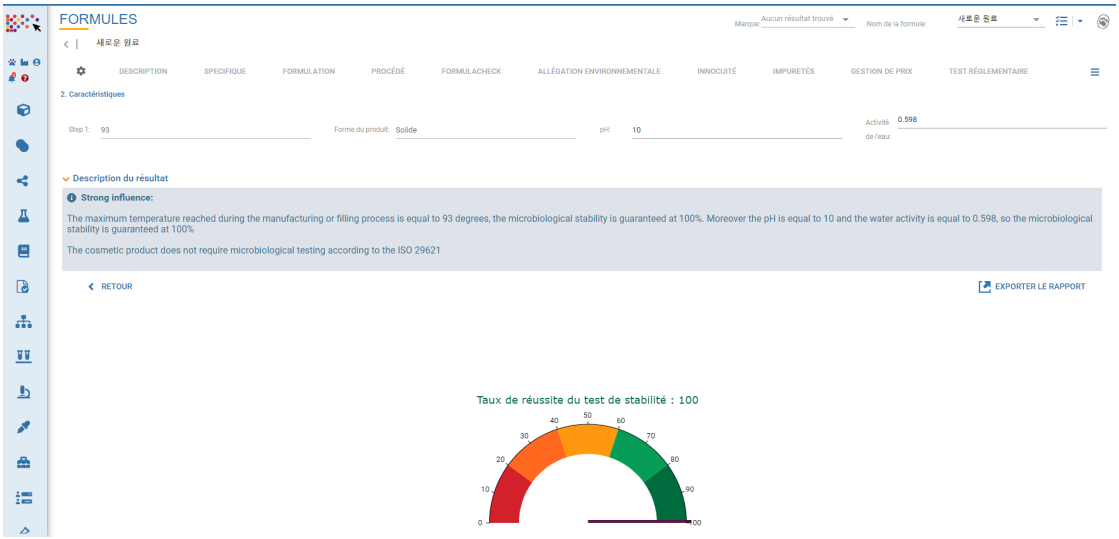


Figure 3.2 – CF interface

To develop this hybrid AI solution, Ecomundo started to study the criteria influencing the risk of microbial contamination. The pH, which measures the acid or basic character of a solution, the water activity that quantifies the free water available to assure growth of microorganisms. Moreover, some raw materials are known to create a hostile environment for the survival of microorganisms like alcohols, inorganic solvents, humectants, inorganic salts, acid and bases. The formulation processes may have an influence on the microbial protection like temperature, incorporation order of ingredients, time of homogenization. And eventually the physical state, if our cosmetic product is a solid, powder, foam etc..

A hybrid system benefits from the reasoning and explanation of the ES, as well as the learning and generalization potential of ML algorithms, which determines an effective and efficient model. This allows users (cosmeticians, formulators, etc..) to understand the prediction, along with the scientific and

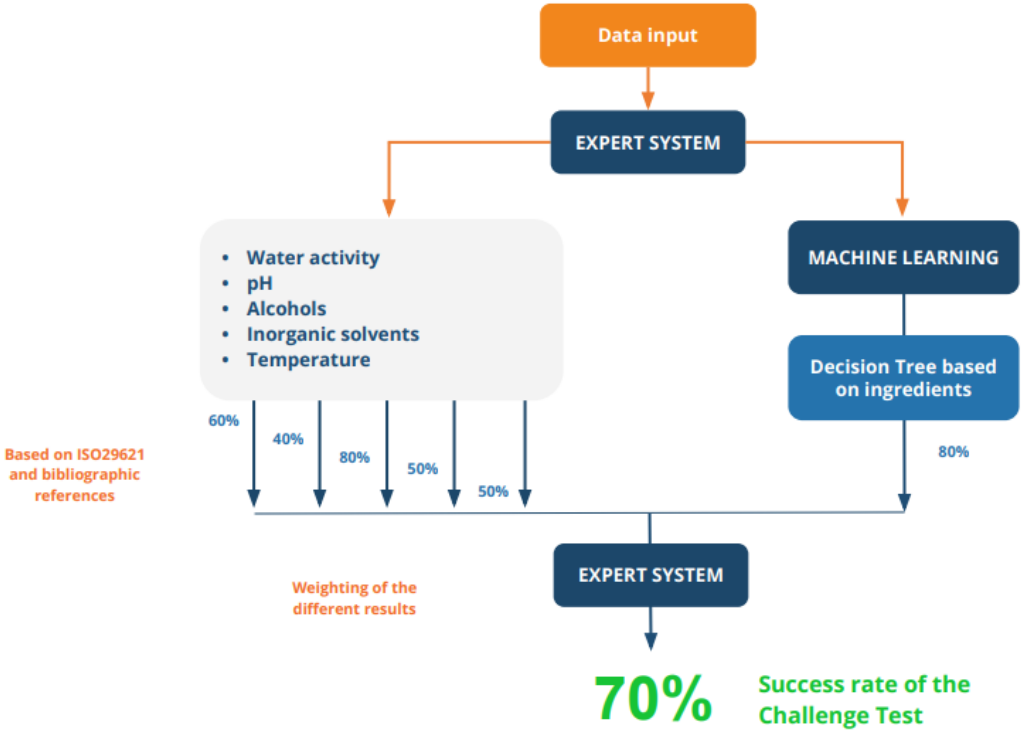


Figure 3.3 – CT flow

regulatory reasons behind it. Indeed, the system provides suggestions for improvement, making this solution a useful and reliable formulation support and optimization tool to achieve the success of the challenge test immediately.

How is the pipeline structured? The data are feeded to the ES which starts to assign success rate of the CT according to the rules codified in regulation, namely the ISO29621 and based on experimental results published in literature. These rules concern the empiric criteria influencing the risk of contamination based on certain values of pH, water activity, temperature etc.

Thereafter, the ES calls the ML to have its success rate. We trained different ML algorithms, but the one that gave the best results is the decision tree applied on the set of ingredients present in the formulas. The result coming from the ML is eventually weighted by the expert system together with the other results, to give finally to the user the success rate of the challenge test. Figure 3.3 visually explains this process.

The ML algorithm chosen is the supervised binary classification, see the decision tree in figure 3.4 , trained on the ensemble of ingredients and their concentration. 3000 formulas furnished by the Group Occitane have been selected and have been divided into 2 sets: a training set which represents 75% and the rest has been used as a testing set. The decision tree at the end gives as answer Yes/No, in other words it is able to say if each formula passes or fails the CT, with a precision of around 80%.

The learning database is composed of formulas, descriptors like pH, water

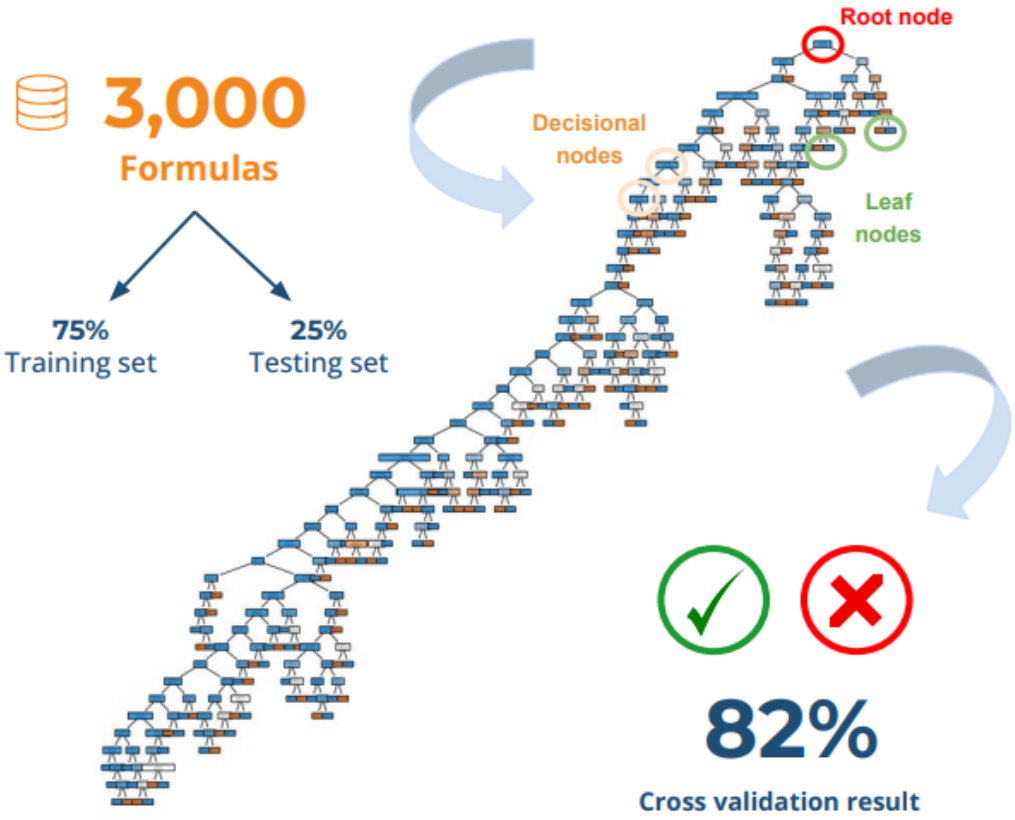


Figure 3.4 – Decision tree

activity and temperature, etc.., which somehow affect the CT results, and of course the CT results. Here, you can find the data conceptual model (DCM) to have a complete vision of the information necessary to train the models.

3.1.3 New solution

The new solution is based on ensemble learning. By utilizing multiple models instead of just the decision tree, we will be using XGBoost, Random Forest and a Neural Network (Not implemented yet). In the next subsection, I am presenting the datasets that have been used, the problems with it and how we managed to overcome them.

3.1.3.1 Datasets

These datasets have been provided by L'Occitane (see figure 3.5):

- **ECO_liste des formules avec désignation:** contains list of codes of chemical formulas as well as the type, for example a cream, oil, scrub etc..
- **ECO_formule avec element tech:** Contains the formula code and its corresponding challenge test result.
- **ECO_formule avec liste INCI:** Contains the formula code, ingredients and their corresponding concentration.
- **ECO_formule avec code protocole:** Contains for each formula code, its corresponding protocol code.
- **ECO_code protocole avec liste des codes étapes:** Each protocol of the last table, has many steps. Each step has its own code.
- **ECO_liste des codes étapes avec les conditions:** On each step, some conditions must be respected. Those conditions are in this table like temperature, time with each condition referred to as a code.
- **ECO_formule avec liste MP:** Table that contains for each formula the raw materials in it and their concentrations.

Remarks and problems about the datasets

- CODEMAT and CODEART are the same thing.
- INCI represents one ingredient.
- Raw Material is composed of multiple INCIs (ingredients)
- Total number of Formulas is 7369, though only 5303 have CT results. Out of the 5303, only around 4000 formulas we have their exact compositions with the concentrations.
- Usually, the tests are positives because many hours are spent on making sure the product passes it. This leaves us with a very unbalanced dataset of around 90% of the data have the label "YES", while only 10% did not pass.

— Not all raw materials have the same INCIs, so we have to put all the INCIs in columns and have it at 0 concentration if doesn't exist in the formula. This has left us with a large dataset column wise and it's very sparse (around 80 percent of the values are zeros).

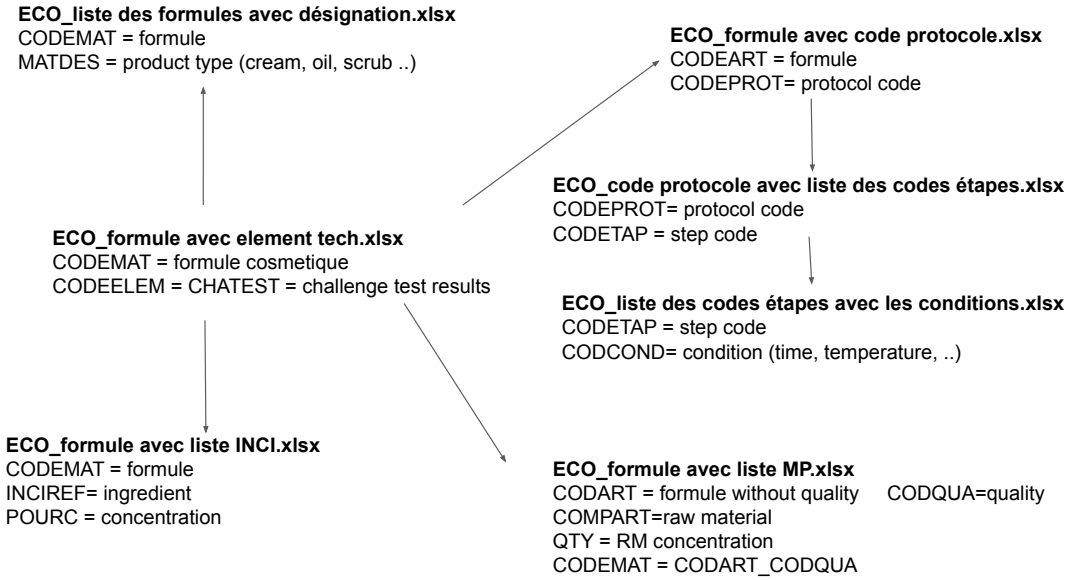


Figure 3.5 – Datasets L'Occitane

To tackle these problems, first we needed to do a startified split on the dataset in order to well train it. Then imperatively use PCA in order to reduce the size of the dataset. (code is not included in the report, will be presented live).

3.1.4 ML Pipeline and Deployment Architecture

Following the figure 3.6, this is how the ML pipeline is structured:

1. We have implemented a robust data collection mechanism to gather the required data from the EcoMundo database using SQL scripts. The collected data is then transformed into a CSV file for further processing.
2. For now, we have used a Google Colab notebook, where we perform extensive preprocessing on the collected data to ensure it is in a suitable format for ML model training. Using the pandas library, we create a structured dataframe. We then employ scikit-learn and TensorFlow libraries to train various ML models. Through optimization and fine-tuning techniques, we enhance the model's performance.
3. To integrate the ML model into the existing Cosmetic Factory software, we have designed and developed an AI application using the Django framework. This application is called when there is no response from

the Rule Engine. Through a REST API, a Java controller in the back-end is called, allowing communication between the ML model and the software.

4. We have implemented a mechanism for real-time inference, enabling the ML model to make predictions in real-time. The ML model's predictions are integrated with the existing software's logic. This process involves making API calls to retrieve answers, which are then sent to the Java controller and displayed in the application's front-end.
5. We have set up a manual periodic fine-tuning process to retrain the ML model using new data from the EcoMundo database. By incorporating user feedback and leveraging the latest data, we continuously improve the model's accuracy and effectiveness.

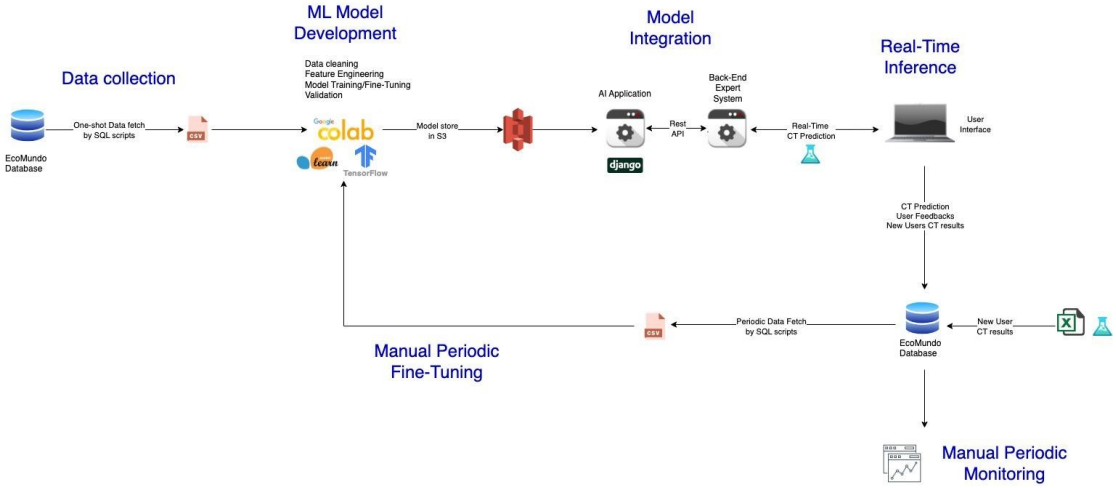


Figure 3.6 – ML Pipeline

3.1.5 Results and Comments

3.1.5.1 Metrics

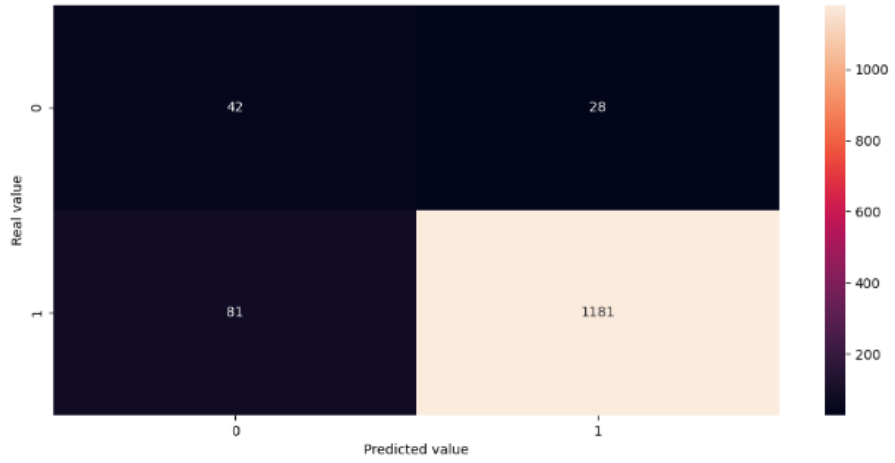
Because we are working on a binary classification, precision alone is not a good metric for how performant the model is. This metric only computes how many times a model made a correct prediction across the entire dataset. Recall on the other hand measures how many of the positive class samples present in the dataset were correctly identified by the model. F1-score is the best for this use case, since it's compatible with unbalanced datasets and it combines both recall and precision.

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

Another Metric that we have used is the ROC AUC score. It reflects on how efficient the model is. The higher the Area Under the Curve, the better model's performance at distinguinshing between the two classes.

The expert actually wanted less false negatives and did no care about the other metrics. If we gave as a result that a formula will pass the challenge, but when its conducted it does not pass, it's not as big of a problem as the other way around: We say it is not going to pass the test but it actually does.

3.1.5.2 Results




```
accuracy  score: 0.9181681681681682 %
auc  score: 0.8216022110593315
              precision    recall  f1-score   support

           0       0.34      0.60      0.44        70
           1       0.98      0.94      0.96      1262

    accuracy                           0.92      1332
   macro avg       0.66      0.77      0.70      1332
weighted avg       0.94      0.92      0.93      1332
```

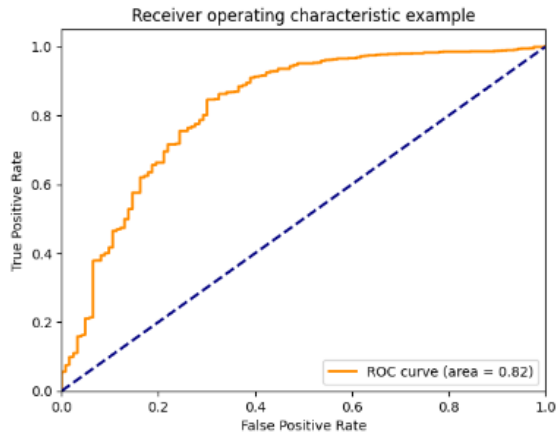


Figure 3.7 – XGBoost Confusion Matrix, AUC

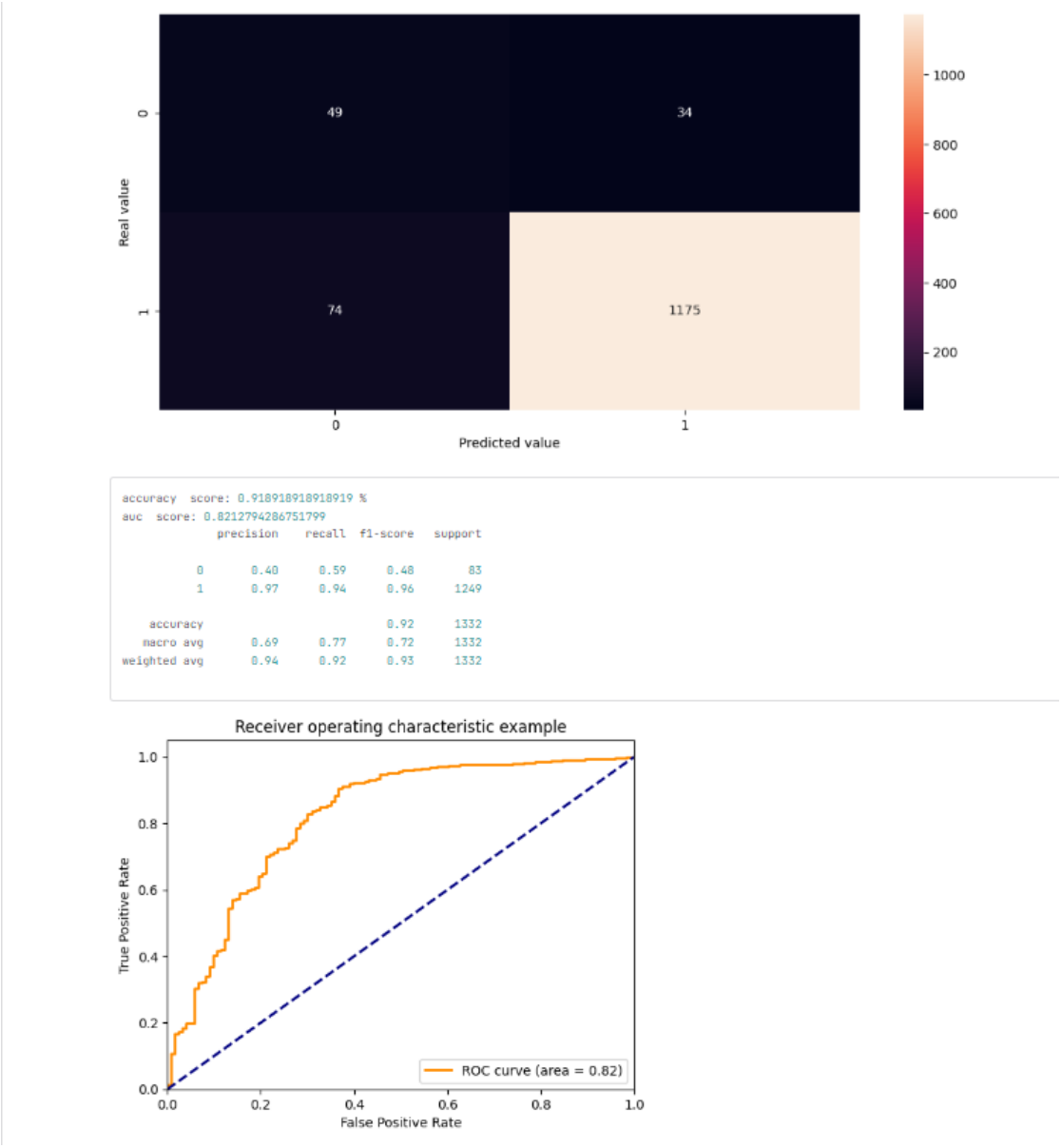


Figure 3.8 – Random Forest Confusion Matrix, AUC

3.1.5.3 Comments

The result that we got on the inference of these two models have been enough to support the Expert system in its prediction. At the end of the day, I suggested that we do not take 50% as the threashold of classification. This means if for example a product had a 60% for example of passing the challenge test, we would show the percentage and our expert suggestion that effecting the challenge test for this product will probably be a loss of time and money in case of failure.

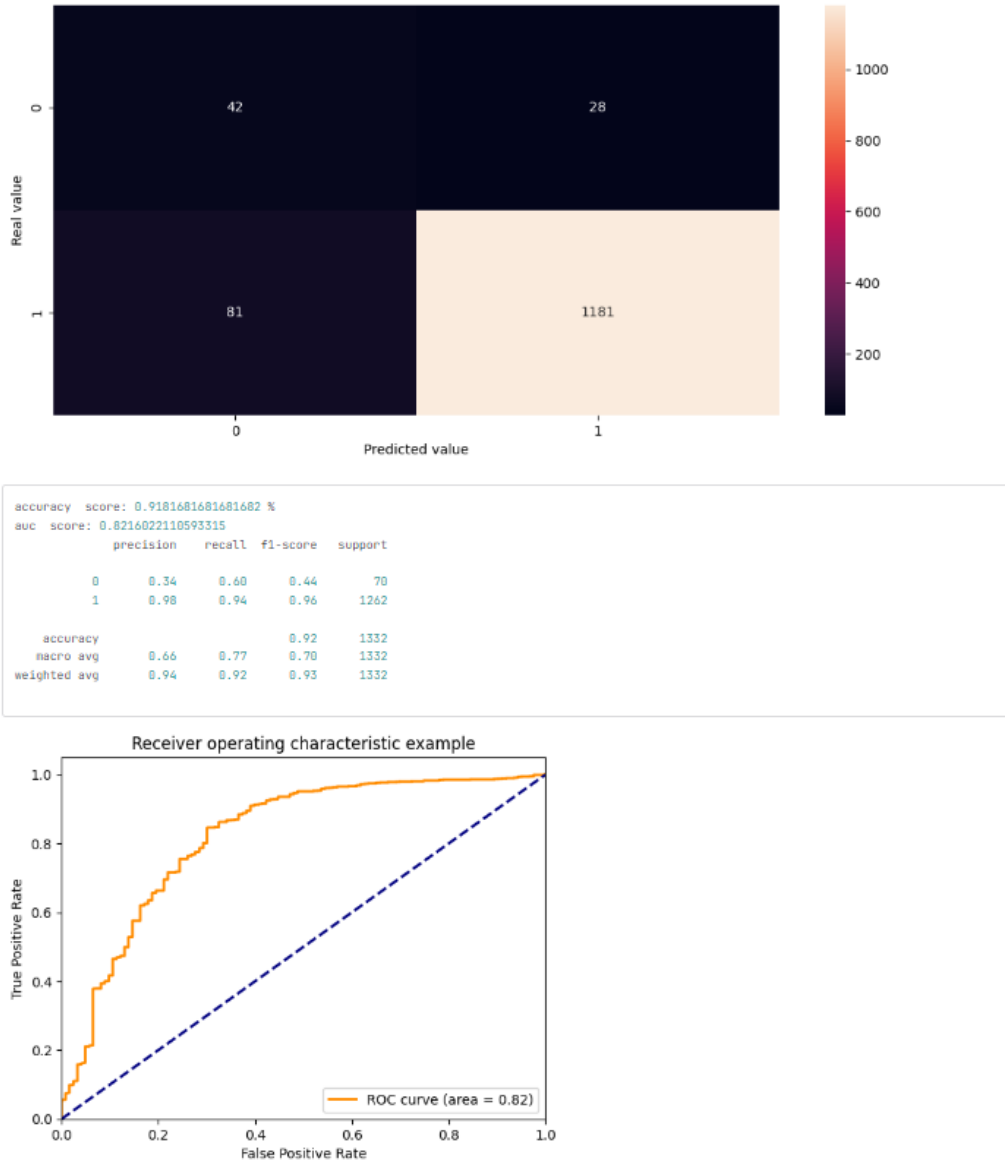


Figure 3.9 – XGBoost Confusion Matrix, AUC

3.2 SDS Reader

3.2.1 Presentation of the project

Currently called PDF Reader, is an application made by Ecomundo specifically for Safety Data Sheets:

3.2.1.1 What is an SDS?

Safety Data Sheet, also known as Material Safety Data Sheet (MSDS), is a very important document that provides informations about the hazards and safe handling of a specific substance/mixture. Its main purpose is to ensure

the safety of individuals who may use or come in contact with the substance in any settings. It it composed of sixteen sections:

- **Identification:** Identification of the substance/mixture on the SDS as well as the recommanded uses. It contains also the contact informations for the suppliers/manifacturers
- **Hazard Identification:** This section identifies the hazards of the substance/mixture presented in the SDS along with their appropriate warning informations
- **Composition/Information on Ingredients:** This section identifies the substance(s) contained in the product, their concentration and their hazard classification.
- **First-Aid Measures:** Descriptions of the initial care that should be ginven by untrained responders to an individual who has been in contact with the chemical.
- **Firefighting Measures:** Recommandations for fighting the fire caused by the chemical.
- **Accidental Release Measures:** This section has recommandations on the appropriate responses to spills, leaks ro relases including containment and cleanup practices to minimize or ideally prevent exposure to the people and the environment.
- **Handling and Storage:** Guidance on the best and safe practices of handling and storing the product.
- **Exposure Controls/ Personal Protection:** In this section, the exposure limits, engineerings controls and personal protective measures are indicated.
- **Physical and Chemical Properties:** This section identifies the physical and chemical properties of the substance/mixture
- **Stability and Reactivity:** This section provides information about the reactivity hazards of the chemical and its stability information.
- **Toxicological Informations:** Informations about the toxicological and health effects.
- **Ecological Information:** This section provides information to evaluate the environmental impact of the substance(s).
- **Disposal Considerations:** Guidance on proper disposal practices, recycling or reclamation of the substance(s) or its container and safe handling practices.
- **Transport Information:** Guidance on classification information for shipping and transporting of hazardous substance(s) by road, air, rail, or sea.
- **Regulatory Information:** Safety, health and environmental regulations specific for the substance/mixture that is not indicated anywhere else on the SDS.
- **Other Information:** Indications about the preparation and revision date of the sds along with other useful information such as the changes that have been made compared to the last version.

3.2.2 How it is currently done

3.2.2.1 Get sections

The old code is a static code that mainly starts by dividing the pdf into paragraphs. then go through each paragraph and detect where the sections are by comparing the words to the words that define sections in the SDSFactory database. For example you can have "SECTION" as the word splitting the document into sections, might also be "RUBRIQUE" if it's in French. It uses all possible words that are available in the database in the language of the SDS.

3.2.2.2 Get Key-Value pairs

Using Regular Expressions, the code then goes through each section and finds all the key value pairs that are mentionned in the PDF file. This is already problematic since there are inconsistencies of how a key value pair is shown in the file: It could be in the form of Key:Value, sometimes it's Key n Value.

3.2.2.3 Match with Ecomundo Template

Using the same static method of looking into the database, the code then goes through all the key value pairs and matches them with the Ecomundo Template of Keys.

3.2.3 Problems with the old way

As it is very obvious, the static way of searching through the database for keywords, and expecting to have the exact same match is far from optimal. If we have an SDS with a different language then we have to recreate a table in the database for that exact same language, then fill it with all possible words for each key. For example the key "Product Name" can be in the pdf as "Product identifier" or just "Name", if we don't have these keys in the database, it will assume this data does not exist in the pdf. We notice as well the workload of refreshing and updating the database frequently in order to keep it updated.

3.2.4 New solution

3.2.4.1 Usage of Large Language Models

The usage of large language models was first proposed as it will allow us to read the pdf and understand it then extract data from it. This method as demonstrated first by using ChatGPT and giving him texts in different languages from an SDS and providing him the keys to extract. Two major problems have arised from the proposition of using this method:

Security Sending PDF files of clients to OpenAI servers in North America for processing is not a straightforward procedure. The Technical directors were worried about data privacy of the clients even if SDS are not considered confidential in most cases. But as we will be using GPT models using the OpenAI api, and as a company, I suggested we contact sales and request to not store our sent data after the processing is done. It is a right that anyone can ask for when purchasing the api key [10].

Cost At the time of the first meeting 13th March 2023, the pricing for gpt-3.5-turbo model was 0.002 USD/1000 Tokens. The commercial team were not convinced of how cheap it is until I presented a proof of concept later on.

3.2.4.2 Open-Source Large Language Models

To tackle these two problems, I proposed to postpone the GPT api purchase while I dig deep into how we can use an opensource model. I started by looking into the models that are Open Source + Available for commercial use:

BERT and its derivatives The BERT model [2] (Bidirectional Encoder Representations from Transformers) was introduced in 2018 by researchers at google. It achieved state-of-the-art performance on many NLP benchmarks.
In my proposition to use open source models, BERT models came in as top results. BERT model was optimized when RoBERTa [7] (Robustly optimized BERT) was introduced in 2019 which was optimized even more with the introduction of DeBERTa [4] (Decoding-enhanced BERT with disentangled Attention).
I started initially by using these models for question answering. The PDF file would be converted into a raw text fille, then fed to the model as context. Then I gave it the keys that I wanted to extract. DeBERTa Model was the best obviously, but still lacked in many areas:

Bad extraction The model was very inconsistent with the extractions, some keys would be perfectly extracted while other were not. The model also does not quite understand the structure of the page only from the raw text given to it as context. Finetuning is always an option with these models, for the lack of time that I had with them on-site, and the deadline that we had for the end of June, we have proceded to the usage of other solutions.

3.2.4.3 GPT

OpenAI has made the model gpt-3.5-turbo available from their api for as cheap as 0.0015 USD/1K tokens (June 2023). To demonstrate the power of the model and how cheap it can be, I did a demo doing extraction of all

key value pairs of an SDS page by page. It costs around 0.01 Euros per pdf of 13 pages. This has captured the attention of Christian FRENEUIL, the Software and Cloud services Sales Director, and pushed this project.

First Version of the stable code: Non-Hybrid This code works as follows (see figure 3.10):

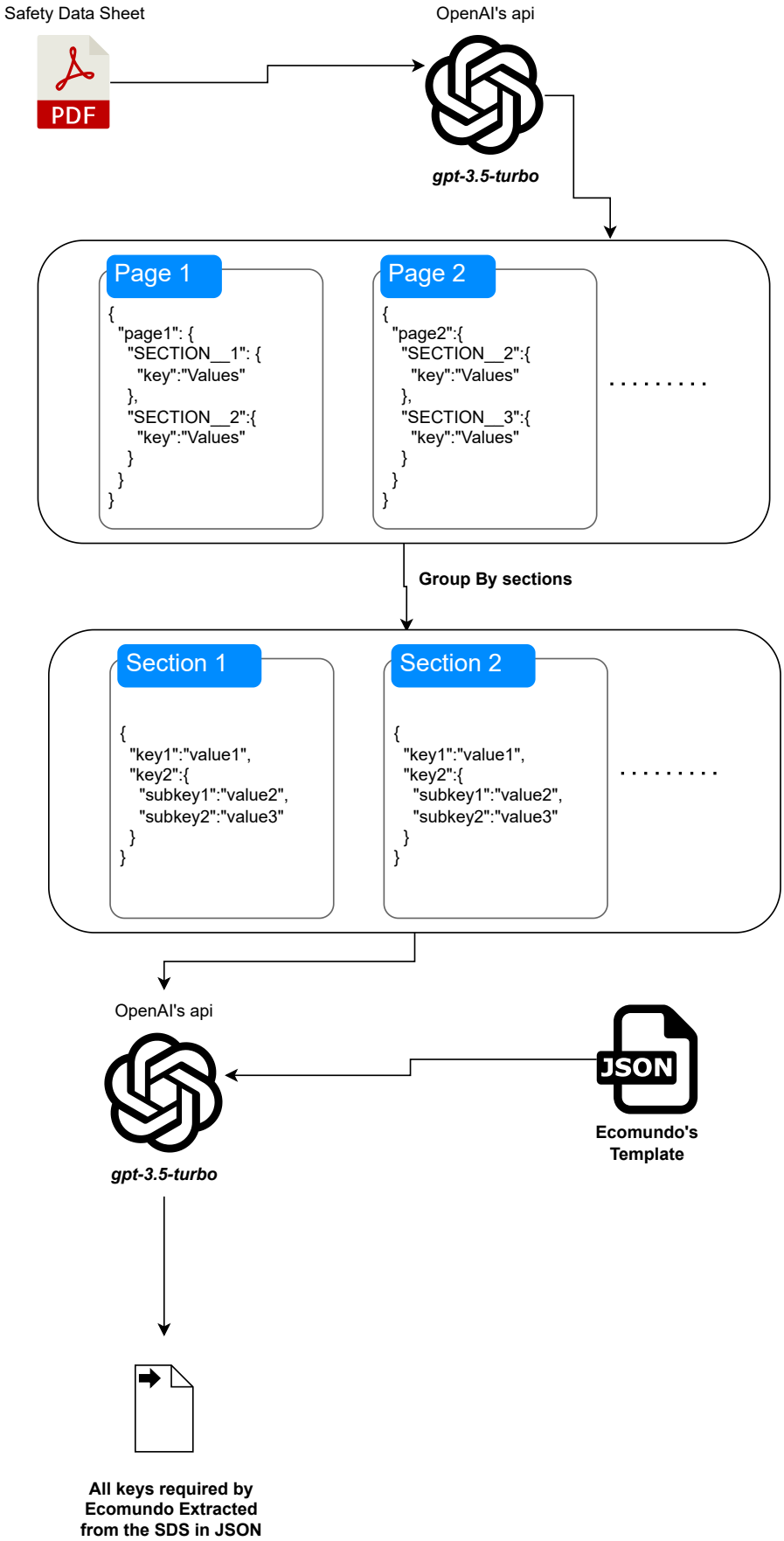
1. Extract all key value pairs from each page. Keep a record of what sections are in that exact page, and save the latest one. For example if a page has Section 1 and Section 2, we save them both in an array, then we save the fact that section 2 was the latest.
2. When we move to the next page, we check whether the page starts direclty with a section, then we assign the key values of that section to the corresponding key, if not we assume it's the Section 2 continued.
3. Group by Sections then send it to another controller in the application that will then match the key value paris from each section into the keys that we have.

Hybrid Version of the code As it is obvious in figure 3.10, we are using OpenAI's api twice. The first extraction is taking an overhead task of detecting sections, that in some cases fail. To overcome this problem, we thought about using the existing PDFReader controller that splits the SDS into sections, and then use them directly for extraction and matching (see figure 3.11). This will give us Two benefits:

1. **Parallel extraction:** We will be able to parallelize the extraction for each section, and the extraction by Ecomundo's template for each section. Which means the complexity of this extraction will be the complexity of the longest Section.
2. **Better Extraction:** Since we are focusing on only one task, the gpt model will work better with a one-goal instruction rather than multiple ones.

Even More improvements To tackle problems relating to either: Size of the context (in our case the raw pdf text) and/or the quality of section extraction, I proposed a further architecture improvement (see figure 3.12).

1. The pdf is loaded into the SDSReader interface, we first extract the text, if the text is empty, this means it's not a digital PDF. We then use PyTesseract OCR to extract the text.
2. Check for the quality of section extraction. If the length of the sections' list is not 16, we assume we did not correclty extract and we use the previous stable code. If it's till not 16 sections, we move to the langchain solution for the entire pdf proposed in section 3.3.
3. Assuming we had 16 sections, before the extraction, we need to check for the context limit of the prompt+the raw text. If we are in the context window of gpt-3.5-turbo, which is 4097 tokens, we are good. If not, we



/2

Figure 3.10 – Non-Hybrid

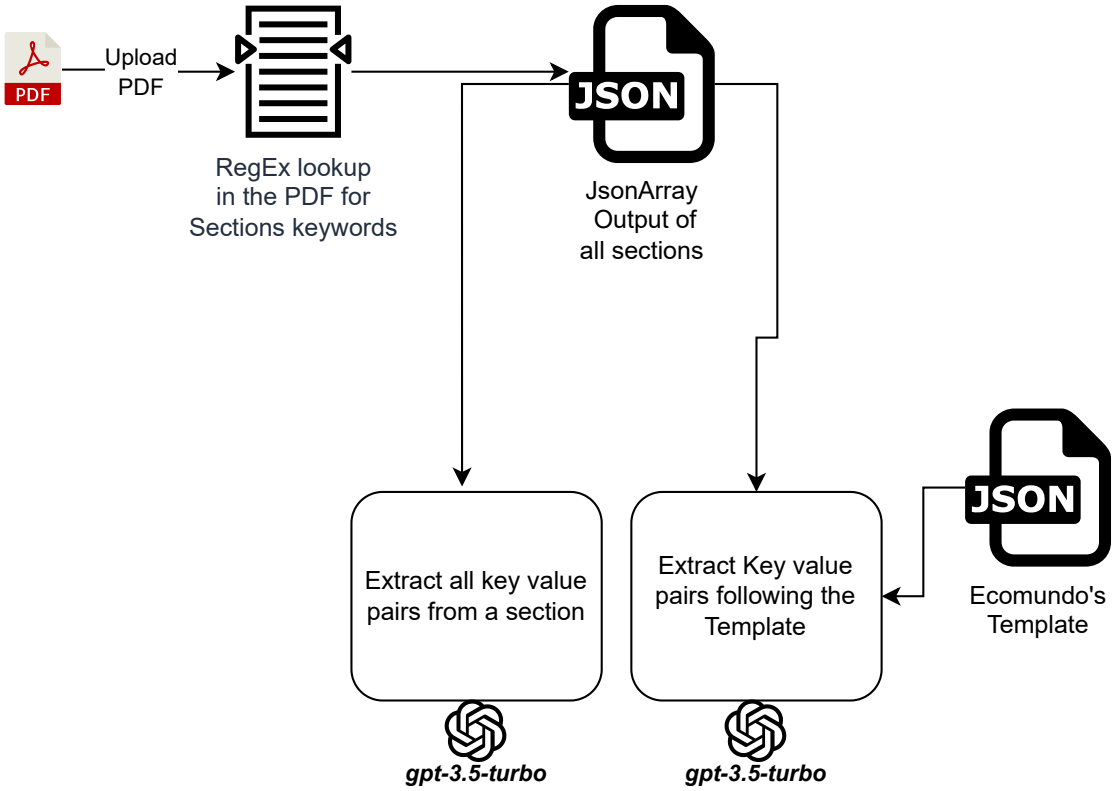


Figure 3.11 – Hybrid

check if it's less than 16K tokens, then we use gpt-3.5-turbo-16k, which costs about 0.003 for input tokens and 0.004 for output tokens. If we are still outside this range, then we use LangChain solution but for only that exact section.

3.2.4.4 Deployment, CI pipeline

Continious Integration As in the section on the CT prediction project, Ecomundo uses GitLab-CI, the YAML file is in appendix **??**. It consists of Three Stages:

1. **lint:** Linting is basically the automated checking of the source code for programmatic and stylistic erros. It starts by installing pyflakes since we are working with python, then runs it on the container.
2. **test:** This part of the pipeline checks whether an api or any other secret key have been left out in the code when developping or testing. Then it runs the SemGrep-SAST which stands for Semantic grep(global regex print) static application security testing. It helps by analyzing the code, finding bugs and detecting security issues.
3. **container-build:** In this stage, we are quiet sure that the code is secure and there are no problems relating to the syntax. In this case, I have already created the dockerfile, so all it has to do is build it and then push it into the docker repository of the IT team.

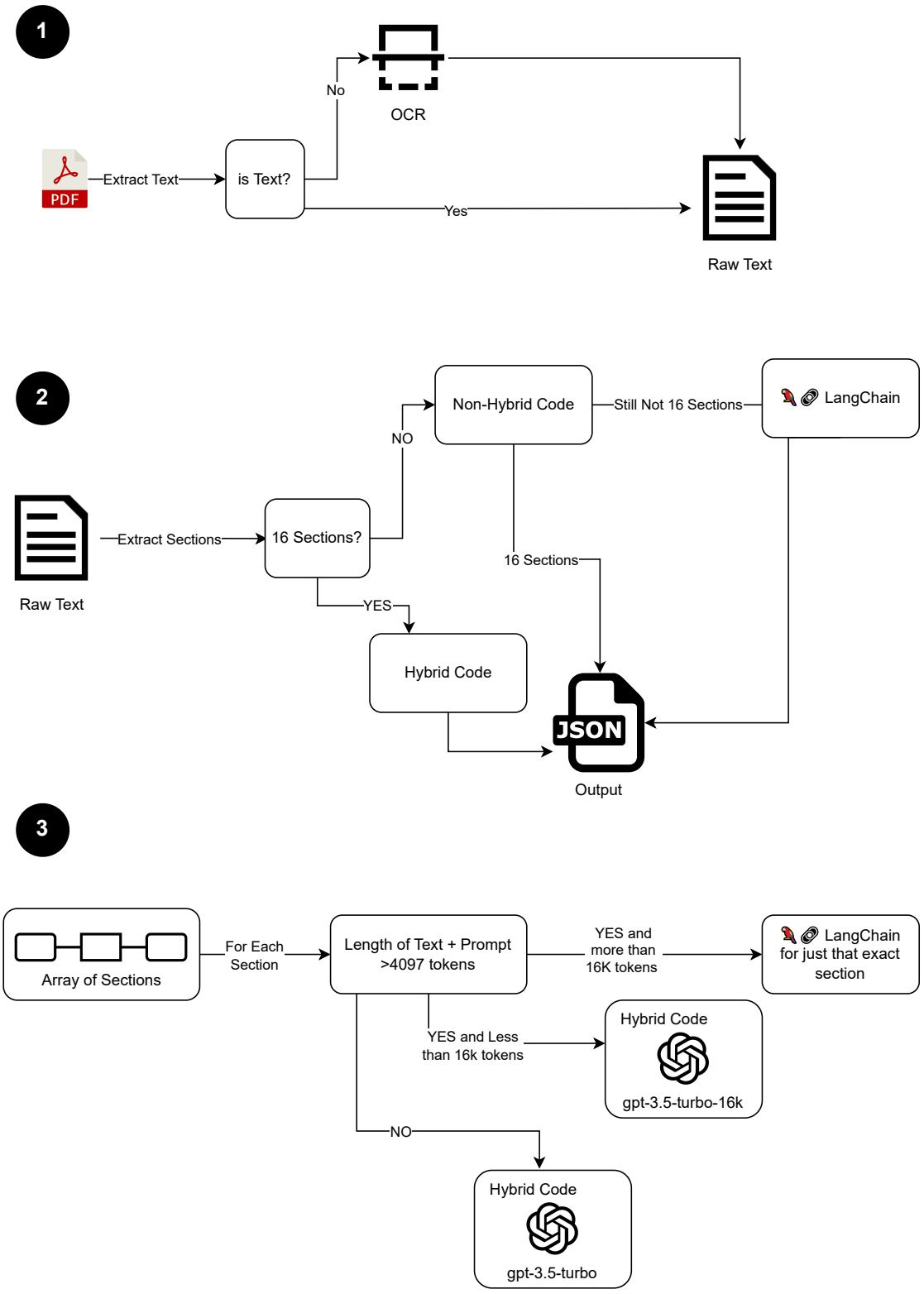


Figure 3.12 – Hybrid++

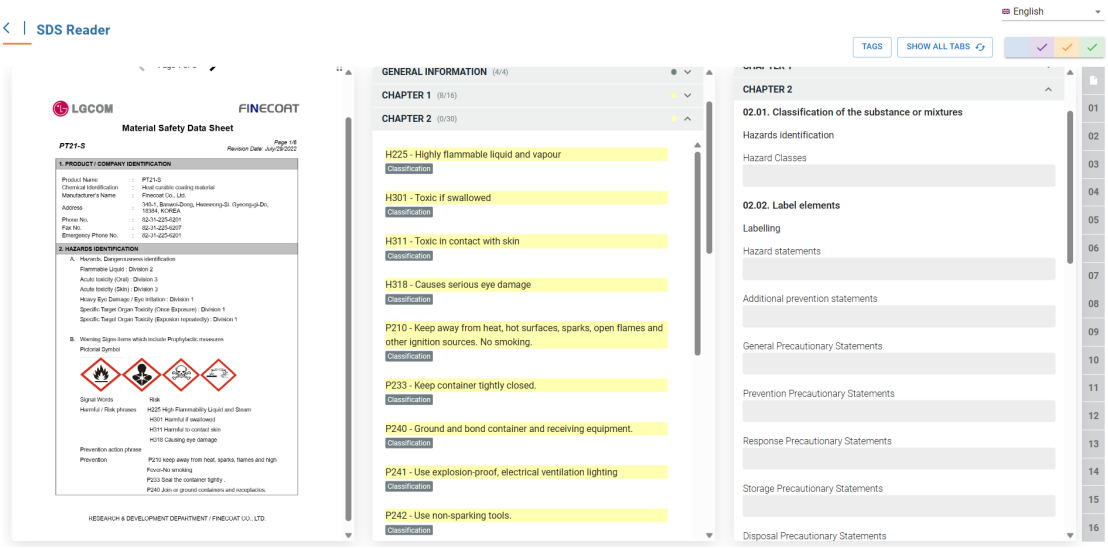


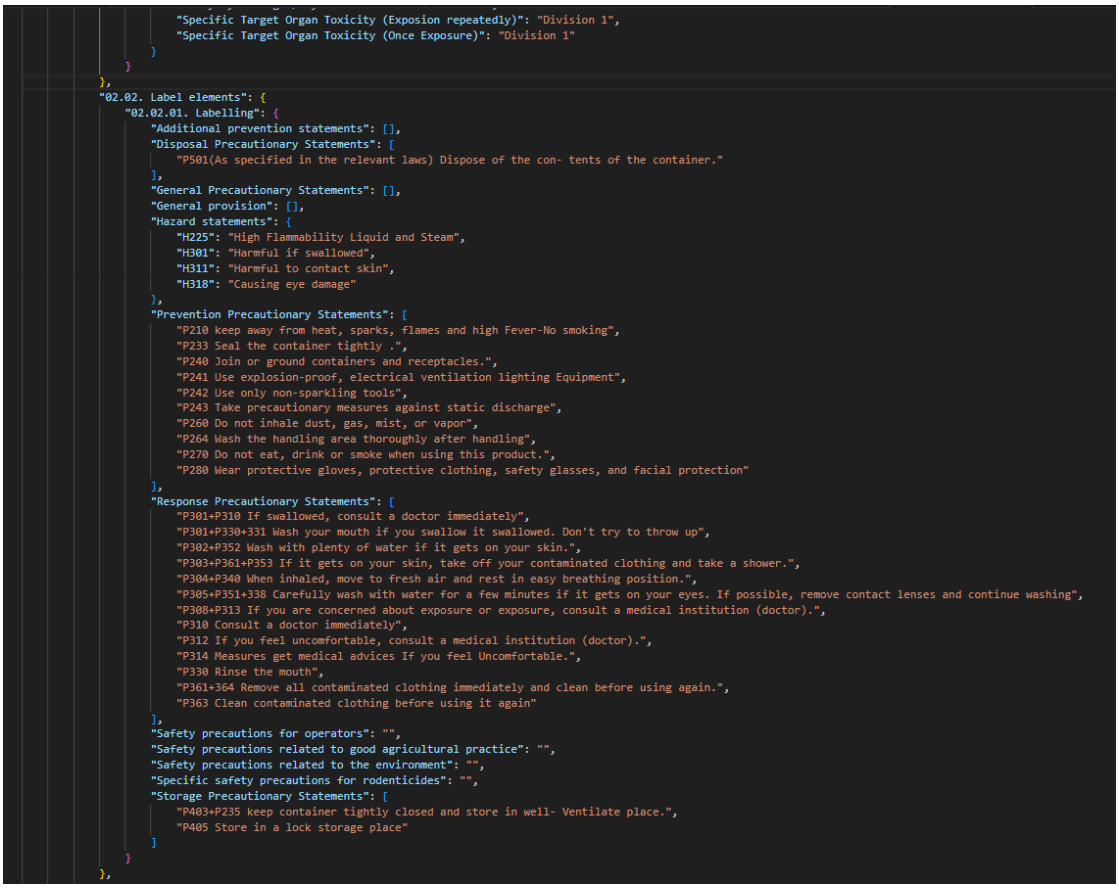
Figure 3.13 – Old SDSReader Output in the application

Deployment The deployment comes after the docker image push into the repository.We use HELM Charts as well as Jenkins to architect and deploy the containers in the kubernetes cluster.

3.2.5 Results and Comments

3.2.5.1 Difference between old solution and new one:

As we see in figure 3.13, the first column is where we should extract all the key value pairs from the pdf, the second one is the matching key value pairs. Chapter 2 is one of the most important sections in any sds since it contains all the precautionary and hazard statements. As we can see, the old code has failed to match any of them. Contrary while using the new code with GPT, figure 3.14 shows the JSON Output direclty from the console that sould be send to the backend and then display it on the front. The extraction was perfect and all the data are there. In appendices 18, 19, 20 are the page of the pdf from where we did the extraction. The execution time is also very fast when calling the api, as we can see in figure 3.15 clocking at a less than 32 seconds time , thanks to the parallel requests for both raw extraction and matched extraction. This test has been done using Postman.



```
            "Specific Target Organ Toxicity (Exposion repeatedly)": "Division 1",
            "Specific Target Organ Toxicity (Once Exposure)": "Division 1"
        }
    }
},
"02.02. Label elements": {
    "02.02.01. Labelling": {
        "Additional prevention statements": [],
        "Disposal Precautionary Statements": [
            "P501(As specified in the relevant laws) Dispose of the con- tents of the container."
        ],
        "General Precautionary Statements": [],
        "General provision": [],
        "Hazard statements": {
            "H225": "High Flammability Liquid and Steam",
            "H301": "Harmful if swallowed",
            "H311": "Harmful to contact skin",
            "H318": "Causing eye damage"
        },
        "Prevention Precautionary Statements": [
            "P210 keep away from heat, sparks, flames and high Fever-No smoking",
            "P233 Seal the container tightly .",
            "P240 Join or ground containers and receptacles.",
            "P241 Use explosion-proof, electrical ventilation lighting Equipment",
            "P242 Use only non-sparkling tools",
            "P243 Take precautionary measures against static discharge",
            "P260 Do not inhale dust, gas, mist, or vapor",
            "P264 Wash the handling area thoroughly after handling",
            "P270 Do not eat, drink or smoke when using this product.",
            "P280 Wear protective gloves, protective clothing, safety glasses, and facial protection"
        ],
        "Response Precautionary Statements": [
            "P301+P310 If swallowed, consult a doctor immediately",
            "P301+P330+331 Wash your mouth if you swallow it swallowed. Don't try to throw up",
            "P302+P352 Wash with plenty of water if it gets on your skin.",
            "P303+P361+P353 If it gets on your skin, take off your contaminated clothing and take a shower.",
            "P304+P340 When inhaled, move to fresh air and rest in easy breathing position.",
            "P305+P351+338 Carefully wash with water for a few minutes if it gets on your eyes. If possible, remove contact lenses and continue washing",
            "P308+P313 If you are concerned about exposure or exposure, consult a medical institution (doctor).",
            "P310 Consult a doctor immediately",
            "P312 If you feel uncomfortable, consult a medical institution (doctor).",
            "P314 Measures get medical advices If you feel Uncomfortable.",
            "P330 Rinse the mouth",
            "P361+364 Remove all contaminated clothing immediately and clean before using again.",
            "P363 Clean contaminated clothing before using it again"
        ],
        "Safety precautions for operators": "",
        "Safety precautions related to good agricultural practice": "",
        "Safety precautions related to the environment": "",
        "Specific safety precautions for rodenticides": "",
        "Storage Precautionary Statements": [
            "P403+P235 keep container tightly closed and store in well- Ventilate place.",
            "P405 Store in a lock storage place"
        ]
    }
},
```


Figure 3.14 – New SDSReader Output in JSON format

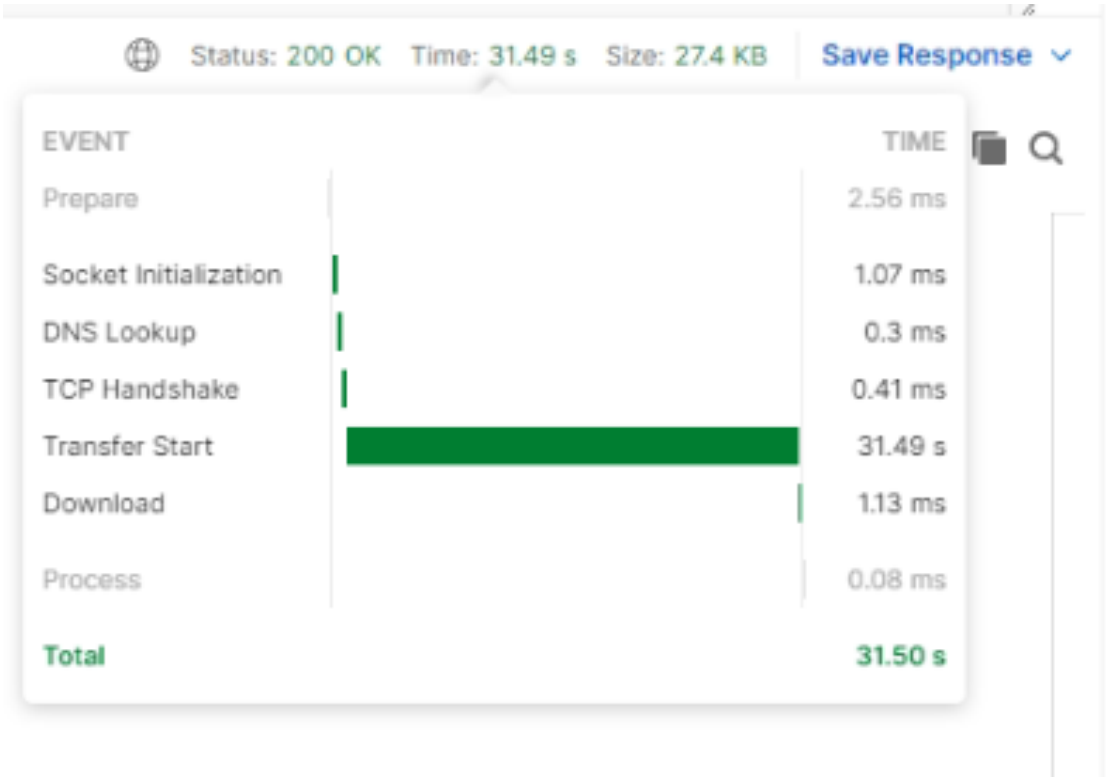


Figure 3.15 – Extraction Time for new SDSReader

3.3 ChatBot: EcoMundo Smart Assistant

3.3.1 Presentation of the project

With the release of ChatGPT, the chatbot paradigm has shifted from simple Question Answering models to a human-to-human interaction. In Ecomundo, the idea of having a chatbot was needed to answer clients' frequently asked questions as well as special cases. The ideal solution would be a chatbot integrated in all of Ecomundo's Applications (Starting first as a beta version with Cosmetic Factory).

With a chatbot powered by gpt-3.5-turbo (or whatever latest and high quality per price range model is available), this project is very doable.

3.3.2 Theoretical solution

The problem with using a gpt model directly, is that it has zero knowledge about our software (Obviously it was never intended to be used directly), has little to zero knowledge about the updated regulations since its training database dates to april 2021. The first solution that comes to mind is to fine tune it with our database of questions already answered by the commercial team as well as the latest regulations and documentation about our software. But this will be a finnacial hit since the model fine-tuned costs 0.12 USD/1K

tokens (6 times more expensive than normal DAVINCI model) + 0.0300 / 1K tokens for training it. We need another solution:

3.3.2.1 Langchain: Context Injection

Langchain as I briefly described in section 2.2 is an opensource framework that allows the users to create agents, chatbots and question documents using Large Language Models. It mainly works using a method called **Context Injection**. This method involves using vectorstores in order to calculate similarity searches between the embedding of the user's question and our context documents. If there is already a chat history with the chatbot, we will create a standalone question that reformulate the questions taking into consideration that history.

The context documents will be splitted into chunks, so the search will actually be between the query and the chunks. The figure 3.16 explains exaclty how this method works:

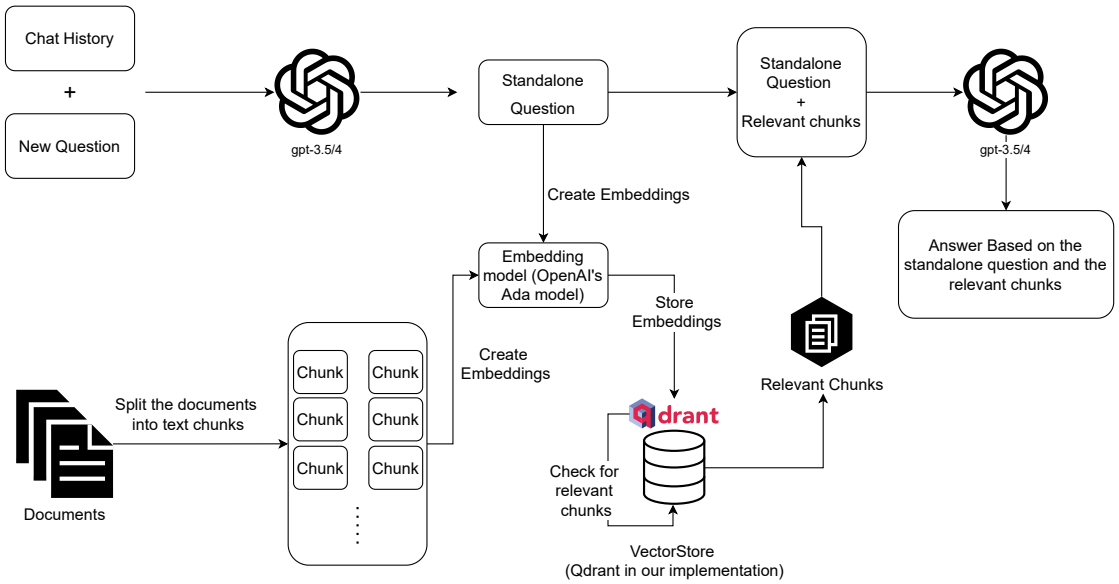


Figure 3.16 – Context Injection using GPT Models

1. We start by taking the client's question and combining it in a comprehensive way with the chat history if it exists, then we generate a standalone question using GPT model. We create using an embedding model like OpenAI's ADA for the standalone question.
2. We then take our context documents and split them into text chunks. We also embedd these chunks using the same model. **This step will be done only once unless we add new documents in the future**.
3. We compute using the cosine similarity (or any other similarity metric) to extract the top chunks that are relevant to our question. **This step is the weakest link in this architecture since it's dependent on how good the similarity metric is.**

4. We combine the standalone Question and the relevant chunks and we send them as a request to GPT model again. This step can be done with Langchain in four ways:
   (a) **Stuffing:** Takes all the context and puts it in one request with the question.
   (b) **Map Reduce:** Sends multiple requests in parallel, each one with a chunk of text. At the end it combines all the answers in one final answer.
   (c) **Refine:** Sends multiple requests in sequences, each one with a chunk of text. After having an answer from one request, it is refined using the next context chunk, if no improvements are possible then the previous answer is preserved.
5. We output the final Answer to the user.

3.3.2.2 Future work

This project was only introduced as a concept, we managed to make a proof of concept when we used it in PDFReader in the case of a very long section or when a document that we was not splitted correctly into sections. We still need to test it again with its main purpose which is a chatbot, figure out how to integrate it into our softwares and obviously how impactful it will be for the clients.

Conclusion

This internship report demonstrates the transformative impact of artificial intelligence (AI) on the chemical regulations industry through the execution of three distinct AI-powered projects. The projects include predicting Challenge Test outcomes in cosmetic products, reading Safety Data Sheets, and creating a chatbot using the Langchain framework. These AI applications have significantly improved product development processes, information extraction accuracy, and customer support, leading to better decision-making, resource optimization, and enhanced customer satisfaction. The report emphasizes the importance of responsible AI usage, including data privacy, transparency, and ongoing model monitoring, to ensure the integrity and reliability of AI-powered systems. The successful adoption of AI methodologies lays the groundwork for continued exploration of AI's role in shaping a more efficient, safer, and customer-centric chemical regulations industry in the future.

Bibliography

[1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[3] Ecomundo. Conformité internationale - Services & Logiciels | EcoMundo. https://www.ecomundo.eu, 2023. Accessed on June 5, 2023.

[4] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention, 2021.

[5] HuggingFace. The AI community building the future. https://huggingface.co/, 2023.

[6] LangChain. Building applications with LLMs through composability. https://github.com/hwchase17/langchain, 2023.

[7] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[8] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[9] OpenAI. OpenAI. https://openai.com/, 2023.

[10] OpenAI. OpenAI Usage Policy. https://openai.com/policies/api-data-usage-policies, 2023.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Appendices

```yaml
include:
  - template: Jobs/SAST.gitlab-ci.yml
  - template: Jobs/Dependency-Scanning.gitlab-ci.yml
  - template: Jobs/Secret-Detection.gitlab-ci.yml
  - template: Docker.gitlab-ci.yml

stages:
  - lint
  - test
  - container-build

.python-req:
  image: python:3.9
  before_script:
    - pip3 install pyflakes
  script:
    - pip install -r requirements.txt

lint-python:
  extends: .python-req
  stage: lint
  allow_failure: true
  script:
    - pyflakes .
sast:
  stage: test


docker-build:
  stage: container-build
  image: docker.io/ecomundoit/worker:1.22
  before_script:
    - echo "${DOCKER_IO_PASSWORD}" | podman login --username "${DOCKER_IO_USER}" --password-stdin docker.io
  script:
    - "[ -f ./build.sh ] && ./build.sh"
    - IMAGE_FULL_PATH="docker.io/ecomundoit/${CI_PROJECT_NAME}:$(echo "${CI_COMMIT_REF_NAME}" | sed -E 's/^master|main$/latest/g')"
    - buildah build --jobs=2 --manifest="${IMAGE_FULL_PATH}" --platform=linux/amd64,linux/arm64/v8 --format docker .
    - buildah manifest push --all --format v2s2 "${IMAGE_FULL_PATH}" "docker://${IMAGE_FULL_PATH}"
  rules:
    - if: $CI_COMMIT_BRANCH
    - if: $CI_COMMIT_TAG

```

Figure 17 – Python CI pipeline

LGCOM FINECOAT

Material Safety Data Sheet

PT21-S

Page 1/6
Revision Date: July/29/2022

1. PRODUCT / COMPANY IDENTIFICATION

Product Name	:	PT21-S
Chemical Identification	:	Heat curable coating material
Manufacturer's Name	:	Finecoat Co., Ltd.
Address	:	340-1, Banwol-Dong, Hwaseong-Si. Gyeong-gi-Do, 18384, KOREA
Phone No.	:	82-31-225-6201
Fax No.	:	82-31-225-6207
Emergency Phone No.	:	82-31-225-6201

2. HAZARDS IDENTIFICATION

A. Hazards. Dangerousness identification

Flammable Liquid : Division 2

Acute toxicity (Oral) : Division 3

Acute toxicity (Skin) : Division 3

Heavy Eye Damage / Eye irritation : Division 1

Specific Target Organ Toxicity (Once Exposure) : Division 1

Specific Target Organ Toxicity (Exposion repeatedly) : Division 1

B. Warning Signs items which include Prophylactic measures

Pictorial Symbol

Signal Words	Risk
Harmful / Risk phrases	H225 High Flammability Liquid and Steam
	H301 Harmful if swallowed
	H311 Harmful to contact skin
	H318 Causing eye damage
Prevention action phrase	
Prevention	P210 keep away from heat, sparks, flames and high Fever-No smoking
	P233 Seal the container tightly .
	P240 Join or ground containers and receptacles.

RESEARCH & DEVELOPMENT DEPARTMENT / FINECOAT CO., LTD.

Figure 18 – Demo Chapter 2

LGCOM FINECOAT

Material Safety Data Sheet

PT21-S

Page 2/6
Revision Date: July/29/2022

	P241 Use explosion-proof, electrical ventilation lighting Equipment P242 Use only non-sparkling tools P243 Take precautionary measures against static discharge P260 Do not inhale dust, gas, mist, or vapor P264 Wash the handling area thoroughly after handling P270 Do not eat, drink or smoke when using this product. P280 Wear protective gloves, protective clothing, safety glasses, and facial protection
Response	P301+P310 If swallowed, consult a doctor immediately P301+P330+331 Wash your mouth if you swallow it swallowed. Don't try to throw up P302+P352 Wash with plenty of water if it gets on your skin. P303+P361+P353 If it gets on your skin, take off your contaminated clothing and take a shower. P304+P340 When inhaled, move to fresh air and rest in easy breathing position. P305+P351+338 Carefully wash with water for a few minutes if it gets on your eyes. If possible, remove contact lenses and continue washing P308+P313 If you are concerned about exposure or exposure, consult a medical institution (doctor). P310 Consult a doctor immediately P312 If you feel uncomfortable, consult a medical institution (doctor). P314 Measures get medical advices If you feel Uncomfortable. P330 Rinse the mouth P361+364 Remove all contaminated clothing immediately and clean before using again. P363 Clean contaminated clothing before using it again
Storage	P403+P235 keep container tightly closed and store in well-Ventilate place.

RESEARCH & DEVELOPMENT DEPARTMENT / FINECOAT CO., LTD.

Figure 19 – Demo Chapter 2

LGCOM FINECOAT

Material Safety Data Sheet

PT21-S

	P405 Store in a lock storage place
Disposal	P501(As specified in the relevant laws) Dispose of the con-Tents of the container.

C. Hazardous which is not included in other dangerous risks.

Health	1
Fire	3
Reactivity	0

3. COMPOSITION/INFORMATION ON INGREDIENTS

Components	CAS No.	Content
Polyester resin	65072-11-9	2-4 wt%
Silicone Dioxide	7631-86-9	2-4 wt%
De-Ionized water	7732-18-5	20-30 wt%
Methyl Alcohol	67-56-1	50-65%

4. FIRST AID MEASURES

Inhalation	:	Remove to fresh air. Give artificial respiration if not breathing.
Skin Contact	:	Wash skin with soap, rinse with water, seek medical attention.
Eye Contact	:	Immediately flush eyes with water for at least 15 minutes, then obtain medical attention.
Ingestion	:	Do not induce vomiting, if vomiting should occur spontaneously keep airway clear, seek medical attention.
Others	:	Never give anything by mouth to an unconscious person. Discard contaminated clothing and shoes immediately.

5. FIRE FIGHTING MEASURES

Extinguishing Media	:	Carbon dioxide, dry chemical media, appropriate foam
Explosion Hazards	:	Keep container cool by spraying with water if exposed to fire. Product may polymerize at high temperature. Polymerization is a highly exothermic reaction and may produce sufficient heat to cause thermal decomposition and/or rupture of container. Vapor may travel considerable distance to source of ignition and flash back
Special Equipment	:	Self-contained breathing apparatus and protective clothing should be provided to fire fighters in building or confined area where this product is stored

6. ACCIDENTAL RELEASE MEAUSRES

RESEARCH & DEVELOPMENT DEPARTMENT / FINECOAT CO., LTD.

Figure 20 – Demo Chapter 2

Résumé

Ce rapport présente les résultats et les retombées de trois projets qui ont utilisé l'intelligence artificielle (IA) pour améliorer les opérations dans l'industrie des réglementations chimiques.

Le premier projet visait à développer une solution alimentée par l'IA pour prédire les résultats des Challenge Tests, qui évaluent l'efficacité des produits cosmétiques pour inhiber la croissance des microorganismes. En analysant la formulation des produits cosmétiques à l'aide d'algorithmes d'apprentissage automatique, la solution fournissait des prédictions précises des résultats des tests. Cette optimisation a permis aux formulateurs d'améliorer leurs formulations et d'augmenter les chances de réussite des tests, réduisant ainsi le temps et les ressources nécessaires pour le développement de produits. Ce projet a été mené en collaboration avec le groupe L'Occitane et intégré dans le logiciel Cosmetic Factory, une solution de gestion du cycle de vie des produits (PLM).

Le deuxième projet était axé sur le développement d'une solution d'IA pour la lecture des fiches de données de sécurité (FDS). La solution existante utilisait un code statique pour rechercher des mots clés dans des fichiers PDF et les faire correspondre avec une base de données. Cependant, cette approche présentait des limites pour traiter différentes langues et variations dans les paires clé-valeur. La nouvelle solution proposait l'utilisation de grands modèles de langage et l'injection de contexte pour améliorer la précision et l'efficacité de l'extraction. Cela a renforcé la capacité à extraire des informations clés des FDS, améliorant la sécurité et la conformité dans l'industrie des réglementations chimiques.

Le troisième projet visait à développer un chatbot alimenté par un grand modèle de langage pour fournir un support et répondre aux questions fréquemment posées. Cependant, le coût de l'adaptation fine du modèle avec la base de données de l'entreprise était prohibitif. En guise d'alternative, le cadre Langchain a été proposé, utilisant l'injection de contexte pour fournir des réponses pertinentes en fonction de la question de l'utilisateur et des documents de contexte disponibles. Cette solution de chatbot a amélioré le support client et l'efficacité pour répondre aux demandes des clients.