NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Improving an Employee Stock Ownership Plan Communication Strategy with GenAI

Development of a GenAI Chatbot for an Employee Stock Ownership Plan in a major aerospace company

Raquel Alexandra Ascensão Rocha

Project Work

presented as partial requirement for obtaining a Master's Degree in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Improving an Employee Stock Ownership Plan Communication Strategy with GenAI

Development of a GenAI Chatbot for an Employee Stock Ownership Plan in a major aerospace company

by

Raquel Alexandra Ascensão Rocha

Project Work presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics, with a specialization in Data Science

Supervised by

Supervisor: Bruno Jardim, PhD, NOVA Information Management School

June, 2025

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Lisbon, 10th of July 2025

DEDICATION

I would like to dedicate this work to my family: to my parents, Aida and Luís, and to my sister Andreia, for supporting me through everything. None of it would have been possible without you.

ACKNOWLEDGEMENTS

I would like to thank my supervisor, Bruno Jardim, for his continuous guidance and support throughout the development of this thesis.

I am also thankful to my team leader, Quentin Ferreira, for believing in my potential and giving me an opportunity even before I had completed my studies.

To my friends and family, thank you for standing by me not only during the writing of this thesis, but throughout the entirety of my studies. Your unwavering support and encouragement means the world to me.

ABSTRACT

This thesis presents the development of a Retrieval-Augmented Generation (RAG) chatbot to support internal communication of the Employee Stock Ownership Plan (ESOP) at a major aerospace company. ESOPs often involve complex documentation and regulations, leading to difficulties in understanding and low employee engagement. To address this, a Generative AI chatbot was built using *Google Vertex AI*, *LangChain*, and *Gemini* large language models. The methodology included cleaning and analysing over 100 ESOP PDF documents, splitting them into chunks using optimal configurations, embedding them into semantic vectors, and testing different retrievers and generative models. Evaluation was conducted using the Retrieval-Augmented Generation Assessment (RAGAS) framework across three metrics: faithfulness, answer relevancy, and noise sensitivity. The best-performing configuration combined hybrid retrieval with the *text-embedding-005* model for embedding and *gemini-2.0-flash-001* for generation. In addition to quantitative metrics, a qualitative user study with ten employees was conducted to assess usability and perceived helpfulness. Results demonstrate the chatbot's potential to enhance information accessibility, reduce dependence on HR teams, and improve overall user experience.

KEYWORDS

Retrieval-Augmented Generation; Large Language Models; Chatbot; Generative AI; Employee Stock Ownership Plan

Sustainable Development Goals (SDG):

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AIML	Artificial Intelligence Markup Language
ALICE	Artificial Linguistic Internet Computer Entity
BERT	Bidirectional Encoder Representations from Transformers
DPR	Dense Passage Retriever
ESOP	Employee Stock Ownership Plan
FAISS	Facebook AI Similarity Search
FAQ	Frequently Asked Questions
GenAI	Generative Artificial Intelligence
GPT	Generative Pre-Trained Transformers
HR	Human Resources
LLM	Large Language Model
LSTM	Long Short-Term Memory
MAS	Multi-Agent Systems
ML	Machine Learning
MTEB	Massive Text Embedding Benchmark
NLP	Natural Language Processing
RAG	Retrieval-Augmented Generation
RAGAS	Retrieval-Augmented Generation Assessment
RNN	Recurrent Neural Networks
Seq2Seq	Sequence-to-Sequence
UI	User Interface
XML	Extensible Markup Language

1. INTRODUCTION

In recent years, organisations have increasingly recognized the importance of employee engagement in creating a better workplace. Among the many tools used to align employee interests with organizational success, the Employee Stock Ownership Plan (ESOP) has been shown to increase employee retention (Guzek & Whillans, 2024). ESOPs allow employees to acquire shares in the company they work for, creating a sense of ownership, improving productivity, and contributing to long-term retention (Blasi & Kruse, 2024). However, while the strategic value of ESOPs is well established, their communication remains a challenge, with research showing that the complexity of legal and financial terminology and the diversity of regulations across countries are meaningful barriers to the adoption and understanding of ESOPs (Guzek & Whillans, 2024).

In parallel, the rise of Generative Artificial Intelligence (GenAI), especially advancements in Natural Language Processing (NLP) and Large Language Models (LLMs), has opened new doors for improving information distribution within organisations (McCorkindale, 2024). Among these innovations, Retrieval-Augmented Generation (RAG) stands out as a standard approach to combine document search with generative models. By retrieving relevant document fragments and feeding them into an LLM to generate contextual answers, RAG chatbots can bridge the gap between static documentation and interactive understanding (Gao, Xiong, Gao, Jia, Pan, Bi, Dai, Sun, & Wang, 2024).

Given this, this project explores the application of a RAG-based GenAI chatbot to improve the communication strategy of a major aerospace company's ESOP. The objective is to design, develop, and evaluate a chatbot capable of responding accurately and understandably to employee questions using only the official ESOP documentation.

The key contributions of this work are:

- A full implementation of a domain-specific chatbot built with *Google Vertex AI* (Google, 2025b) and *LangChain* (LangChain, 2025c);
- An empirical comparison of different embedding, chunking, retrieval, and generation configurations;
- A systematic evaluation using the Retrieval-Augmented Generation Assessment (RAGAS) (Es et al., 2025) framework to assess answer faithfulness, relevance, and noise sensitivity;
- A qualitative human evaluation with employees to assess the chatbot's usability, clarity, and overall user experience.

This thesis aims to fill a gap in both academic literature and enterprise practice, where while there are a lot of enterprise applications of GenAI, including in HR (Human Resources) departments, no known use cases of chatbots tailored specifically to ESOP communication are

currently published. Not only this, but it also further contributes to the growing body of work on responsible and effective applications of GenAI in internal organizational contexts.

2. BACKGROUND

This section introduces the concept of ESOP, and provides a comprehensive view into chatbots, their history and how they have become a phenomenon in the current organisational context. The aim is to provide the context for the subsequent chapters by underlining the growing importance of chatbots in organisations and understanding how they can provide support for an ESOP in an organisation.

2.1. EMPLOYEE STOCK OWNERSHIP PLAN

An ESOP is a workplace benefit that allows employees to own stock of the company they work for, aligning their financial interests with the company's performance (Blasi & Kruse, 2024), aiming to, not only fight inequality, but also to incentivize employee productivity, cooperation and teamwork (Blasi et al., 2017). These were first implemented in the United States of America in 1974 to address shortfalls in Social Security, to incentivise economic growth and to redistribute corporate profits to motivate employees (Freeman, 2007). The point was, if employees felt like they also had a stake in the company's success, they would have more motivation to be productive.

However, employees often perceive ESOPs as complex and difficult to understand, which can lead to misinformation, low participation, and reduced trust in the program. Research has shown that ineffective communication regarding ESOPs can result in employees feeling disengaged or unaware of their benefits, consequentially reducing the program's effectiveness. Therefore, proper communication strategies must be in place to overcome these challenges. According to Klein (1987), organisations with well-structured ESOP communication strategies report higher levels of employee satisfaction and engagement. Such communication makes employees feel more informed and connected to the ESOP, which in turn, improves their sense of ownership and engagement with the organisation.

In recent years, AI has been increasingly integrated into HR processes to improve efficiency, automate repetitive tasks, and improve employee engagement. GenAI chatbots facilitate real-time interactions, providing employees with instant access to information while reducing the administrative burden on HR departments (Nawaz et al., 2024). Given this trend, integrating AI-powered chatbots for ESOP communication could reduce misinformation and increase employee engagement, aligning with the current digital transformation efforts within HR departments.

2.2. NATURAL LANGUAGE PROCESSING

NLP is a branch of AI focused on enabling machines to understand, interpret, and generate human language. NLP facilitates the interaction between humans and machines, making it applicable to a wide range of tasks, including machine translation, text categorisation, spam

filtering, and conversational AI (Khurana et al., 2023). Chatbots have become a main application of NLP, allowing users to communicate with a machine using natural language.

2.2.1 Seq2Seq Models

One of the key challenges in NLP is handling tasks that require transforming one sequence of text into another, such as translating sentences between languages or summarising large documents. To address this, Sequence-to-Sequence (Seq2Seq) models were developed, which enable machines to process input text and generate coherent, contextually relevant output (Sutskever et al., 2014). These models have become fundamental in many NLP applications, particularly those requiring structured text generation.

Seq2Seq models are composed of two main components:

- **Encoder**: Processes the input sequence and encodes it into a fixed-length vector representation (the context vector). This vector captures the essence of the input sequence.
- **Decoder**: Takes the context vector as input and generates the output sequence token by token. Each token generation depends on the previously generated tokens and the context vector.

While Seq2Seq models achieved impressive results, their reliance on fixed-length vectors derived from the encoder's final hidden state limited their ability to handle long and complex sequences. A hidden state is a vector produced by the encoder at each step, capturing information about the current input token and the preceding context. This led Bahdanau et al., (2016) to come up with the attention mechanism. The attention mechanism revolutionised Seq2Seq models by allowing the decoder to focus on relevant parts of the input sequence at each step, instead of relying on a single fixed-length vector. Attention calculates a weighted sum of the encoder's hidden states, where the weights represent the relevance of each input token to the current decoding step (Bahdanau et al., 2016). The process involves:

1. Computing a score for each input token based on its alignment with the decoder's current state.
2. Applying a *softmax* function to normalise these scores into attention weights.
3. Generating a context vector as the weighted sum of the encoder's hidden states.

This innovation improved the performance of Seq2Seq models, especially for long sequences, and paved the way for the Transformer architecture.

2.2.2 Transformers

Transformers, introduced by Vaswani et al. (2017), revolutionised NLP architecture by eliminating the need for sequential processing through self-attention mechanisms, unlike

their predecessors, which relied on recurrent neural networks (RNNs). This innovation addressed the previous models' inability to efficiently capture long-range dependencies and their inherent computational inefficiency due to sequential processing constraints (Hochreiter & Schmidhuber, 1997; Sutskever et al., 2014)

Transformers have become the foundational architecture for various NLP tasks, including machine translation, summarisation, question answering, and text generation (Vaswani et al., 2017; Radford et al., 2018). Furthermore, their flexibility has enabled them to expand beyond NLP into other domains such as computer vision (Dosovitskiy et al., 2021), protein structure prediction (Jumper et al., 2021), and time-series forecasting (Zerveas et al., 2020).

Like Seq2Seq models, Transformers are based on an encoder-decoder architecture. In the encoder, the self-attention mechanism allows the model to dynamically attend to all tokens in the input sequence, enabling it to capture global dependencies regardless of their positional distance (Vaswani et al., 2017). The feed-forward network then applies non-linear transformations independently to each token's representation, refining the contextual information generated by the attention mechanism. These components are connected using residual connections and layer normalisation, which stabilise training and improve optimisation, as can be observed in Figure 1.

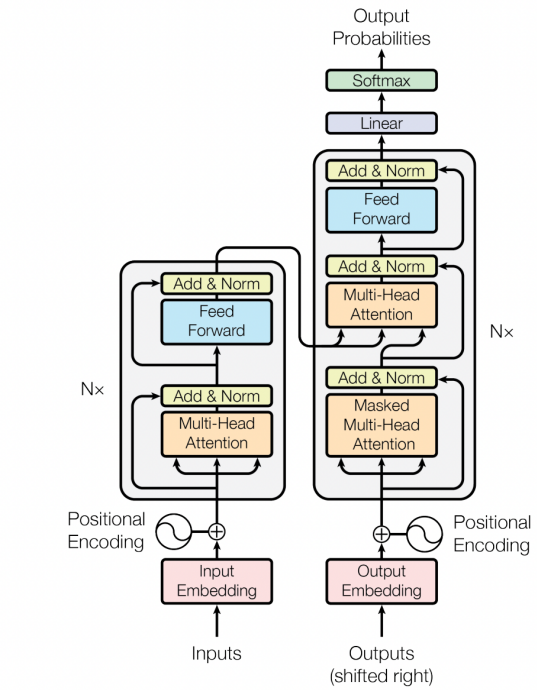

Figure 1 - The Transformer Model Architecture (Vaswani et al., 2017)

The decoder shares many components with the encoder but includes additional mechanisms tailored for sequence generation. One of the most important is masked self-attention, which ensures that each token can only attend to those that come before it. This maintains the correct order during autoregressive decoding, allowing the model to generate the output one

token at a time, without access to future positions. For example, when predicting the second word, it should only use the first word for context, and when predicting the third word, it should consider only the first and second words. This sequential dependency ensures that the model doesn't "cheat" by looking ahead to future tokens that it hasn't predicted yet.

Furthermore, cross-attention allows the decoder to focus on relevant parts of the encoder's output, aligning the input and output sequences effectively. Transformers also employ positional encodings to address the lack of inherent sequential order in self-attention. These encodings, added to the input embeddings, enable the model to differentiate between positions in the sequence (Vaswani et al., 2017).

2.2.3 Retrieval-Augmented Generation

The concept of RAG was first introduced in the seminal paper by Lewis et al. (2021), titled "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." This work proposed a framework that integrates a neural retriever with a pre-trained generative model to improve the model's ability to generate accurate and contextually relevant text based on documents or other type of information. The retriever accesses a large knowledge base to find passages relevant to a given query, which are then used by the generative model to produce responses grounded in external knowledge (Lewis et al., 2021).

One of the biggest challenges in generative models, particularly LLMs, is the phenomenon of hallucinations. Hallucinations occur when a model generates information that is factually incorrect, implausible, or not grounded, despite appearing coherent or convincing. These errors arise because generative models rely heavily on the statistical patterns learned during pre-training, which may not always align with factual or up-to-date information. Hallucinations are particularly problematic in knowledge-intensive tasks, where accuracy and reliability are critical (L. Huang et al., 2024). In these scenarios, it's ideal to store the necessary information in a database, so that when specific questions arise, the relevant information can be retrieved and presented to the LLM, providing a more efficient and reliable approach (Jeong, 2023).

The RAG framework consists of two main components: the Retriever and the Generator. The RAG framework addresses the hallucination problem by incorporating a retriever, which ensures that the generative component has access to external, contextually relevant, and up-to-date information. The retriever is a crucial component that identifies relevant documents or passages from an external knowledge base, such as informative documents. It employs a dense passage retriever (DPR) (Karpukhin et al., 2020), which encodes both queries and passages into a shared vector space using dense embeddings, and proceeds to perform similarity searches in this shared space. This allows the retriever to efficiently identify the top-k relevant passages for a given query, which provide factual grounding for the generative model, reducing its reliance on potentially incomplete internal representations, therefore mitigating the risk of hallucinations.

The generator in the RAG model is typically a Seq2Seq model, which takes the retrieved passages as input, concatenates them with the input query, and generates a response that combines the query with the retrieved information. By grounding its responses in these retrieved passages, the RAG model significantly mitigates the risk of hallucinations, producing outputs that are more aligned with verified external knowledge (Lewis et al., 2021).

By integrating retrieval and generation, RAG combines the strengths of information retrieval systems and generative models, ensuring that responses are both contextually coherent and factually grounded. This dual approach improves the accuracy of knowledge-intensive tasks while significantly reducing the frequency of hallucinations, thereby enhancing trust in the model's outputs.

2.2.3.1. Word Embeddings

Word embeddings play a key role in RAG systems by providing a dense and continuous representation of words in a vector space. These embeddings capture both semantic and syntactic relationships between words, enabling models to perform better in a wide range of applications such as sentiment analysis, translation, and information retrieval (Almeida et al., 2023). Since RAG relies on retrieving semantically relevant text chunks before generating an answer, word embeddings serve as the foundation for semantic search and contextual understanding.

In 2024, Google DeepMind introduced Gecko, a new text embedding model, powered by a two-step distillation process to transfer knowledge from LLMs into a retriever. In this context, passages refer to short segments of text extracted from a larger document or corpus. These passages are typically a few sentences or paragraphs long and are used as candidates for retrieving information in response to a query. In Gecko's training process, passages served as the context that the model evaluated to determine how relevant they were to a given query. During training, these passages were categorised as positive (relevant), negative (irrelevant), or hard negatives (irrelevant examples that are superficially similar to the query). The model learned to generate embeddings that differentiate between these categories, improving its ability to retrieve the most relevant passages in real-world applications (Lee et al., 2024).

Compared to other models, Gecko offers significant advantages in efficiency and performance. On the Massive Text Embedding Benchmark (MTEB) (*Mteb (Massive Text Embedding Benchmark)*, 2025), Gecko with only 256 embedding dimensions outperforms models with 768 dimensions. Its larger version, with 768 dimensions, achieves competitive results with models that are seven times larger and have embeddings five times the size. This combination of compactness and strong performance demonstrates Gecko’s ability to conglomerate the capabilities of LLMs into a lightweight model that excels in retrieval, semantic similarity, and other NLP tasks (Lee et al., 2024).

Google also released two more English embedding models, *text-embedding-004* and *text-embedding-005* (Google, 2025b). Although no information on their architecture is publicly

available, it is known that these two models outperformed Gecko, with a higher performance on MTEB.

2.2.4 Chatbots

Chatbots are programs that can simulate a human conversation (Z. Han, 2023), interacting with users using spoken, written, and visual languages. The concept of a chatbot was first proposed in 1950, when Alan Turing introduced the "Turing Test" in his paper "Computing Machinery and Intelligence" (Turing, 1950). This test aimed to evaluate whether a machine could have an intelligent conversation indistinguishable from that of a human, being able to fool a person into believing it was human (Al-Amin et al., 2024).

Chatbots can be of two types (Pillare et al., 2023):

- **Rule-based chatbots**: based on pre-defined rules
- **AI-powered chatbots**: based on machine learning (ML) algorithms

2.2.4.1. History of Chatbots

The earliest well-known chatbots were ELIZA and Artificial Linguistic Internet Computer Entity (ALICE), both of which are what is called a rule-based chatbot. As described by Al-Amin et al. (2024), early chatbots functioned on a basic text-based interaction, using pattern-matching techniques. This meant that the responses were generated from predefined templates based on a set of rules: the chatbot would search the input for keywords, and, if a match was found, the chatbot would respond with the corresponding output template. ELIZA, developed by Dr. Joseph Weizenbaum from the Artificial Intelligence Laboratory at MIT in the 1960s, ran the famous "DOCTOR" script, which worked with set rules for interactions between a psychiatrist and a patient (Bassett, 2019).

A conversation between ELIZA and a user can be observed in Figure 2l:

User: Men are all alike.
ELIZA: In what way?
User: They're always bugging us about something or other.
ELIZA: Can you think of a specific example?
User: Well, my boyfriend made me come here.
ELIZA: Your boyfriend made you come here.
User: He says I'm depressed much of the time.
ELIZA: I am sorry to hear you are depressed

Figure 2 - A conversation between ELIZA and a user (Weizenbaum, 1966)

In 1995, another milestone was achieved in the history of chatbots with the creation of ALICE, developed by Dr. Richard Wallace. Wallace was inspired by the work done by Weizenbaum on

ELIZA and created a new, more sophisticated form of pattern-matching rules, using a new technology called Artificial Intelligence Markup Language (AIML) (Al-Amin et al., 2024). AIML derives its structure and syntax from XML (Extensible Markup Language), which is a standardised markup language used to define, store and transport data across different systems and platforms, providing a universal framework for encoding documents in both human and machine-readable formats. The way this happens is through the usage of tags to describe data and its structure (Salminen & Tompa, 2001).

Chaudhary & Gupta (2023) explain that AIML provides a structured framework similar to XML, however, specifically tailored for defining user input patterns and corresponding responses. By simplifying the process of scripting chatbot dialogues, ALICE allowed for more complex conversations, while AIML provided the basis for further advances in chatbot technology.

With the introduction of deep learning in the late 90s, such as the invention of the long short-term memory (LSTM) model in 1997, machines began to comprehend written, visual, or audio information, and the field of ML experienced significant progress (Al-Amin et al., 2024). These advancements laid the groundwork for more sophisticated chatbot systems, moving beyond purely rule-based responses toward models that could learn patterns from data.

In 2001, Active Buddy created Smarter Child, a new chatbot that follows a more hybrid approach between rule-based and AI-powered chatbots, by including a few simple NLP elements. It ended up being one of the first widely famous chatbots, designed to interact with users via messaging platforms such as MSN Messenger, and to provide information on a range of different topics. With its exposure to the public on widely used platforms and its ability to answer a wide range of questions, SmarterChild showed the potential of chatbots in supporting average tasks and becoming a part of the public's daily life (Al-Amin et al., 2024).

Later, in the 2010s, voice assistants gained prominence, such as Amazon's Alexa or Apple's Siri, introducing the concept of a personal assistant to a larger consumer base. These virtual personal assistants were integrated into devices such as speakers or smartphones and were able to understand when they are being summoned, comprehend human requests and respond both via voice and behaviour (X. Huang & Marcus, 2021). These technologies provided more intuitive and personalised user experiences through the developments in NLP and ML.

2.2.4.2. Generative AI Chatbots

The most impactful breakthrough in chatbot technology happened in the last decade, with an unprecedented speed, with the emergence of Generative Pre-trained Transformers (GPT). GPT models are built on the previously mentioned Transformer architecture, introduced by Vaswani et al. in 2017 in the famous paper "Attention Is All You Need", which revolutionised NLP by replacing older RNNs and LSTM models. These advancements originated what is nowadays called GenAI, enabling chatbots to generate original, human-like responses instead

of relying on pre-defined rules. GenAI chatbots can "understand and produce natural language, learn from experience and portray emotions" (Hoyer et al., 2020).

The release of *ChatGPT* (*ChatGPT*, 2025) in 2022 by OpenAI is considered a milestone not only in the history of chatbots but in the history of AI. *ChatGPT* is an NLP system that can understand the context of a conversation and then generate responses resembling humans, originally based on the deep learning model *GPT-3* which was trained on 499 billion tokens. It can also generate answers in multiple languages and styles (Deng & Lin, 2023).

The impact caused by the release of *ChatGPT* was tremendous, with over a million users in its first week (Holmström & Carroll, 2024), and the number of organisations looking to automate and facilitate their processes with GenAI chatbots grew exponentially.

2.2.4.3. Organizational Chatbots

Due to the recent developments and advancements in GenAI and the ongoing digital transformation of businesses, chatbots are now providing support in the daily activities of many businesses. Various industries are taking advantage of this technology, such as banking, telecommunications and travel (Bălan, 2023). Chatbots can improve productivity by reducing workload for employees, automating repetitive work, and generating new insights not caught by humans to improve decision-making. They are also capable of processing and learning from large amounts of data, allowing them to constantly improve themselves, which helps companies improve their services and response to market changes (Wang et al., 2022).

According to Gkinko & Elbanna, (2023), from an organisational point of view, chatbots can be externally facing when they are designed to interact with external users, such as clients or customers, and provide information such as guiding purchases, troubleshooting issues or answering questions, usually being optimised for accessibility and user engagement. They can also be internal facing, when they are designed for organisational usage, supporting employees in tasks such as accessing HR services, automating repetitive processes or gathering information about internal processes and systems. The focus is to enhance productivity and reduce dependency on staff from other departments, such as HR, IT or Financial.

Wang et al., (2022), in their study "How does artificial intelligence create business agility? Evidence from chatbots" suggest that the usage of internal facing chatbots can improve productivity. The study divides chatbot usage into two types: routine and innovative, and both showed positive results when it comes to business agility and sustainability. Gkinko & Elbanna, (2023) also argue that AI chatbots improve not only productivity but also employees' workplace experience.

Replacing human chat agents or Frequently Asked Questions (FAQ) pages with chatbots is becoming recurrently common (Adam et al., 2021), given that chatbots can provide consistent, 24/7 responses, allowing users to have access to the needed information at any

time. A study conducted by Brandtzaeg & Følstad in 2017, shows that users resort to them mostly due to productivity, reporting the “ease, speed and convenience of using chatbots”, and give especial importance to time saving and easy access to information, instead of having to contact a human chat agent or search for the information in a FAQ. The study also mentions that this happens mainly in Western countries, where there is a bigger concern about spending time productively.

3. RELATED WORK

Given the previously mentioned developments in ML and NLP, organisations have witnessed an increase in the adoption of chatbots. Internal facing chatbots, designed for employee usage, are being used to optimise workflows, enhance resource accessibility, and boost both employee productivity and employee satisfaction. This section examines some of the found case studies of organisations using chatbots.

Afzal et al. (2024) developed an HR support chatbot to answer employees' inquiries, allowing domain experts to focus on other valuable tasks. The database was composed of an FAQ dataset and a collection of previous iterations. RAG was the used framework, with which the retriever fetched relevant HR documents from a knowledge base comprising approximately 50,000 unique articles. To improve retrieval accuracy, two primary methods were explored: DPR using Bidirectional Encoder Representations from Transformers (BERT), which fine-tuned the *bert-base-uncased* model (*Google-Bert/Bert-Base-Uncased · Hugging Face*, 2024) to retrieve pertinent articles based on user queries, and OpenAI's Vector Search that generated embeddings for articles using the *text-embedding-ada-002* model (*Model - OpenAI API*, 2025) and used similarity search to find relevant content. Additionally, the retriever incorporated filters to ensure that it only considered articles aligned with user-specific attributes like country, region, and employment status.

When it comes to text generation, the *long-t5-local-base* model (Guo et al., 2022), containing 296 million parameters, was fine-tuned with training configured with a learning rate of 1e-4, a batch size of 8, over 5 epochs. OpenAI's *GPT-4* and *GPT-3.5* models (*Models - OpenAI API*, 2025) generated answers based on the user query and the retrieved articles, and the prompt engineering was iterative, involving qualitative analysis and evaluations by HR experts to improve responses according to the company's requirements.

The chatbot's performance was assessed through:

- **Retriever Evaluation**: accuracy was measured by the percentage of times the retriever returned the correct article for a given question
- **Human Evaluation**: domain experts evaluated 100 samples across different models, and rated dimensions such as readability, relevance, truthfulness and usability on a Likert scale
- **Reference-Based Metrics**: N-gram-based metric and embedding-based metrics were employed to assess similarity between generated responses and ground truth answers
- **Reference-Free Metrics**: LLM-based evaluations were explored to provide insights without relying on predefined reference responses.

The evaluation results demonstrated that *GPT-4* outperformed all other models, including *LongT5* and *GPT-3.5*, across multiple metrics. Human evaluators rated *GPT-4* responses as the most readable, relevant, and truthful, making it the most reliable choice for HR-related

question answering. The *LongT5* model, despite being fine-tuned, showed lower performance in truthfulness and usability, indicating that fine-tuning alone was insufficient to match *GPT-4*'s reasoning capabilities. In terms of retrieval accuracy, the DPR fine-tuned with BERT achieved the highest accuracy in selecting relevant HR documents, while OpenAI's *text-embedding-ada-002* performed well but required query transformations (e.g., HyDE and Multi-Query methods) to improve relevance.

Focusing on the efficiency of RAG systems, Kaif et al. (2024) integrated RAG with Google's *gemini-1.5-pro* model (*Gemini Models | Gemini API*, 2025) with a methodology starting with parsing and chunking text from PDF documents into pieces, and each chunk then transformed into embeddings, using Google's *embedding-001* model (*Gemini Models | Gemini API*, 2025). The embeddings serve as the foundation for the similarity-based retrieval using the *Facebook AI Similarity Search* (FAISS) library (Matthijs Douze, 2025), indexing the embeddings and supporting high-speed similarity searches. The FAISS index is implemented through *LangChain*, a framework that integrates retrieval and generation processes (LangChain, 2025c). Upon receiving a query, the system retrieves relevant document chunks via FAISS and uses the *gemini-1.5-pro* model to generate grounded responses. The study shows that, by using a RAG system, the contextual understanding of the model is improved, and reduces costs by retrieving and using fewer tokens than using the full context with every call of the model, as can be observed in Figure 3 (Kaif et al., 2024).

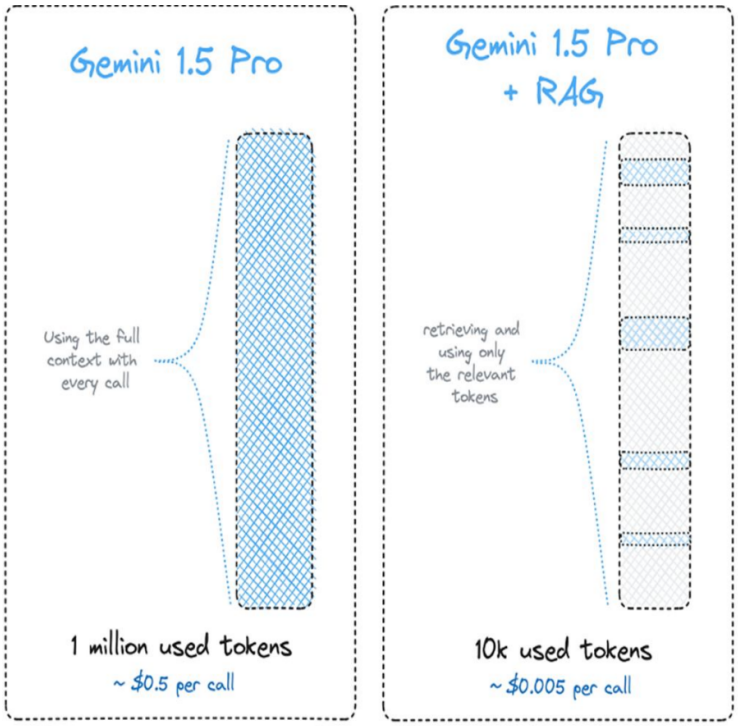

Figure 3 - Cost comparison of base *Gemini* model vs. *Gemini* with RAG (Kaif et al., 2024).

In NVIDIA, Akkiraju et al. (2024) developed three internal chatbots, which is explained in their paper "*FACTS About Building Retrieval Augmented Generation-based Chatbots*". In this paper,

the authors introduce the FACTS framework, which addresses five critical dimensions: Freshness, Architectures, Cost, Testing, and Security. Their primary objective was to increase employee productivity by developing effective enterprise chatbots powered by GenAI.

One of the built chatbots, NVHelp, was designed to assist employees with IT and HR-related queries. The model architecture was multi-agentic, using a Multi-Agent System (MAS), which supports multiple LLMs depending on the task requirements. Its architecture was centred around hybrid search, which combined keyword and semantic search techniques to ensure response relevance. Keyword search, powered by Elasticsearch (*Elasticsearch*, 2025), was used to handle keyword-based queries, while semantic search utilised vector embeddings, using models like FAISS and OpenAI's embedding models. By integrating both retrieval strategies, the system achieved improved precision and recall, making sure that the retrieved information was relevant and accurate. Additionally, documents within the chatbot's knowledge base were pre-processed using chunking and metadata enrichment techniques, allowing for more effective retrieval and improved context understanding.

Once the chatbot retrieved relevant information, it processed user queries using a combination of different LLMS to generate responses. NVHelp was tested using *GPT-4* from OpenAI and *Llama3-70B* from Meta (Meta AI, 2025) both of which were evaluated based on response accuracy, coherence, and computational efficiency. *GPT-4* was established as a strong baseline due to its high-quality generative capabilities, whereas *Llama3-70B* was considered a cost-effective alternative with a favourable trade-off between response quality and latency. The chatbot's query processing pipeline involves retrieving relevant knowledge, reranking the retrieved results, and generating a final response using the selected LLM.

To assess the chatbot's effectiveness, the developers conducted an evaluation using the RAGAS framework, which enabled automated testing and benchmarking against a dataset of approximately 245 IT and HR-related queries. The results demonstrated that the integration of RAG techniques significantly enhanced response accuracy, as the retrieval component provided the LLM with highly relevant contextual data. This implementation provides a relevant precedent for the development of an ESOP-specific chatbot, where the same challenges (e.g.: frequent regulatory updates, individual-specific eligibility conditions, and the need for clarity in complex financial terminology) are present. By adopting similar retrieval mechanisms, the chatbot developed in this thesis aims to deliver accurate and transparent support for employees who are interested in the ESOP.

During the research conducted for this Literature Review, a variety of organizational chatbot use cases were identified, including in HR, using mostly OpenAI models, but also *Llama*. Notably, *Gemini* has yet to gain relevance in these contexts, indicating a gap in the current research and presenting an opportunity to examine its applicability and potential advantages in organizational chatbot implementations.

Although numerous HR chatbots exist, no documented implementations specifically designed to provide information on ESOPs were identified, which shows a research and application gap. While chatbots are widely adopted in topics such as customer support, healthcare, finance, and HR, their use in addressing ESOP-related inquiries remains unexplored. Existing HR chatbots tend to focus on general topics such as payroll assistance, benefits management, or policy FAQs, yet none appear tailored to the unique regulatory, financial, and contextual complexities associated with ESOPs.

These complexities pose unique challenges for technical implementation, particularly in ensuring the accuracy and personalization of chatbot responses. Addressing these challenges not only increases the value of the ESOP chatbot but also strengthens the generalisability of this research to other high-stakes, information-dense use cases within organizational settings. This thesis, therefore, aims to bridge the identified gap by developing a domain-specific chatbot that uses LLMs and RAG to deliver transparent, accessible, and compliant ESOP-related information.

4. METHODOLOGY

This section presents the methodology used in the development of the ESOP Chatbot. Given that the chatbot uses a RAG system, which integrates external knowledge retrieval with LLMs, the applied methodology was based on the steps needed for the implementation of this type of system. Jeong, (2023) presents the development of a RAG-based system following a structured pipeline, comprising the following steps:

1. **Data Collection and Preparation**: gathering relevant data sources to ensure the model has access to accurate and comprehensive information;
2. **Text Chunking**: dividing the collected data into smaller text fragments, known as chunks, to facilitate efficient retrieval;
3. **Embedding**: the chunked texts are transformed into vector representations that capture their semantic meaning;
4. **Construction of a Vector Database**: storing the embedded vectors within a high-dimensional space to enable similarity searches;
5. **Information Retrieval:** retrieving the most relevant text chunks in response to a user query, ensuring that the generated responses are contextually relevant and factually grounded;
6. **Text Generation**: synthesises the retrieved information into a response, allowing the LLM to provide accurate and well-structured answers.

The steps Text Chunking, Embedding, Information Retrieval, and Text Generation were tested multiple times with different parameters or models to understand what the best choices for this use case were, allowing the chatbot to reach the most of its potential with the resources and data available.

Although the original methodology does not incorporate it, this thesis introduces an additional evaluation step using RAGAS, a specialized framework for evaluating RAG systems (Es et al., 2025), and human evaluation. This step enables an assessment of the chatbot's response accuracy and contextual relevance, marking an important addition to understanding the overall effectiveness and value of the developed solution. An overview of the pipeline can be seen in Figure 4.

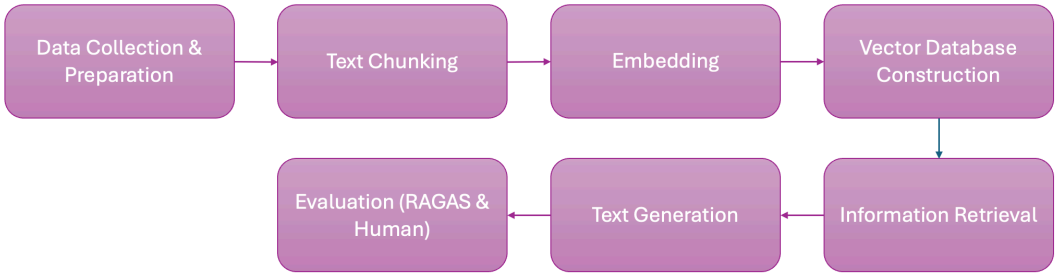

Figure 4 - Methodology pipeline for the ESOP Chatbot Development

4.1. DATA ANALYSIS AND PREPARATION

The data to be used by the RAG system was completely unstructured, consisting of 110 informative PDF files for the ESOP in different countries and languages, previously provided by the responsible team for the program. These documents were used in 2024 to inform employees about the ESOP, meaning that the chatbot would not introduce new information to the public, but help users find what they need faster, instead of searching through each document.

To better understand the content of the PDF files, a word count of all the files was performed, revealing that the PDFs ranged from 88 words to 11715, and, on average, they contained 1842 words, which can be considered a low value compared to the maximum one. To find which were the most common words in the files, a word cloud was generated, which can be found below in Figure 5.

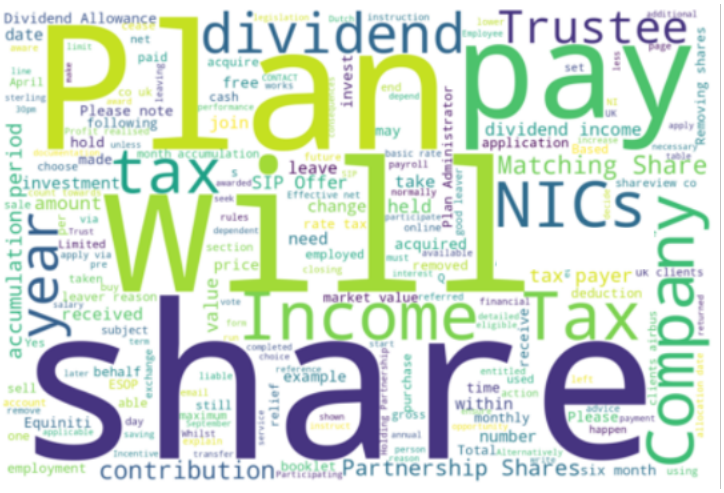

Figure 5 - Word Cloud derived from PDF files

Most of the words shown in the word cloud are expected, such as “Plan”, “Share”, “Pay” and “Company”, which are included in the common syntax of an ESOP, while words such as “tax” and “dividend”, although less frequent, also align with the topic of the documents at hand. Although not a lot of specific insights can be retrieved from this word cloud, it suggests that the files are relevant for the topic. When it comes to stop words, these are essential for the fluency, grammar and meaning of natural language in LLMs and RAG systems, therefore should not be removed.

For this task, many HTML tags had to be removed, mostly “\\n” and “\\t”, given that these would be shown in the word cloud and the word count statistics. However, this step had to be undone after the generation of this information, due to the importance of word boundaries in the text chunking to avoid division of sentences.

Upon a manual analysis of the PDF files, it was found that the same files were translated into different languages, leading to a need to thoroughly read them and understand which

contained the same content. The duplicated files in a different language would be removed from the dataset, given that the duplicated information could confuse the model. In Appendix A, a table with the titles of the files, the number of files and repeated languages can be found. After this analysis, out of 110, only 62 were found to be useful and to be kept in the RAG system.

After the data cleaning, the mean of words per document is lower than previously (1570), suggesting that the files that were removed contained more words than the previous average.

When it comes to the word cloud in Figure 6, the words “share”, “plan”, “pay” and “will” are still in the spotlight, suggesting that the content of the text did not vary much with the data cleaning, which makes sense given that only repeated documents were removed.

Figure 6 - Word Cloud after Data Cleaning

4.2. TEXT CHUNKING

The next step in Jeong’s (2023) methodology was text chunking, a process which consists of splitting the data into smaller fragments (chunks) to allow for more efficient and accurate retrieval.

In this step, *Langchain*'s (2025) *RecursiveCharacterTextSplitter* was used. *LangChain* is a framework designed to support the development of applications using LLMs, allowing for simplification of the various stages of the application lifecycle, including development, production and deployment (LangChain, 2025c). The *RecursiveCharacterTextSplitter* allows for splitting text based on a fixed character count, defined by the developer, while ensuring that no information is lost at the boundaries of each chunk (Narimissa & Raithel, 2024). This can be guaranteed in two ways: with a chunk overlap, which as the name states, allows for a set number of characters to be overlapped so no context is lost between sentences, and through word separators, such as “\n” or “\\n”, that help the splitter avoid splitting words or sentences in between chunks. This is the reason why the HTML tags removal step had to be undone after the word statistics.

To identify the best chunking configuration, eight combinations of chunk size and chunk overlap were tested, starting with a size of 200 and overlap of 20% (40), and gradually increasing until a size of 1500 and overlap of 50% (750). The sizes 200, 500, 1000 and 1500 were tested, with overlaps of 20% and 50%.

The tested configurations are detailed in Table 1.

Table 1 - Tested Chunk Sizes and Overlap

Chunk Size	Chunk Overlap
200	20% (40)
200	50% (100)
500	20% (100)
500	50% (250)
1000	20% (200)
1000	50% (500)
1500	20% (300)
1500	50% (750)

These attempts allowed not only to determine the most effective chunk size, but also the best overlap percentage to preserve semantic continuity and improve retrieval performance, which represents an extremely important step given that poorly sized chunks can lead to loss of context, low retriever accuracy and reduced answer quality.

4.3. EMBEDDING AND VECTOR DATABASE

To develop this chatbot, the platform used was *Google's Vertex AI* due to organisational guidelines. *Vertex AI* is an AI development platform that provides access to the most recent *Gemini* models, providing a fully managed infrastructure, facilitating model deployment, scaling and monitoring (Google, 2025b). However, a significant limitation of this approach is that it restricts model selection, as only Google's proprietary generative models can be utilised, eliminating the possibility of experimenting with alternative architectures. Despite this constraint, as of the development of this chatbot, *Vertex AI* provides access to three multilingual embedding models: *text-embedding-004*, *text-embedding-005* and *textembedding-gecko@003* (Gemini, 2025a).

All three models have the same dimensions, outputting 768-dimensional dense vectors and accepting 2048 input tokens. On the MTEB leaderboard, the *text-embedding-004* model scores 59,06 on retrieval tasks, the *text-embedding-005* model scores 58,77 (*MTEB Leaderboard - a Hugging Face Space by Mteb*, 2025), while *textembedding-gecko@003* scored

55,7 (Gemini, 2025b). At the time of writing, the first two models ranked in the MTEB leaderboard in 8th and 7th place, respectively, as shown in Table 2.

Table 2 - Embedding Models' information on dimensions and MTEB scores

Embedding Model	Input Tokens	Output Size	Score on Retrieval Tasks	MTEB Leaderboard
textembedding-gecko@003	2048	768-dimensional dense vectors	55,7	N/A
text-embedding-004	2048	768-dimensional dense vectors	59,06	8th
text-embedding-005	2048	768-dimensional dense vectors	58,77	7th

To store the generated embeddings, a vector database was required. Unlike traditional relational databases, which rely on exact-match queries, vector databases store data as high-dimensional vectors and support similarity searches, allowing for more effective retrieval of semantically relevant content. They also enable the storage and retrieval of unstructured data, such as PDF documents, making them particularly well-suited for RAG applications (Y. Han et al., 2023). In this implementation, *ChromaDB* was selected as the vector database, integrated through *LangChain* to facilitate efficient retrieval operations (LangChain, 2025b). A collection was created within *ChromaDB* for each PDF file, enabling structured indexing and retrieval of document content.

4.4. INFORMATION RETRIEVAL

In the next stage, a retriever was incorporated to identify and retrieve the most relevant chunks in response to a question. The retriever plays a crucial role in the RAG framework by narrowing down the search space, ensuring that only the most contextually relevant information is passed to the model. Without it, the chatbot would generate responses based on an excessive amount of data, leading to inefficiency and, possibly, inaccurate answers.

Two retrieval methods were tested: the *ChromaDB* vector space as retriever, and an ensemble retriever containing a *BM25* retriever (LangChain, 2025a) and a FAISS vector space. Both methods were used inside *LangChain*'s *RetrievalQA* function (LangChain, 2025e), which allows the developer to choose which to use. The impact of these retrievers was evaluated on the results of the chatbot, both being tested with the different chunk sizes and models mentioned above.

4.5. TEXT GENERATION

After preparing the embeddings and configuring the retriever, the next step involved querying the LLM to generate responses. A critical aspect of optimising LLM performance is prompt engineering, which ensures the model interprets queries correctly and delivers high-quality responses, which is why a detailed prompt was designed.

To establish a clear role, the chatbot was designated as "Ask ESOP," a knowledgeable policy expert explicitly instructed to provide precise and reliable answers based solely on the provided ESOP documents. The prompt also includes clear instructions that constrain the model to base its response solely on the retrieved documents and discourage any speculation or hallucination. This ensures that the model's output remains grounded in verifiable enterprise content, aligning with the RAG methodology even without the explicit inclusion of the query text in the visible prompt. The prompt used was the following:

"*You are Ask ESOP, a knowledgeable and helpful expert on [redacted]'s Employee Stock Ownership Plan (ESOP).*

Your task is to provide concise, accurate, and conversational answers to the following [question], using only the provided ESOP documents as your source of information.

Avoid phrases like 'The documents state' or 'According to the documentation.'. Integrate the information smoothly into your responses.

Important Guidelines:

- *Accuracy First: Your responses must be directly supported by the provided ESOP documents. Do not guess, speculate, or hallucinate information.*
- *Avoid Financial Advice: Do not offer personalized financial advice.*"

By incorporating these prompt guidelines, the responses remained reliable, transparent, and aligned with organisational policies while reducing risks associated with misinformation.

Moreover, the chatbot incorporates memory through *LangChain*'s *ConversationBufferMemory* (LangChain, 2025d), which stores the dialogue history between user and assistant across turns. This memory is passed into each interaction with the LLM, allowing it to maintain context inside the conversation.

To build the chatbot's interactive interface, Streamlit (Streamlit, 2025) was used, an open-source framework that allows the quick development of a simple User Interface (UI).

4.6. EVALUATION

The evaluation and assessment of GenAI chatbots, specifically, RAG-based, has been proven to be a challenging task, given that their results are most times subjective and context specific. While, for example, classification models expect defined outputs, generative models return

open-ended and natural language responses. In the case of RAG, this becomes more complex since not only does the response have to be evaluated, but also the retrieval quality (whether the correct information is selected). The RAGAS framework, presented in the paper "RAGAS: Automated Evaluation of Retrieval Augmented Generation" by Es et al. (2025), was created specifically for evaluating RAG pipelines through a list of metrics that assess the different mentioned dynamics.

4.6.1 Evaluation with RAGAS

The evaluation process followed a specified order, starting with the evaluation of different combinations of chunking size and overlap, and choosing the one providing the best results. Then, these chunking settings were applied, and nine combinations of three embedding and three generative models were tested. All these tests were performed with both hybrid and semantic search, to comprehend the effect of each retriever on the results. This evaluation strategy allowed for choosing the best combination of retrieval, chunking, embedding and generative models, maximising the potential of the chatbot to the available resources. The evaluation was performed on 10 example questions provided by the business, but, unfortunately, a ground truth to perform a semantic similarity evaluation was not provided.

The provided questions were the following:

- *Is the salary deduction the only accepted method of payment for ESOP?*
- *Can I cancel my ESOP subscription even though the subscription period has ended?*
- *Will I perceive dividends on my ESOP stock investment and how will they be taxed as a German tax resident?*
- *When will my shares be allocated onto my Société Générale bank account?*
- *What difference is there between the PEG and the direct ownership in France?*
- *Can I be eligible for the ESOP 2024 offer if I was hired on the 1 st of January 2024?*
- *Can I subscribe to a package for which the price is higher than 3 months of my salary?*
- *When will I be able to sell my shares?*
- *Are there any taxes or benefit in kind in France, which may add to the cost of the subscribed packages?*
- *Who can I contact in order to help me to subscribe to ESOP?*

When it comes to the evaluation strategy, the following RAGAS metrics were used (RAGAS, 2025):

- **Faithfulness**: measured how factual the response was, given the retrieved context. The results range from 0 to 1, with a higher score meaning a better result. This metric was calculated by identifying all the claims in the response, checking if each claim was found in the retrieved context and using the formula:

$$Faithfulness = \frac{Number\ of\ claims\ in\ the\ response\ supported\ by\ the\ retrieved\ context}{Total\ number\ of\ claims\ in\ the\ response}$$

- **Answer Relevancy**: measured how relevant the response was, compared to the question input by the user: the higher the score, the better the alignment with the user input. This metric was calculated by generating a set of three artificial questions based on the response, computing the cosine similarity between the embedding of the user input and the embedding of each synthetic question, and, finally, getting the average of these three similarity scores.

$$Answer\ Relevancy = \frac{1}{N}\sum_{i=1}^{N} \cos\left(E_{gi}, E_o\right)$$

- **Noise Sensitivity:** evaluated how many errors were made by providing incorrect responses or using irrelevant chunks. The score goes from 0 to 1, and a lower value equals less noise. This metric was calculated based on how much of the generated response was supported by the relevant retrieved context and, ideally, the entire answer should be grounded in that context.

$$Noise\ Sensitivity = \frac{Total\ number\ of\ incorrect\ claims\ in\ the\ response}{Total\ number\ of\ claims\ in\ the\ response}$$

Using these metrics, it was possible to evaluate the RAG's different components, such as the retriever, through the faithfulness and noise sensitivity metrics, and the generator through the answer relevancy metric.

4.6.2 Human Evaluation

While automatic metrics provide valuable insights into the performance of GenAI, human evaluation remains a critical component of the assessment process, especially in corporate contexts. Automatic evaluations can approximate quality, but only real users can assess whether the system is genuinely helpful, trustworthy, and easy to use in practical scenarios. For this reason, a human evaluation phase was conducted to complement the model-based evaluation and capture user feedback on the chatbot's real-world effectiveness in answering questions regarding the ESOP.

The human evaluation was carried out using a structured user feedback survey, targeting employees who interacted with the chatbot. The survey aimed to collect both quantitative insights about the system's performance, usefulness, clarity, and trustworthiness. Participants were asked to evaluate various aspects of the chatbot after interacting with it, and their responses were anonymised to ensure honest and unbiased feedback.

To obtain comprehensive feedback, the survey was structured around five core dimensions (Table 3):

Table 3 - Metrics and example questions for human evaluation

Metric	Example Questions
Task Success	Did the chatbot resolve the user's query?
Response Quality	Relevance, faithfulness, clarity, tone
Efficiency	Was the response quick and concise?
User Experience	Was the experience useful and trustworthy?
Information Quality	Was the content complete, correct, and readable?

The questions were aimed to receive feedback in a Likert scale, with the options ranging from 1 (strongly disagree) to 5 (strongly agree):

1. *The chatbot answered my questions correct and efficiently.*
2. *The information provided by the chatbot appeared factually accurate.*
3. *The responses were clear, concise, and easy to understand.*
4. *I felt confident relying on the chatbot's answers.*
5. *The tone and language used by the chatbot was appropriate and professional.*
6. *The chatbot did not answer my questions.*
7. *The chatbot helped clarify my understanding of the ESOP.*
8. *Using the chatbot saved me time compared to asking HR directly or to reading the provided documentation.*
9. *I felt uncomfortable using this chatbot.*
10. *If this chatbot were made public, I would use it for my ESOP inquiries.*
11. *I would recommend this chatbot to my colleagues.*

A particularly high or low rating in any of the core dimensions can indicate the strengths or weaknesses of the system. For instance, a high average score on "*The chatbot did not answer my questions*" may suggest a need to improve the retrieval component of the system, such as by refining the embedding model or adjusting the similarity threshold.

5. RESULTS AND DISCUSSION

As mentioned previously, the performance of the chatbot was tested on two different retrievers, eight different chunk combinations, and on nine pairs of embedding and generative models. Each combination was run three times to make sure the results were not a matter of chance, given that answers can vary in different runs.

5.1. RAGAS RESULTS

The automatic evaluation was done using the RAGAS framework, based on ten example questions provided by the team responsible for the ESOP, allowing to get a comprehensive view on the impact of each of the parameters on the chatbot answers.

5.1.1 Chunking Results

Starting with semantic search, which used a *ChromaDB* vector space, the results for the chunking size were as presented in Table 4 and Figure 7. To maintain consistency across all evaluation metrics, where higher values indicate better performance, the Noise Sensitivity score was inverted. Originally, lower Noise Sensitivity values signified better results, in contrast to the other metrics, and, by inverting the scale, all metrics could be interpreted uniformly, enabling a clear comparison.

Table 4 - Results per Chunking Size and Overlap with Semantic Search using RAGAS

Chunking	RAGAS Metrics			Average
	Faithfulness	Answer Relevancy	Noise Sensitivity (Inverted)	
Chunk Size = 200 Chunk Overlap = 40 (20%)	0.66	0.36	0.61	0.54
Chunk Size = 200 Chunk Overlap = 100 (50%)	0.63	0.41	0.84	0.63
Chunk Size = 500 Chunk Overlap = 100 (20%)	0.88	0.63	0.68	0.73
Chunk Size = 500 Chunk Overlap = 250 (50%)	0.76	0.64	0.72	0.71

Chunking	RAGAS Metrics			Average
	Faithfulness	Answer Relevancy	Noise Sensitivity (Inverted)	
Chunk Size = 1000 Chunk Overlap = 200 (20%)	0.90	0.76	0.55	0.74
Chunk Size = 1000 Chunk Overlap = 500 (50%)	0.95	0.81	0.50	0.75
Chunk Size = 1500 **Chunk Overlap = 300 (20%)**	**0.91**	**0.73**	**0.63**	**0.76**
Chunk Size = 1500 **Chunk Overlap = 750 (50%)**	**0.94**	**0.81**	**0.53**	**0.76**

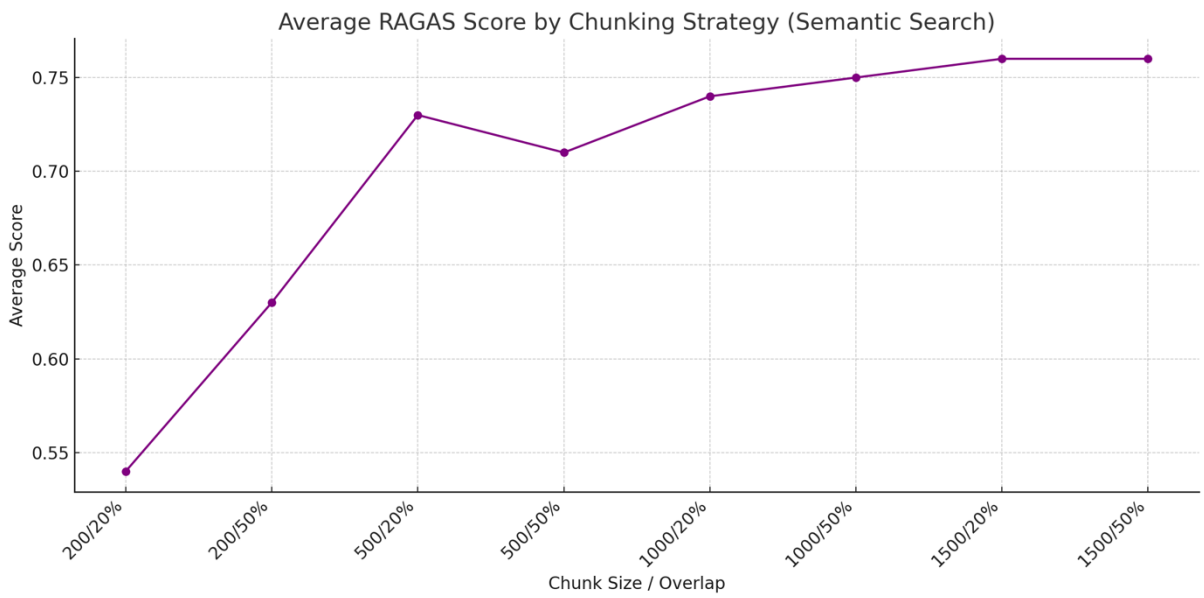

Figure 7 - Average RAGAS Score by Chunking Strategy (Semantic Search)

As it can be observed on the table and on the graph, there's a clear performance gain as the chunk size increases, particularly from 500 upward. The optimal chunking size appears to be 1500 with either 20% or 50% overlap, both reaching the highest average score of 0.76.

In Figure 8 the chunking results can be observed per metric, showing that faithfulness and answer relevancy improve with higher chunks, while noise sensitivity (inverted, so higher is better) peaks early (at 200/50%), and shows quite some fluctuation.

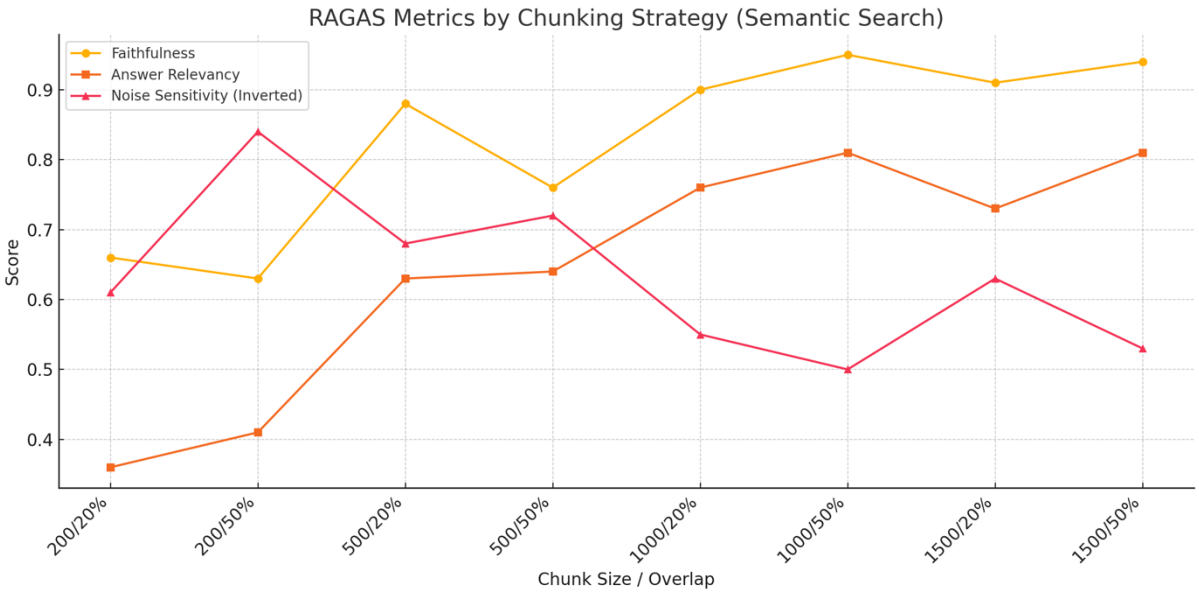

Figure 8 - RAGAS Metrics by Chunking Strategy (Semantic Search)

The best result was obtained with a chunk size of 1500, with very similar results with either a chunk overlap of 20% or 50%. However, while the first delivered less noise sensitivity, the second combination had a higher faithfulness and answer relevancy, which made it the best choice to move forward and test the different models.

The results with hybrid search are presented in Table 5 and Figure 9.

Table 5 - Results per Chunking Size and Overlap with Hybrid Search using RAGAS

Chunking	**RAGAS Metrics**			**Average**
	Faithfulness	**Answer Relevancy**	**Noise Sensitivity (Inverted)**	
Chunk Size = 200 Chunk Overlap = 40 (20%)	0.63	0.43	0.61	0.56

Chunking	RAGAS Metrics			Average
	Faithfulness	Answer Relevancy	Noise Sensitivity (Inverted)	
Chunk Size = 200 Chunk Overlap = 100 (50%)	0.72	0.44	0.59	0.58
Chunk Size = 500 Chunk Overlap = 100 (20%)	0.73	0.74	0.61	0.69
Chunk Size = 500 Chunk Overlap = 250 (50%)	0.63	0.57	0.7	0.63
Chunk Size = 1000 Chunk Overlap = 200 (20%)	0.61	0.57	0.75	0.64
Chunk Size = 1000 Chunk Overlap = 500 (50%)	0.8	0.73	0.63	0.72
Chunk Size = 1500 Chunk Overlap = 300 (20%)	0.84	0.72	0.65	0.74
Chunk Size = 1500 **Chunk Overlap = 750 (50%)**	**0.9**	**0.77**	**0.62**	**0.76**
Chunk Size = 200 Chunk Overlap = 40 (20%)	0.63	0.43	0.61	0.56

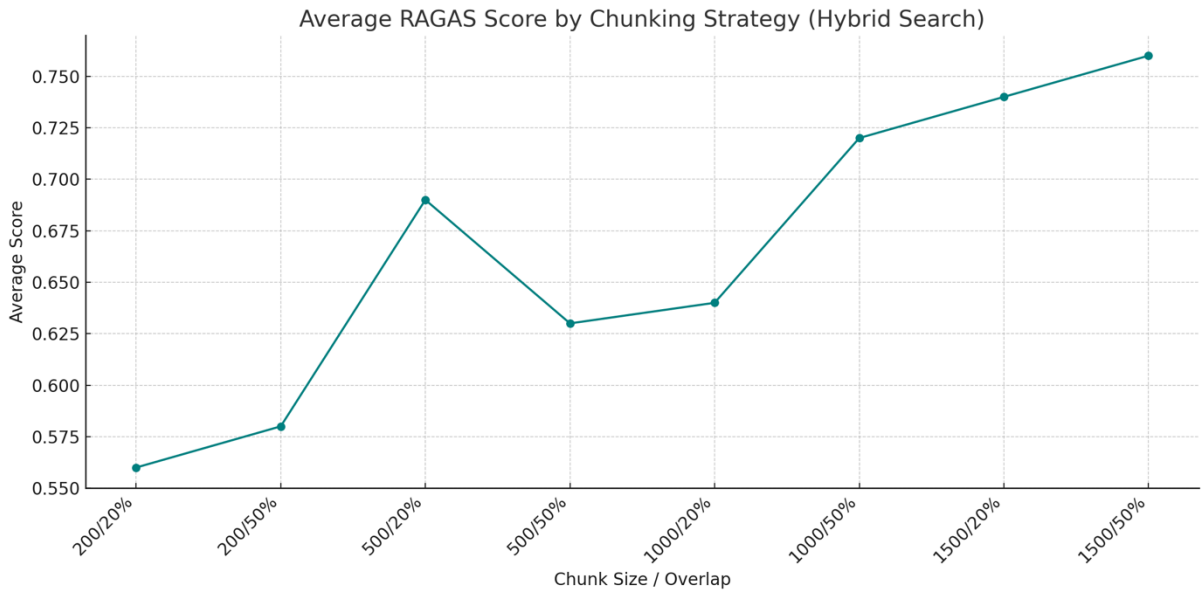

Figure 9 - Average RAGAS Score by Chunking Strategy (Hybrid Search)

Regarding chunking, the results did not show a significant change between semantic and hybrid search, with the best result being delivered, once again, by a chunk size of 1500, with an average of the different metrics of 0.76.

When it comes to the results per metric, once again faithfulness and answer relevancy consistently improve with larger chunk sizes, while noise sensitivity peaks with a chunk size of 1000 (Figure 10).

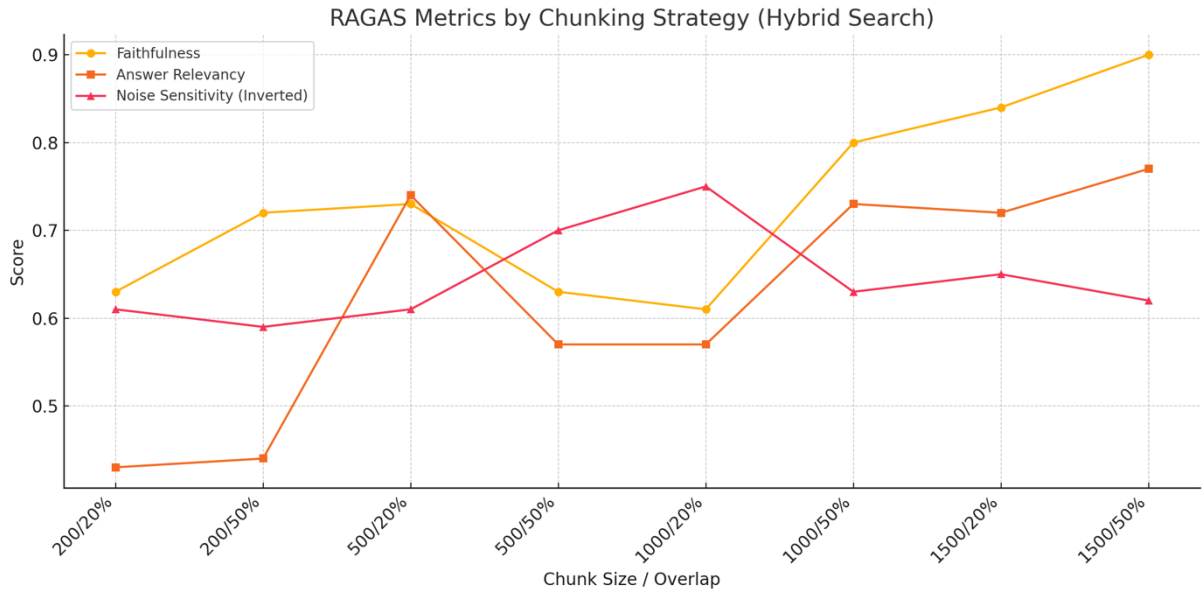

Figure 10 - RAGAS Metrics by Chunking Strategy (Hybrid Search)

5.1.2 Model Results

Given that the previous results did not show a significant difference between the two retrieval methods, the test continued with both, with the application of the 1500 chunking size and 50% overlap on the nine model combinations. The model results using semantic search can be found below in Table 6 and Figure 11.

Table 6 - Results per Embedding and Generative Model with Semantic Search using RAGAS

Models	RAGAS Metrics			Average
	Faithfulness	Answer Relevancy	Noise Sensitivity (Inverted)	
Embedding: textembedding-gecko@003 Generation: gemini-1.5-pro-002	0.88	0.59	0.53	0.67
Embedding: textembedding-gecko@003 Generation: gemini-1.5-flash-002	0.95	0.60	0.59	0.71
Embedding: textembedding-gecko@003 Generation: gemini-2.0-flash-001	0.82	0.67	0.58	0.69
Embedding: text-embedding-004 Generation: gemini-1.5-pro-002	0.90	0.55	0.62	0.69
Embedding: text-embedding-004 Generation: gemini-1.5-flash-002	0.9	0.65	0.63	0.73

Models	RAGAS Metrics			Average
	Faithfulness	Answer Relevancy	Noise Sensitivity (Inverted)	
Embedding: text-embedding-004 Generation: gemini-2.0-flash-001	0.91	0.67	0.66	0.75
Embedding: text-embedding-005 Generation: gemini-1.5-pro-002	0.95	0.71	0.59	0.75
Embedding: text-embedding-005 Generation: gemini-1.5-flash-002	0.92	0.64	0.61	0.72
Embedding: text-embedding-005 **Generation: gemini-2.0-flash-001**	**0.94**	**0.82**	**0.59**	**0.78**

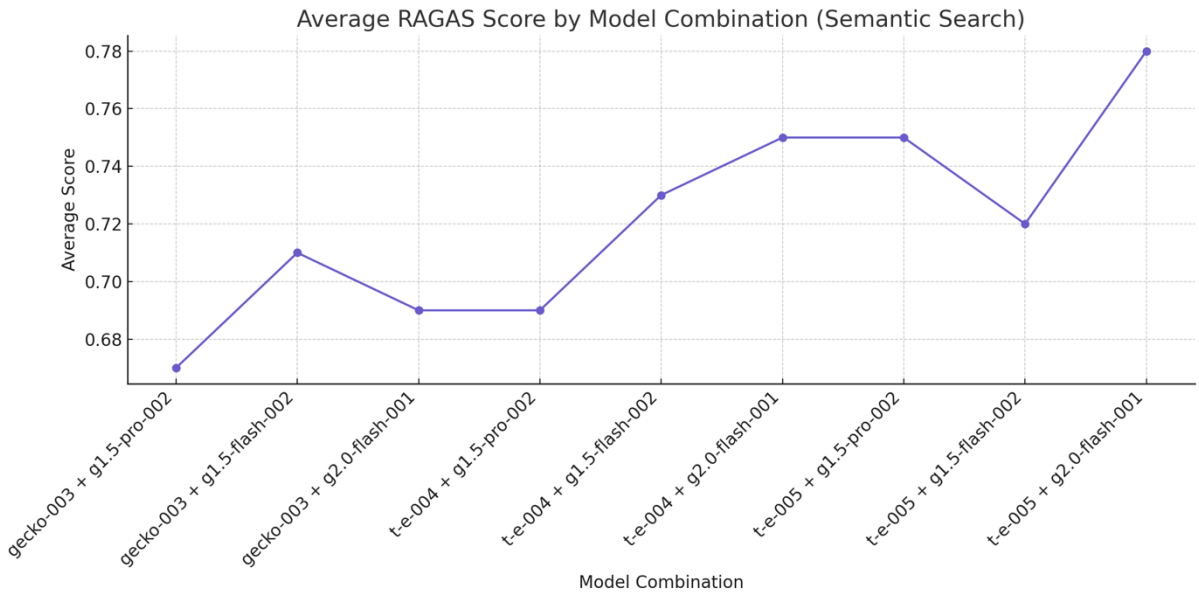

Figure 11 - Average RAGAS Score by Model Combination (Semantic Search)

When it comes to the combination of embedding and generative models with semantic search, the best result is *gemini-2.0-flash-001* (Google, 2025a) with *text-embedding-005*. While this combination gets the best score with answer relevancy and very high score in faithfulness, its noise sensitivity is still moderately high, meaning that, while the responses generated by this configuration were generally more accurate, relevant, and well-formulated, the system still included some unsupported or irrelevant information in its answers.

In Figure 12 the model results per metrics can be observed, showing that faithfulness generally increases with newer embeddings and generation models, answer relevancy highly peaks with the final combination (*text-embedding-005* and *gemini-2.0-flash-001*), while noise sensitivity remains generally stable across different combinations, with a mild improvement with *text-embedding-004* and *gemini-2.0-flash-001*.

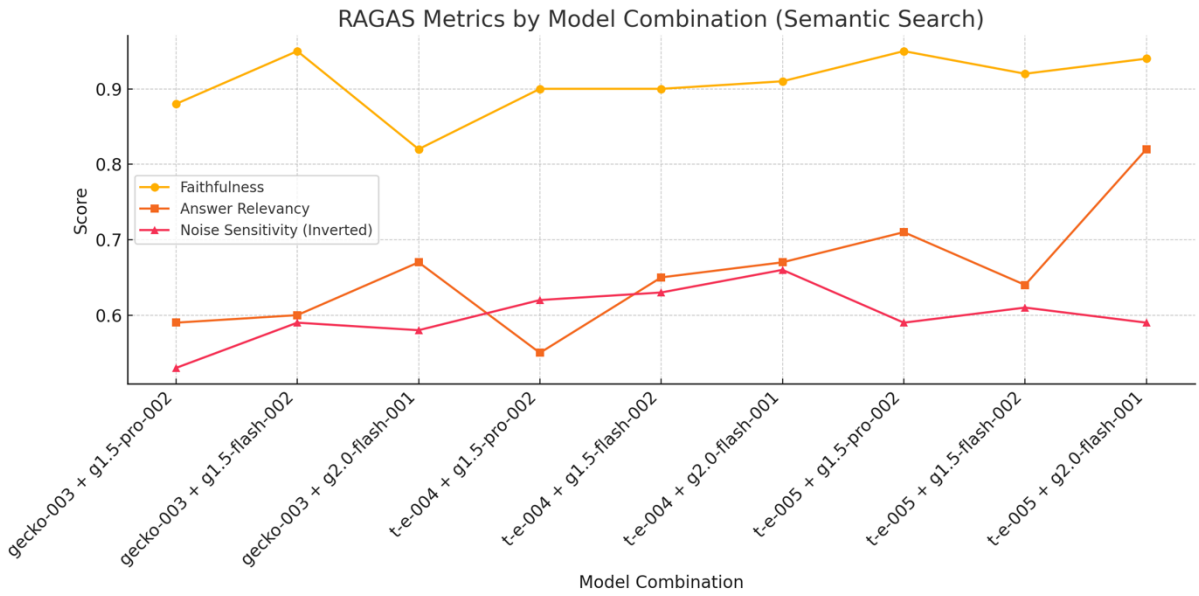

Figure 12 - RAGAS Metrics by Model Combination (Semantic Search)

The model comparison results with hybrid search can be found below in Table 7 and Figure 13.

Table 7 - Results per Embedding and Generative Model with Hybrid Search using RAGAS

Models	RAGAS Metrics			Average
	Faithfulness	Answer Relevancy	Noise Sensitivity (Inverted)	
Embedding: textembedding-gecko@003 Generation: gemini-1.5-pro-002	0.88	0.59	0.64	0.70
Embedding: textembedding-gecko@003 Generation: gemini-1.5-flash-002	0.95	0.64	0.50	0.70
Embedding: textembedding-gecko@003 Generation: gemini-2.0-flash-001	0.81	0.69	0.49	0.66

Models	RAGAS Metrics			Average
	Faithfulness	Answer Relevancy	Noise Sensitivity (Inverted)	
Embedding: text-embedding-004 Generation: gemini-1.5-pro-002	0.81	0.52	0.54	0.62
Embedding: text-embedding-004 Generation: gemini-1.5-flash-002	0.87	0.62	0.46	0.65
Embedding: text-embedding-004 Generation: gemini-2.0-flash-001	0.95	0.74	0.48	0.72
Embedding: text-embedding-005 Generation: gemini-1.5-pro-002	0.92	0.70	0.57	0.73
Embedding: text-embedding-005 Generation: gemini-1.5-flash-002	0.94	0.71	0.51	0.72
Embedding: text-embedding-005 **Generation: gemini-2.0-flash-001**	**0.99**	**0.8**	**0.61**	**0.80**

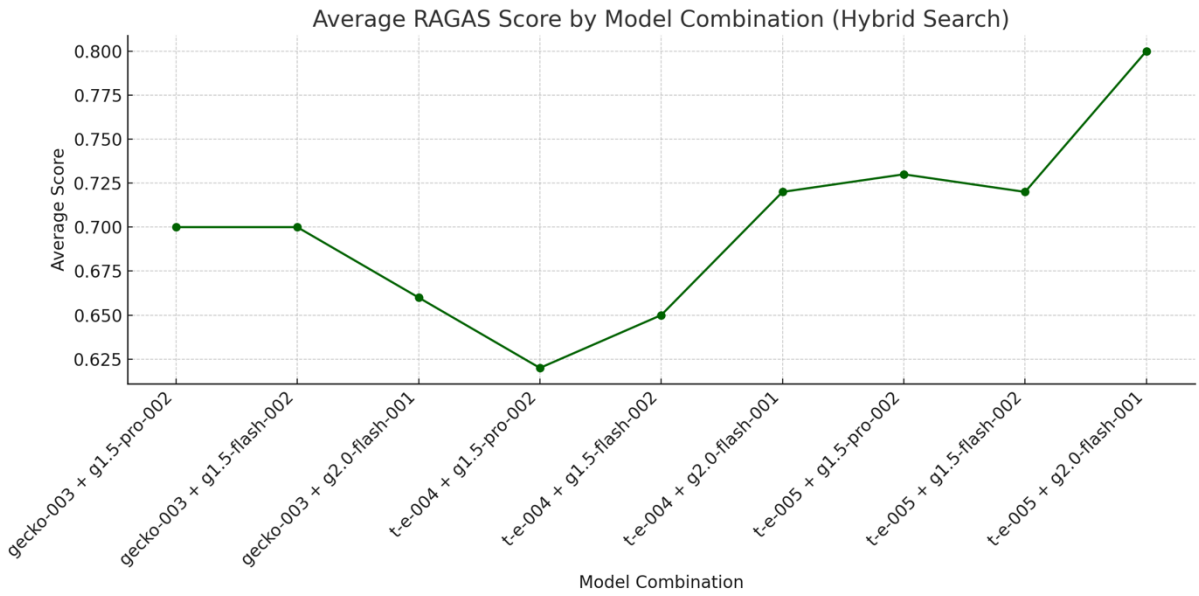

Figure 13 - Average RAGAS Score by Model Combination (Hybrid Search)

Slightly higher than the semantic search, the best result among all combinations was achieved using hybrid search, with the embedding model *text-embedding-005* and generative model *gemini-2.0-flash-001*. Although this combination had the highest overall average across the evaluation metrics, it still showed a relatively high noise sensitivity (0.39, not inverted). This could happen due to the same information being repeated for the same countries, given that the rules are mostly the same for all countries, but each country contains its own file.

In Figure 14, the model results per metric show that faithfulness peaks with *text-embedding-005* and *gemini-2.0-flash-001*, answer relevancy improves consistently with newer generation models, while noise sensitivity is more variable, and has generally low scores.

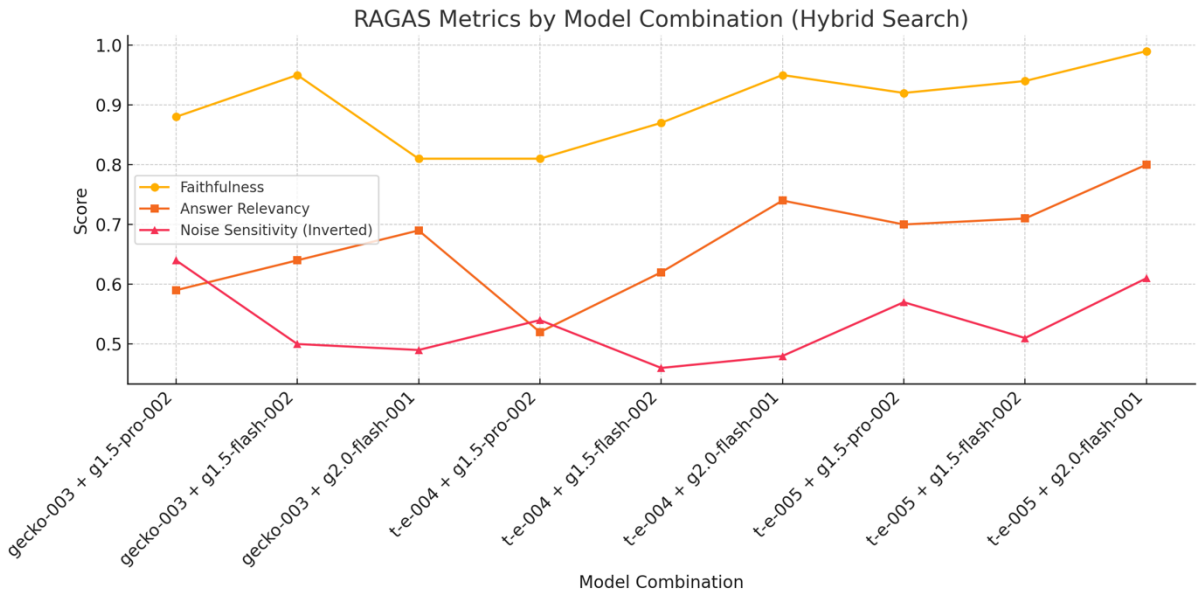

Figure 14 - RAGAS Metrics by Model Combination (Hybrid Search)

5.2. HUMAN EVALUATION RESULTS

Although the RAGAS framework offers a comprehensive and objective evaluation of the chatbot's performance, understanding user experience remains essential for assessing its usability and value. To complement the technical evaluation, a user study was conducted involving ten employees from the company. After interacting with the chatbot and testing it with various questions, participants were asked to respond to eleven questions about their experience.

Of the eleven survey questions, nine received highly positive feedback, with participants consistently selecting scores of 4 or 5 out of 5 (if the question was negative, 1 or 2), as can be observed in Table 8.

Table 8 - Survey Results

Question	Score (%)				
	1	2	3	4	5
The chatbot answered my questions correct and efficiently.			20	50	30
The information provided by the chatbot appeared factually accurate.				60	40
The responses were clear, concise, and easy to understand.				40	60

Question	Score (%)				
	1	2	3	4	5
I felt confident relying on the chatbot's answers.		10		20	70
The tone and language used by the chatbot was appropriate and professional.					100
The chatbot did not answer my questions.	80	20			
The chatbot helped clarify my understanding of the ESOP.				30	70
Using the chatbot saved me time compared to asking HR directly or to reading the provided documentation.					100
I felt uncomfortable using this chatbot.	100				
If this chatbot were made public, I would use it for my ESOP inquiries.				20	80
I would recommend this chatbot to my colleagues.				20	80

Notably, the statement "I felt uncomfortable using this chatbot" received only scores of 1, indicating that all users felt at ease during their interaction with the system. However, the question "The chatbot answered my questions correctly and efficiently" returned slightly more varied results, with 20% of the participants rating it 3 out of 5, 50% selecting 4, and 30% selecting 5. This variation may be attributed to the fact that the chatbot was deployed locally during testing, which made it quite slow and affected its response efficiency. Despite this limitation, 80% of the responses still rated the chatbot positively (4 or 5), suggesting that efficiency concerns, while present, are not a major issue. The statement "I felt confident relying on the chatbot's answers" also received a lower rating from 10% of the participants, which may reflect concerns related to model hallucination, response limitations, or an inherent lack of trust in GenAI by those particular users. Aside from this outlier, the remaining responses were positive, consisting exclusively of 4s and 5s.

Importantly, all 10 participants stated they would recommend the chatbot to a colleague, underlining both the perceived value of the tool and its potential for internal adoption within the organization.

5.3. EXAMPLE OF CHATBOT ANSWERS AND UI

As mentioned previously, a UI was developed for the Ask ESOP chatbot using *Streamlit*, which was used to support the users during testing. Figure 15 provides a brief preview of the implemented interface, which was intentionally designed with a simple and intuitive layout to support ease of use during testing and evaluation.

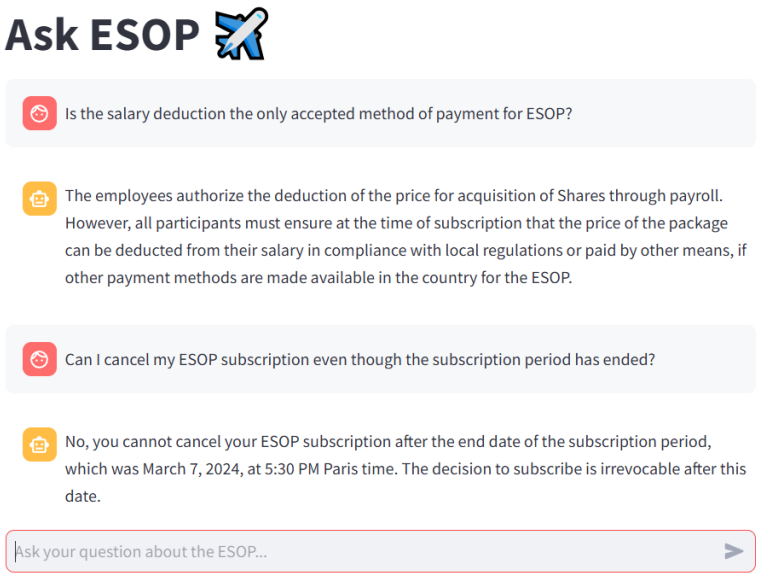

Figure 15 - Ask ESOP UI

Table 9 presents a list of provided questions and the corresponding responses generated by the model with the final configuration (hybrid search retriever with *text-embedding-005* and *gemini-2.0-flash-001).*

Table 9 - User questions and the chatbot's answers

Question number	User Question	Chatbot Response
1	Is the salary deduction the only accepted method of payment for ESOP?	*The employees authorize the deduction of the price for acquisition of Shares through payroll. However, all participants must ensure at the time of subscription that the price of the package can be deducted from their salary in compliance with local regulations or paid by other means, if other payment methods are made available in the country for the ESOP.*

Question number	User Question	Chatbot Response
2	Can I cancel my ESOP subscription even though the subscription period has ended?	*No, you cannot cancel your ESOP subscription after the end date of the subscription period, which was March 7, 2024, at 5:30 PM Paris time. The decision to subscribe is irrevocable after this date.*
3	Will I perceive dividends on my ESOP stock investment and how will they be taxed as a German tax resident?	*As a German tax resident participating in the Employee Share Ownership Plan (ESOP), you will be taxed on any dividends distributed by the company during the three-year lock-up period. These dividends will likely be subject to a withholding tax in the Netherlands and then taxed in Germany, your country of residence.*
4	When will my shares be allocated onto my Société Générale bank account?	*You should be able to check your share allocation on your account from April 20, 2024. Depending on your choice of financial institution, any newly acquired shares will appear on your account by the end of April 2024 at the latest.*
5	What difference is there between the PEG and the direct ownership in France?	*In France, ESOP 2024 offers two investment options: the PEG Formula and the Direct Shareholding Formula. If you choose the PEG Formula, you subscribe to shares through the FCPE ESOP Relais. The subscription price for PEG is the average of share closing prices on Euronext over 20 trading days before February 14, 2024, and your investment will be locked up for approximately 5 years, until May 31, 2029.* *If you choose the Direct Shareholding Formula, your investment is subject to the terms and conditions of ESOP 2024. The subscription price is equal to the closing price of the share on Euronext on February 14th, 2024. However, if the*

Question number	User Question	Chatbot Response
		closing price of share on March 13th, 2024 is lesser, then the subscription price retained to make your investment will be equal to the closing price of share on March 13th, 2024. Your shares are delivered in a share account with Société Générale and have a mandatory holding period of 3 years, until March 14, 2027.
6	Can I be eligible for the ESOP 2024 offer if I was hired on January 1st, 2024?	*No, to be eligible for the ESOP 2024 offer, you must be employed at least since December 31st, 2023 and still employed on the Date of Grant of shares on March 14, 2024.*
7	Can I subscribe to a package for which the price is higher than 3 months of my salary?	*No, you cannot. You should select a package that does not exceed three months of your maximum salary deductions. The firm is acting as a responsible company, and you should not invest more than your available earnings for the coming three months.*
8	When will I be able to sell my shares?	*You will not be able to sell or transfer the shares acquired under the 2024 ESOP for a period of three years from the date of registration in the accounts. Thus, you can sell the Shares after March 14, 2027 if you wish.*
9	Are there any taxes or benefits in kind in France which may add to the cost of the subscribed packages?	*Yes, in France, you make a taxable benefit corresponding to the difference between the value of all shares delivered to you (Subscription Shares and Matching Shares) and the subscription price that you paid. This benefit is considered employment income and subject to income tax at prevailing progressive rates currently up to 49.5% (including social security contributions) for 2024. Also, due to the fact that the shares are subject to a mandatory holding period of three years, a tax-free discount up to 13.5% of the fair market value of the shares should apply.*

Question number	User Question	Chatbot Response
10	Who can I contact in order to help me subscribe to ESOP?	*You can subscribe to the ESOP either via the internet on [website redacted] or, depending on your country of residence, by calling the Toll-Free Number. The France Toll-Free Number is [phone number redacted], Germany is [phone number redacted], Spain is [phone number redacted], and the UK is [phone number redacted]. If you are unable to access the web application, please call the multilingual call centre. The Other countries number is [phone number redacted].*

A first look at the chatbot's responses indicates that the majority are complete, accurate, and appropriate. Specifically, answers to questions 2, 3, 5, 6, 7, 8, 9, and 10 demonstrate a high level of precision, directly addressing the users' questions with concise and relevant information. This reflects a strong overall performance of the system in terms of both comprehension and retrieval accuracy.

However, responses to questions 1 and 4 exhibit some shortcomings related to contextual relevance. In the case of question 1, which explicitly inquired whether salary deduction is the only accepted payment method for the ESOP, the chatbot refrained from providing a direct yes or no answer. Instead, it offered a more general explanation, referencing that the method is possible and advising users to consult local regulations. While factually correct, this indirect approach does not align with the intent of the question. Similarly, for question 4, which asked for the specific receiving date of shares when using the Société Générale platform, the chatbot responded with a general statement about potential variability depending on the financial institution, again failing to directly address the user's request.

On the other hand, the chatbot's performance on question 5 shows its potential for delivering high-quality responses. The user asked for the difference between PEG and direct ownership in France, and the chatbot produced a well-structured, two-paragraph explanation, clearly defining the characteristics of each model. This response quality was predominant across the 10 examples, which shows the effectiveness of the chatbot in providing useful, accurate and comprehensive questions, with a few limitations in some cases.

6. CONCLUSIONS

The main goal of this thesis was to address a gap in communication of an ESOP program within a large aerospace organization. Despite the power of ESOPs for employee engagement, their complex and technical nature can lead to low participation and misunderstandings, drifting employees away from partaking. Through LLMs and a RAG system, this work demonstrates how a domain-specific chatbot can improve information accessibility, reduce dependence on HR teams, and improve the overall employee experience with an ESOP.

The development process followed the methodology provided by Jeong (2023), starting with the analysis and cleaning of over 100 documents across multiple languages, ultimately being narrowed down to 62. Various chunking sizes were tested to optimize the information retrieval, with a chunk size of 1500 and a 50% overlap returning the best results. After, embedding and generative models were tested using RAGAS, a specialized evaluation framework for RAG pipelines. The final model configuration, using the *text-embedding-005* embedding model with *gemini-2.0-flash-001* generative model, and a hybrid search retrieval, achieved the highest average performance across the faithfulness, answer relevancy and noise sensitivity metrics.

However, technical evaluation alone cannot capture the full picture of user experience. A study involving ten employees revealed that the chatbot was perceived as helpful, trustworthy, and easy to use, proving its potential as a communication tool in the organisation. The chatbot was able to provide accurate, concise and user-friendly responses to complex and legal questions. Traditionally, these types of questions would require a subject matter expert to answer, but the chatbot was able to handle it.

When compared with the existing literature explored, this project builds upon many of the same foundational techniques, such as chunking text to facilitate retrieving, hybrid and semantic search, and the use of RAG to factually ground LLMs. However, while many existing chatbots focus on broader HR or IT support topics, such as FAQs (e.g., Afzal et al., 2024), this thesis addresses a more complex topic, the ESOP, introducing unique challenges around factual accuracy and context interpretation. Moreover, while many related works relied on OpenAI's *GPT-4* or Meta's *Llama* models, this project contributes with a different perspective by exploring Google's *Gemini* models. By conducting a systematic evaluation with these models, it adds to the growing body of knowledge on how different LLMs behave in enterprise chatbot development. Finally, in addition to the automated RAGAS assessment used in some prior work, this project incorporated a human evaluation phase, which helped assess its real-world applicability.

In conclusion, this thesis demonstrates the potential of combining GenAI with domain-specific knowledge to improve employee understanding and engagement in complex programs like ESOPs. Through the described methodology, it's possible to create chatbots that not only reduce the load on HR teams but also deliver, on demand, reliable and useful information

about challenging topics. While there are areas for improvement, this work lays a solid foundation for future enhancements, paving the way for more accessible internal communication tools in corporative organisations.

7. LIMITATIONS AND FUTURE WORKS

Despite its overall success, the chatbot still demonstrated areas for improvement. One recurring issue was the occasional inclusion of irrelevant or unsupported content, as reflected in its noise sensitivity score, which may be attributed to overlapping information across country-specific documents, causing confusion to the retrieval system. Additionally, the platform constraints imposed by the organization restricted model selection, preventing comparisons with non-Google models like *GPT-4* or *LLaMA-3*, impeding any understanding of how other models might have impacted the chatbot's performance.

Future developments could include multilingual deployment, enabling support for employees across different regions and aligning responses with localized ESOP regulations. Additionally, the introduction of deeper personalization, where user-specific attributes, such as country, role, or employment status, are dynamically integrated into the retrieval and generation process, could further increase the relevance and precision of the chatbot's answers.

Another promising direction would be to extend the chatbot into a MAS, instead of relying on a single general-purpose agent, multiple specialized agents could be each responsible for handling different, and personalized tasks, such as eligibility verification, tax-related inquiries, or subscription process guidance. This architecture would not only improve the chatbot's ability to manage more complex, multi-turn interactions but also align with growing research on agent-based approaches in GenAI. It would allow to automatize the eligibility verification and the subscription processes, allowing users to complete their ESOP subscription directly through the chatbot without needing to access an external platform.

Addressing these limitations would improve the usefulness and generalizability of the chatbot, strengthening its role as a reliable and time-saving solution.

BIBLIOGRAPHICAL REFERENCES

Adam, M., Wessel, M., & Benlian, A. (2021). AI-based chatbots in customer service and their effects on user compliance. *Electronic Markets*, *31*(2), 427–445. https://doi.org/10.1007/s12525-020-00414-7

Afzal, A., Kowsik, A., Fani, R., & Matthes, F. (2024). *Towards optimizing and evaluating a retrieval augmented QA chatbot using LLMs with human in the loop* (No. 2407.05925). arXiv:2407.05925. https://doi.org/10.48550/arXiv.2407.05925

Akkiraju, R., Xu, A., Bora, D., Yu, T., An, L., Seth, V., Shukla, A., Gundecha, P., Mehta, H., Jha, A., Raj, P., Balasubramanian, A., Maram, M., Muthusamy, G., Annepally, S. R., Knowles, S., Du, M., Burnett, N., Javiya, S., … Boitano, J. (2024). *FACTS about building Retrieval Augmented Generation-based chatbots* (No. arXiv:2407.07858). arXiv. https://doi.org/10.48550/arXiv.2407.07858

Al-Amin, M., Ali, M. S., Salam, A., Khan, A., Ali, A., Ullah, A., Alam, M. N., & Chowdhury, S. K. (2024). *History of generative artificial Intelligence (AI) chatbots: Past, present, and future development* (No. arXiv:2402.05122). arXiv. http://arxiv.org/abs/2402.05122

Almeida, F., & Xexéo, G. (2023). *Word embeddings: A survey* (No. arXiv:1901.09069). arXiv. https://doi.org/10.48550/arXiv.1901.09069

Bahdanau, D., Cho, K., & Bengio, Y. (2016). *Neural machine translation by jointly learning to align and translate*. arXiv. https://doi.org/10.48550/arXiv.1409.0473

Bălan, C. (2023). Chatbots and voice assistants: Digital transformers of the company–customer interface—A systematic review of the business research literature. *Journal of Theoretical and Applied Electronic Commerce Research*, *18*(2), 995–1019. https://doi.org/10.3390/jtaer18020051

Bassett, C. (2019). The computational therapeutic: Exploring Weizenbaum's ELIZA as a history of the present. *AI & SOCIETY*, *34*(4), 803–812. https://doi.org/10.1007/s00146-018-0825-9

Blasi, J., & Kruse, D. (2024b, April 8). Employee ownership and ESOPs: What we know from recent research. *Aspen Institute*. https://www.aspeninstitute.org/publications/employee-ownership-and-esops-what-we-know-from-recent-research-3/

Blasi, J., Kruse, D., & Freeman, R. (2017). Having a stake: Evidence and implications for broad-based employee stock ownership and profit sharing. *Third Way*. https://www.thirdway.org/report/having-a-stake-evidence-and-implications-for-broad-based-employee-stock-ownership-and-profit-sharing

Brandtzaeg, P. B., & Følstad, A. (2017). Why people use chatbots. In I. Kompatsiaris, J. Cave, A. Satsiou, G. Carle, A. Passani, E. Kontopoulos, S. Diplaris, & D. McMillan (Eds), *Internet Science* (pp. 377–392). Springer International Publishing. https://doi.org/10.1007/978-3-319-70284-1_30

ChatGPT. (2025). https://chatgpt.com

Chaudhary, S., & Gupta, P. (2023). AIML in the modern era: Advancements, innovations, and future prospects. *International Journal of Research and Analytical Reviews (IJRAR)*, *10*(4). https://www.ijrar.org/papers/IJRAR23D1420.pdf

Deng, J., & Lin, Y. (2023). The benefits and challenges of ChatGPT: An overview. *Frontiers in Computing and Intelligent Systems*, *2*(2), 81–83. https://doi.org/10.54097/fcis.v2i2.4465

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). *An image is*

worth 16x16 words: Transformers for image recognition at scale (No. arXiv:2010.11929). arXiv. https://doi.org/10.48550/arXiv.2010.11929

Elasticsearch: The official distributed search & analytics engine. (2025). Elastic. https://www.elastic.co/elasticsearch

Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2025). *Ragas: Automated evaluation of retrieval augmented generation* (No. arXiv:2309.15217). arXiv. https://doi.org/10.48550/arXiv.2309.15217

Freeman, S. F. (2007). *Effects of ESOP adoption and employee ownership: Thirty years of research and experience* (Organizational Dynamics Working Paper Series). University of Pennsylvania. https://www.researchgate.net/publication/45598844_Effects_of_ESOP_Adoption_and_Employee_Ownership_Thirty_years_of_Research_and_Experience

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2024). *Retrieval-Augmented Generation for Large Language Models: A survey* (No. arXiv:2312.10997). arXiv. https://doi.org/10.48550/arXiv.2312.10997

Gemini. (2025a). *Embeddings | Gemini API*. Google AI for Developers. https://ai.google.dev/gemini-api/docs/embeddings

Gemini. (2025b). *Textembedding-gecko@003—One API 200+ AI Models | AI/ML API*. https://aimlapi.com/models/textembedding-gecko-003-api

Gemini models | Gemini API. (2025). Google AI for Developers. https://ai.google.dev/gemini-api/docs/models

Gkinko, L., & Elbanna, A. (2023). The appropriation of conversational AI in the workplace: A taxonomy of AI chatbot users. *International Journal of Information Management*, *69*, 102568. https://doi.org/10.1016/j.ijinfomgt.2022.102568

Google. (2025a). *Gemini 2.0 Flash | Generative AI on Vertex AI*. Google Cloud. https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-0-flash

Google. (2025b). *Vertex AI platform*. Google Cloud. https://cloud.google.com/vertex-ai

Google-bert/bert-base-uncased · Hugging Face. (2024, March 11). https://huggingface.co/google-bert/bert-base-uncased

Guo, M., Ainslie, J., Uthus, D., Ontanon, S., Ni, J., Sung, Y.-H., & Yang, Y. (2022). *LongT5: Efficient Text-To-Text Transformer for Long Sequences* (No. arXiv:2112.07916). arXiv. https://doi.org/10.48550/arXiv.2112.07916

Guzek, J., & Whillans, A. (2024). Overcoming Barriers to Employee Ownership: Insights From Small and Medium-Sized Businesses. *Compensation & Benefits Review*, *57*(1), 64–81. https://doi.org/10.1177/08863687241272556

Han, Y., Liu, C., & Wang, P. (2023). *A comprehensive survey on vector database: Storage and retrieval technique, challenge* (No. arXiv:2310.11703). arXiv. https://doi.org/10.48550/arXiv.2310.11703

Han, Z. (2023). The applications of chatbot. *Humanities and Social Sciences Education and Teaching (HSET)*, *57*. https://doi.org/10.54097/hset.v57i.10011.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Holmström, J., & Carroll, N. (2024). How organizations can innovate with generative AI. *Business Horizons*, S0007681324000247. https://doi.org/10.1016/j.bushor.2024.02.010

Hoyer, W. D., Kroschke, M., Schmitt, B., Kraume, K., & Shankar, V. (2020). Transforming the customer experience through new technologies. *Journal of Interactive Marketing*, *51*(1), 57–71. https://doi.org/10.1016/j.intmar.2020.04.001

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2024). A Survey on hallucination in Large Language Models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 3703155. https://doi.org/10.1145/3703155

Huang, X., & Marcus, P. M. (2021). *Chatbot: Design, architecture, and applications* [University of Pennsylvania, School of Engineering and Applied Science]. https://www.cis.upenn.edu/wp-content/uploads/2021/10/Xufei-Huang-thesis.pdf

Jeong, C. (2023). A Study on the implementation of Generative AI services using an enterprise data-based LLM application architecture. *Advances in Artificial Intelligence and Machine Learning*, *03*(04), 1588–1618. https://doi.org/10.54364/AAIML.2023.1191

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kaif, M., Sharma, S., & Rana, Dr. S. (2024). Gemini MultiPDF Chatbot: Multiple Document RAG Chatbot using Gemini Large Language Model. *International Journal for Research in Applied Science and Engineering Technology*, *12*(5), 24–30. https://doi.org/10.22214/ijraset.2024.61195

Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. (2020). *Dense passage retrieval for open-domain question answering* (No. arXiv:2004.04906). arXiv. https://doi.org/10.48550/arXiv.2004.04906

Khurana, D., Koli, A., Khatter, K., & Singh, S. (2023). Natural language processing: State of the art, current trends and challenges. *Multimedia Tools and Applications*, *82*(3), 3713–3744. https://doi.org/10.1007/s11042-022-13428-4

Klein, K. J. (1987). Employee stock ownership and employee attitudes: A test of three models. *Journal of Applied Psychology*, *72*(2), 319–332. https://doi.org/10.1037/0021-9010.72.2.319

LangChain. (2025a). *BM25 | 🦜🔗 LangChain*. https://python.langchain.com/docs/integrations/retrievers/bm25/

LangChain. (2025b). *Chroma | 🦜🔗 LangChain*. https://python.langchain.com/docs/integrations/vectorstores/chroma/

LangChain. (2025c). *Introduction | 🦜🔗 LangChain*. https://python.langchain.com/docs/introduction/

LangChain. (2025d). *LangChain*. ConversationBufferMemory — 🦜🔗 LangChain Documentation. https://python.langchain.com/api_reference/langchain/memory/langchain.memory.buffer.ConversationBufferMemory.html

LangChain. (2025e). *RetrievalQA — 🦜🔗 LangChain documentation*. https://python.langchain.com/api_reference/langchain/chains/langchain.chains.retrieval_qa.base.RetrievalQA.html

Lee, J., Dai, Z., Ren, X., Chen, B., Cer, D., Cole, J. R., Hui, K., Boratko, M., Kapadia, R., Ding, W., Luan, Y., Duddu, S. M. K., Abrego, G. H., Shi, W., Gupta, N., Kusupati, A., Jain, P., Jonnalagadda, S. R., Chang, M.-W., & Naim, I. (2024). *Gecko: Versatile text embeddings distilled from Large Language Models* (No. arXiv:2403.20327). arXiv. https://doi.org/10.48550/arXiv.2403.20327

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2021). *Retrieval-Augmented Generation for knowledge-intensive NLP tasks* (No. arXiv:2005.11401). arXiv. https://doi.org/10.48550/arXiv.2005.11401

Matthijs Douze. (2025). *Faiss*. GitHub. https://github.com/facebookresearch/faiss/wiki/Home

McCorkindale, T. (2024). *Generative AI in Organizations: Insights and Strategies from Communication Leaders | Institute for Public Relations*. Institute for Public Relations. https://instituteforpr.org/ipr-generative-ai-organizations-2024/

Meta AI. (2025). *Models and libraries*. Meta AI. https://ai.meta.com/resources/models-and-libraries

Model—OpenAI API. (2025). https://platform.openai.com

Models—OpenAI API. (2025). https://platform.openai.com

MTEB Leaderboard—A Hugging Face Space by mteb. (2025). https://huggingface.co/spaces/mteb/leaderboard

Mteb (Massive Text Embedding Benchmark). (2025, February 19). https://huggingface.co/mteb

Narimissa, E., & Raithel, D. (2024). *Exploring information retrieval landscapes: An investigation of a novel evaluation techniques and comparative document splitting methods* (No. arXiv:2409.08479). arXiv. https://doi.org/10.48550/arXiv.2409.08479

Nawaz, N., Arunachalam, H., Pathi, B. K., & Gajenderan, V. (2024). The adoption of artificial intelligence in human resources management practices. *International Journal of Information Management Data Insights*, *4*(1), 100208. https://doi.org/10.1016/j.jjimei.2023.100208

Pillare, T., Chaoudhari, M., & Hiwrale, V. (2023). A survey paper on chatbot. *International Research Journal of Modernization in Engineering Technology and Science*. https://www.irjmets.com/uploadedfiles/paper//issue_10_october_2023/45644/final/fin_irjmets1698841142.pdf

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by Generative Pre-Training*. OpenAI. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

RAGAS. (2025). *Metrics—Ragas*. https://docs.ragas.io/en/stable/concepts/metrics/

Salminen, A., & Tompa, F. Wm. (2001). Requirements for XML document database systems. *Proceedings of the 2001 ACM Symposium on Document Engineering*, 85–94. https://doi.org/10.1145/502187.502201

Streamlit. (2025). *Streamlit Docs*. Streamlit Documentation. https://docs.streamlit.io/

Sutskever, I., Vinyals, O., & Le, Q. V. (2014a). *Sequence to sequence learning with neural networks* (No. arXiv:1409.3215). arXiv. https://doi.org/10.48550/arXiv.1409.3215

Turing, A. M. (1950). I.—Computing machinery and intelligence. *Mind*, *LIX*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention is all you need* (No. arXiv:1706.03762). arXiv. https://doi.org/10.48550/arXiv.1706.03762

Wang, X., Lin, X., & Shao, B. (2022). How does artificial intelligence create business agility? Evidence from chatbots. *International Journal of Information Management*, *66*, 102535. https://doi.org/10.1016/j.ijinfomgt.2022.102535

Zerveas, G., Jayaraman, S., Patel, D., Bhamidipaty, A., & Eickhoff, C. (2020). *A Transformer-based framework for multivariate time series representation learning* (No. arXiv:2010.02803). arXiv. https://doi.org/10.48550/arXiv.2010.02803

APPENDIX A

File	Number of files	Languages
Main Steps – ESOP 2024	6	English, French (2x), German, Portuguese and Spanish
Factsheet with Infographics	4	English, Mandarin, French and German
European Information Note	6	Dutch, English, French, German, Italian and Spanish
European Information Note for Poland	2	English and Polish
European Information Note for Portugal	2	English and Portuguese
International Plan Rules	7	Brazilian Portuguese, English, French, German, Polish, Portuguese and Spanish
Privacy Policy	7	Brazilian Portuguese, English, French, German, Polish, Portuguese and Spanish
Australian Supplement	1	English
Employer Statement - Denmark	2	English and Danish
Remuneration supplement – France – PEG Formula	2	English and French
Country Supplement – France – Excluding PEG	2	English and French
Tax Note Employee - Australia	1	English
Tax Note Employee - Belgium	3	English, Dutch and French
Tax Note Employee - Brazil	2	English and Portuguese
Tax Note Employee - Canada	2	English and French
Tax Note Employee - Chile	2	English and Spanish

Tax Note Employee – Colombia	2	English and Spanish
Tax Note Employee - Denmark	1	English
Tax Note Employee – Finland	1	English
Tax Note Employee - Germany	2	English and German
Tax Note Employee – Hong Kong	1	English
Tax Note Employee – Hungary	1	English
Tax Note Employee - India	1	English
Tax Note Employee - Indonesia	1	English
Tax Note Employee - Ireland	1	English
Tax Note Employee - Italy	2	English and Italian
Tax Note Employee - Japan	1	
Tax Note Employee - Malaysia	1	English
Tax Note Employee - Mexico	2	English and Spanish
Tax Note Employee - Morocco	2	English and French
Tax Note Employee - Netherlands	2	Dutch and English
Tax Note Employee – New Zealand	1	English
Tax Note Employee – North Macedonia	1	English
Tax Note Employee – Norway	1	English
Tax Note Employee - Philippines	1	English
Tax Note Employee - Poland	2	English and Polish

Tax Note Employee - Portugal	2	English and Portuguese
Tax Note Employee - Qatar	1	English
Tax Note Employee – Romania	1	English
Tax Note Employee – Saudi Arabia	1	English
Tax Note Employee - Singapore	1	English
Tax Note Employee - Slovakia	1	English
Tax Note Employee – South Africa	1	English
Tax Note Employee – South Korea	1	English
Tax Note Employee – Spain	2	English and Spanish
Tax Note Employee - Sweden	1	English
Tax Note Employee - Switzerland	3	English, French and German
Tax Note Employee - Taiwan	1	English
Tax Note Employee - Thailand	1	English
Tax Note Employee - Tunisie	1	French
Tax Note Employee - Turkey	1	English
Tax Note Employee – UAE	1	English
Tax Note Employee - UK	1	English
Tax Note Employee – USA	1	English
Tax Note Employee – Uruguay	1	English
How to manage my shares after subscription period	1	English

Bank contact details 2024	1	English
ESOP 2024 QA	1	English
German Employees	2	English and German
Process for eligible minor employee	1	English
2024 ESOP Brochure	1	English

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa