جامعة ابن طفيل
ⵜⴰⵙⴷⴰⵡⵉⵜ ⵉⴱⵏ ⵟⴼⵉⵍ
Ibn Tofaïl University
Faculty of Science

Ibn Tofail University
Faculty of Science, Kenitra

End-of-Studies Project Thesis

Master's Degree in Artificial Intelligence and Virtual Reality

Development of a Multi-Agent System for ESG Report Generation

Host institution: Version28 LTD

Prepared by: *Mr. Wassim EL BAKKOURI*

Supervised by: *Mrs. Fadoua GHANIMI (ENSC KÉNITRA, UIT)*

Mr Reyadh Mallam Hassam (Version28 LTD)

Defended on 2024, in front of the jury composed of:

- Mrs Raja TOUAHNI (FS KÉNITRA ,UIT)
- Mrs Fadoua GHANIMI (ENSC KÉNITRA ,UIT)
- Mr Rochdi MESSOUSSI (FS KÉNITRA ,UIT)
- Mr Anass NOURI (FS KÉNITRA ,UIT)
- Mr Tarik BOUJIHA (FS KÉNITRA ,UIT)

Academic year 2023/2024

Dedication

“

To my family, for your boundless love, support, and encouragement To my parents, for your sacrifices and teaching me the value of hard work and kindness To my siblings, for your uplifting words and faith in me And to my teachers, for your guidance and inspiration throughout my academic journey, Thank you all for being a crucial part of my journey.

”

- **Wassim EL BAKKOURI**

Acknowledgements

First and foremost, I thank Allah, the Almighty, for granting me the strength and patience to complete this work. I extend my heartfelt thanks to my supervisor,**Fadoua Ghanimi**, for her invaluable guidance, patience, and encouragement. Her critical insights were instrumental in structuring this work and enhancing the quality of various sections.

I also wish to express my deep gratitude to my CEO, **Pete Melhuich**, at **Version28**, for his immense support, meticulous guidance, and the wealth of advice and information he provided with unmatched patience and professionalism.

My sincere thanks go to **Reyadh Mallam Hassam**, my supervisor, for his constant support and invaluable contributions to this project. I also want to thank **Helen Yu** and **Milan Hawkins** for their continuous encouragement and assistance throughout this journey.

A special thanks to **Ms. Raja Touahn**i, the Master Coordinator, for her unwavering support and guidance throughout my academic journey.

Lastly, I would like to extend my deepest appreciation to all my teachers, whose dedication, knowledge, and inspiration have been a vital part of my learning experience.

Lastly, I would like to thank everyone who contributed, directly or indirectly, to the completion of this work.

Abstract

During my tenure at Version28 from February to July 2024, I led the development of an advanced system for generating Environmental, Social, and Governance (ESG) reports, leveraging sophisticated Large Language Models (LLMs) alongside Retrieval-Augmented Generation (RAG) techniques. This project aimed to improve the extraction and structuring of ESG data, thereby facilitating compliance with global standards such as the Global Reporting Initiative (GRI).

The initiative began with the design and implementation of a scalable microservice architecture on Microsoft Azure, which enabled the seamless integration of LLMs into the RAG pipeline. My role primarily involved directing the creation of this pipeline, as well as the development of a multi-type knowledge base to support precise and efficient data retrieval.

During the months of February and March, my focus was on optimizing the system's data ingestion processes and refining the ESG metadata module. This module was crucial for organizing and structuring the extracted data according to established ESG criteria, ensuring that the information was both relevant and accurate. Additionally, I implemented monitoring and evaluation protocols to ensure that the system's performance adhered to the highest academic and industry standards, thereby supporting continuous improvement and adaptation.

In conclusion, this project resulted in the development of a robust solution that significantly enhanced the capacity of businesses to produce detailed ESG reports. It underscores my expertise in integrating LLMs with advanced data retrieval methods to address complex challenges in the field of ESG reporting.

Durant mon stage chez Version28, de février à juillet 2024, j’ai dirigé le développement d’un système avancé pour la génération de rapports Environnementaux, Sociaux et de Gouvernance (ESG). Ce projet s’appuyait sur des modèles de langage avancés (LLM) et des techniques de génération augmentée par récupération (RAG) afin d'améliorer l'extraction et la structuration des données ESG, facilitant ainsi la conformité avec des normes internationales telles que la Global Reporting Initiative (GRI).

Le projet a débuté par la conception et la mise en œuvre d'une architecture de microservices évolutive sur Microsoft Azure, permettant l’intégration fluide des LLM dans le pipeline RAG. Ma mission a principalement consisté à superviser la création de ce pipeline ainsi que le développement d'une base de connaissances multi-types pour garantir une récupération de données précise et efficace.

Entre février et mars, je me suis concentré sur l’optimisation des processus d’ingestion de données et sur l’amélioration du module de métadonnées ESG. Ce module était crucial pour organiser et structurer les données extraites selon des critères ESG prédéfinis, assurant ainsi que les informations soient pertinentes et précises. De plus, j'ai établi des protocoles de suivi et d’évaluation pour garantir que la performance du système réponde aux normes académiques et industrielles les plus élevées, favorisant ainsi une amélioration continue et une adaptation efficace.

En résumé, ce projet a permis de développer une solution robuste qui renforce significativement la capacité des entreprises à produire des rapports ESG détaillés. Il met en avant mon expertise dans l’intégration des LLM avec des méthodes avancées de récupération de données pour résoudre des défis complexes dans le domaine du reporting ESG.

Table of Contents

Figures list

List of Abbreviations and Acronyms

Term	Description
SIC code	Standard Industrial Classification
ESG	Environmental, social, and governance
MVS	Minimum viable solution
MVP	Minimum viable product
ORM	Object Relational Mapper
VM	Virtual Machine
BERT	Bidirectional transformers for language understanding
LLM	Large language model
RAG	Retrieval augmented generation
AI	Artificial Intelligence
API	Application Programming Interface
NLP	Natural Language Processing

General Introduction

This report explores the development steps and implementation details behind **i-Prove**, an innovative AI-driven platform designed to assist businesses in generating detailed ESG (Environmental, Social, and Governance) reports. The platform focuses on automating the extraction, structuring, and reporting of ESG data using sophisticated technologies such as **Large Language Models (LLMs)** and **Retrieval-Augmented Generation (RAG)** techniques. With the growing regulatory demands for transparency and compliance in ESG practices, i-Prove offers small and medium-sized enterprises (SMEs) and startups a comprehensive, efficient solution for their reporting needs.

The development of i-Prove began with the design of a **scalable microservice architecture**, hosted on Microsoft Azure, which allowed for seamless integration of LLMs into the RAG pipeline. This architecture was crucial in ensuring system stability and flexibility, enabling the platform to handle large volumes of data while delivering accurate, contextual responses. The system's **data ingestion processes** were optimized to efficiently capture and organize ESG-related data, with particular attention given to the ESG metadata module. This module plays a critical role in aligning the data with established standards such as the **Global Reporting Initiative (GRI)**, ensuring the reports produced are relevant, structured, and compliant.

Throughout the project, multiple AI-driven features were implemented, including **automated report generation** powered by sophisticated algorithms, which dramatically reduced the time and effort required for businesses to compile their ESG reports. The platform also integrates data from a wide range of sources, providing a holistic view of a company’s ESG performance. The **customization options** built into i-Prove allow businesses to tailor their reports to meet specific stakeholder demands and reporting frameworks, enhancing the flexibility and relevance of the disclosures.

This report will provide a comprehensive overview of the **i-Prove development**, focusing on the challenges faced during the ESG report generation process and how the platform’s architecture and AI features addressed these complexities. The combination of **LLMs** and **RAG** techniques marks a significant advancement in ESG reporting, offering businesses a transformative tool to streamline their sustainability initiatives while ensuring compliance with global standards.

Chapter 1:
The host company

1. Core business

Figure 1:Version28 logo

Version28 is a Venture Capital Accelerator based in London, UK, primarily focusing on supporting innovative businesses in their growth journey. The company offers a comprehensive suite of services aimed at helping these businesses secure funding and achieve sustainable success.

2. Services Offered:

Figure 2:Version28 Banner

Fractional Services:

Version 28 provides a range of fractional services, which essentially means offering part-time expertise from experienced professionals. These services can cover various business functions, such as finance, marketing, or operations, to help startups and SMEs scale efficiently.

Tailored Services:

Recognizing that every business is unique, Version28 offers customized solutions tailored to the specific needs of its clients. These services could include strategic planning, business development, or operational improvement.

Financial Planning:

A crucial aspect of business growth is sound financial management. Version28 provides financial planning services to help businesses make informed decisions about their finances, including budgeting, forecasting, and investment analysis.

Debt Financing:

In addition to equity-based funding, Version28 also offers debt financing options. This

can be a valuable source of capital for businesses looking to expand operations, invest in new projects, or manage cash flow effectively.

ESG Reporting:
Recognizing the growing importance of sustainability, Version28 offers ESG reporting services to help businesses assess and communicate their environmental, social, and governance impact.

3. Understanding ESG

ESG stands for Environmental, Social, and Governance. It represents a set of criteria used to evaluate a company's or investment's dedication to sustainable and ethical practices. ESG assessments focus on:

- **Environmental**: Evaluates how a company manages its environmental impact, including carbon footprint, waste management, and resource utilization.
- **Social**: Assesses the company's treatment of employees, customers, and communities, including labor practices, diversity and inclusion, and community engagement.
- **Governance**: Reviews the quality of a company's leadership, ethical practices, transparency, and governance structures.

4. Significance of ESG

ESG considerations are vital as they reflect a company's commitment to sustainability and ethical practices, which can influence its long-term success and risk management. Investors increasingly view strong ESG performance as indicative of resilience, effective risk management, and sustainable returns. Furthermore, a focus on ESG can enhance a company's reputation, attract top talent, and foster stakeholder trust.

5. Applications of ESG

ESG criteria are utilized in various areas:

- **Investment Analysis**: Investors use ESG metrics to assess potential investments, aligning their portfolios with sustainable and ethical values.
- **Corporate Reporting**: Companies disclose their ESG performance through sustainability or integrated annual reports, demonstrating their dedication to responsible practices.

- **Regulatory Compliance**: ESG reporting frameworks and standards, such as the Global Reporting Initiative (GRI), the Sustainability Accounting Standards Board (SASB), and the Task Force on Climate-related Financial Disclosures (TCFD), provide guidelines for reporting ESG performance.

6. Developing an ESG Report

Creating an ESG report involves several critical steps:

- **Data Collection**: Collect quantitative and qualitative data on environmental impacts, social practices, and governance structures, including metrics on energy usage, diversity, and governance practices.
- **Materiality Assessment**: Identify key ESG issues relevant to the company's operations and stakeholders, evaluating the impact and significance of these factors.
- **Stakeholder Engagement**: Engage with stakeholders such as employees, customers, investors, and community members to gather insights and expectations regarding ESG performance.
- **Framework Selection**: Choose an appropriate reporting framework or standard, such as GRI or SASB, to structure the report effectively.
- **Data Analysis and Reporting**: Analyse the collected data to assess performance against ESG criteria and prepare the report, including an overview of policies, performance metrics, achievements, challenges, and future goals.
- **Verification and Assurance**: Seek third-party verification or assurance to enhance the credibility and accuracy of the ESG report, if applicable.
- **Publication and Communication**: Publish the ESG report and communicate it to stakeholders through various channels, including company websites, investor relations platforms, and sustainability forums.

7. Challenges in ESG Report Development:

Despite the increasing emphasis on ESG reporting, the process of developing an ESG report remains a significant challenge for many organizations. The meticulous nature of data collection, the complexity of conducting a thorough materiality assessment, and the necessity of engaging diverse stakeholders all contribute to the considerable time and effort required. Additionally, the need to adhere to various reporting frameworks and standards, coupled with the demand for accurate data analysis and third-party verification, further complicates the process. As a result, creating a comprehensive and credible ESG report can be a resource-intensive task, often straining organizational capacities and delaying timely disclosures.

8. Solution: An AI-Driven Solution for ESG Reporting

Figure 3:I-prove logo

In today's rapidly evolving business landscape, the integration of Environmental, Social, and Governance (ESG) principles has become a cornerstone of responsible corporate practice. With increasing regulatory scrutiny and complexity surrounding ESG reporting, businesses are seeking innovative solutions to streamline their reporting processes. i-Prove, a cutting-edge AI-assisted ESG report writer and management platform, emerges as a beacon for small and medium-sized enterprises (SMEs) and startups aiming to enhance their ESG practices. Meticulously designed to support these businesses in implementing impactful ESG transformations, i-Prove offers essential tools and resources to help organizations mitigate 'greenwashing' and foster genuine corporate citizenship. Backed by a team of industry leaders in ESG, corporate governance, and sustainability, the platform provides expert guidance throughout the ESG certification process, ensuring that businesses can achieve excellence in their ESG endeavors.

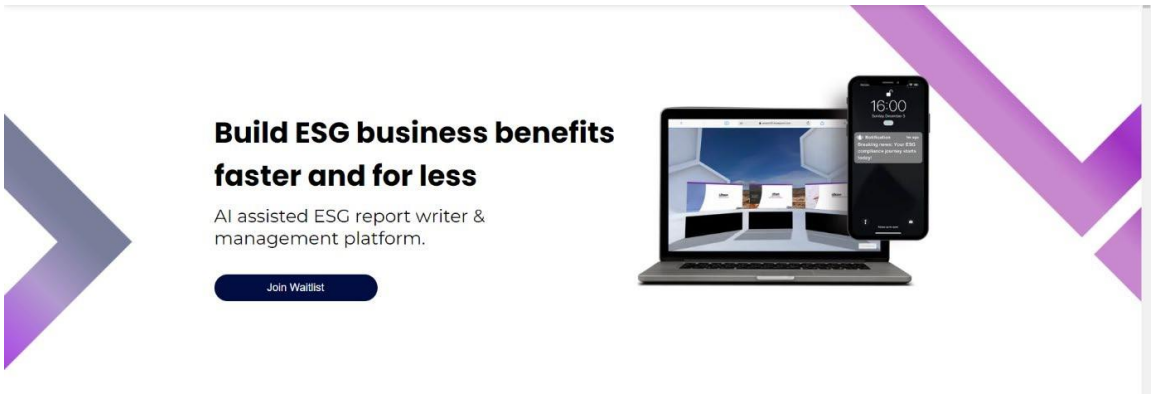

Figure 4:I-prove banner

Chapter 2:
Project Introduction

1. Background

ESG reporting has gained importance as businesses adopt sustainable practices, but creating comprehensive reports remains challenging, particularly for SMEs and startups. Current AI-powered ESG tools are often basic and fail to address the complexity of ESG data. However, AI has the potential to revolutionize the process by automating data analysis, report generation, and compliance, reducing the time and resources required. As AI advances, it will make ESG reporting more efficient, accurate, and impactful, allowing companies to focus on sustainability and long-term value creation.

2. State of the Art ESG Reporting Tools

The ESG reporting landscape has evolved with the advent of sophisticated platforms like Credible ESG, Novisto, and Diginex, each offering unique capabilities:

a. Credible ESG :

Figure 5:Credibl logo

Strengths: Provides a comprehensive suite for ESG data management and reporting, facilitating alignment with global standards.
Weaknesses: Limited in advanced AI integration, primarily relying on conventional data analysis methods.

a. Novisto:

Figure 6:Novisto logo

Strengths: Excels in ESG data collection, analysis, and customizable reporting, empowering businesses to make informed sustainability decisions.
Weaknesses: While robust, it lacks cutting-edge AI features that could enhance

automation and insight generation.

Figure 7:Diginex logo

b. Diginex:

Strengths: Leverages blockchain technology to ensure data integrity, offering a transparent and compliant ESG reporting solution.
Weaknesses: The platform's AI capabilities are relatively basic, limiting its potential to streamline the reporting process.

While these tools mark significant progress in ESG reporting, their AI functionalities remain underdeveloped. This gap underscores the need for more advanced AI-driven solutions, such as i-Prove, to further enhance the efficiency and effectiveness of ESG reporting, particularly in reducing time and resource demands.

3.Our solution:

The proposed solution aims to address the critical challenges faced by small and medium-sized enterprises (SMEs) and startups in generating comprehensive Environmental, Social, and Governance (ESG) reports. This solution leverages advanced artificial intelligence (AI) technologies to create a sophisticated ESG reporting tool that enhances the efficiency and effectiveness of the reporting process.

4.Key Features :

- **Automated Data Analysis**: The AI system automates the collection and analysis of relevant ESG data, significantly reducing the time required to compile reports. This automation enables businesses to focus on strategic initiatives rather than manual data entry.
- **Comprehensive Report Generation**: The tool facilitates the generation of detailed ESG reports that align with international standards, including the Global Reporting Initiative (GRI) framework. It ensures that all necessary disclosures are included, promoting transparency and compliance.
- **AI-assisted Recommendations**: After generating a report, users receive tailored advice on improving their ESG performance over specified periods (e.g., 6 months, 1 year). This feature encourages continuous improvement and helps businesses demonstrate progress in their ESG initiatives.
- **Data Visualization Dashboard**: The platform includes an interactive dashboard that visualizes the total progress of ESG reporting. Users can see the split between Environmental, Social, and Governance components,

allowing for quick identification of focus areas.

- **Customizable Survey Functionality**: The tool allows users to conduct surveys to gather data on employee commuting and homeworking energy consumption. This feature helps organizations accurately assess their carbon footprint related to staff activities.
- **Material Topic Assessment:** Our solution supports single and double materiality assessments, leveraging AI to evaluate both financial and non-financial impacts of ESG issues. This AI-driven approach ensures a holistic understanding of material risks and opportunities, providing deeper insights and more accurate evaluations.
- **Benchmarking and Scoring**: Users can benchmark their ESG performance against industry standards and peer companies. The software calculates ESG scores based on user-provided data and predefined criteria, offering insights into areas for improvement.
- **Integration with Other Standards**: Future phases of the project will include the integration of additional ESG frameworks and standards, such as the European Sustainability Reporting Standards (ESRS) and the Sustainability Accounting Standards Board (SASB) criteria, ensuring comprehensive compliance and reporting capabilities.

5. Project Planning

To ensure the successful development and deployment of our AI-driven ESG reporting tool, the project has followed a structured plan, progressing from the initial proof-of-concept to the current stage of MVP testing. The planning process has been divided into the following key phases:

a. Minimal Viable Solution (MVS)

The first phase involved creating a minimal viable solution to demonstrate the effectiveness of the AI in ESG reporting. This stage focused on implementing core functionality, specifically GRI questions answering using GenAI techniques such as RAG and agentic RAG, to prove that the AI could handle the complexities of ESG data.

Objective:
Validate the core AI components and their ability to process and analyze ESG data accurately.
Testing and Validation:
Rigorous testing was conducted to ensure the AI functions as expected, with adjustments made based on performance results.

c. Minimum Viable Product (MVP)

Building on the MVS, the next step was the development of a Minimum Viable Product (MVP) that included essential features necessary to attract early users and

potential investors. The MVP provided a functional version of the tool, enabling businesses to start generating ESG reports with AI assistance.

Objective:
Develop the core features necessary for the MVP, including automated data analysis, report generation, AI-assisted recommendations, and a basic data visualization dashboard.
Funding Strategy:
The MVP was used to raise funding, showcasing the potential of the solution to investors and stakeholders.
Testing and Debugging:
The MVP underwent thorough testing to identify and resolve any bugs or issues, ensuring a stable and reliable product.

d. User Feedback and Iteration

Having completed the MVP, we are currently in the process of testing it with a select group of clients. This stage is crucial for understanding user needs and refining the product before broader deployment.

Objective:
Collect detailed feedback from early users on the functionality, usability, and effectiveness of the MVP.
Iteration:
Based on user feedback, necessary improvements and adjustments will be made to the tool to enhance user experience and performance.

e. Planning and Developing the Production Version

With insights gained from MVP testing, the next phase involves planning the full production version of the tool. This includes defining all specifications, features, and integration capabilities needed to support a wide range of ESG reporting requirements.

Objective:
Create a comprehensive plan for the production version, outlining all features, integrations, and technical specifications.
Development:
Begin the development of the production version, ensuring it meets the needs of SMEs and startups while maintaining scalability and reliability.

f. Gantt diagram:

Gantt charts are essential tools in project planning and management, offering a visual representation of a project's timeline, tasks, and milestones. Developed by Henry L. Gantt in the early 20th century, these charts facilitate the organization and sequencing of project activities, allowing for a clear understanding of task dependencies and time allocation. In the context of project planification, Gantt charts play a critical role in outlining the scope, objectives, and deliverables by enabling project managers to schedule tasks, allocate resources, and monitor progress against planned timelines. This systematic approach to project management ensures that projects are executed efficiently, with potential delays and resource constraints identified and addressed proactively.

Task	Duration
Company integration	7 days
Development of the MVS	21 days
Test and validate the MVS	7 days
Plan the MVP	7 days
Build the MVP	88 days
Test and validate the MVP	2 weeks
Total	133 days

Table 1: Project tasks

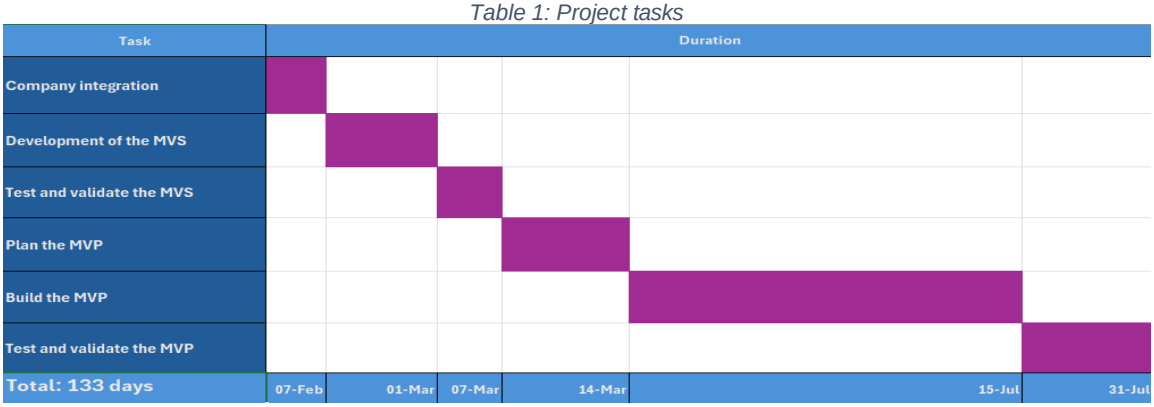

Figure 8:Gantt diagram

Chapter 3:
State of the Art in AI Answering Applications

1. Introduction

The rapid advancements in Artificial Intelligence (AI) have significantly transformed the landscape of automated communication, particularly in the development of chatbots and answering systems. AI-driven chatbots are increasingly being integrated into various sectors, enhancing user experience by providing accurate, context-aware responses. A critical innovation in this domain is the Retrieval- Augmented Generation (RAG) technique, which combines the strengths of retrieval- based methods with generative models. RAG enables chatbots to not only pull relevant information from vast document repositories but also generate coherent and contextually appropriate answers. This approach is particularly valuable in applications that require the synthesis of information from multiple sources, ensuring that users receive precise and informative responses. As AI continues to evolve, the role of RAG and similar techniques in advancing the capabilities of answering applications is poised to expand, offering more sophisticated and reliable tools for automated communication.

1. Current Capabilities

AI answering systems have made significant progress due to advancements in machine learning and natural language processing. Key developments include:

Enhanced Understanding and Contextualization:
Modern AI systems use advanced algorithms like transformer models to better understand and process natural language, leading to more accurate responses.

Integration with Knowledge Bases:AI is increasingly integrated with large data sources, enabling it to retrieve and synthesize information, making it more capable of answering complex questions.

Real-Time Interaction: Improved processing power allows AI to provide instant responses, which is crucial for areas like customer service.

Adaptive Learning and Personalization:AI systems now use adaptive learning to improve over time, offering personalized responses based on user preferences and past interactions.

Multi-Modal Capabilities:AI can handle various input types such as text, voice, and images, allowing it to provide more versatile and comprehensive answers.

2. Introduction to Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG) represents a significant advancement in AI, merging the advantages of both generative and retrieval-based models. This approach enhances the accuracy and contextual relevance of responses by retrieving pertinent information from extensive corpora and utilizing it to generate informed replies. Unlike purely generative models, which may produce responses based on limited training data, RAG ensures that answers are grounded in factual and up-to-date information.

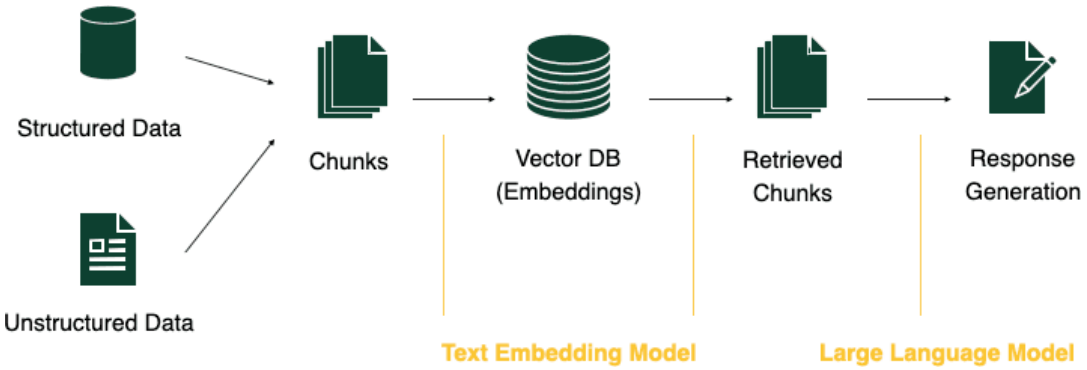

Figure 9:RAG pipeline

RAG models consist of two core components: the retriever and the generator. The retriever scans a large database to identify and extract relevant documents or sections based on the user's query. The generator then uses this retrieved information to create a coherent and precise response, leveraging the context provided by the retrieved content.

3. Core Technologies and Frameworks for RAG

a. Embedding Techniques

Embedding techniques [1] like the mixedbread-ai/mxbai-embed-large-v1 convert textual data into dense vector representations, capturing the meaning and context of text in numerical form. These embeddings map words, phrases, or sentences into high-dimensional spaces where semantically similar items are placed closer together. By passing the text through neural network layers, often using transformers like BERT, embeddings capture not only word meanings but also their relationships within the surrounding context. This helps distinguish polysemous words based on their usage.

The resulting vectors, which are often reduced in size but retain key semantic information, are useful for tasks like semantic search. In this process, the system compares the vectors of a query and documents to identify the most similar or relevant results. Dense vectors allow efficient similarity calculations, providing faster and more accurate retrieval than traditional keyword-based searches.

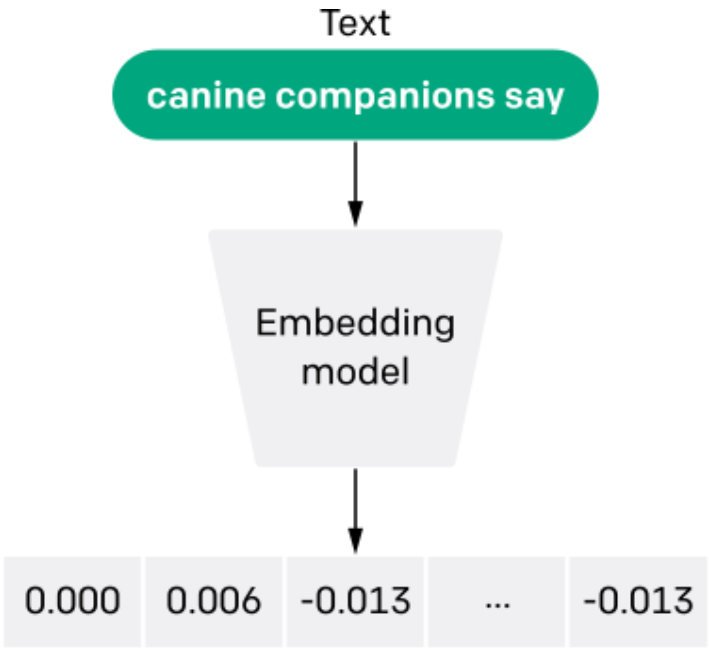

Figure 10:Embedding pipeline

b. Vector Databases

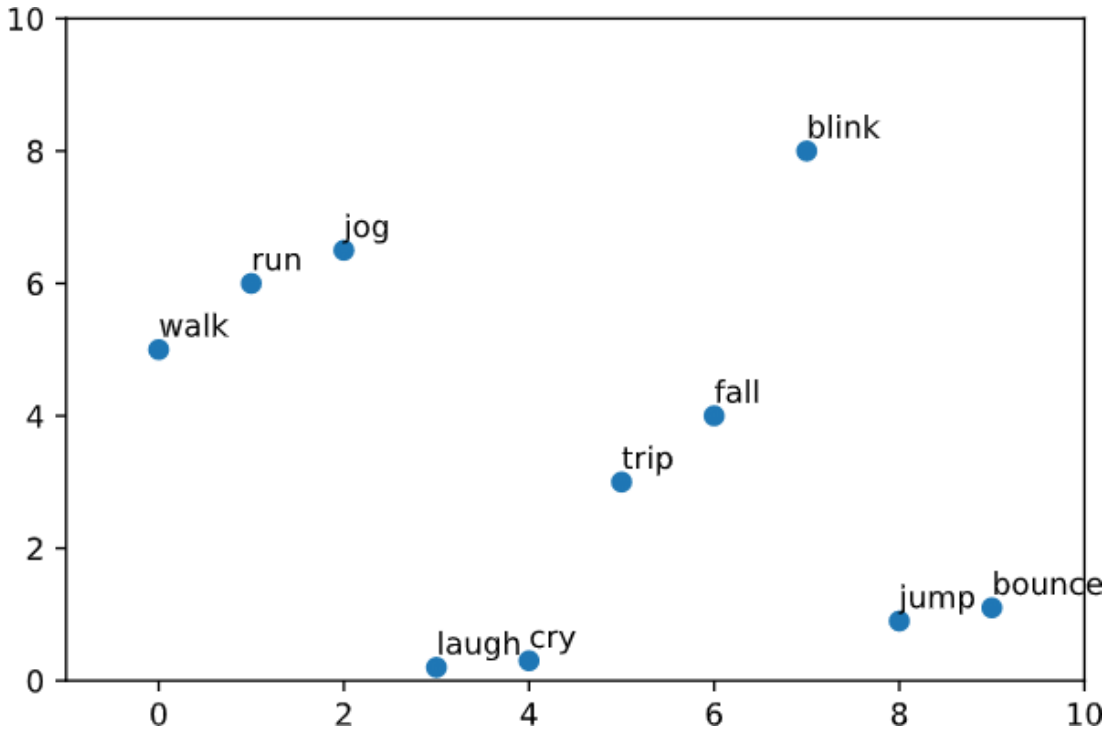

Figure 11:Embedding vector space

Vector databases like Qdrant, FAISS, and Pinecone store, manage, and retrieve dense vector representations of data, such as text or images, for tasks like similarity search. These vectors, generated by embedding models (e.g., BERT), are stored along with metadata and organized using indexing techniques, such as Approximate Nearest Neighbor (ANN) search, to ensure fast retrieval in large datasets.[4]

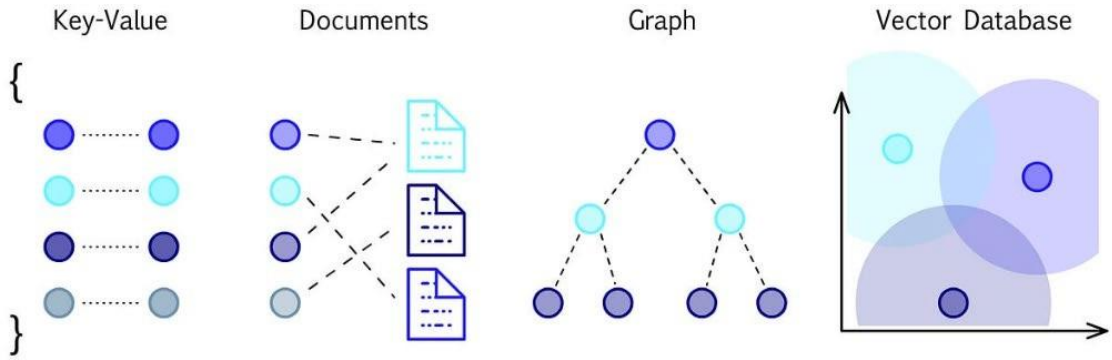

Figure 12:Vector database types

Qdrant uses HNSW indexing for proximity searches in high-dimensional spaces, while

Pinecone focuses on distributed storage and scalability for vast amounts of data. When a query is embedded into a vector, the database compares it to stored vectors using similarity metrics like cosine similarity or Euclidean distance. This allows efficient retrieval of the most relevant results, even in systems handling millions or billions of vectors. Vector databases also support scalability, working across multiple nodes or clusters, making them ideal for large-scale applications like search engines, recommendation systems, and Retrieval-Augmented Generation (RAG) systems.

c. Libraries and Frameworks

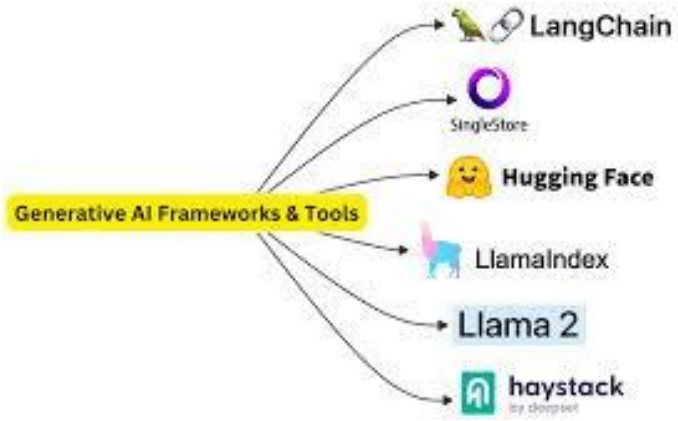

Figure 13:GenAI frameworks

Libraries and frameworks such as Langchain, Haystack, and Hugging Face Transformers are instrumental in developing and deploying RAG-based question-answering systems. They provide pre-trained models and user-friendly interfaces, streamlining the integration of RAG technologies into practical applications.

d. Key Techniques and Methodologies

In Retrieval-Augmented Generation (RAG) systems, effective retrieval techniques are crucial for improving the quality of generated responses. **Dense Passage Retrieval (DPR)** remains a widely adopted technique, encoding both queries and documents into dense vectors for efficient retrieval.

Key advancements include **Cross-Encoder Reranking**, where a secondary model refines initial results by directly comparing queries and retrieved passages, and **Hybrid Retrieval**, which blends DPR with sparse methods like BM25 for broader relevance. **Hard Negative Sampling** introduces challenging non-relevant documents during training to sharpen retrieval accuracy.

Further, **Knowledge Distillation** reduces model complexity without sacrificing precision, distilling knowledge from larger models into more efficient ones. These methods enhance RAG systems by improving both speed and precision in document

retrieval, leading to more accurate and contextually relevant responses.

Lastly, **Late Interaction Mechanisms**, like the ColBERT model, allow for token-level interaction between query and document embeddings, enhancing fine-grained matching. This combination of advanced techniques ensures that RAG systems deliver high-quality and timely responses by leveraging optimal retrieval strategies.

e. Evaluation Metrics

In evaluating RAG-based question-answering systems, the following metrics are commonly used:

1. **BLEU (Bilingual Evaluation Understudy)**: BLEU measures how similar the generated text is to a set of reference texts. It evaluates precision, focusing on the overlap of n-grams (sequences of words) between the generated and reference outputs, rewarding high similarity.

2. **ROUGE (Recall-Oriented Understudy for Gisting Evaluation)**: ROUGE evaluates both recall and precision by measuring the overlap of n-grams between generated text and references, particularly for summaries. It includes ROUGE-N (for unigrams, bigrams), ROUGE-L (longest common subsequence), etc.
3. **F1 Score**: The F1 Score is the harmonic mean of precision and recall. Precision refers to how many generated answers are correct, while recallmeasures how many relevant answers are retrieved. This metric balances the two, providing a well-rounded assessment of a model's accuracy.

These metrics offer complementary perspectives on how well a RAG system performs, from similarity to human-like responses to relevance and coverage of correct answers.

4. Integration of LLMs and Vector Stores in RAG

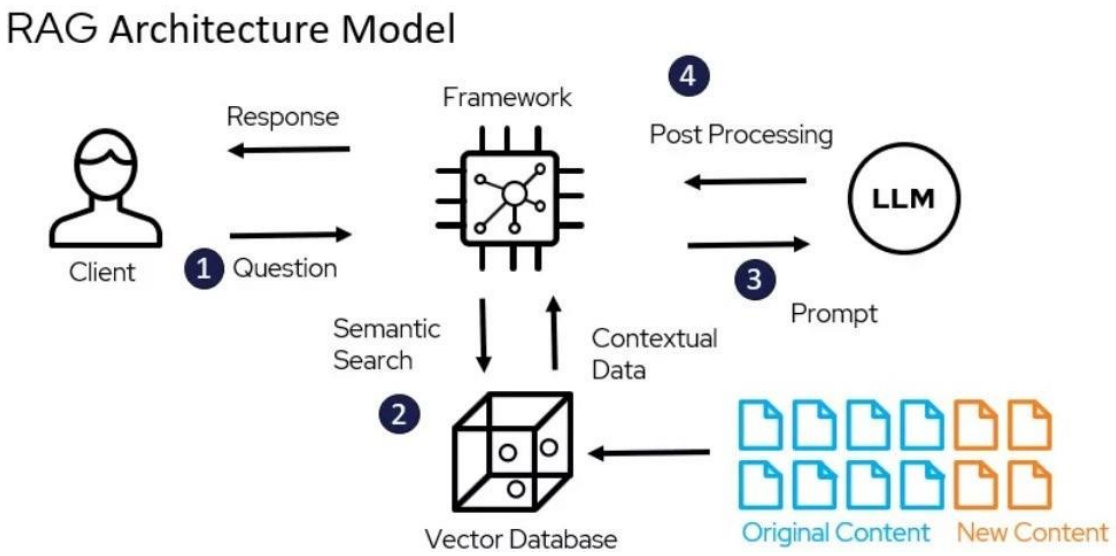

Figure 14:9. Integration of LLMs and Vector Stores in RAG

The combination of **Large Language Models (LLMs)** and **vector stores** in Retrieval-Augmented Generation (RAG) systems is transformative because it merges the generative power of LLMs with the precision of vector-based retrieval. Vector stores, which use embeddings to represent documents and queries as dense vectors, enable fast and accurate semantic searches by retrieving the most relevant information from vast data sources. The LLM then generates responses based on this retrieved data, allowing for contextually rich and accurate answers, overcoming limitations like hallucinations and irrelevant outputs.

This integration allows RAG systems to handle complex, open-ended queries by retrieving only the most relevant chunks of information while maintaining the generative flexibility of LLMs, ensuring the responses are grounded in factual data. Together, this pairing enhances the quality and accuracy of generated content, improving the system's reliability across diverse applications such as document summarization, question answering, and more.

5. Role of Large Language Models (LLMs)

LLMs, such as GPT-3, GPT-4, and their variants, are renowned for their capacity to generate coherent and contextually appropriate text [2]. These models are trained on vast datasets, enabling them to understand and generate text based on a wide range of topics. In a RAG framework, LLMs are responsible for generating responses after relevant information has been retrieved. They use their deep understanding of language to produce accurate and contextually rich answers based on the input query and the retrieved context.

In language models like GPT-3, GPT-4, and their variants, parameters such as temperature and top_k play a crucial role in shaping the generated text and can significantly influence the quality and diversity of the answers. Temperature controls the randomness of the model's predictions. A lower temperature (close to 0) makes the model more deterministic, leading to more focused and predictable responses. Conversely, a higher temperature introduces more randomness, allowing the model to explore less obvious or creative word choices, but at the risk of producing less coherent responses. For example, with a low temperature, the model might consistently generate the most probable answer, whereas with a high temperature, the model may provide a wider variety of responses, which might include more

imaginative or less relevant content.

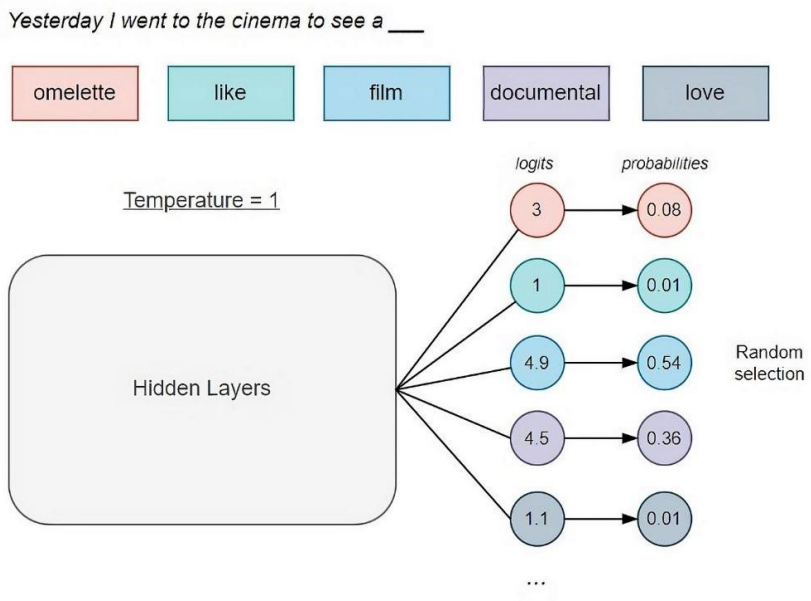

Figure 15: LLM Architecture

Top_k, on the other hand, limits the model's word selection to the top k most probable tokens during each step of the generation process. When top_k is small,

the model is restricted to selecting from only the most probable words, leading to more conservative and often more coherent answers. As top_k increases, the model is allowed to choose from a broader set of words, introducing more diversity and creativity into the responses, but potentially at the cost of precision. Balancing

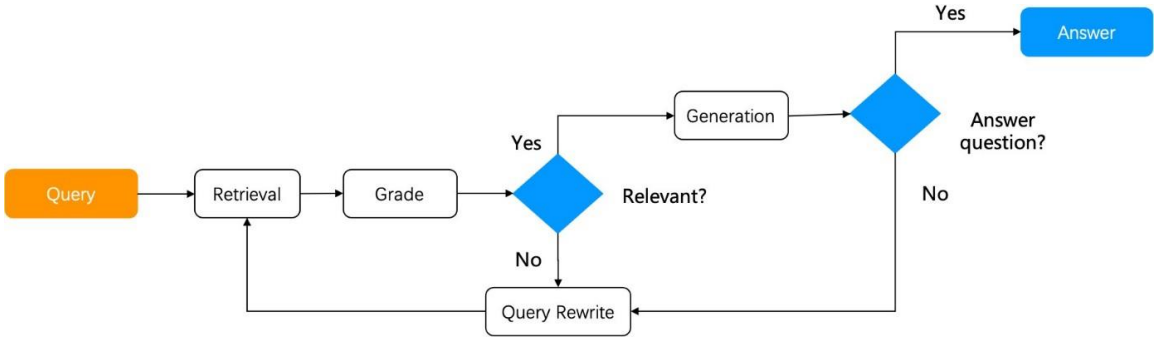

Figure 16.Graph agent pipeline

temperature and top_k is essential to control the trade-off between creativity and coherence, allowing users to fine-tune the model's behavior based on the needs of a particular task. For instance, in a Retrieval-Augmented Generation (RAG) framework, adjusting these parameters helps ensure that the generated responses remain accurate and contextually aligned with the retrieved information.

6. Graph Multi-LLM Agent System:

The multi-LLM agent system, as detailed in the article, combines a graph-based data structure with multiple large language models (LLMs) to enhance the performance of Retrieval-Augmented Generation (RAG) applications. The system operates by:

Graph-Based Data Organization: Information is structured in a graph to facilitate efficient query decomposition and context retrieval.

Multi-LLM Usage:A primary LLM handles core tasks, while an assistant LLM supports additional contextual understanding.

Parallel Context Retrieval: Queries are broken down and processed in parallel, with context re-ranked for relevance.

Answer Generation and Refinement: The main LLM generates responses, which are then refined for accuracy and adherence to guidelines.

This setup improves response accuracy and efficiency by leveraging both graph-based and multi-LLM approaches.

7. Azure OpenAI Service:

Figure 17:Azure OpenAI

The **Azure OpenAI Service** is a robust platform offered by Microsoft Azure that provides access to OpenAI's advanced models, including powerful large language models (LLMs) and embedding models, at a production level. This service is designed to integrate seamlessly into enterprise applications, offering the scalability, reliability, and security necessary for deploying sophisticated AI solutions in real- world scenarios.

The Azure OpenAI Service provides access to advanced models like GPT-3 and GPT-4 for tasks such as text generation, translation, and question-answering. It also offers embedding models for semantic search and contextual understanding. Built on Microsoft's cloud infrastructure, it ensures scalability, high performance, and production-ready deployments. The service prioritizes security and compliance, featuring encryption and secure access controls, making it suitable for industries with strict regulations like finance and healthcare. It integrates seamlessly with other Azure services, supports model fine-tuning for customization, and offers a developer- friendly API for easy integration into applications.

8. Semantic Chunking for RAG Using GPT-2 Tokenizer

Semantic Chunking is a technique employed in Retrieval-Augmented Generation

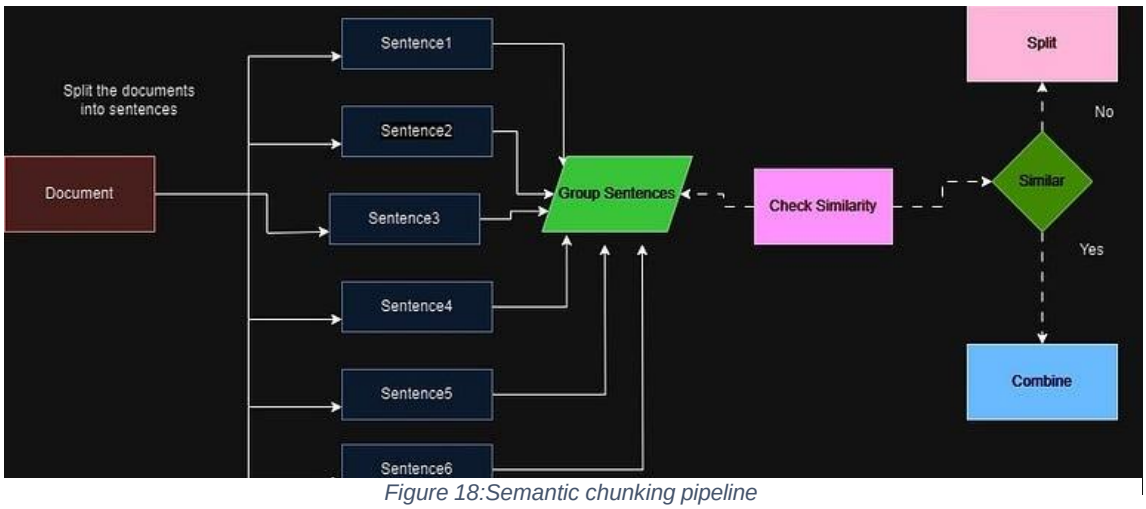

Figure 18:Semantic chunking pipeline

Semantic chunking in Retrieval-Augmented Generation (RAG) systems involves segmenting documents into smaller, coherent units, such as paragraphs or sections, to improve relevance and coherence in text generation. This method is especially useful for handling large texts or datasets, enabling the system to retrieve and process contextually relevant chunks rather than entire documents. Using the GPT-2 tokenizer, these systems efficiently analyze and handle textual data. By focusing on specific chunks, RAG systems enhance information retrieval and generate more accurate, contextually relevant responses to queries.

Role of GPT-2 Tokenizer in Semantic Chunking

The GPT-2 tokenizer is vital in the semantic chunking process for RAG systems, performing key functions like:

Tokenization: It converts text into smaller units (tokens) like words or subwords, enabling the system to process text at a granular level.

Chunk Creation: The tokenizer segments text into manageable chunks based on token limits, ensuring they fit within the model's constraints.

Context Preservation: Using Byte Pair Encoding (BPE), it creates subword tokens that preserve the semantic context of the text.

Efficient Processing: Tokenizing and chunking allow the system to process and retrieve relevant information faster.

Improved Retrieval and Generation: This combination enhances the system’s ability to retrieve relevant chunks and generate accurate, contextually appropriate responses.

9. The Hyde Technique in RAG

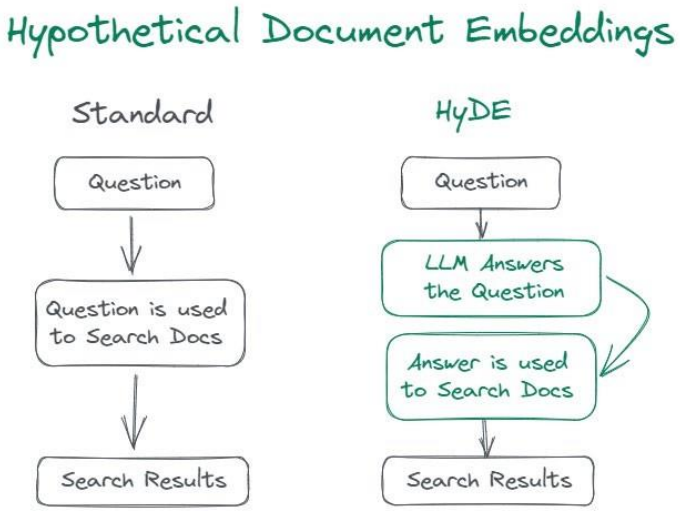

Figure 19:Standard vs Hyde retriever

The **Hyde Technique** is an advanced method used in Retrieval-Augmented Generation (RAG) systems to enhance the retrieval process by leveraging hypothetical document embeddings. This technique aims to improve the efficiency and accuracy of retrieving relevant information, which in turn enhances the overall performance of RAG systems.

a. Concept of Hyde Technique

The Hyde Technique enhances retrieval by generating hypothetical document embeddings based on queries.

Hypothetical Embeddings: The system creates embeddings for hypothetical documents that represent information relevant to the query but may not exist in the document corpus, simulating the presence of such information.

Improved Retrieval Precision: This approach improves precision by better matching queries with semantically relevant documents.

Contextual Enrichment: Hypothetical embeddings broaden the retrieval process, capturing related contexts that may be missed by focusing solely on actual documents.

Enhanced Document Representation: Unlike static embeddings, hypothetical embeddings dynamically represent potential relevant scenarios, providing more flexibility.

Integration with Models: The technique augments traditional retrieval models,

bridging gaps between document content and the query's information needs.

b. Advantages of the Hyde Technique

The Hyde Technique enhances retrieval accuracy by generating hypothetical document embeddings, enabling the system to find relevant information that may be missed otherwise. This approach broadens contextual coverage, capturing a wider range of scenarios for more comprehensive retrieval. It is also highly adaptable and can be integrated with various retrieval and generation models, making it a versatile tool for improving RAG systems.

10. DistilBERT: Efficiency in Fine-Tuning for Small Tasks

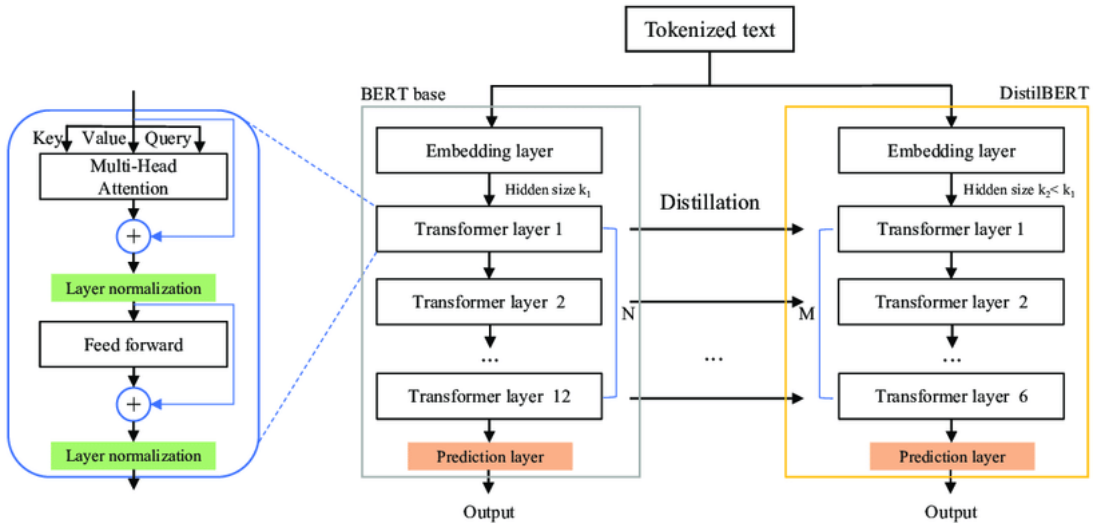

Figure 20:DistillBert architecture

DistilBERT is a streamlined version of BERT (Bidirectional Encoder Representations from Transformers), developed to offer a more efficient alternative while maintaining much of the original model's performance. It is particularly effective for fine-tuning on smaller tasks, providing a balance between computational efficiency and model accuracy. Here's why DistilBERT is well-suited for fine-tuning on smaller tasks compared to using a large language model (LLM):

a. Model Size and Efficiency

DistilBERT is designed to be significantly smaller than its predecessor, BERT, by reducing the number of layers while retaining the essential features of the model. Specifically, DistilBERT uses only 6 layers compared to BERT's 12 (in the base model), which results in a reduced number of parameters and a smaller model size. This compact design translates to faster training and inference times, making

DistilBERT more efficient for fine-tuning on smaller datasets and less resource-intensive compared to large LLMs.

b. *Retained Performance*

DistilBERT retains much of BERT's performance through knowledge distillation, where the smaller model mimics the larger one. This allows it to perform well on tasks like text classification, sentiment analysis, and named entity recognition, while reducing computational demands. Its efficiency makes it suitable for smaller tasks that don't require a full large language model.

Chapter 4:
Requirements Analysis

1. Introduction

Requirements analysis is a foundational phase in software development, serving as the cornerstone for the successful design, implementation, and deployment of a system. This chapter is dedicated to a meticulous examination of the requirements for the AI-driven ESG reporting tool, which aims to revolutionize how businesses approach sustainability and governance reporting. The analysis will be divided into two key sections: functional requirements, which define the specific behaviors and functionalities the system must exhibit, and non-functional requirements, which outline the performance criteria, usability standards, and other quality attributes essential to the system's success. Through this rigorous analysis, the chapter establishes a comprehensive framework that will guide the development process, ensuring that the final product not only meets but exceeds the expectations of its users and stakeholders.

2. Functional Requirements

Automated Data Collection from Companies House:The system must automatically retrieve and analyze company data from Companies House. This includes extracting relevant information such as company activities, sectors, services, and industries in which the company might be involved.

Company Data Analysis and Classification: The tool must analyze the retrieved data to categorize and predict the company's potential activities, sectors, services, and industries. This analysis should be based on the information available from Companies House and other relevant data sources.

Document Analysis and Management: The system should support the upload and analysis of company documents. It must be able to extract relevant ESG-related information from these documents and integrate it with the data collected from Companies House.

GRI Framework Question Answering: The tool must answer all questions from the Global Reporting Initiative (GRI) framework using the analyzed company data. The responses must be accurate, credible, and include sources to avoid hallucination. For the MVP, the system will focus on answering 8 key GRI questions.

Custom ESG Report Generation: The system should generate a custom ESG report that includes the analyzed company data, answers to the selected GRI framework questions, and other relevant company information. The report must be formatted according to industry standards and include sources for all data used.

User Document Management: Users must be able to manage their uploaded documents within the system. This includes adding, updating, or removing documents as needed.

Dynamic Company Data Management: The system should allow users to update or modify the company data at any time. Changes in the data should be reflected in the analysis and the subsequent ESG report generation.

3. Non-Functional Requirements

1. **Performance and Scalability:**The system must handle large datasets and generate ESG reports quickly, scaling efficiently as the user base grows without compromising performance.
2. **Security and Data Privacy**: Ensure user data is secure through encryption, secure authentication, and compliance with data privacy regulations like GDPR, giving users control over their data.
3. **Reliability and Availability**: The system should be highly reliable with minimal downtime and a target uptime of 99.9%, ensuring critical functionalities remain accessible.
4. **Usability and User Experience**: Provide a user-friendly interface with intuitive navigation, clear instructions, and a responsive design that works across devices, ensuring a seamless user experience.
5. **Maintainability and Extensibility**: The system should be maintainable with well-documented code and modular architecture, allowing for future updates, new features, and third-party integrations.
6. **Compliance**: Ensure compliance with ESG reporting standards (GRI, SASB, TCFD) and AI ethics, making reports explainable and free from bias.
7. **Interoperability**: Integrate seamlessly with other business tools like financial software and cloud storage to fit within existing workflows.

4. Conclusion

This chapter outlines the key requirements for developing an AI-driven ESG reporting tool designed for SMEs and startups. The functional requirements focus on analyzing company data, generating accurate ESG reports, and allowing users to manage their data. The non-functional requirements ensure the tool is secure, scalable, user-friendly, and compliant with industry standards. By addressing both functional and non-functional needs, the project aims to deliver a robust, high-performance solution that supports businesses in their sustainability efforts while adapting to future challenges. These requirements will guide the development process, ensuring an innovative and valuable tool.

Chapter 5:
Implementation and Demonstration

1. Implementation

a. Introduction

The implementation phase of the project is crucial in transforming the conceptual framework into a functional AI-driven ESG reporting tool. This chapter outlines the key components of the implementation process, detailing the system architecture that supports the tool's operations and providing a comprehensive explanation of each service that contributes to its overall functionality.

The system architecture serves as the backbone of the application, ensuring seamless integration of data analysis, report generation, and user interaction features. Each service within the architecture is designed to address specific functional requirements[3], working together to deliver a cohesive and efficient solution. By carefully implementing these components, the project aims to achieve the desired outcomes, meeting the needs of small and medium-sized enterprises (SMEs) and startups in generating accurate and credible ESG reports.

This chapter will delve into the technical aspects of the implementation, providing insights into how the system is constructed and how each service contributes to the tool’s overall effectiveness.

b. System Architecture

The **Manager** in the i-Prove system is a central Django API, containerized using Docker and hosted on an Azure Virtual Machine (VM). This API orchestrates communication between various microservices, serving as the backbone of the application by managing the inputs and outputs of components such as the **Answers Generator**, **Chunks Classifier**, **Evaluator**, **Report Generator**, **Vector Store**, and **Web Crawler**. It ensures seamless data flow and coordination among these services to produce accurate and consistent outputs.

In addition to managing microservices, the Manager is responsible for persisting data in the **Database**, making it a critical component in the data management process. The Docker containerization allows for a consistent and isolated environment, ensuring that the Manager can operate efficiently across different deployment stages.

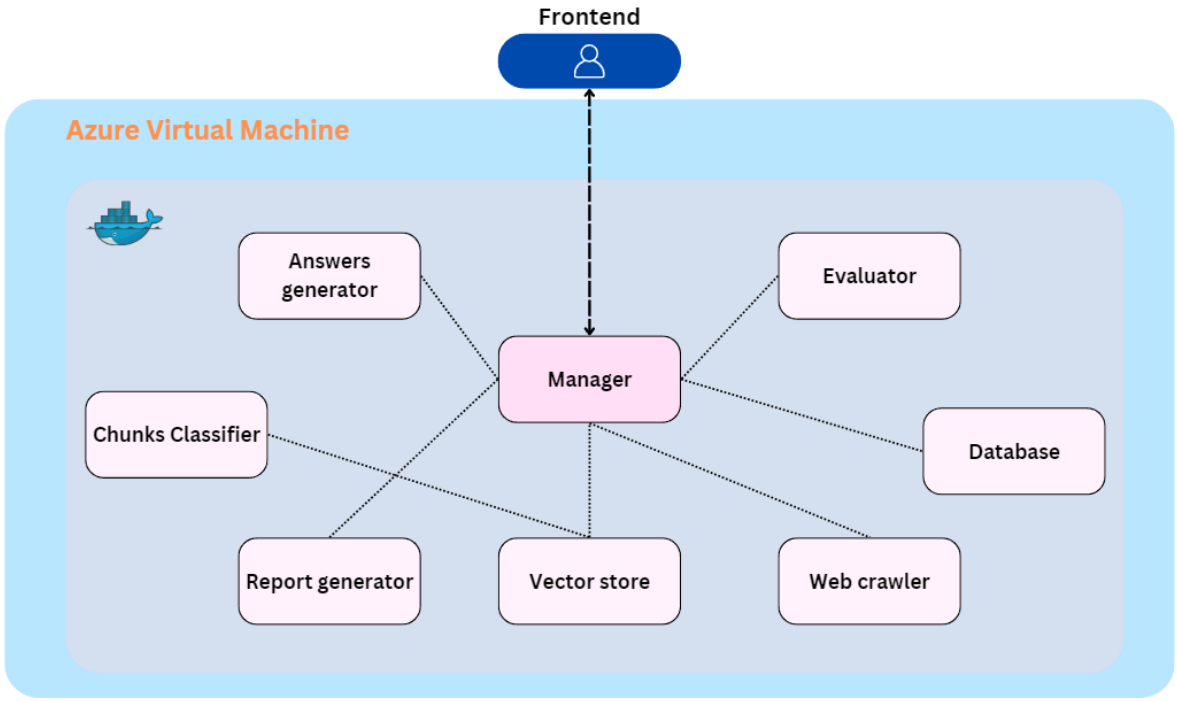

Figure 21:Global system architecture

c. Manager:

The Manager in the i-Prove system is a central Django API that orchestrates the communication between various microservices. This API serves as the backbone of the application, managing the inputs and outputs of different services, such as the Answers Generator, Chunks Classifier, Evaluator, Report Generator, Vector Store, and Web Crawler. It ensures that the data flows seamlessly between these components, coordinating their operations to produce the desired outcomes. In addition to managing the microservices, the Manager is responsible for saving data in the Database, making it a crucial part of the data management process. The Django

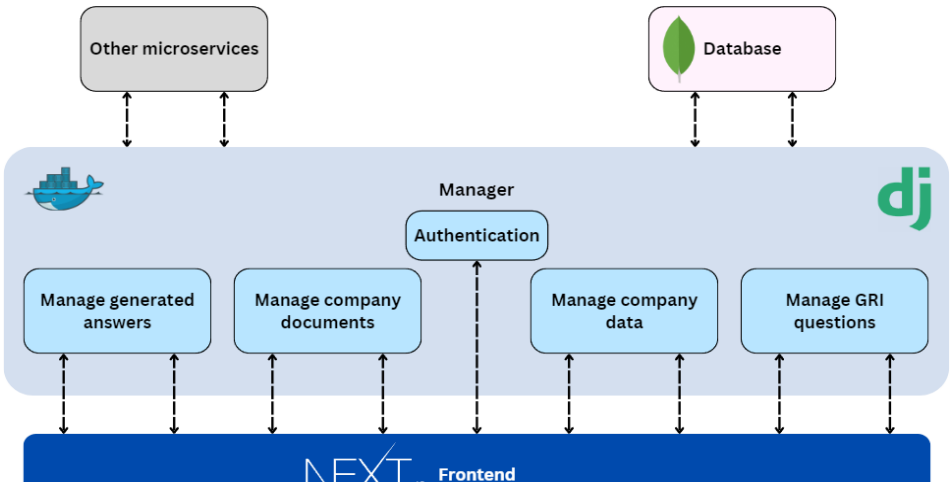

Figure 22:Manager

API's design allows it to handle complex interactions and data exchanges, maintaining the integrity and consistency of the application's operations.

Authentication

This section manages all authentication-related processes, including user registration, login, and logout. By delegating these tasks to a separate URL configuration, the application ensures that users are properly authenticated before accessing other endpoints. This helps maintain secure access to the system's resources.

Company Management

Create Company**:**This endpoint allows users to create new company profiles within the system. When a user submits the necessary details for a company, the system validates the input data, stores it in the database, and returns a confirmation of the creation. This ensures that new companies are accurately represented in the system.

Company Detail**:** This endpoint provides detailed information about a specific company. Users can retrieve company data by its unique identifier, allowing them to view comprehensive details about the company, including its activities, documents, and other related information.

Update Company**:**This feature allows users to update the details of an existing company. By submitting updated information, users can modify company profiles, ensuring that the system reflects the most current and accurate data.

Delete Company**:**This endpoint enables users to remove a company profile from the system. When a company is deleted, the system ensures that all related data is properly managed and that the deletion process is handled securely.

Generate ESG Report**:**This functionality is responsible for generating the ESG reports for a company. The report generation process involve communicating with multiple microservices, such as the Answers Generator and Report Generator, to compile data and produce a comprehensive ESG report. Users can specify the company and any additional parameters, such as the cover image, for a tailored report.

Generate All Stock Answers**:**This endpoint automates the generation of stock answers (products, services, sectors, industries) for all questions associated with a company. The system interacts with the Answers Generator microservice to produce standard responses that can be used in various contexts, streamlining the process of ESG reporting.

Read Website: This feature enables the system to crawl an entire company's website. By interacting with the Web Crawler microservice and the vector store, the system stores the website's content, which can then be utilized by the Answers Generator for enhancing the company's profile or answering specific

. Documents Management

Create Document:This endpoint allows users to upload and create new documents associated with a company. Once a document is uploaded, the system stores it in the database and makes it accessible for future reference or analysis.

Document List:This functionality provides a list of all documents stored in the system, either for all companies or filtered by a specific company. Users can view and manage documents, ensuring that they have access to all necessary files.

Delete Document:This feature enables users to delete a specific document from the system. By removing documents that are no longer needed, users can maintain an organized and relevant document repository.

Question and Answer Management

Question List: This endpoint provides a list of all questions stored in the system. Users can view existing questions, which may be used in various contexts such as ESG reporting or company assessments.

Question Detail:This endpoint retrieves detailed information about a specific question. Users can access the question's content, its associated metadata, and any related answers.

Answer List:This endpoint provides a list of all answers associated with a specific company. Users can view the responses to questions, which may be used for reporting or evaluation purposes.

Answer Detail:This functionality retrieves detailed information about a specific answer. Users can access the content of an answer, its associated question, and any relevant metadata.

Create Answer**:**This endpoint allows users to submit an AI generated answer to a specific question for a company. The system stores the answer in the database, making it available for future reference or analysis.

Submit Manual Answers:This feature enables users to manually submit answers to questions. By providing responses directly, users can ensure that the answers are tailored to the company's specific context.

Delete Answer:This endpoint allows users to delete a specific answer from the system. By removing answers that are no longer relevant, users can maintain a clean and accurate dataset.

Answer Module Questions:This functionality generates a list of answers for questions within a specific module(ESG). Users can view the responses associated with a particular section, which may be used for modular evaluations or reporting.

Answer by Stock Answer:This feature allows users to generate an answer for a question using a predefined stock answer. By selecting a standard response, users can quickly populate answers, ensuring consistency across different evaluations.

Answer All Questions:This endpoint automates the process of answering all questions associated with a company. The system communicates with the Answers Generator microservice to produce responses, streamlining the evaluation process.

Database:

The database service in this architecture is a MongoDB instance hosted on Atlas, a cloud-based managed database service. MongoDB Atlas provides a fully managed environment, ensuring high availability, scalability, and security for the database operations. In this setup, the database is responsible for storing all the data related to the application, including company information, generated answers, documents, and GRI questions.

The interaction with the MongoDB service is facilitated through Mongoose, an Object-Relational Mapping (ORM) library for MongoDB and Node.js. Mongoose provides a structured and organized way to define the schema of the documents stored in MongoDB, ensuring data validation and integrity. It also offers powerful features like middleware, virtuals, and query builders, making database interactions efficient and more intuitive.

Through Mongoose, the application can perform various CRUD (Create, Read, Update, Delete) operations on the database, enabling seamless data management. This includes the creation and updating of company profiles, storing documents, managing GRI questions, and saving generated answers. The database service is a critical component in the architecture, ensuring that all application data is persistently stored and easily retrievable when needed, thus supporting the overall functionality and performance of the system.

Azure Virtual VM

The application is deployed on an Azure Virtual Machine (VM), which serves as the

core infrastructure for hosting the application in the cloud. Azure VMs provide the flexibility of virtualization without the need to manage physical hardware, allowing the application to run in a scalable and secure environment. The VM acts as a dedicated server with configurable resources like CPU, memory, and storage, tailored to meet the specific needs of the application.

In this setup, the application is containerized using Docker, which packages the application and its dependencies into a portable container. This containerization ensures that the application runs consistently across different environments, eliminating issues related to differing system configurations.

The Docker container, which houses the entire application stack including the Django backend, is deployed on the Azure VM. This setup allows for streamlined deployment, management, and scaling of the application. Azure VM provides the necessary computing power and network connectivity, while Docker ensures that the application runs smoothly with all its dependencies encapsulated. This combination of Azure VM and Docker containers offers a robust, flexible, and scalable solution for hosting and managing the application in the cloud.

d. Web crawler

The Web Crawler microservice, developed using FastAPI, is a crucial component of our data collection and analysis infrastructure. Its primary function is to systematically crawl a specified company's website, beginning from an initial URL and traversing through internal links within the same domain. This thorough crawling process allows the microservice to gather comprehensive web content across multiple pages, which is essential for creating a detailed and up-to-date repository of company information.

Once the content is collected, the microservice performs a series of cleaning and processing tasks to ensure the data is accurate and usable. The cleaned HTML content is then stored in a vector store, facilitating efficient search and retrieval operations. This storage mechanism allows the system to manage large volumes of data effectively and supports the seamless integration of this information into various analytical processes.

In addition to storing the web content, the microservice plays a pivotal role in generating detailed stock answers related to the company's offerings. By analyzing the collected data, the system can generate insightful responses about the company's products, industries, and services. These tailored answers are crucial for providing accurate and relevant information, enhancing the overall effectiveness of the reporting and analytical tools in use.The integration with FastAPI ensures that the

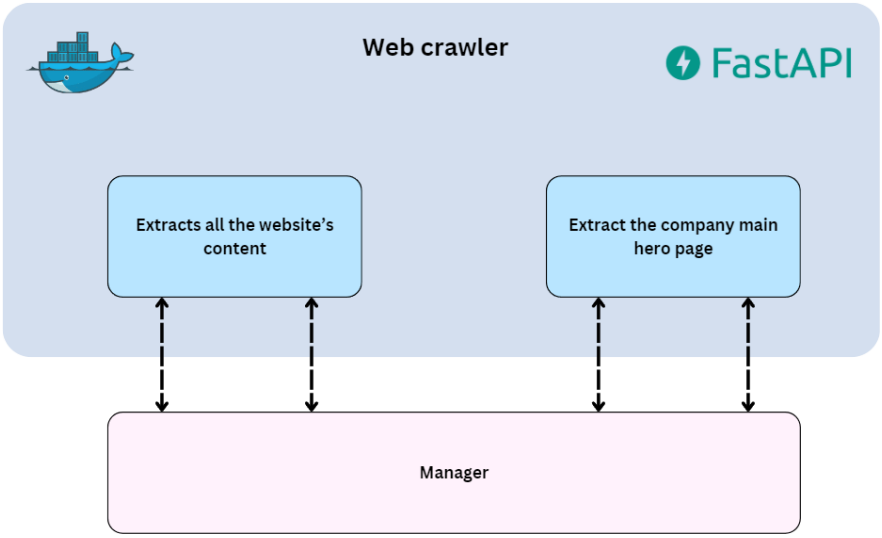

Figure 23: Web crawler microservice architecture

Web Crawler microservice operates with high performance and responsiveness. It provides robust endpoints for initiating the crawling process and retrieving specific pieces of company information, allowing for smooth interaction within the broader application framework. This design ensures that the microservice can handle complex data requirements efficiently, supporting the generation of actionable insights and comprehensive company profiles.

e. Vector store

The Vector Store microservice, built with FastAPI, is a powerful tool designed to manage and utilize vector-based text chunk storage efficiently. This service integrates advanced technologies to enhance the organization, retrieval, and management of text chunks, making it a key component in analyzing and responding to GRI questions.

Text Chunk Management

The service allows for the addition of text chunks to the vector store. Text from uploaded files is processed and split into semantic chunks using the GPT-2 tokenizer. Each chunk is then embedded using the Azure OpenAI text-embedding-3-

large model and stored in a FAISS vector store. This approach ensures that text is organized into meaningful chunks, making it easier to retrieve relevant information.

Querying Text Chunks

The service supports querying for relevant text chunks based on user inputs. It employs FAISS for similarity search and utilizes Maximal Marginal Relevance (MMR) reranking to refine the results, ensuring that the most pertinent text chunks are retrieved. Queries are embedded using the text-embedding-3-large model, and the search retrieves the top 5 text chunks, which is a configurable hyperparameter for optimizing search results.

Text Chunk Deletion

Users can delete specific text chunks from the vector store. The deletion process can be customized based on various criteria, ensuring that outdated or unnecessary chunks are removed efficiently.

Text Chunk Classification

The service classifies text chunks using a fine-tuned BERT model into categories such as:

- Compliance and Legal Policies
- HR Policies
- Operational Policies
- Health and Safety Policies

This classification simplifies querying by categorizing chunks according to their content, which is particularly useful for answering GRI questions. The classification ensures that relevant text chunks are easily accessible based on their category.

Vector Store Management

The service can create new vector stores if one does not already exist for a given company and type. It also supports loading existing vector stores from disk. Two vector stores are maintained: one for company documents and another for website content, allowing for efficient and localized management of text chunks.

By combining the GPT-2 tokenizer for semantic text splitting, the Azure OpenAI text-embedding-3-large model for embeddings, and FAISS's MMR reranking for similarity search, the Vector Store microservice provides a comprehensive and scalable solution for managing and querying text chunks. The integration of a fine-tuned BERT model for chunk classification further enhances the service's ability to organize

and retrieve relevant information based on GRI questions.

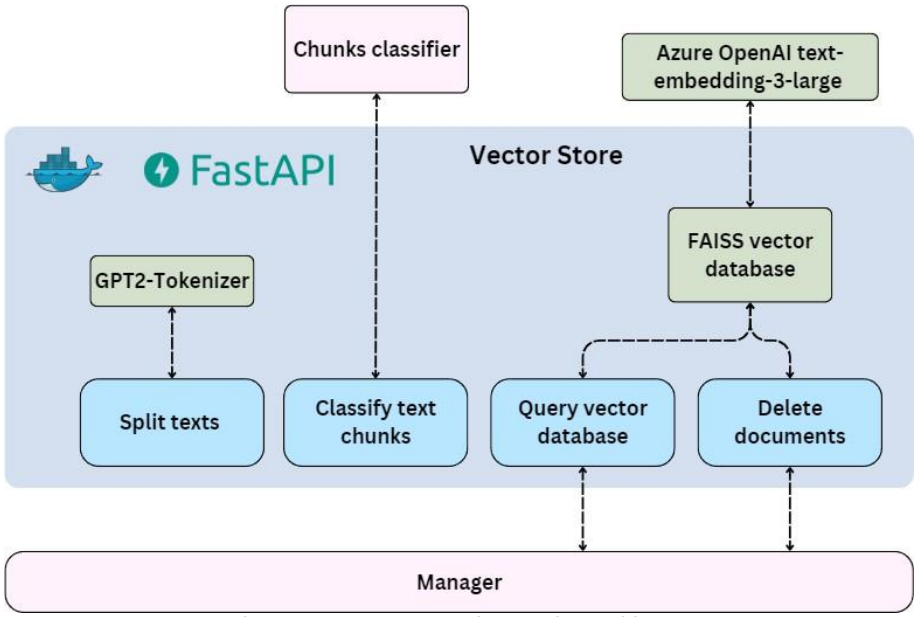

Figure 24:Vector store microservice architecture

f. Text chunks classifier:

The Chunks Classifier microservice is designed to categorize text chunks into predefined policy classes using a fine-tuned DistilBERT model. The model was optimized and deployed with ONNX Runtime for efficient inference. This service is particularly useful for classifying company documents into categories such as compliance, HR, operational, and health and safety policies, helping to automate document processing and reduce the cost of using more computationally expensive models like GPT-4.

Model Architecture:

The microservice uses a fine-tuned version of DistilBERT, a lightweight transformer model optimized for sequence classification tasks.

The model was fine-tuned on synthetic data generated from real company documents, classified using GPT-4 to ensure high-quality training data.

Predefined Policy Classes:

The service classifies text chunks into the following categories:

- Compliance and Legal Policies
- HR Policies
- Operational Policies
- Health and Safety Policies

Efficient Inference:

The model is deployed using ONNX Runtime (ORT) for optimized and efficient inference, allowing for quick classification of large volumes of text.

Cost Reduction:

By using this fine-tuned DistilBERT model, the service reduces the reliance on GPT-4, significantly lowering the operational costs associated with text classification tasks.

Custom Tokenization:

The service employs the DistilBertTokenizer to preprocess text data, including handling token type IDs, which are necessary for sequence classification in transformer models.

Hyperparameters:

The model was fine-tuned using the following hyperparameters:

Training Epochs: 10

Batch Size (Train/Eval): 16

Warmup Steps: 100

Weight Decay: 0.01

Logging: Every 10 steps, with logs stored in the ./logs directory

Evaluation Strategy: Performed at the end of every epoch

Model Saving Strategy: The best model is saved at the end of training based on validation performance

Performance Results:

After 10 epochs of training, the model achieved the following metrics:

Training Loss: 0.1794

Validation Loss: 0.3897

Accuracy: 87.83%

Precision: 88.08%

Recall: 87.83%

F1 Score: 87.62%

Example:

Text chunk: In response to the increasing regulatory requirements and the evolving legal landscape, our company has developed a comprehensive compliance and legal framework to ensure that all our operations adhere to the highest standards of governance. We have implemented stringent policies that cover anti-bribery, anti-corruption, and data protection regulations. Regular audits are conducted to monitor adherence, and mandatory training sessions are held for all employees to stay updated with the latest legal requirements. Our commitment to compliance is further demonstrated by our zero-tolerance approach towards any violations, ensuring that our business operations remain transparent and lawful at all times.

Class: Compliance and Legal Policies

g. Answers generator

The Answers Generator microservice is a crucial component designed to generate, refine, and retrieve answers based on specific company information. It uses a graph AI agent to efficiently process queries and deliver accurate, context-aware answers.

To optimize performance and cost-effectiveness while avoiding token limits during system overload, the microservice utilizes two Large Language Models (LLMs) deployed by OpenAI and Azure:

Main LLM: A fast and cost-efficient model (GPT-3.5 Turbo) with settings focused on precision (temperature 0, top_k 0).

Assistant LLM: A slower but more adaptable model (GPT-4o) with settings optimized for lighter tasks (temperature 0.7).

i. The Graph AI Agent

- **Query**: The question to be answered.
- **Guidance**:GRI-specific guidance to inform the LLM how the final answer should be structured.
- **Company Information**: A paragraph containing essential company details such as size, name, and address.

The graph is decomposed into the following nodes:

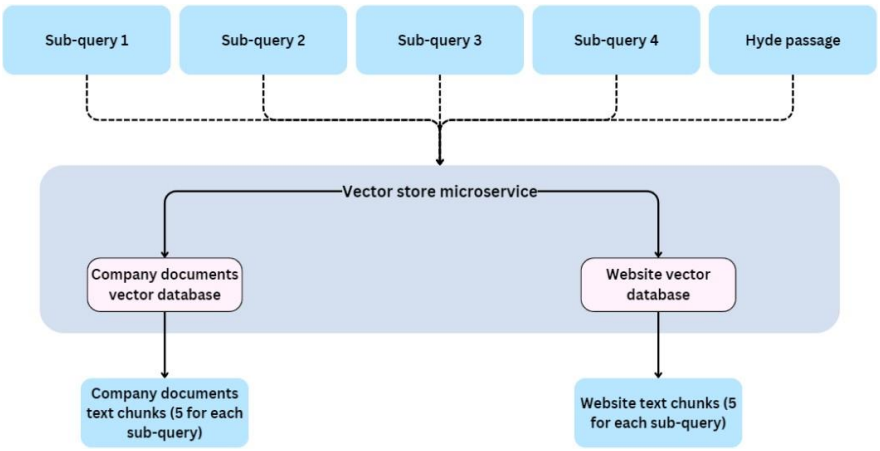

Figure 25:Query decomposition pipeline

Query Decomposition

To break down the main query into four sub-queries related to the primary question.
Example:
Main Query: Report a statement from the highest governance body or most senior executive of the organization about the relevance of sustainable development to the organization and its strategy for contributing to sustainable development.

Sub-Queries:

1. What is the organization's vision and strategy for managing its impact on the economy, environment, and people over the short, medium, and long term?
2. How are the organization’s purpose, business strategy, and business model designed to avoid negative impacts and create positive ones?
3. What are the organization’s short- and medium-term goals for contributing to sustainable development, and how do these goals align with global standards?
4. What significant events, achievements, and challenges related to sustainable development occurred during the reporting period, and how did the organization perform against its sustainability goals and targets?

This step is crucial in the retrieval process as it helps fetch related text chunks from the vector store database, even if they are not directly similar to the main query. The following prompt is used to achieve the best results:

```
"""You are an ESG auditor generating a report for a client. You have been tasked
with answering the following question: "{query}".

to get the best results,you should decompose the query into smaller 4 sub-queries
that will be used to search in company documents and websites vector store.

Use the guidance as a reference to generate the sub-queries.



Use these sub-queries to thoroughly search the company's documents and website
vector store for comprehensive and relevant information.



query; {query}

guidance; {guidance}



{format_instructions}"""
```

Generate the Hyde Passage

Figure 26:Hyde passage node

The Hyde Passage is a generated paragraph similar to the expected fetched text chunk. It acts as an answer to the query and helps retrieve the correct text chunks by their similarity to the targeted information.

- **Example**:

Query: Report a statement from the highest governance body or most senior executive of the organization about the relevance of sustainable development to the organization and its strategy for contributing to sustainable development.

-

Hyde Passage: Our organization is dedicated to leading sustainable practices with a holistic approach to managing economic, environmental, and social impacts. In the short term, we focus on operational efficiency and waste reduction. Medium-term goals include establishing sustainable supply chain practices, while our long-term aim is to achieve net-zero carbon emissions and social equity leadership. Aligned with the UN SDGs, we target a 20% carbon emission reduction and a 30% increase in

renewable energy use short-term, and a 50% reduction in landfill waste and 40% boost in community engagement mid-term. Despite supply chain challenges, we surpassed targets by reducing our carbon footprint by 15%. Sustainability is embedded in our strategy to drive innovation and long-term growth.

Retrieval Nodes:

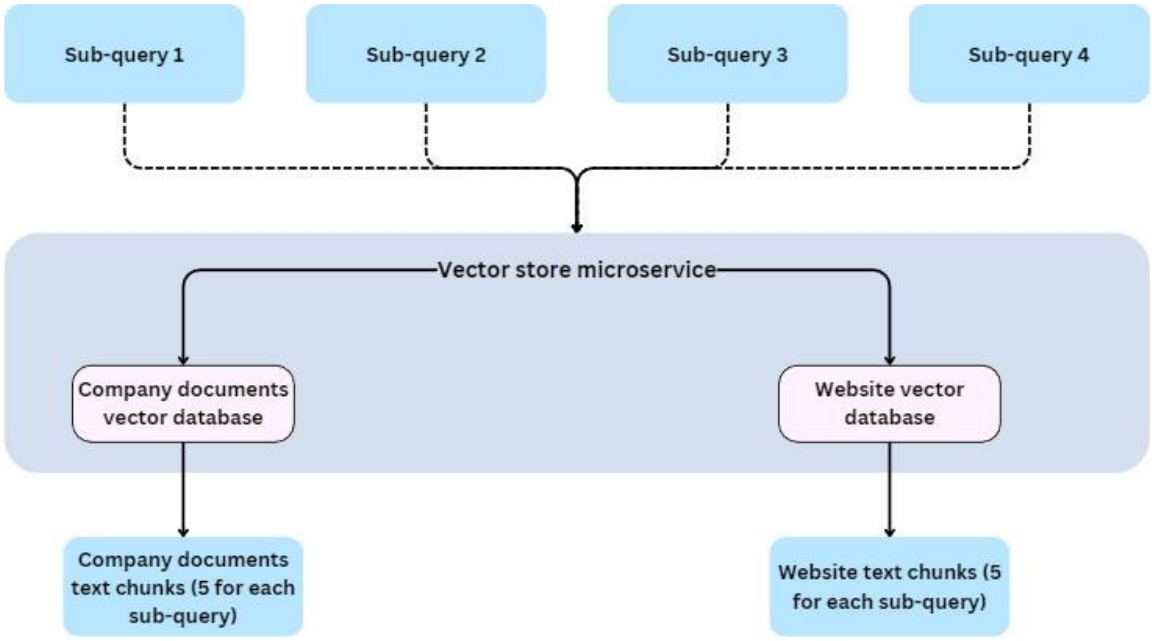

Figure 27:Retrieval nodes

The retrieval process is conducted in parallel from both the documents vector store and the website vector store. For all five sub-queries (including the Hyde Passage), the vector store fetches the top 5 text chunks from each source, resulting in a total of 25 text chunks combined from both the website and company documents.

Re-ranking Node:

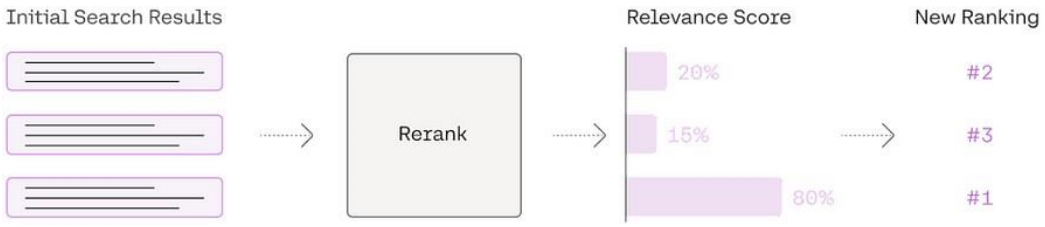

Figure 28: Reranking pipeline

Given the limited context window of the LLM, it is essential to select only the most

relevant and similar text chunks. To achieve this, we employ an open-source cross-encoder architecture (ms-marco-MiniLM-L-12-v2) to predict the softmax similarity between the main query and the text chunks. The text chunks are then sorted, and the top 5 that are most likely to contain the answer are selected.

Generating the Answer Node:

In this node, we utilize the main LLM (GPT-3.5 Turbo) to generate an answer based on the ranked context. The goal is to craft a precise and context-aware response to the question.

Answer revision Node:

While the answer generation process typically produces a satisfactory response, this node is designed to ensure that the final answer adheres to GRI standards and is free from any hallucinations or misinformation. If necessary, the node refines the answer and incorporates additional relevant information from the context. If a good answer cannot be generated, the node will return the following message: "The AI engine could not find a relevant answer to the question. Please type your answer or use the Template Answer function below."

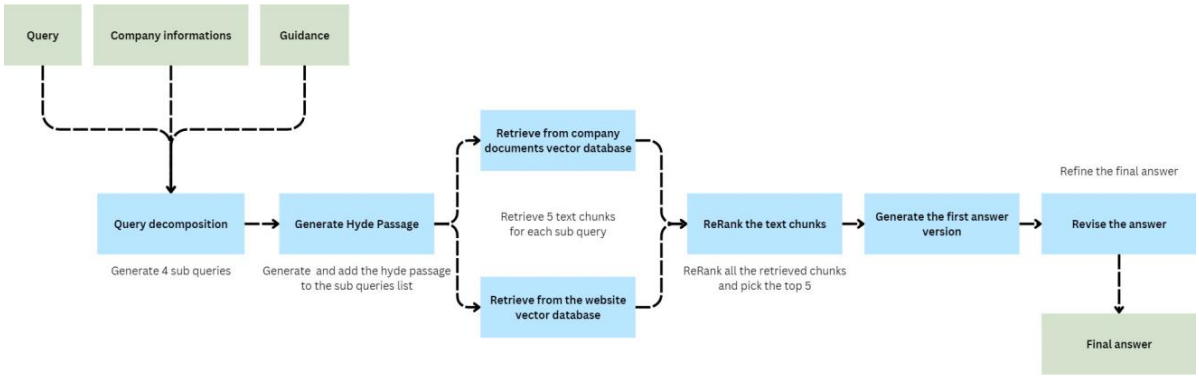

Figure 29:answer revision node

ii. Other Features:

Answer Using Stock Answer: In some cases, the AI agent may not find the most relevant source to answer a question, resulting in no answer being provided. To address this, users are given the option to select from five predefined profiles that represent varying levels of company commitment to sustainable development:

1. Very Low: This profile indicates that the company does not currently have a policy (e.g., a waste management policy).
   Example:
   "Currently, our organization is in the early stages of recognizing the importance of sustainable development. While we understand its relevance,

we are still in the process of developing a comprehensive strategy to manage our impacts on the economy, environment, and people. In the coming years, we aim to establish a clear vision and strategic priorities to better align our business practices with sustainable development goals."

2. Low: This profile reflects an organization that acknowledges sustainable development but is in the early stages of implementing relevant strategies.
   Example:
   "Our organization recognizes the importance of sustainable development and its relevance to our long-term success. While we have begun incorporating sustainable practices into our operations, we are still developing a robust strategy. In the short term, our focus is on identifying key areas where we can reduce our negative impacts on the environment and society. We aim to align our efforts with international guidelines and develop medium-term strategic priorities that support sustainable development."

3. **Moderate:** This profile indicates that the organization has a defined strategy and is actively working on sustainability initiatives.
   Example:
   "Our organization is committed to integrating sustainable development into our business strategy. We have outlined a short- and medium-term vision to manage our impacts on the economy, environment, and people. Our business strategy aims to prevent negative impacts and achieve positive outcomes, particularly in areas such as environmental sustainability and community engagement. We are actively aligning our priorities with global sustainability frameworks and have identified key areas for improvement and investment over the next few years."

4. **High:** This profile represents a company with a strong commitment to sustainable development, with well-established practices and goals.
   Example:
   "Sustainable development is a core element of our organization's strategy. Our short-, medium-, and long-term vision focuses on managing our impacts on the economy, environment, and people, including their human rights. Our business strategy is designed to prevent negative impacts and create positive outcomes across our operations and supply chain. We have set strategic priorities that align with authoritative intergovernmental instruments and are actively addressing broader trends such as climate change and social inequality. Our recent achievements include significant reductions in carbon emissions and enhanced labor practices across our supply chain."

5. **Very High:**This profile reflects an organization that has deeply embedded sustainable development into its core operations and strategy.

Example:

"Sustainable development is a fundamental aspect of our organization's strategy. Our short-, medium-, and long-term vision emphasizes managing our impacts on the economy, environment, and people, including their human rights. Our business strategy is meticulously designed to prevent negative impacts and foster positive outcomes across our operations and supply chain. We have established strategic priorities that align with authoritative intergovernmental instruments and are actively addressing broader trends such as climate change and social inequality. Our recent accomplishments include significant reductions in carbon emissions and improved labor practices throughout our supply chain."

After the user selects the stock answer that best fits their company's profile, the AI agent will create a customized answer using the company's specific information with the following prompt:

Customization Prompt:

```
"Your task involves reformulating an answer related to the ESG GRI framework,
ensuring the response is tailored specifically to the company in question. The
reformulation should adopt an affirmative and reassuring tone, utilize British
English, and avoid the use of personal pronouns. Where relevant, incorporate the
company name. Do not introduce new information; only reformulate the answer to
make it more positive and specific to the company's context. It is crucial that
the reformulated answers not only adhere to GRI standards but also reflect the
company's commitment to these guidelines, showcasing its efforts in sustainability
and ethical practices.
Company Information: {company_info}
Reformulate the following answer, making sure it is positive, specific to the
company's context, and aligns with GRI standards.
Question: {question}
Answer: {stock_answer}
Reformulated Answer:"""
```

Batch Answering: The AI agent can also be utilized in a batch mode, allowing multiple questions to be answered simultaneously. This can be done at the module level (E-S-G) or for all questions at once, streamlining the process and improving efficiency.

h. Report generator

Figure 30: Generated ESG report example

The ESG Report Generator is a FastAPI-based microservice designed to generate custom Environmental, Social, and Governance (ESG) reports in PDF format. This service utilizes Jinja2 templating and WeasyPrint for HTML-to-PDF conversion. The generated reports are customized based on company-specific data, including an introduction, module-based answers, and company branding elements like logos and cover images.

Template Rendering: The service uses Jinja2 to render HTML templates, allowing for dynamic generation of report content. The template directory can be customized, and the service supports auto-escaping for HTML and XML files to ensure security.
Customizable Branding: The service allows for the inclusion of company-specific

branding elements. Users can select a cover image from a predefined set and include the company logo in the report. These assets are loaded dynamically based on the company's selection.

PDF Generation: The final ESG report is generated as a PDF using WeasyPrint. The service ensures that the report is formatted correctly and includes all relevant sections, such as the company introduction, answers categorized by ESG modules, and visual elements like the cover image and logo.

The microservice is designed to be integrated into a larger ESG reporting system, where users can input company-specific data, and the service will generate a professional and customized PDF report. The service is ideal for companies needing to produce ESG reports that align with corporate branding and regulatory standards.

- **Template Directory:** Users can specify a custom directory for HTML templates.
- **Output Directory:** The directory for saving generated PDF files can be customized.
- **Cover Images and Logos:** Users can select from different cover images and include their company logo in the report.

This microservice streamlines the process of creating tailored ESG reports, ensuring consistency, professionalism, and alignment with corporate and regulatory expectations.

i. Evaluation microservice:

This microservice is a web application responsible for calculating evaluation metrics for generated answers and their associated context. Each metric is calculated using a Large Language Model (LLM), ensuring high precision in the assessment process. The key metrics include:

- **ESG Report Quality**: Assesses the quality of the generated answer in alignment with ESG standards.
- **Report Clarity and Accessibility**: Evaluates how clear and accessible the generated answer is for users.
- **Answer Relevancy**: Measures how relevant the answer is to the given context.
- **Faithfulness**: Assesses the accuracy and truthfulness of the answer in relation to the source data.

- **Context Relevancy**: Evaluates whether the provided context is relevant to the question.

This system ensures that generated reports meet high standards of quality and reliability.

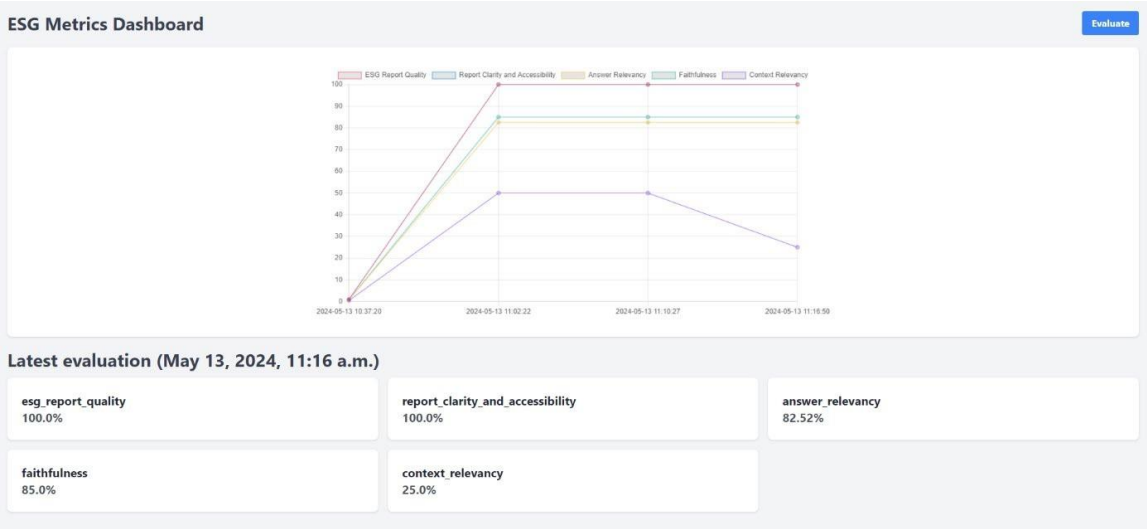

Figure 31: Evaluation dashboard page

2. Demonstration

a. Introduction:

This chapter provides a detailed overview of the i-Prove platform's frontend, highlighting its key functionalities and user interaction points. While the backend has been thoroughly explained in previous sections, this demonstration will focus on how the frontend, developed by a team member, leverages AI-driven capabilities to generate comprehensive ESG reports.

The i-Prove platform enables users to input company data, manage profiles, upload relevant documents, and generate ESG reports based on predefined frameworks. Through a series of screenshots, the platform's core functionalities will be illustrated, showing how users navigate the interface, interact with the system, and receive ESG reports tailored to their specific company data.

This section aims to showcase the seamless integration between the AI-powered backend and the frontend, emphasizing the platform's ease of use, responsiveness, and ability to deliver accurate ESG reports. The user interface, designed with simplicity and efficiency in mind by the team member, ensures that all interactions

are intuitive while maintaining the system's powerful capabilities.

b. My Company page

After logging into the portal, the user is directed to the My Company page. This page is responsible for creating and managing the company's profile in the database, including key details such as the legal name, country of operation, SIC code, annual revenue, website, and other relevant information.

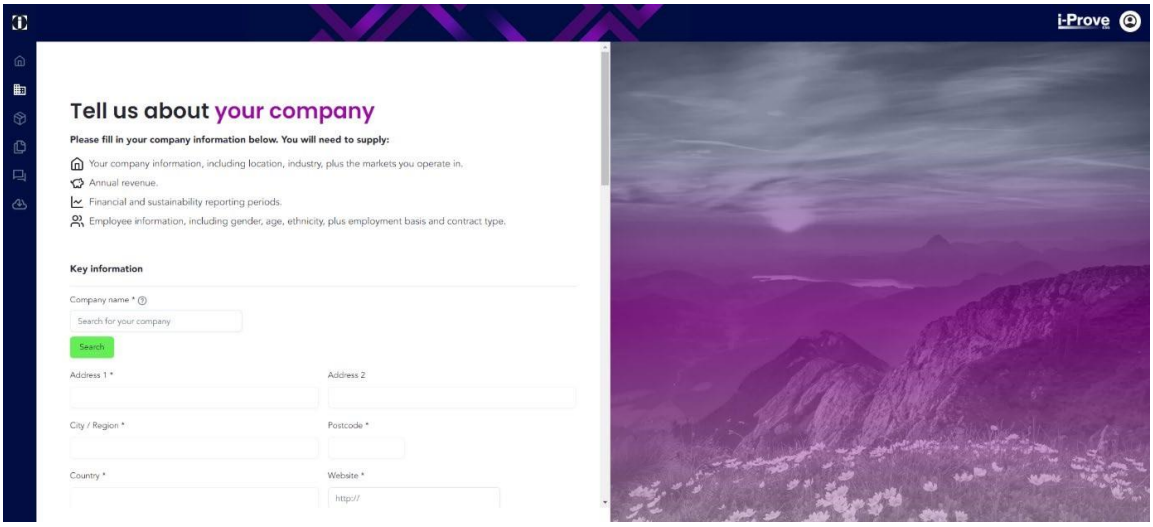

Figure 32:My company page

c. Activities page:

The **Activities** page is responsible for generating the company's activities, including products, services, sectors, and industries, based on the company's website and SIC code.

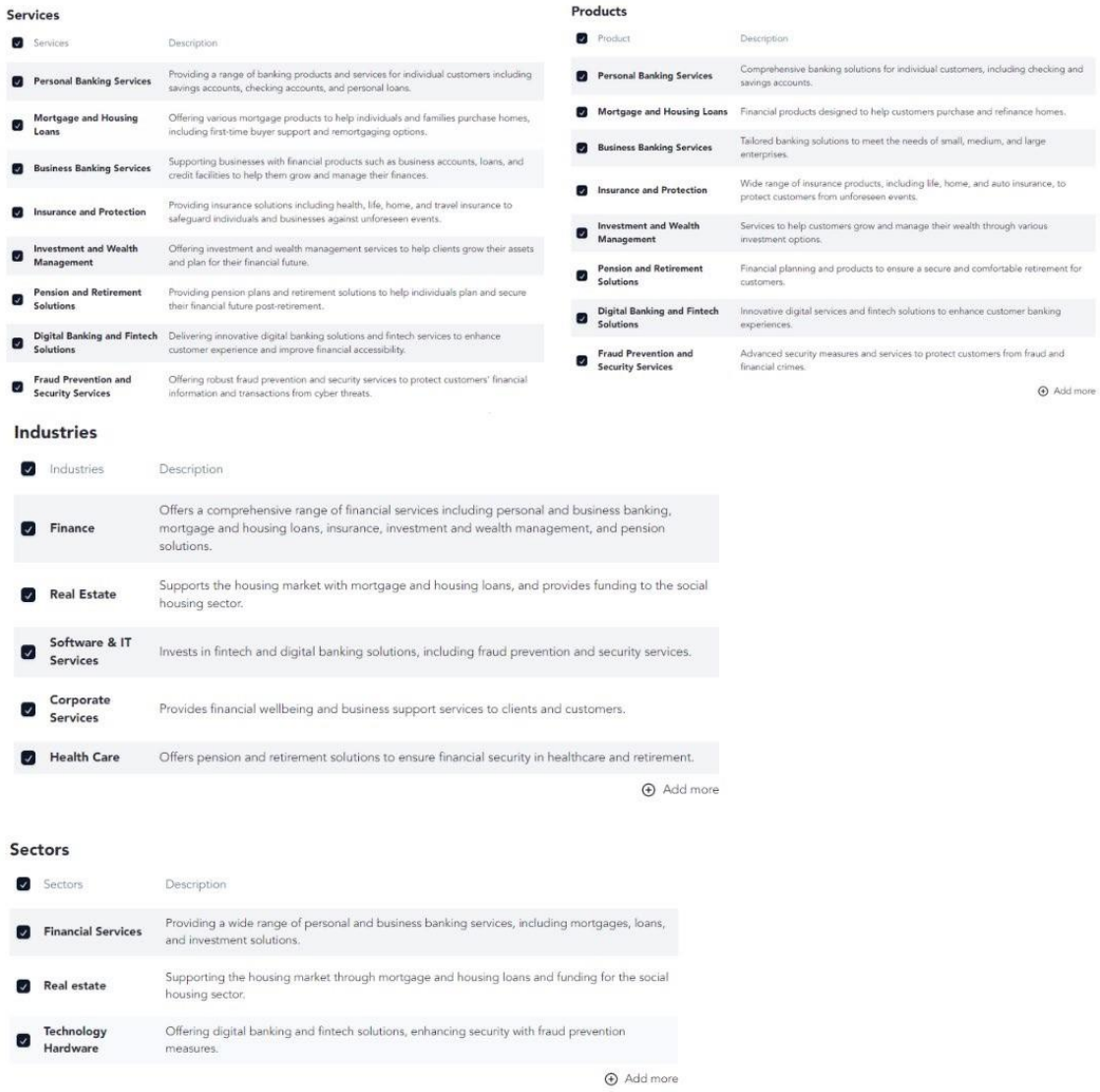

Figure 33:Activities page

b. Assets page:

This page is crucial for ensuring the quality of the generated reports, as it collects important company policies such as Modern Slavery Statements, Sustainability Reports, Waste Management Reports, and others. In addition, the user can provide the company's website, which can be scraped to extract relevant content for use as a source of information. If the backend detects a policy file on the website, the frontend will prompt the user to upload the file to enhance the accuracy of the generated ESG

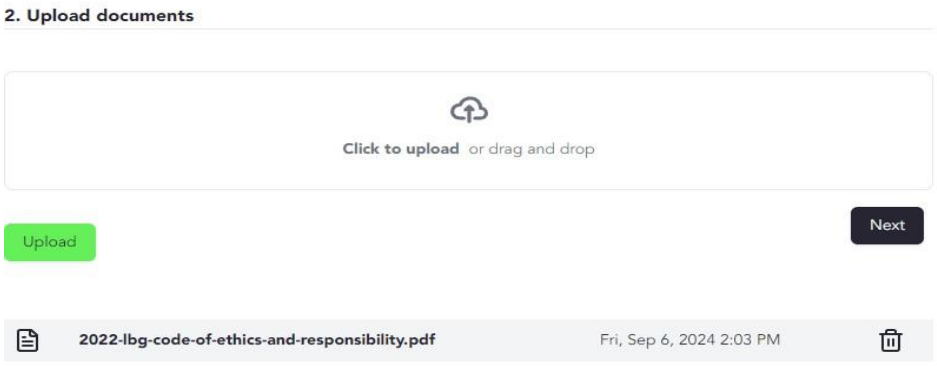

Figure 35:Documents upload section

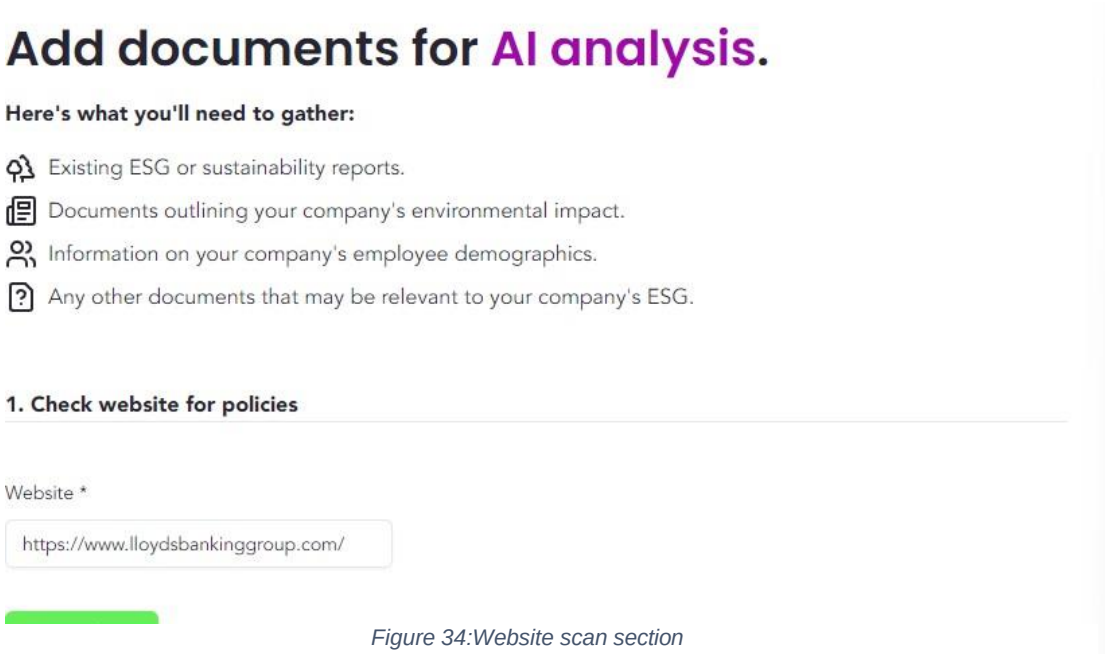

Figure 34:Website scan section

The following documents have been found on your website.

- 2023-lbg-modern-slavery-statement.pdf
- the-mortgage-business-plc.pdf
- 2023-lbg-sustainability-report.pdf
- 2023-lbg-sustainability-metrics-basis-reporting.pdf
- 2023-lbg-sustainability-reporting-framework-index.pdf
- 2022-lbg-environmental-sustainability-report.pdf
- 2022-lbg-social-sustainability-report.pdf
- 2022-lbg-esg-performance-review.pdf
- 2022-lbg-esg-reporting-criteria.pdf
- 2022-lbg-esg-reporting-framework-index.pdf
- local-low-carbon-accelerator-report.pdf
- lbg-climate-report-2021-interactive-final.pdf
- lbg-esg-report-2021-interactive-final.pdf
- lbg-esg-appendices-interactive-final.pdf
- esg-datasheet-2021-final.pdf
- 2021jul-lbg-esg-investor-presentation.pdf
- lbg-esg-interactive-210223.pdf
- 2021feb-lbg-esg-investor-presentation.pdf
- 2020-lbg-hy-esg-approach.pdf
- lbg-esg-awards-and-ratings-210222-p5.pdf
- lbg-responsible-business-update-april-2019.pdf

Figure 36:Extracted Company documents

d. Questions page

Once all the necessary data has been collected from the user, they can proceed to answer the 8 testing GRI questions based on the provided information. For each question, the user has the option to either use the AI-powered system to generate an answer, input the answer manually, or select from pre-defined stock answers. This flexibility ensures that the user can tailor their responses based on the specific needs of their company.

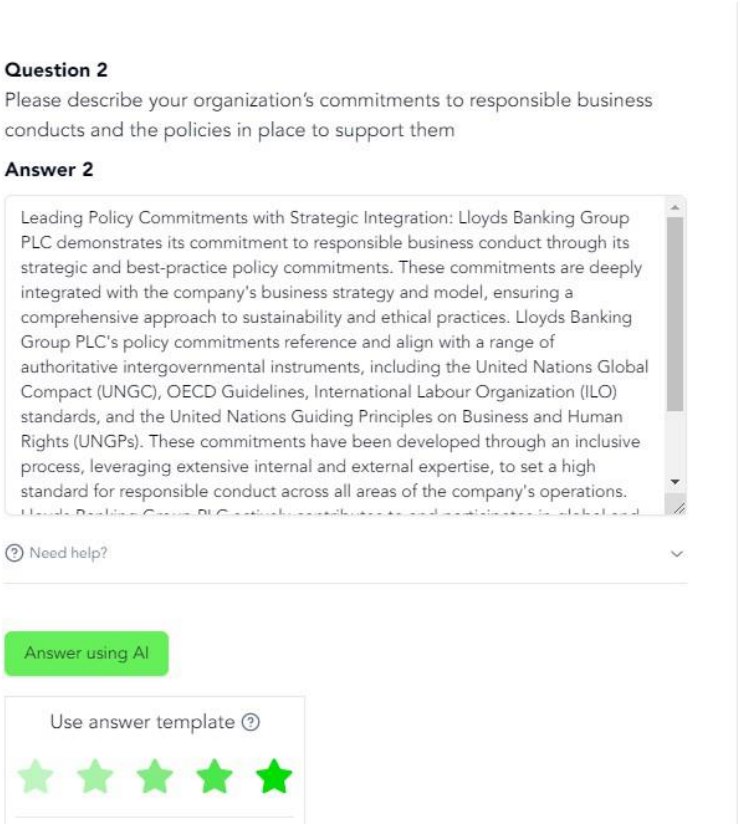

Figure 37: I-prove MVP question 2

e. *Report generation page*

Once all the answers are answered,the user can proceed to generate his ESG report,before he got a version of this file he can choose a cover image to be customized to his brand style company,and after all of this he can get a ESG report using i-prove

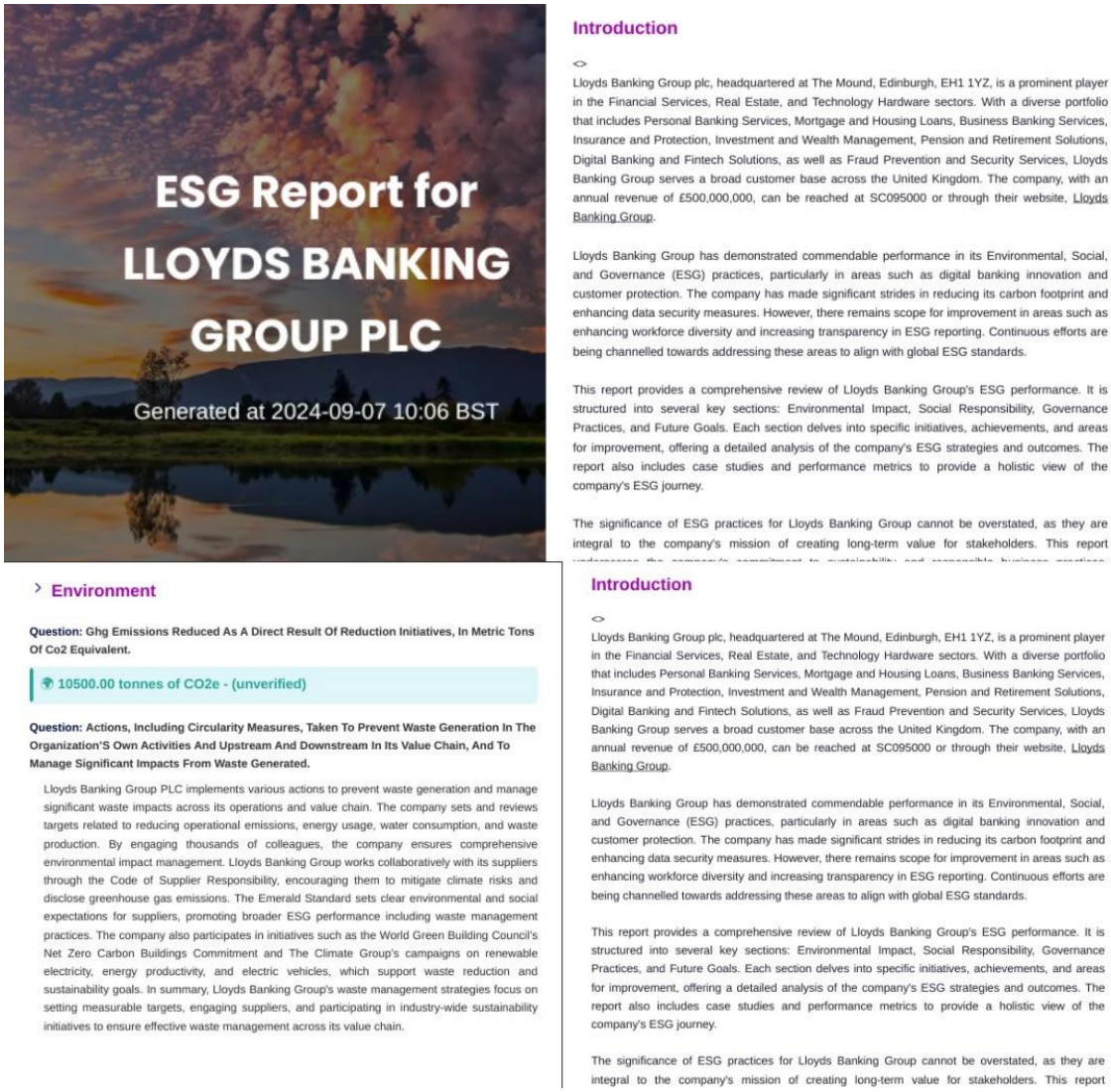

Figure 38:Generated ESG report example

Conclusion and Perspectives

1. Overview

The **i-Prove** platform marks a significant advancement in automating Environmental, Social, and Governance (ESG) reporting through AI-driven solutions. This project's completion represents the culmination of efforts to build an innovative system that bridges complex ESG frameworks with cutting-edge AI technology. The main focus of this project was the creation of a **Minimum Viable Product (MVP)** that demonstrates the potential to streamline the ESG reporting process for businesses, allowing them to meet regulatory requirements efficiently and effectively.

The i-Prove MVP successfully integrates advanced AI techniques, including large language models (LLMs) and retrieval-augmented generation (RAG), into a functional backend capable of processing ESG-related data and generating structured reports. Key achievements of the project include:

Backend Development: The backend system was architected using microservices to ensure scalability and robustness, supporting both current requirements and future expansion. Each service, such as the answers generator, chunks classifier, and web crawler, was built with specific tasks in mind, ensuring high accuracy and efficiency.

ESG Report Generation: The AI models used in the backend allow the platform to interpret company data, generate relevant content based on GRI questions, and create comprehensive ESG reports. This process significantly reduces the time and effort traditionally required for companies to compile their ESG documentation.

Frontend Integration: Although the primary focus was on the backend, the platform's user interface was also developed to be intuitive and user-friendly, enabling easy interaction for non-technical users. It allows users to input company data, manage profiles, and track progress through a streamlined dashboard.

2. Lessons Learned

This project provided an invaluable learning experience, both technically and professionally. Several key areas of knowledge and skill development were crucial throughout the process:

Generative AI and ESG Reporting: Working with **Generative AI** models like GPT-3.5 and GPT-4 offered deep insights into their application for business purposes. The task of generating context-aware, reliable responses for ESG questions was a complex challenge that involved fine-tuning models to ensure accuracy and relevance. Through this, a solid understanding of **Retrieval-Augmented Generation (RAG)** and document embedding techniques was gained, along with experience in optimizing model outputs for real-world business applications.

UK Business Environment: Collaborating with UK-based companies offered an in-depth view of how businesses in the region operate, particularly around compliance and reporting standards like ESG. Understanding the needs and challenges of these companies shaped the development of the i-Prove platform, ensuring that the solution catered to both the regulatory and operational needs of its users.

Deploying AI Applications: Managing the deployment of an AI-driven platform was a critical aspect of the project. From ensuring that the backend system could handle scaling demands to configuring a cloud-based architecture on Microsoft Azure, the deployment experience provided essential skills in **cloud infrastructure management**, **containerization** using Docker, and overall platform stability.

Collaboration and Team Dynamics: Working closely with a diverse team across multiple disciplines—AI developers, data scientists, product managers—highlighted the importance of effective communication, project management, and collaboration. This experience in working within a cross-functional team fostered strong professional growth, especially in the areas of **agile methodologies** and stakeholder management.

3. Transitioning from MVP to Production

While the MVP has successfully demonstrated the platform's potential, this marks only the initial phase of the project. The next phase will focus on transitioning the MVP into a **production-ready version**. Key areas of focus for the production phase include:

Enhanced Features: The production version will introduce new features such as multi-language support, more detailed customization for ESG reporting, and integrations with other compliance frameworks like SASB and TCFD.

Scaling for Larger Enterprises: The MVP was designed primarily for small to medium-sized enterprises (SMEs). The production version will scale the platform's capabilities to cater to larger enterprises, with enhanced data handling capabilities and support for more complex datasets.

Improved AI Models: Fine-tuning the AI models used in the system will be a priority, focusing on improving the accuracy of ESG answers, reducing model biases, and ensuring better performance for different industries. Additionally, integrating newer AI models and incorporating **real-time updates** will be part of the platform's evolution.

User Experience: Based on feedback from the MVP testing phase, the user

interface will be refined for better user interaction and navigation, ensuring that users with minimal technical knowledge can easily access the platform's capabilities.

Security and Compliance: As the platform moves to production, further enhancements will be made to ensure compliance with data privacy regulations like **GDPR**. Additional layers of security, such as encryption for sensitive data and multi-factor authentication, will also be implemented to safeguard user data.

4. Future Perspectives

Looking ahead, the i-Prove platform has the potential to become a leading tool in the ESG reporting industry. As global focus on sustainability and responsible governance grows, tools like i-Prove that streamline compliance and reporting processes will become essential for businesses across sectors. Some future perspectives for the platform include:

Adapting to Global Standards: As new ESG frameworks emerge globally, the platform will continuously evolve to integrate these standards, ensuring businesses can remain compliant across multiple jurisdictions.

AI-Driven Insights: Beyond report generation, i-Prove aims to offer businesses deeper insights into their ESG performance. By leveraging AI to analyze trends and forecast future ESG risks or opportunities, companies can use the platform not just for reporting but as a strategic tool for long-term planning.

Expansion to Other Reporting Domains: While currently focused on ESG, the architecture of the i-Prove platform has the potential to expand into other regulatory reporting domains, such as financial disclosures, supply chain transparency, and corporate governance, further broadening its scope.

5. Conclusion

In conclusion, the **i-Prove** project has provided an excellent foundation for automating ESG reporting through AI. The MVP has successfully showcased the platform's capabilities, and with the upcoming transition to production, the system is poised for significant growth. Personally, this project has been a transformative experience, enhancing technical skills in AI deployment, refining knowledge of ESG standards, and providing practical insights into how businesses operate within the UK regulatory landscape.

As i-Prove continues to evolve, it holds the promise of transforming ESG reporting for companies worldwide, offering them an efficient, reliable, and scalable solution to

meet their sustainability and governance goals. The lessons learned and achievements gained throughout this project will undoubtedly shape the next phases of development and beyond.

References

1. Bibliography

1. “Designing Data-Intensive Applications" by Martin Kleppmann
2. "AI Superpowers: China, Silicon Valley, and the New World Order" by Kai-Fu Lee
3. "Building Microservices: Designing Fine-Grained Systems" by Sam Newman
4. "The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling" by Ralph Kimball

2. Webography

- https://www.i-prove.org/
- https://medium.com/
- https://www.globalreporting.org/
- https://www.gov.uk/

Abstract

This report describes the development of an innovative platform that helps businesses generate ESG (Environmental, Social, and Governance) reports automatically. Using generative AI, the platform reduces the time and effort required by automatically collecting and organizing data while adhering to global standards. It also offers customization options to meet stakeholder needs. Through advanced AI features, the reports are accurate, contextual, and aligned with companies' sustainability initiatives.

Ce rapport décrit le développement d'une plateforme innovante qui aide les entreprises à générer des rapports ESG (Environnement, Social, Gouvernance) de manière automatisée. Utilisant l'intelligence artificielle générative, la plateforme réduit le temps et les efforts nécessaires en collectant et organisant automatiquement les données, tout en respectant les normes mondiales. Elle propose également des options de personnalisation pour répondre aux besoins des parties prenantes. Grâce à des fonctionnalités avancées d'IA, les rapports sont précis, contextuels et alignés sur les initiatives de durabilité des entreprises.