MSc. Business Data Science

Aalborg University

Master thesis

Leveraging LLMs to generate Business Plans

Authors: Huba Kalny,
Imola Szilagyi
Supervisor: Hamid Bekamiri

Number of characters (with spaces): 132.559

Number of physical pages: 74

Date: 02.06.2025.

Abstract

In today's competitive business environment, well-structured business plans are essential for startups to be successful and secure funding, yet many small businesses lack the resources to create them effectively. To provide a convenient solution for the stated problem, this research investigates how Large Language Models (LLMs) can be utilized to generate business plans and examines the benefits and limitations compared to traditional planning methods. A multi-agent system was developed using the CrewAI framework, employing six specialized agents for different business plan sections, with Google's Gemini 2.5 Pro for content generation and Meta's Llama 3.3 70B for initial evaluation. The system was tested using both expert-provided fictional company data and synthetic data generated from existing business plans using GPT 4.1. Evaluation employed both human expert judgment and LLM-as-a-Judge techniques across five metrics including relevance, completeness, correctness, consistency, and clarity. Results demonstrate that LLMs can generate structurally coherent business plans with relevant content across all sections, offering significant advantages in speed, cost-effectiveness, and accessibility for startups. However, the generated plans can be utilized as drafts only, since limitations include hallucination tendencies, inconsistencies between sections, and lack of deep business insights that human expertise provides. The research concludes that AI-generated business plans serve best as starting points and still requires human review and refinement, representing a complementary tool rather than a complete replacement for traditional business planning methods.

Acknowledgements

We would like to express our gratitude to our supervisor, Hamid Bekamiri, for their continuous guidance, feedback, and support throughout the entire process. His expertise was determining in shaping the direction and quality of our work.

We also extend our thanks to Peter Thomsen and Brian Balslev Andersen for their great ideas, the time they invested in providing generous assistance in the evaluation process and for providing access to the data that made this research possible. Their knowledge in business planning provided us with valuable contributions that strengthened the relevance and depth of our study.

Contents

1. Introduction

A well-written business plan could be a fundamental component for the success of startups and small-scale companies. As the importance of a business plan has only increased lately due to the business world being more and more competitive, having newer and newer technological advancements, and the increasing need for structured and well-informed decision-making. In this environment, having a solid business plan serves both strategic and informational purposes. Such plan not only guides entrepreneurs through the initial phases of their ventures but could also be helpful in securing funding (*Abdullah, 2020*).

Since these factors in the business world are not to be ignored, it is crucial for entrepreneurs and business owners to create a plan and thus, gain a competitive advantage. Still, numerous research shows that startups and small businesses do not use business planning in their early stages, because they do not feel the need for it, or they simply lack resources to do it effectively. However, this lack of adequate planning is one of the main reasons for the failure of small businesses (*Scarborough and Cornwall, 2018*).

This highlights the importance of having a business plan, which is like an entrepreneur's roadmap on the beginning of the way into building a successful business. It describes the directions to take towards the goals. Those direction are stated in the business plan to be followed for the company over multiple years (*OpenStax, 2019*).

Moreover, having a good business plan proves that the entrepreneur took the time, did the necessary research, and studied the business opportunities. Business plans are comprehensive, as they summarize a company's vision, mission, financial projections, market strategies, and operational frameworks. They are important for securing external funding, as investors and financial institutions require a clear, data-backed strategy before committing capital. Without a well-structured plan, startups often struggle to attract investors, as a lack of clear goals and financial forecasts increases perceived risk (*Scarborough and Cornwall, 2018*).

As previously mentioned, securing external funding is one of the primary functions of a business plan. The reason for that is investors, and banks often require a well-developed and clearly structured business plan before considering any form of commitment. As a strong business plan demonstrates to potential investors that the company has a clear strategy, understands its market, and has evaluated potential risks. Without a convincing plan, businesses struggle to attract funding, making it difficult to launch or expand operations. On the other hand, a well-written plan significantly increases the likelihood of attracting the necessary capital to launch or grow the venture (*McKeever, 2016*).

The other key function of a business plan is that it provides tools to the business owners to translate their ideas and goals into actionable steps and allows them to turn their vision into a well-functioning company. (*Scarborough and Cornwall, 2018*).

Having a business plan is a great advantage in the business world from many aspects, however creating a solid one takes time, effort and determination from the business owners. Despite this, research shows that the benefits outweigh the costs in the long run. In today's digital age, where businesses operate in global markets, a business plan also plays a vital role in demonstrating credibility to external stakeholders. Whether dealing with suppliers, partners, or government agencies, a well-prepared business plan establishes legitimacy and professionalism *(Parsons, 2024).*

While traditional business plans have primarily relied on manual research, intuition, and industry experience, the fast advancement of data-driven decision-making has significantly transformed how businesses operate. In an era of rapid technological development, businesses are starting to use machine learning and artificial intelligence in their operations to enhance decision-making. Traditional business strategies are starting to get transformed by data-driven solutions, where AI-generated insights enable more accurate and much faster results. This has resulted in the development of AI-driven business tools that enable decision-makers to optimize operational performance, enhance customer experience, and increase efficiency (*Mahabub et. al., 2025*). These data-driven solutions could increase the success rate of startups, if used strategically, as these solutions are more cost-effective alternatives than traditional methods. This idea of using data, and AI-driven business planning is the main motivation behind our thesis. Our goal is to merge traditional business planning methods with AI and machine learning to create an application that simplifies the early stages of company development. By combining business perspective with technological advancements, we aim to provide startups and small businesses with a clearer, more effective path to success.

This research employs Large Language Models – LLMs - to generate business plans for companies, indicating the need to provide detailed explanations of the advancements in machine learning leading to the development of these models, which will be detailed in the literature review section. LLMs such as OpenAI's GPT, Google's Gemini, and Meta's Llama models enhance the technological transformation of company operations by automating complex tasks like drafting business plans, creating a tailor-made document for the firms.

These models help create customized business plans tailored to each company's specific needs and goals; however, to get accurate results, the company must clearly explain their requirements.

This idea of generating plans for the specific companies could greatly boost the early stages of development of smaller scale companies, because it is inexpensive, takes much less time and requires minimal human intervention. Naturally it is advised that a human decision maker, such as a company owner, or executive reviews and if necessary, modifies these generated business plans, however the companies, who utilise this solution, could save time, money and resources on the overall process of writing a well-constructed business plan.
It is important to emphasize further that human judgment remains essential when reviewing the generated business plans. While the outputs from language models may not fully match what a company is looking for, they can serve as valuable drafts or starting points. These prototypes can then be refined and developed further, saving time and effort compared to creating a business plan from scratch. The concept of generating these plans is aimed at startups and small businesses, which might not prioritize creating a business plan due to the significant time and effort involved or additional funding to make one properly. With the help of generative AI and the knowledge base of these models, the application has the potential to offer game changing value to companies by shortening and making the business planning process straightforward.

1.1 Problem formulation and research questions

We will present our two research questions here, which we would attempt to provide an answer for in our research and development of our application.

A) How can Large Language Models (LLMs) be used to generate effective business plans?

With the rise of artificial intelligence, LLMs have become useful tools for generating content and analysing information. In this case, LLMs can help us create business plans through an automated process by using data derived from company owners. This could save time and effort for businesses, especially startups and small companies. However, it is important to understand how well LLMs can generate business plans, what their strengths and weaknesses are, and how they can be improved with human input. Therefore, our next research question is the following:

> *B) What are the benefits and limitations of AI-generated business plans compared to traditional business planning methods?*

Startups and small businesses often struggle with creating the business plan or their enterprise in the traditional way because it requires time, expertise, and resources. On the other hand, AI-generated business plans offer a promising alternative by automating the process, reducing costs, and improving efficiency. However, AI-driven solutions may also have limitations compared to conventional methods, such as a lack of human intuition, deep understanding of the enterprise, or the ability to properly capture a company's unique vision. Therefore, this research aims to identify the benefits and limitations of AI-generated business plans compared to the traditional business planning methods.

2. Literature review

2.1 The Business Plan

In the book The Entrepreneur's Manual, Richard M. White, Jr. states that business plans are essential (it is like a road map) for creating successful business: "You identify your origin, select a destination, and plot the shortest distance between the two points."
According to McKeever (2016) "A business plan is a written statement that describes and analyzes your business and gives detailed projections about its future."

To be more specific, the business plan describes in detail the business vision, which includes its long-term aspirations and what it aims to achieve in the future. It also outlines the mission, which defines its core purpose and the value it provides to customers, as well as the goals, which specify the objectives the business aims to accomplish within a defined timeframe. Additionally, it covers the necessary financial, material, and human resources, including funding, equipment, and workforce required to operate and grow the business. The plan also details the development strategies, outlining the planned actions and approaches to achieve growth and success in the market (*Abdullah, 2020*).

To explain it further, the business plan is a detailed guide that builds on 3 main things: idea assessment, feasibility analysis, and the business model. While these early steps help to test a business idea, the business plan explains how to put it into action. It provides a step-by-step plan for running and growing the business. Together with a strong business model, it helps turn an idea into a successful and sustainable company *(Scarborough & Cornwall, 2018).*

Although often used interchangeably, the concepts of a business model and a business plan refer to different aspects of a business. The business model serves as the foundation of the business plan explaining how a company makes money and delivers value to customers. It defines what the business offers, who its customers are, and how it generates revenue (*Osterwalder & Pigneur, 2010*), while a business plan is a broader strategic document that explains how the business will operate, grow, and succeed over time, typically including financial forecasts, market analysis, and operational plans (*McKeever, 2016*).

A business plan not only helps entrepreneurs organize their ideas but helps them test if their idea is realistic before launching the business. It allows them to understand how money will flow through the business, and the possible ways to use it. The process also helps the entrepreneur gain clarity and confidence, helping better decisions during the early stages of the start-up (*McKeever, 2016*).

Another main function of the business plan is to attract lenders and investors. It is essential for entrepreneurs and businesses to secure loans from banks and attract potential investors. "A business plan must prove to potential lenders and investors that a venture will be able to repay loans and produce an attractive rate of return by providing proof that an entrepreneur has evaluated the risk involved in the new venture realistically and has a strategy for addressing it" *(Scarborough & Cornwall, 2018).* Business plan is important in cases of completing mergers and acquisitions as well. With its help evaluating will be easier from both seller's and buyer's side. The business plan demonstrates to potential stakeholders that you have thoroughly considered the future of the business, including the strategies for its growth and the pathways through which it will be developed and expanded. Moreover, a well-written business plan attracts skilled workers too. The business plan convinces them to take the risk of joining the enterprise by demonstrating that it is positioned for long-term viability and growth, indicating stability and opportunities for development in the future for the business and its employees *(Abdullah, 2020).*

Regardless of the size of companies, those that engage in business planning outperform those that do not. Studies show that entrepreneurs who create a business plan early are 2.5 times more likely to start their business than those who don't. Without proper planning, many small businesses struggle and fail, revealing the importance of a solid business plan *(Scarborough & Cornwall, 2018).*

2.1.1 Components of the Business Plan

A business plan is made up of several key components, each playing a crucial role in defining a company's direction and potential for success. Like every business venture, every business plan is unique. While business plans can vary depending on industry and purpose, they typically include the following key components:

The Executive Summary is the first section of a business plan, providing an overview of the entire document. It clearly states the company's needs and objectives within one -page typically. It highlights what the product is, who it's for, and why it matters. It includes the mission (what the business aims to achieve in the future) along with clear objectives that outline what success looks like in the short term. It also covers the key factors that will contribute to the success of the business *(Georgetown University Law Center, 2020)*.

The Company Summary outlines the core details of the business, including who owns it and how it was founded, and where it operates. It tells about the basic start-up details such as initial funding, key resources, and early development efforts (*McKeever, 2016*).

The Products & Services section outlines what the business offers and how it meets customer needs. It explains what makes the product stand out from competitors, and the technology behind it. Finally, it tells in detail about how the product was designed and developed, including the process of turning the idea into a working solution *(Cambridge Judge Business School, 2020)*.

The Market Analysis section evaluates the industry, target market, and competitive field by identifying key trends that affect the business. It breaks the market into segments based on different factors and defines the specific group the product is targeting. Analyzing the competition is also an important part, evaluating what makes the business stand out. Lastly, it examines customer behaviour and buying patterns too *(Cambridge Judge Business School, 2020)*.

The Strategy and Implementation Summary details how the business is planning to succeed in the market. It contains competitive analysis to understand the position of the business and how it stands out. It explains the marketing and sales strategies, and forecasts sales, which also gives an idea of expected growth and revenue. Lastly, key milestones help track progress, from product launch to future development goals *(Georgetown University Law Center, 2020)*.

The Operations and Management section explains that the business has a clear structure for daily operations and appropriate leadership. It introduces the management team, outlines any skill or experience gaps that still need to be filled, and introduces a plan for hiring and staffing. It also describes how the business will operate in practice *(Georgetown University Law Center, 2020).*

The Financial Plan section provides a clear picture of the business's financial performance and future growth potential while providing insights into its overall viability. It starts with the key assumptions used in the projections, followed by a break-even analysis to show when the business is expected to cover its costs and become profitable. It contains forecasts for profit and loss, cash flow, and the balance sheet, helping to predict future income and expenses. Lastly, business ratios are used to indicate financial health and efficiency over time *(Georgetown University Law Center, 2020).*

The Appendices section includes supporting documents that strengthen your business plan and provide evidence for key claims. Depending on your business, this may include detailed financial documents (like sales forecasts, profit and loss statements, balance sheets, cash flow forecasts, break even analysis, and funding plans), legal agreements (contracts, licenses, permits), market research (surveys, reports), resumes of key team members, product images or prototypes, and any other materials relevant to your business (*McKeever, 2016*).

It is important to highlight, that for every business plan there is always a different audience, that one should never lose sight of during the process of creating it. It is needed to tailor the business plan in a way to meet the needs of target audiences *(Cambridge Judge Business School, 2020).*

2.1.2 Traditional Business Plan creation

Traditionally, business plans are created manually by entrepreneurs or consultants through an iterative process. They collect market data, develop financial models, and write narrative sections to produce the final plan. As Nakajima & Sekiguchi (2025) describe, "business planning is the process by which entrepreneurs gather and analyze information on business opportunities, assess key challenges, identify risks and strategies, forecast financial conditions, and document these elements in a formal plan." This meticulous process takes human expertise, time, effort and requires significant knowledge in the different areas of the business plan. It requires expertise in market research, finance, accounting, strategic management, marketing strategy and entrepreneurial leadership.

The traditional business planning process typically begins with extensive research and data collection. The representatives of the company gather information about the industry, market, competitors, and potential customers. After the data and information collection phase, they must develop a comprehensive and competitive business strategy. This means companies need to develop a strategy on how to operate their business, based upon the information collected and analysed in the earlier phase. In the next phase companies need to create their operations plans, where they collect and list all their key activities and personnel. Companies need to present a projection for the future based on their data collected from the markets and their implemented business strategy. These projections are mostly financial projections regarding volumes of sales, profits and a break-even analysis. The final step of this exhaustive and comprehensive process is writing the business plan into a complete document and being able to present it to external organizations (*Barrow et al., 2012*).

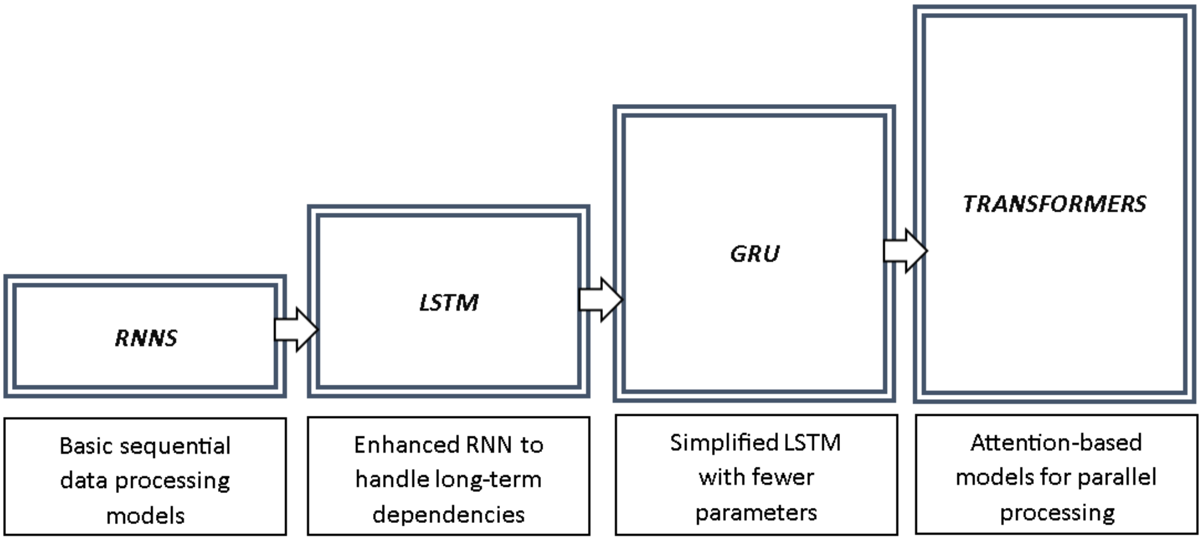
The act of planning positively correlates with venture performance, especially when plans are written, regularly updated, and used to guide strategic decisions (*Brinckmann et al., 2010*). Moreover, a well-structured business plan helps to reduce uncertainty and mitigate risks and improves the entrepreneur's preparedness, especially in scenarios when the plan has to be presented to external organizations or bodies to apply for funding or in an acquisition situation (*Delmar & Shane, 2003*). The study conducted by Brinckmann et al. in 2010 "confirmed that business planning increases the performance of both new and established small firms, yet different factors moderate the strength of the relationship. In samples with more established small firms, business planning has a stronger positive effect on performance than it does in samples consisting only of new firms." One of the reasons behind this might be that a well-established small business may already have stable operations, customer data, and financial history. So, when it uses a business plan to set goals or secure funding, the plan is built on solid ground, leading to better results. On the other hand, a brand-new startup may not have prior information regarding structures and procedures, so while planning is helpful, it has less immediate impact on performance because the foundation is still being tested. These findings explains that while planning is valuable at all stages, its impact becomes more pronounced as businesses grow and stabilize (*Brinckmann et al., 2010*).

2.2 Technological background

The appearance and fast development of machine learning algorithms and LLMs has transformed how businesses utilise these new technologies for their operations and automation. This section explores the key technical foundations behind the proposed application, which generates custom business plans using an agentic system structured within the CrewAI framework. It is crucial to get an understanding of all the concepts present in the project, to develop an application, which outputs high-quality, accurate customised business plans.
Firstly, the following section introduces the Transformer architecture, a breakthrough in natural language processing – NLP - that enables models like GPT to generate coherent and contextually appropriate text. This is followed by an overview of LLMs more broadly and a discussion of prompt engineering as a critical technique for effectively interacting with such models. Finally, the literature review examines agentic AI systems, explaining their structure, operational mechanisms, and their potential utility in the automated generation of business plans.

2.2.1 Attention mechanism and the Transformer architecture

Transformers are a type of neural network (*Vaswani et al.*, 2017). They were originally known for their strong performance in machine translation and are now a de facto standard for building large-scale self-supervised learning systems (*Xiao & Zhu*, 2023). It is a sequence-to-sequence model, meaning it processes an input sequence and generates a corresponding output sequence. The model learns to capture the relationships and dependencies between elements in the input sequences, enabling it to generate contextually appropriate responses. It is considered to be more flexible to adapt to different tasks, than the traditional neural networks such as recurrent neural networks, particularly long short-term memory and gated recurrent units, which have been widely used for sequence modelling tasks like language modelling and machine translation. These models process sequences step by step, making it difficult to capture long-range dependencies and limiting parallelisation. The attention mechanism addresses these limitations by allowing models to focus on relevant parts of an input sequence at each step, irrespective of their distance. The Transformer model eliminates recurrence entirely, using only attention mechanisms to model global dependencies (*Vaswani et al., 2017*).



RNNS
LSTM
GRU
TRANSFORMERS
Basic sequential data processing models
Enhanced RNN to handle long-term dependencies
Simplified LSTM with fewer parameters
Attention-based models for parallel processing

Figure 1: The comparison of NLP models

Source: Data based on findings from Bayat and Isik (2023).

The core concept of the attention mechanism is to focus selectively on relevant past information. Word embeddings face a significant challenge with ambiguous words, words with multiple meanings. For example, a word “goal” would be assigned the same vector in traditional embeddings regardless of whether it's used to refer to a score in a sports context or the goal of an individual in life. The attention mechanism effectively distinguishes the usage of words based on context, thereby transforming standard word embeddings into contextualized embeddings that vary according to the sentence they appear in (*Vaswani et al., 2017*).

Attention assigns different importance levels to different elements in an input sequence when generating an output. This is achieved through the computation of attention scores, which determine how much influence each input element should have on a given output position. One form of attention is self-attention, which enables a model to relate every token in a sequence to every other token. Self-attention is a special type of attention mechanism that enables a model to weigh the importance of different words in a sequence when encoding a particular word. Instead of processing tokens sequentially, self-attention considers all tokens simultaneously, computing their relationships with each other. How does it work essentially? Given an input sentence each word is first converted into an embedding vector, representing its meaning in a high-dimensional space. Each word’s embedding is transformed into three vectors: query, which represents what the word is looking for; key, which represents the meaning of the word, and value, which holds the actual content of the word. The attention score between two words is computed as the dot product of their query and key vectors. This determines how much focus one word should have on another. These scores are scaled down using the softmax function, which normalises them into probabilities. The final representation of each word is obtained by computing the weighted sum of the value vectors based on the attention scores. Words that are more relevant to a given word receive higher attention weights (*Vaswani et al., 2017*). We can look at a simple example: “*She poured water into the cup because it was empty.*”

When processing the word "it", the model needs to determine whether it refers to "water" or "cup". Through self-attention, the model assigns higher weights to "cup" based on context, improving its understanding of the sentence.

The Transformer consists of an encoder-decoder structure, where the encoder processes input sequences and the decoder generates the output sequence. Both encoder and decoder are composed of multiple identical layers stacked on top of each other. The following figure illustrates the architecture of the Transformer model:

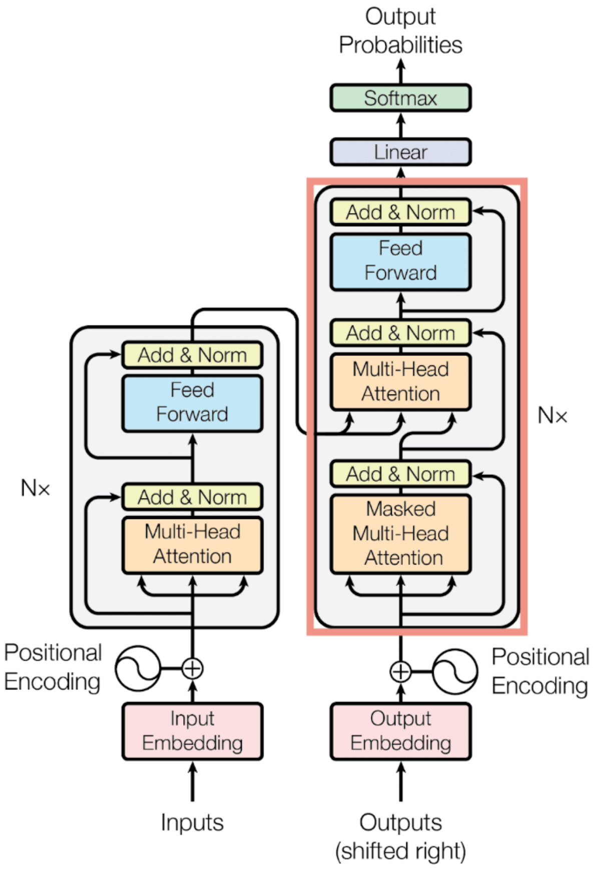


Figure 2: The Transformer model architecture (*Vaswani et al.,* 2017)

Each encoder layer receives input embeddings along with positional encodings, which add order information to the sequence, as Transformers process entire sequences simultaneously rather than sequentially. The decoder follows a similar structure to the encoder but includes an additional mechanism for handling the encoder's output. Instead of using a single attention mechanism, the Transformer employs multi-head attention, where multiple sets of query, key, and value matrices are computed in parallel. Each head learns different aspects of the relationships between words, allowing the model to capture diverse contextual information. The outputs from all attention heads are then concatenated and passed through a fully connected layer (*Xiao & Zhu*, 2023).

2.2.2 Generative Pre-Trained Models

One of the most significant implementations of the Transformer is the Generative Pre-trained Transformers (GPT) models, which have advanced the field of NLP. These models are built on the Transformer architecture, introduced by Vaswani et al. in 2017. GPTs are called generative because they can generate coherent, human-like text. OpenAI introduced the first GPT in 2018, a relatively small model by today's standards, but many other developers and companies followed it, creating newer and newer models rapidly (*Naveed et al.,* 2024).
GPT models, first introduced by OpenAI, are built on the Transformer decoder architecture. Unlike full Transformer models, which have both an encoder and a decoder, GPT relies solely on the decoder stack to generate text in an autoregressive manner (*Topal et al.,* 2021). Their goal is predicting the next word in a sequence given the context. This task forces the model to learn grammar, factual knowledge, reasoning abilities, and even some common sense. They utilize attention mechanisms twice during training: initially, masked multi-head attention, where only a part of a target sentence is revealed, and the model should predict the masked word, and later, multi-head attention, like encoders. GPT models undergo a two-stage training process: the model is trained on a large corpus of text using unsupervised learning. It predicts the next token in a sentence based on previous context, learning syntactic and semantic patterns. The model is further trained on domain-specific or task-specific data with supervised learning, adapting it to specific use cases. The model is trained using backpropagation and gradient descent, adjusting billions of parameters to minimize the prediction error across the training data (*Brown et al.*, 2020). These models can generate human-like text merely in seconds, therefore their application is widespread from customer service chatbots, to

programming assistants, however they proved to be useful in assisting people's lives in easy, everyday tasks as well, and they are able to help businesses achieve their goals too.

2.2.3 LLMs

LLMs are deep neural networks trained on massive text corpora. Modern LLMs are almost in all cases based on the Transformer architecture (*Vaswani et al., 2017*). In principle, an LLM processes text by mapping each token to a high-dimensional representation, then repeatedly applying multi-head self-attention and feed-forward layers to capture context. OpenAI's GPT-4, is explicitly described as "a Transformer-based model pre-trained to predict the next token in a document". In training, LLMs are trained on next-word prediction over billions of sentences, empowering them with broad world knowledge and language skills (*OpenAI, 2024*). Training LLMs involves two primary phases: pre-training and fine-tuning. Training is performed using a large corpus of high-quality data. During training, the model iteratively adjusts parameter values until it can correctly predict the next token from the previous sequence of input tokens. It does this through self-learning techniques which teach the model to adjust parameters to maximize the likelihood of the next tokens in the training examples (*Brown et al., 2020*). Once trained, LLMs can be easily adapted to perform multiple, different tasks using relatively small sets of supervised data, a process known as fine tuning. One of the most known examples of fine tuning of these models is few-shot learning. It means that by providing a few relevant training examples, the model performance significantly improves in that specific area. This technique essentially guides the models to a desired way of generating an output by showing it what is good and what is bad through examples. Another possible way to fine tune the models is to train a base model with additional data relevant to a specific application. This enables the models to gain extended knowledge in a specific field of study and becoming more reliable to a certain application (*Brown et al., 2020*).

2.2.4 Prompt engineering

Prompt engineering is a relatively new skill that appeared after the fast development of LLMs. It refers to the art and science of crafting inputs, which are called prompts, to guide these models to produce accurate, relevant, and useful outputs. Given the models' general-purpose nature, the way we phrase questions or tasks has a huge impact on performance (*Wei et al.,* 2022). This skill is key in guiding the language models to the desired output, and we also employ different prompt engineering skills in our project as well.

Simply put, a prompt is just the input text written by the user or the developer who provides it to a language model in order to achieve a desired output from it. However, unlike traditional programming where outcomes are deterministic, language models are probabilistic. Slight differences in phrasing, structure, or even punctuation can lead to vastly different results (*Wei et al.,* 2022). With the help of the different techniques, which are available to every developer, they can guide the model to solve complex tasks, such as multi-step reasoning or breaking down hard tasks into smaller, more understandable chunks. In our project, prompt engineering is one of the most crucial concepts that we must master, because the main part of designing a system of autonomous AI agents, such as with the help of CrewAI, is writing the prompts of the different agents to achieve our goals with the system. These agents need concrete guidance and tasks on how to behave, what to generate and how to structure the output, because without prompt engineering these models are very general, and are not able to provide concrete solutions to specific problems at the required level.

We need to address three important prompting techniques that we also employ in our project and will be discussed later in detail in the methodology section. The first prompting technique is called zero-shot prompting. It involves giving the model a task with no prior examples. The prompt must contain enough detail for the model to understand what is required, relying on the model's broad general knowledge. Zero-shot prompting is simple, fast, and effective in easier tasks, where reasoning is not required. But it may not always yield optimal performance, when the desired output is dependent on reasoning or when the solution would require a step-by-step explanation (*Brown et al.,* 2020). An example for a zero-shot prompt would be any kind of simple translation task, like: "*Translate this sentence from English to Spanish: I am happy.*"

Another important technique is called few-shot prompting. It enhances accuracy of the possible output by including examples within the prompt itself. This allows the model to see the desired format, tone, or logic through analogy from the examples. This technique is very useful in cases where the user would like to achieve a certain format or structure within the output of the model, so they include examples of this structure in the input of the model itself, so it can derive the necessary information and generate a better, more accurate output (*Brown et al.,* 2020). For instance, in our case, providing example sentences in the prompt showing how a coherent and holistic flow of text looks like to allow the model to mimic and structure the tone.

The third prompting technique that we will mention here is the chain-of-thought method. This technique involves encouraging the model to reason step by step and include intermediate steps in the prompt. Chain-of-thought prompting dramatically improves performance on complex reasoning tasks such as arithmetic calculations, tasks related to logical thinking, and even common-sense questions. This technique was first introduced by Wei et al. in 2022, and it revolutionised the solution of reasoning-heavy tasks by LLMs. With this technique we can tell the models exactly how to structure their responses step by step and give reasoning to each step and explain how it got to the final answer or solution (*Wei et al.,* 2022).

In the project all three types of the mentioned techniques are used throughout developing our agentic system, because they are necessary in order to achieve a highly customised, accurate and precise solution to a difficult and complex problem like generating a business plan for a certain company. However, the emphasis will be on the chain-of-thought technique, because building the complex parts of the business plan requires this approach to achieve better, more desired outputs from the agents.

2.2.5 Agentic AI Systems

An agentic system consists of AI agents—autonomous computational entities capable of perceiving their environment, making decisions, and performing actions to achieve defined goals (*Russell & Norvig, 2022*). In multi-agent systems, these agents communicate and coordinate their activities. Agentic AI builds upon the developments of LLMs by employing LLMs as their central cognitive units that integrate with external tools. This integration enables LLM-based agents to execute actions, solve complex problems, and interact dynamically with their environments. In these systems, the complex tasks are distributed to different specialised agents, each agent responsible for a certain area of expertise, tackling the whole complex problem more accurately and efficiently than a single-agent system would (*Tran et al., 2025*). Writing a business plan is indeed a complex task, therefore it is ideal for us to make use of this type of multi-agentic system to achieve higher accuracy in the final product. In these types of models, it is vital for the AI agents to work together collaboratively to achieve a common end goal, in our case, to write a comprehensive business plan. Each agent can have a specific goal and role in the system, which enables them to focus their knowledge on different aspects of the user queries (*Tran et al., 2025*).

2.3 AI in Business Planning

Integrating artificial intelligence into business planning could mean strategic advantage particularly for startups and small businesses, as traditional business planning methods require extensive time, resources, and expertise from many different aspects of the business, which can be barriers for emerging enterprises. Using AI and machine learning for generating business plans is more efficient and cost-effective. This approach allows startups and small businesses to have a higher chance to obtain loans, attract investors, and make appropriate strategic decisions, ultimately increasing their chances of long-term success, while reducing the time and costs associated with manual business plan creation. By using AI in business planning, achieving early drafts of business plans can become easily and widely accessible to non-technical users such as startup owners. They will be able to make further refinements based on the AI-generated drafts, and this reduces costs and saves time for them in the business planning process. Additionally, AI tools can analyse market data or financial trends in real-time, ensuring that business strategies are based on the most up-to-date information. This minimizes risks associated with outdated or inaccurate assumptions and improves decision-making, reducing human error.

There could be drawbacks of solely relying on AI to write a complete business plan. One of the main concerns of AI-generated content is that it often contains “hallucinations”, which means that the content is not based on facts, merely the models made-up that information while generating their output. Brown et al. (2020) found that even GPT-3 struggles to produce error-free output on certain tasks, so the output must be carefully verified by humans. Ultimately it is not the AI that should be the final decision-maker, rather than the actual humans behind the firms. They could utilise the AI-powered solutions in planning their business strategies and developing a business plan with the help of AI agents, but at the end, humans should make the final touches and build a high-level strategy and understanding of the business.

Overall, this approach offers a practical solution for companies that may otherwise struggle with creating the business plan, allowing them to allocate their resources more effectively and increase their chances of long-term success as the AI-generated business plans require only minimal human intervention, primarily for fine-tuning and adding specific elements based on the company's objectives and strategic goals.

2.4 The Evaluation of AI Generated Content

Evaluating AI-generated content, particularly in creative applications, presents challenges that originates from the subjectivity of creativity, the limitations of traditional evaluation metrics, and the complexities of evaluating large bodies of text generated by LLMs. Recent academic efforts have tried to address these challenges through innovative methodologies. Creative content, such as analyses, poetry, or in fact, business plans, often lacks a singular "correct" form, making objective evaluation difficult. In many cases, it is somewhat easier to evaluate a result based on subjectivity. However, in a scientific study, objectivity must be the key, when evaluating the results. “Traditional automatic metrics, such as BLEU, ROUGE and METEOR are widely used for Natural Language Generation - NLG - evaluation, but they have been shown to have relatively low correlation with human judgments, especially for open-ended generation tasks” (*Liu et al., 2023*). This limitation is particularly pronounced in tasks requiring a deep understanding of context, tone, and stylistic elements.

The SummEval project, presented by Fabbri et al. (2021) not only critiqued existing evaluation metrics but also provided a comprehensive benchmarking dataset, including human annotations across multiple dimensions of summary quality. These annotations included coherence, consistency, fluency and relevance. This resource has been a pioneer in facilitating the development and testing of new evaluation methodologies that seek to bridge the gap between automated metrics and human judgment.

Liu et al. (2023) introduced G-Eval, a new approach that leverages GPT-4's capabilities through chain-of-thought prompting and a structured form-filling paradigm. Empirical results demonstrated a high correlation with human assessments in summarization tasks, surpassing previous models. However, the study also noted potential biases, particularly that LLM evaluators favour outputs generated by similar models, raising concerns about objectivity (*Liu et al., 2023*).

The "LLM-as-a-judge" framework has emerged as an alternative, leveraging LLMs to mimic human reasoning for evaluation purposes. Current frameworks struggle to adapt to different text styles, including various answer and ground truth formats, reducing their generalization performance across diverse applications. The evaluation scores produced are often skewed (they may not accurately reflect the true quality or performance) and hard to interpret, showing a low correlation with human judgment. However, the research conducted by Zheng et al. (2023) showed that closed-source LLMs, like GPT-4 can evaluate AI-generated text comparably to humans. Moreover, the authors in Cao et al. (2025) propose a dynamic multi-agent system that automatically designs personalized LLM judges tailored for various NLG applications The results demonstrated that the multi-agent LLM Judge framework not only enhances evaluation accuracy compared to existing methods but also produces evaluation scores that better align with the human judgment (*Cao et al., 2025*).

2.5 The Research Gap in the Literature

Despite significant advancements in both business planning research and Large Language Model (LLM) technology, there remains a gap in the literature concerning their integrated, practical application, especially in multi-agent contexts. Existing work in this field is sparse: for example, a recent study, called BizChat (*Romero Lauro et al., 2025*) is one of the first to develop an LLM-driven tool for drafting business plans. However, BizChat employs a single LLM agent with a guided interface, not a coordinated multi-agent system. Their application helps small businesses draft a business plan for them, and improve it based on AI suggestions through a chat interface (*Romero Lauro et al., 2025*). Beyond BizChat, remains a practical knowledge gap in the literature, which means that there is a lack of application-focused research demonstrating how LLMs can generate business plans in a multi-agent agentic system in real world scenarios. The present study addresses this practical gap by employing a comprehensive questionnaire, which was developed by academic experts at Aalborg University to gather essential data about several aspects of the companies, which is used to generate the plans for them. Furthermore, the questionnaire guides the business owners, who may lack the clarity about their exact goals or do not know how to approach the complex problem of writing a business plan effectively.

Due to the lack of systematic evaluations assessing the effectiveness, accuracy, and usability of such systems in entrepreneurial contexts, an empirical gap can also be addressed. This thesis tries to assess and evaluate the system with the help of the academic experts, thereby providing a measure of how accurately LLM-powered AI agents can generate business plans based on the information gathered from the business owners through the questionnaire.

3. Methodology

This section presents the research approach and system design methodology employed in this study. The methodology encompasses a systematic examination of the research process, from initial data collection to final evaluation methods. This research is structured around three fundamental components: data collection and testing procedures, system development and architecture, and evaluation framework.

The methodology begins with a comprehensive overview of the research design and the structure of the study. This is followed by an examination of the technical implementation of the multi-agent system, detailing the architectural decisions that form the foundation of the application's functionality. The final component presents the evaluation methodology, outlining the metrics and assessment criteria employed to measure the system's performance and effectiveness.

3.1 Research Design

According to Kerlinger (1986) research design is the plan, structure, and strategy of investigation conceived to obtain answers to research questions and to control variance. Building on this foundation, our research adopts a tailored research design that aligns with its purposes.

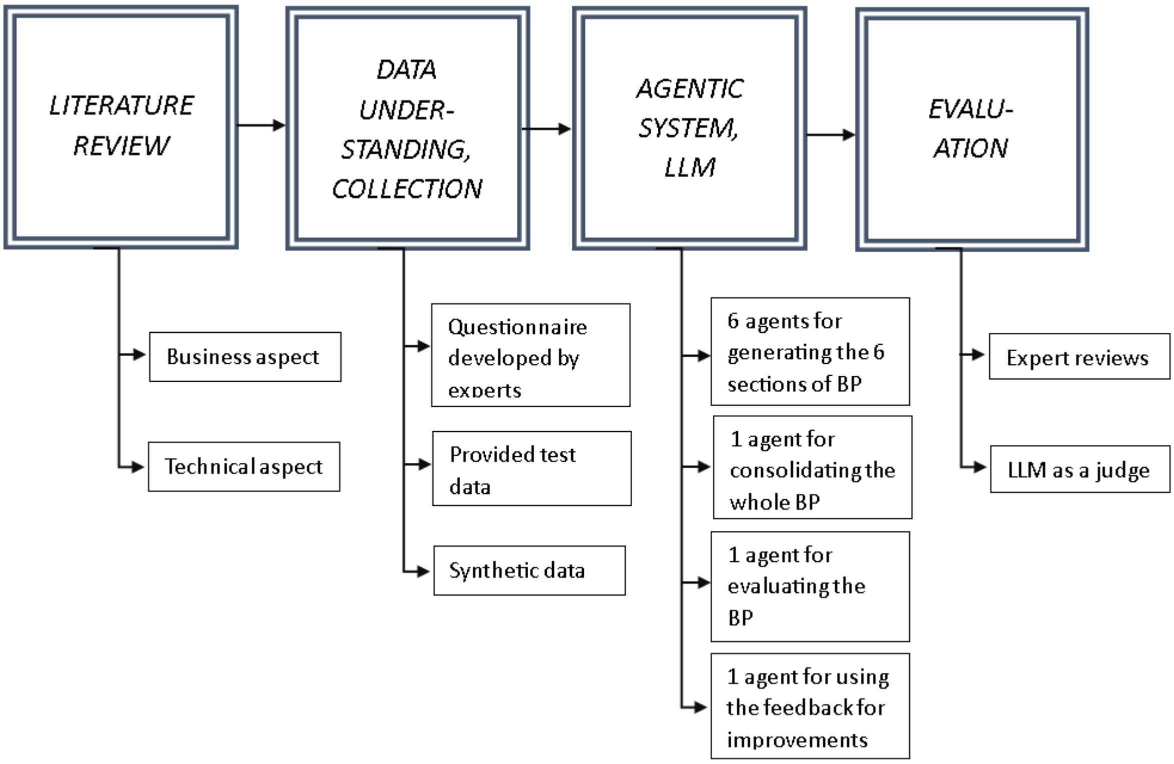

LITERATURE REVIEW
DATA UNDER-STANDING, COLLECTION
AGENTIC SYSTEM, LLM
EVALU-ATION
Business aspect
Technical aspect
Questionnaire developed by experts
Provided test data
Synthetic data
6 agents for generating the 6 sections of BP
1 agent for consolidating the whole BP
1 agent for evaluating the BP
1 agent for using the feedback for improvements
Expert reviews
LLM as a judge

Figure 3: The approach to the research process and its components

The research methodology started with a systematic collection and analysis of academic research papers about business planning, LLMs, and agentic AI systems. Given the limited academic literature on the application of LLM-powered agentic systems in business plan generation, this study aimed to address this research gap. The literature review contains two primary domains: business aspects and technological aspects. The business domain analysis focused on business plan definitions, components, structure, and traditional methodologies for creating effective business plans. The technological analysis examined the evolution of LLMs, their operational principles, foundational concepts, and their potential to help business plan creation. Additionally, the review included an examination of multi-agent system architecture, AI agent definitions, and the techniques employed to direct these agents toward generating comprehensive business plan drafts.

Following the comprehensive review of academic research, qualitative research methods were employed to gather additional information and data specific to the implementation of LLM-powered agentic systems for business planning. The research methodology prioritized insights from business planning experts, as qualitative methods are particularly suited for understanding experiences, opinions, and expectations (*Jain, 2023*). Three structured interviews were conducted with external academic experts from the university, providing insights into the research concept and methodological approach. These experts contributed their expertise in business planning and shared a previously developed questionnaire, which served as the foundation for the upcoming analysis and application development. The questionnaire was designed to collect comprehensive data about companies seeking to create business plans. During the initial development phase, testing was conducted using fictional data provided by the academic experts. Additionally, the experts provided access to authentic business plans for comparative analysis.

The subsequent phase of the research focused on developing a functional multi-agent system designed to generate business plans based on company-specific information. Through extensive analysis of AI agent documentation and relevant frameworks, it was concluded that optimal performance could be achieved by assigning specialized agents to distinct sections of the business plan. This approach was based on the principle that agents demonstrate significantly better performance when assigned specific, focused tasks rather than general responsibilities. Consequently, the business plan generation process was divided into six specialized sections, each managed by a dedicated agent. The system architecture incorporated an evaluation layer comprising two distinct agents: an evaluator agent responsible for comprehensive document assessment and feedback generation based on predefined criteria, and a refinement agent tasked with implementing the recommended changes. This architecture facilitated an efficient, coordinated workflow with integrated feedback mechanisms.

The final phase of the research focused on result evaluation, a component of equal importance to system development and plan generation. Following analysis of existing literature on AI-generated content evaluation, the methodology incorporated expert human judgment for quality assessment, in collaboration with the academic experts. To provide additional evaluation perspectives, the LLM-as-a-judge method was implemented, utilizing a different Large Language Model than the one employed for generation. This dual approach enabled comparative analysis between human and LLM evaluations, using consistent metrics across both assessment methods. The evaluation process consisted of two rounds of scoring by both experts and the LLM, with feedback from the initial round informing system improvements for enhanced accuracy and quality in the following generation.

This methodology followed an iterative design process, with each version undergoing comprehensive evaluation, ensuring continuous improvement and refinement of the system.

3.2 Data Collection

The data collection process is implemented through a comprehensive, multi-section questionnaire developed by academic experts from AAU Business School: Peter Thomsen, Assistant Professor, and Brian Balslev Andersen, PhD Fellow. The questionnaire employed in this study comprises 57 questions designed to gather comprehensive information necessary for business plan generation. The implementation incorporates conditional logic in the structure of the questions to enhance user experience, where questions are dynamically presented based on previous responses. This adaptive approach ensures that users are only presented with relevant questions, optimizing the data collection process.

The questionnaire architecture contains various question types, including text input fields, multiple-choice selections, and multi-select options. These are systematically categorized based on their functional role in business plan generation, with detailed specifications provided in the appendix.

The questionnaire is designed to capture all essential aspects of a business, including basic company information (for example name, year of establishment, legal structure, funding sources, mission and vision, and rationale for founding). It collects market and customer information too (about business sector, target markets, customer segments, value propositions, and competitive landscape). The questionnaire contains questions regarding operational and strategic information as well (including key resources, activities, partnerships, and cost structure). Last, but not least, questions for team structure and funding are incorporated too (more specifically team composition, skills, and funding requirements).

The implementation of conditional logic within the questionnaire serves to enhance both relevance and efficiency, ensuring that users are presented with a streamlined, contextually appropriate set of questions based on their specific business circumstances.

3.2.1 Primary data

The primary data used in this research was collected through a structured questionnaire designed in collaboration with domain experts from the AAU Business School. The data set consists of completed responses for a fictional company, generated during earlier research phases by the experts. Although the company itself is hypothetical, the responses reflect realistic business scenarios and were developed to ensure coherence, completeness, and relevance across key sections of a business plan. This expert-generated data was used for evaluating the system’s ability to transform user input into an appropriate business plan.

3.2.2 Synthetic data

To further validate the system's capabilities, a different testing methodology was implemented utilizing synthetic data generation. This approach employed OpenAI's GPT-4.1 model to simulate user responses by analysing and responding to the questionnaire based on existing business plan documentation. The synthetic data generation process involved providing a complete business plan to the model and prompting it to respond to the questionnaire from the perspective of the business owner. The generated synthetic responses were then processed through the multi-agent system. This system employs an architecture incorporating two different advanced language models: Google's Gemini 2.5 Pro is used for generation tasks and Meta's Llama 3.3-70B-versatile for the evaluation layer.

This process was helpful to assess how well the system can handle structured inputs gained from existing plans, and whether the generated outputs align with the source material.

3.3 Applications and Methods

The system was developed using an agile methodology, with 2 cycles of prototyping, testing with data, and refinement. Agile methodology is an alternative to traditional project management, it supports teams in responding to changing needs and uncertainties through short, iterative work phases called sprints. Therefore, it allows for more flexibility, continuous feedback, and gradual improvement throughout the development process (*Abrahamsson et al., 2002*). The system was built over cycles, during which we gathered feedback from the experts. Using this feedback, each cycle involved refining both the user interface and the functionality of the LLM-based agents. This approach enabled us to iteratively improve the accuracy, usability, and coherence of the generated business plans so that the final system produces the expected results for its users.

3.3.1 Multi-agent Agentic System Architecture

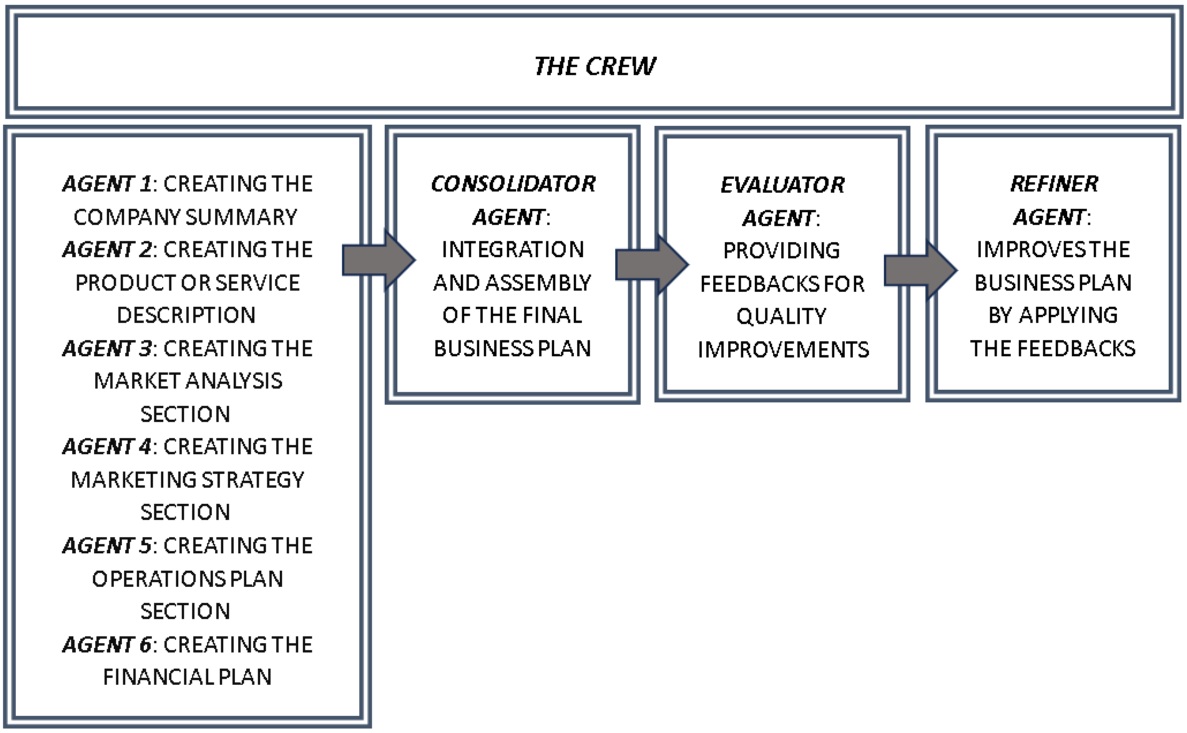

THE CREW
AGENT 1: CREATING THE COMPANY SUMMARY
AGENT 2: CREATING THE PRODUCT OR SERVICE DESCRIPTION
AGENT 3: CREATING THE MARKET ANALYSIS SECTION
AGENT 4: CREATING THE MARKETING STRATEGY SECTION
AGENT 5: CREATING THE OPERATIONS PLAN SECTION
AGENT 6: CREATING THE FINANCIAL PLAN
CONSOLIDATOR AGENT: INTEGRATION AND ASSEMBLY OF THE FINAL BUSINESS PLAN
EVALUATOR AGENT: PROVIDING FEEDBACKS FOR QUALITY IMPROVEMENTS
REFINER AGENT: IMPROVES THE BUSINESS PLAN BY APPLYING THE FEEDBACKS

Figure 4: The architecture of the multi-agent agentic system

The system implements a complex multi-agent architecture that leverages the CrewAI framework to implement a team of specialized AI agents. The system is designed with a modular and hierarchical structure, where each agent is responsible for a specific aspect of business plan creation, working in collaboration to produce a comprehensive and coherent document. The architecture consists of nine specialized agents, each with distinct roles and expertise. Six different agents are responsible to cover six different parts of the business plan as Figure 3 illustrates. The final stages of the process are handled by three agents. The Consolidator, responsible for merging different sections into a cohesive document. Its role is merely a technical one, to provide a clearer document structure to the next agent, which is the Evaluator. It reviews the business plan's quality and coherence based on predefined criteria and the human feedback that we received from our academic experts. The Refiner, who ensures the actionable feedback from the Evaluator agent is applied to the document; it outputs the final version of the generated business plan.

The system employs a sophisticated dual-model approach utilizing two different LLMs to optimize performance. For content generation tasks, the system employs Google's Gemini 2.5 Pro model, configured with a temperature of 0.1 and high reasoning effort to ensure precise and consistent output. This model excels at creative content generation while maintaining factual accuracy and professional tone. The evaluation and refinement tasks are handled by Meta’s Llama 3.3 70B versatile model, which operates with a medium reasoning effort and the same low temperature setting, making it particularly effective for critical analysis and quality checking tasks. This strategic distribution of tasks between the two models allows for consistency during content generation, and versatility during the critical assessment and evaluation layer.

The workflow is implemented as a sequential process, where each agent's output serves as context for subsequent agents. This design ensures that each section builds upon previous work of the previous agents, maintaining consistency throughout the document.

The system implements a context-aware architecture where each task receives relevant context from previous tasks. This context passing mechanism is crucial for coherence and consistency across the whole document.

Error handling and validation are implemented at multiple levels throughout the system. The architecture includes input validation at the crew level, task-specific validation within each agent, and final quality assurance through the evaluation and refinement process.

This architecture aims to enable the system to generate comprehensive business plans while maintaining consistency, coherence, and professional quality across all sections. The modular design allows for easy extension and modification of individual components without affecting the overall system functionality, making it adaptable to various business planning requirements and future enhancements.

3.3.2 Prompt engineering

The system's effectiveness in generating high-quality business plans is significantly enhanced through the implementation of advanced prompt engineering techniques. Drawing from the work of Wei et al. (2022) on Chain-of-Thought (CoT) prompting, the system employs a structured approach to guide the model through complex reasoning processes. This methodology has been shown to significantly improve performance on tasks requiring multi-step reasoning and logical deduction. The implementation of Chain-of-Thought prompting in the business plan generation process follows the framework established by Kojima et al. (2023), which demonstrates that step-by-step reasoning can be elicited even in zero-shot scenarios. This approach enables the model to break down complex business planning tasks into logical components and maintain consistency throughout the document generation process.

The system also incorporates Few-Shot learning techniques, as demonstrated in the work of Brown et al. (2020), to provide the model with examples of well-written parts of text. This technique is mainly used to guide the model in generating coherent flow of texts. These examples serve as templates that demonstrate professional tone and language usage, appropriate level of detail and specificity, logical flow and section transitions.

The systematic approach to prompt design contributes significantly to the system's aim to produce coherent, well-structured, and contextually appropriate business plans. The effectiveness of this methodology is further supported by recent research in the field of prompt engineering, supported by Brown et al. (2020) and Wei et al. (2022).

3.4 Evaluation Approach

The evaluation process for the application's results is structured as an iterative, multi-phase methodology that integrates both human expertise and large language model assessment to ensure a comprehensive evaluation of generated business plans. This dual approach is designed to maximize objectivity, and the overall quality of the evaluation. The data which was used to serve as the basis for the generated business plans was partly provided by the evaluator experts themselves, and the other data source was synthetic data. This synthetic data means answers to the questionnaire based on a real business plan, generated by GPT-4.1, which was prompted to act as the business owner and provide answers to serve as a synthetic data source.

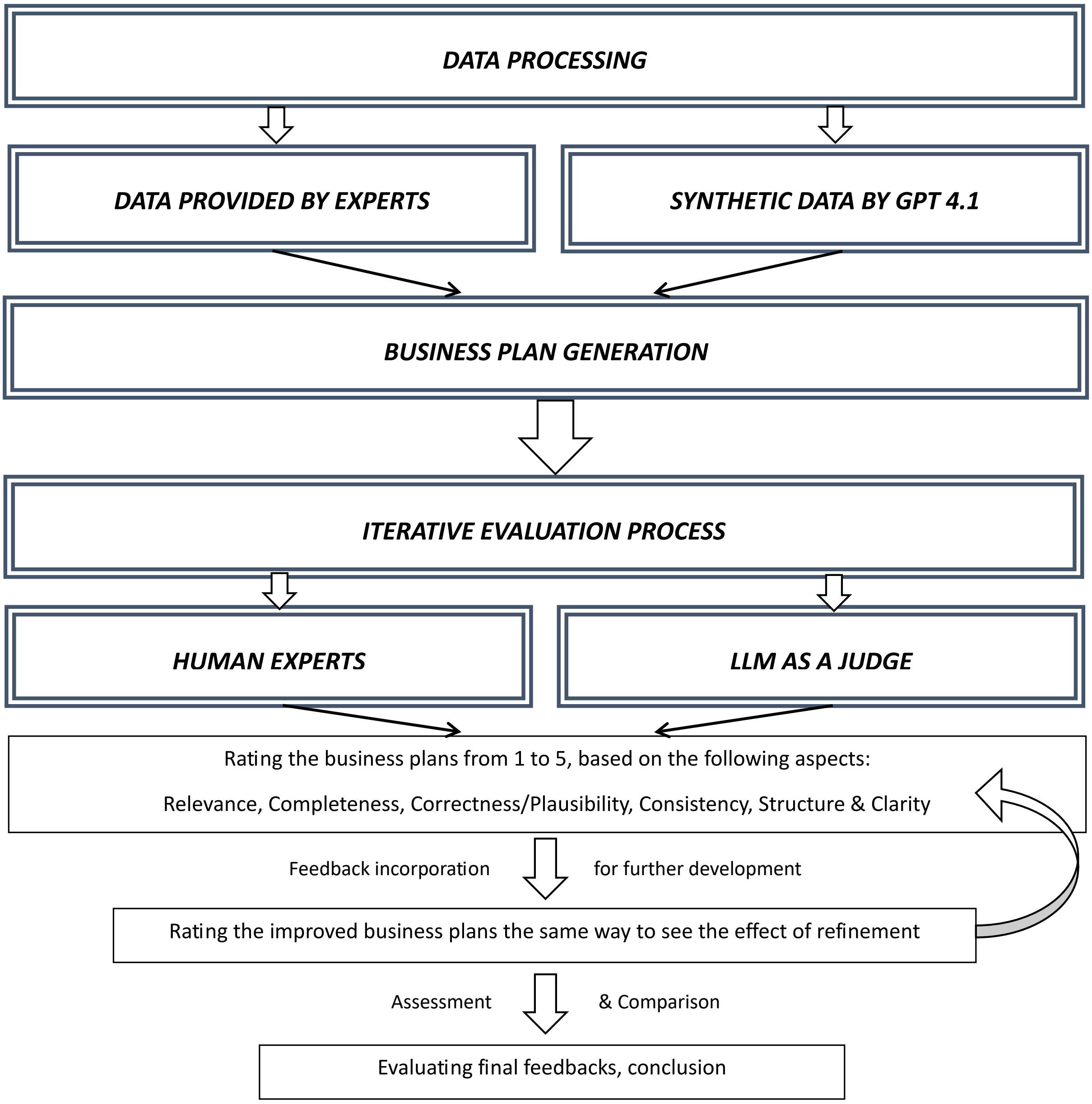

DATA PROCESSING
DATA PROVIDED BY EXPERTS
SYNTHETIC DATA BY GPT 4.1
BUSINESS PLAN GENERATION
ITERATIVE EVALUATION PROCESS
HUMAN EXPERTS
LLM AS A JUDGE
Rating the business plans from 1 to 5, based on the following aspects:
Relevance, Completeness, Correctness/Plausibility, Consistency, Structure & Clarity
Feedback incorporation
for further development
Rating the improved business plans the same way to see the effect of refinement
Assessment
& Comparison
Evaluating final feedbacks, conclusion

Figure 5: The System Evaluation Approach

At the core of this process lies the principle of iterative evaluation, where each business plan produced by the system undergoes multiple rounds of assessment and refinement. The process starts with the parallel involvement of two distinct evaluative agents: two human experts and a large language model acting as an autonomous judge. The human experts apply their contextual knowledge, practical experience, and understanding of business planning conventions during the evaluation phase. In contrast, the LLM, selected to be different from the generative model to avoid bias, offers a scalable, consistent, and replicable means of assessment, and is able to produce comparable judgment to humans (*Zheng et al., 2023*).

The initial phase involves both parties independently rating the first batch of generated business plans. The assessment is conducted using a standardized scale, ranging from one to five, across a set of well-defined criteria.

Criterion	*Scoring Range*	*Description*
Relevance	1-5	Is the content appropriate for the section?
Completeness	1-5	Does it include all key elements?
Correctness / Plausibility	1-5	Is the information reasonable and accurate?
Consistency	1-5	Are all sections logically aligned?
Structure & Clarity	1-5	Is the section well-written and easy to understand?

Table 1: Assessment criteria for evaluating the generated business plans

These criteria contain relevance, completeness, correctness or plausibility, consistency, structure, and clarity as shown in Table 1. Each dimension is critical to the overall quality of a business plan. The use of such multidimensional metrics is supported by contemporary research in automated content evaluation and used by industry experts in evaluating NLG tasks (*Fabbri et al., 2021*). The metrics used in this study are derived from Miller & Tang (2025), however, to make the evaluation clearer for the experts, some of the metrics were renamed and slightly modified. In Miller & Tang (2025) the researchers mention relevance, clarity, coherence, accuracy and efficiency. This study kept relevance, clarity and restructured the remaining metrics. From the original research, accuracy became correctness, efficiency became completeness and coherence became consistency. These modifications were done to provide a more specialised matrix of metrics for the expert evaluators, which captures the nature of the research better, considering the specialised field of the study.

Upon completion of the initial ratings, the feedback generated by both human experts and the LLM is systematically incorporated into the development process. This feedback-driven refinement phase is essential for iterative improvement, as it enables the identification and possibility to correct the generated business plans. The system is updated based on the insights, and a new, improved set of business plans is subsequently produced. The update process typically involves employing different prompting techniques or completely rewriting sections in the task definitions of agents.

The improved business plans are then subjected to a second round of evaluation, following the same well-defined criteria and rating procedures as the previous phase. This repetition serves a dual purpose: it not only measures the effectiveness of the refinements implemented but also provides a basis for comparative analysis between the original and improved outputs. The ability to assess the impact of iterative changes is crucial for validating the efficiency of the system's enhancement and for ensuring that progress is both measurable and meaningful.

Following the second round of assessment, the results from both human and LLM evaluators are systematically compared and analysed. This comparative assessment phase is also crucial in identifying areas of indifference between human and machine judgments, therefore providing insights into the reliability and validity of LLM-based evaluation methods. This iterative method of feedback incorporation and system updates is an effective way to measure and adjust the agentic system to produce better results on each iteration based on the available feedback from the evaluators.

The evaluation incorporated a third round of assessment; however, it must be noted that this round was conducted based on a different dataset, suggested by the human evaluators, who recommended a new dataset, constructed by themselves, which contained higher quality input data than the previous batch.

The evaluation methodology employed in this study employs an iterative approach that leverages the strengths of human expertise and advanced language models to validate and further improve AI-generated business plans.

4. Implementation

This section presents the development and deployment of a business plan generation system with the usage of LLMs to help create comprehensive business plans. The implementation details the modular architecture that processes user inputs through a series of specialized AI agents, each designed to handle specific aspects of a business plan.

The system's core functionality is built around a questionnaire-based interface that captures important business parameters, which are then processed through an AI-powered agentic system. The system employs LLMs as their brain to generate detailed sections of the business plan, ensuring consistency in coherence, style, and content, while maintaining professional standards.

The development process incorporated several key technical considerations, including the integration of multiple AI agents for different aspects of content generation using LLMs, the implementation of a structured data processing pipeline, and the implementation of an evaluation mechanism. The system's architecture was designed to be scalable and adaptable, allowing for future enhancements and the incorporation of additional business sectors and planning requirements.

4.1 Data processing

The data processing phase of the system employs a multi-layer architecture from user input to ready-to-use data format for the Large Language Model. The system utilizes a questionnaire-based approach implemented through a user interface, where users provide different business information through structured questions ranging multiple aspects including company overview, market analysis, operational strategy, marketing and customer segmentation characteristics. The data collection mechanism is implemented through a frontend interface that presents users with different forms of input fields, each targeting specific business aspects. This interface is implemented in a Python framework, called Streamlit, which enables developers to create web interfaces easily and intuitively through Python code.

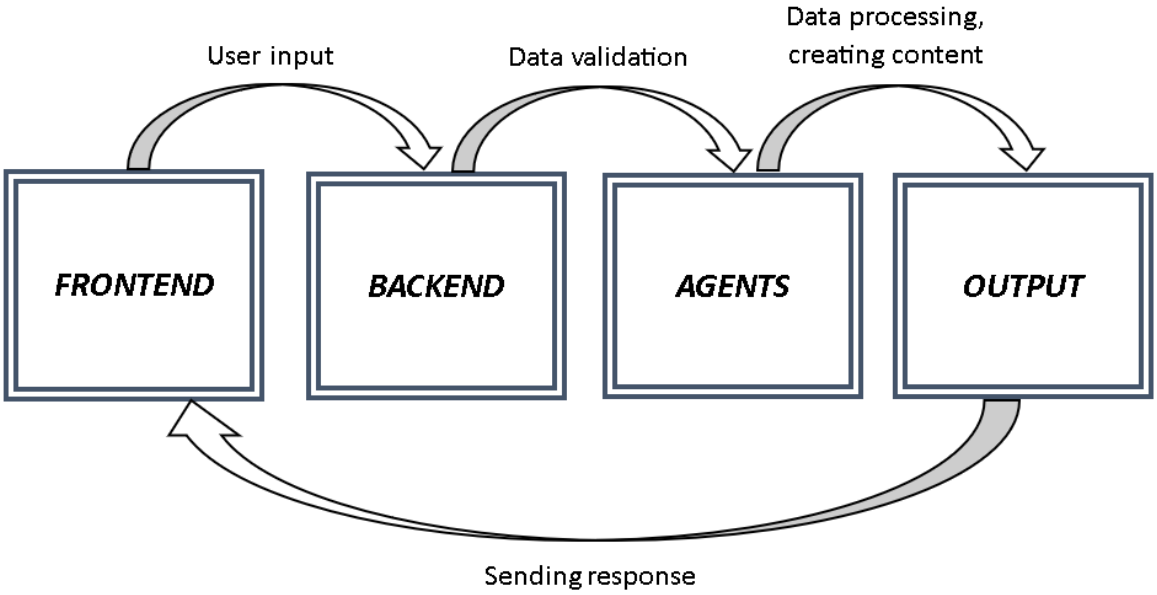


Figure 6: The flow of data in the application

The figure above illustrates an overview of the flow of information in the application from user inputs to the generated business plan. After the data is collected and submitted through the frontend interface, it is passed to the backend layer. This layer performs data validation and instantiates the multi-agent agentic system on the backend server to start the process of generating a response back to the user interface. During the data validation phase, the system ensures that all the necessary components of the required user inputs are present, the types of the inputs are valid and match the required values. These validations are performed using Pydantic data models.

The backend layer contains the agentic system itself. The validated input data is passed to the agents of this system, who process it and base their responses on the information contained in them.

The system maintains a clear data flow throughout the pipeline, enabling tracking of how input data flows through each pipeline stage. The sequential architecture ensures that outputs from earlier stages serve as contextual inputs for subsequent agents, creating a coherent narrative flow across all business plan sections. This approach eliminates manual interventions while reducing errors and enhancing the speed of the data handling process.

The final stages of the pipeline implement consolidation and refinement processes, where individual section outputs are merged into a single document. The system applies a pre-defined evaluation mechanism, which is based on multiple criteria and consistency checks to ensure the final output meets professional standards. The entire pipeline operates as a stateless system, processing data in-memory without persistent storage, ensuring privacy and security.

4.2 The Multi-agent Agentic System

The agentic system is at the core of this system, leveraging a collaborative network of specialized AI agents to automate and facilitate the process of business plan generation. This system is designed to resemble the structure and workflow of a high-performing human organization, where each agent assumes a specialised role and contributes its expertise to the final, collaborative outcome.

At its foundation, the agentic system is composed of multiple autonomous agents, each equipped with domain-specific knowledge and capabilities. The architecture of the system is intentionally modular, which allows for clear separation of logical sections and facilitates scalability and maintainability. This modularity ensures that each agent focuses on one aspect of the business plan while the system as a whole maintains coherence and consistency across all outputs. The reason for this is that even though the different sections are generated separately, the architecture ensures that information flows through the system between agents, enabling a more coherent, and holistic flow of generated output.

A defining characteristic of this agentic approach is its ability to support both autonomy and collaboration. Agents are capable of making independent decisions within their areas of responsibility, yet they also communicate and coordinate with one another to ensure that their outputs align with the goal of the system. This tries to mirror the dynamics of effective human teams, where specialized members work both independently and collectively to achieve complex goals. This agentic system is implemented in the Python framework called CrewAI, which supports the exact purpose of creating collaborative teams of AI agents to achieve a common goal.

4.2.1 Agents and LLMs

The following sections explore in detail the core concepts and implementations of the agentic system, therefore the information about the concepts is based on the official CrewAI documentation, which details them meticulously and thoroughly. While implementing the system, the documentation was followed in accordance with best practices recommended by the authors.

In the agentic system, agents serve as the fundamental building blocks, employing the principles of modularity, specialization, and intelligent collaboration. Each agent is an entity designed to autonomously perform a pre-defined task, leveraging the capabilities of LLMs to process information, make context-aware decisions and generate content based on the task and input data available. The agent concept is central to the CrewAI framework, which is used in this application.

Within CrewAI, agents are defined by a set of core attributes: role, goal, and backstory. The role specifies the agent's function within the broader system, such as market analyst, product designer, or financial expert. The goal defines the specific objective the agent is intended to achieve, providing direction and focus for its operations. The backstory creates a persona for the agent, giving depth to them and influencing how they approach problems and how they collaborate with other agents in the system.

In the following table a definition of an agent is presented as an example from the application implementation:

Role	Goal	Backstory
Business Strategist specialising in corporate strategy and business modelling	Develop a compelling and structured company overview, description, and vision based on the information provided by the user in the questionnaire.	With over 15 years of experience in corporate strategy and entrepreneurship, you have helped startups and Fortune 500 companies refine their business models. You excel at identifying market gaps and crafting compelling business visions that inspire stakeholders. You believe that a well-defined strategy is the foundation of every successful enterprise, and you approach every business plan with analytical thinking and creative problem-solving.

Table 2: Definition of an Agent taken from the implementation of the application

In the system each agent operates as an autonomous unit, capable of making independent decisions within its domain of expertise. However, agents are not isolated; they are designed to interact, share information, and coordinate their actions as part of a larger group, called the crew. This crew-based organization tries to mimic the mentioned dynamics of effective human teams, where members with complementary skills collaborate to solve complex problems. The CrewAI framework enables these interactions through processes, managing task assignments, sequencing, and the integration of agent outputs into a coherent whole. These processes can follow two types of patterns, sequential or hierarchical, depending on the requirements of the application. The application uses a sequential pattern, where the order of tasks matter, and they are executed one after the other, while passing information and context between each other. In this study, the agentic structure is created according to the task of generating business plans. The system contains 6 specialized agents, each responsible for a distinct section of the business plan.

The first agent is responsible for the company summary section in the business plan; the second one needs to create a detailed product/service description for the company. The third agent creates the market analysis section; the fourth agent generates the marketing and sales strategy. The fifth and sixth agents are responsible for the operational and financial sections in the business plan, respectively. Additional agents handle the consolidation, evaluation, and refinement of the final document. Their purpose is to make sure that the output is both comprehensive and professionally structured. This evaluation layer in the application acts as an additional refinement feature, and it is distinct from the final evaluation of the outputs made by human experts or GPT-4.1.

At the core of these agents are their engines, which are LLMs, that perform the core computational work. LLMs are integrated into each agent as the primary text generation engine. When an agent receives a task, this task contains a prompt that encapsulates the relevant context, objectives, and any necessary background information to perform the task. The LLM processes this prompt, and with the help of advanced natural language understanding generates a response that fulfils the agent's goal.

The use of LLMs within agent-based systems offers several advantages. Firstly, it enables agents to handle open-ended, creative, and context-aware tasks that would be challenging for traditional rule-based systems. Certain LLMs are optimized for generating fluent and coherent text and adapting their outputs to specific requirements. Secondly, the integration of LLMs allows for fast iteration and testing. New agents can be introduced with minimal engineering, simply by defining their roles, goals, and prompts, and equipping them with the appropriate

LLM. Thirdly, the prompts of agents can be enhanced with different prompt engineering techniques as mentioned during the literature review of this research. Lastly, LLM-powered agents can be easily updated or enhanced as new models become available.

In practice, the interaction between agents and LLMs is managed through a combination of prompt engineering and tool integration. Tool integration can allow agents to use the LLM's capabilities with specialized functions, such as web searches, web scraping, custom, user defined tool functions.

The collaborative nature of the agentic system is further enhanced by CrewAI's support for inter-agent communication. This enables the system to tackle complex, multi-stage problems that require multiple different expertise and perspectives. The result is a flexible, adaptive, and highly capable agentic system that can generate business plans of good quality, tailored to the specific needs and contexts of its users.

The specific LLMs that are used in this application are utilised through CrewAI's implementation, but parametrization and certain level of modern customization is available. The models used are Google's Gemini 2.5 Pro model for content generation tasks, creating the different sections of the business plans and Meta's Llama 3.3-70B-versatile model for the evaluation and refinement tasks. Google's Gemini 2.5 Pro excels at tackling complex problems, generating structured content especially over long contexts, such as a business plan (*Google Gemini API documentation*). The Llama 3 model is used through the Groq provider, leveraging its speed and efficiency to rapidly evaluate and refine the generated content. Both models' temperature level (that regulates the creativity of the models) is decreased purposefully, to produce a more deterministic and factual response, which is necessary in contexts such as a professional document like the business plan. Temperature is a hyperparameter of models, and its values in the context of LLMs control the randomness in the generated text. Lower values indicate more deterministic and predictable text, while higher values allow for more diversity and creativity in the generated document, often in the form of additions, suggestions and implications (*Wang et al., 2023*).

4.2.2 Tasks

In the architecture of this agentic system, tasks serve as the fundamental units of work that enable the collaborative efforts of specialized agents. Within the CrewAI framework, tasks are not only isolated documents; they are designed for collaboration; therefore, they are contextrich documents that control the flow of information, coordinate agent actions, and shape the quality and coherence of the system's output. The structuring and sequencing tasks are essential for reaching the full potential of multi-agent collaboration, especially in complex objectives such as the task of generating a business plan.

At their core, tasks in CrewAI are defined as discrete, goal-oriented assignments that must be completed by an agent in the system. Tasks provide all the necessary details for execution of a specific description about a problem. Each task is associated with a specific agent that executes it. The CrewAI framework enables tasks to be executed either sequentially or in parallel, depending on the dependencies and logical flow of the overall process. In this implementation the tasks are ordered, which means that they are executed sequentially one after the other. However, an additional important aspect of task execution is prestent in the implementation. To achieve context-awareness and information passing between agents and tasks, context passing is implemented in the application structure. When defining a task's context attributes in CrewAI, it is possible to specify which other task's outputs are to be considered as information and input to the current task. This feature lets the information pass from other parts of the system to the current task, therefore enabling the flow of relevant data throughout the system. During the sequential task execution, subsequent tasks have access to previous tasks' outputs. This results the final document to be more coherent and context-aware due to the context sharing between tasks.

A well-structured task in CrewAI is characterized by several key attributes. Firstly, it must have a clear description that communicates the intended outcome to the agent. This description should provide sufficient context and outline any constraints that must be followed. Secondly, tasks should be given an expected output format, which describes to the agents how should the generated output look like for the end users. It can contain length limits, formatting requirements and output file types. Thirdly, each task should define the agent responsible for its execution, ensuring that it is handled by one with the right expertise. Lastly, tasks may include references to variables, which are pieces of information collected from user input in this implementation. They are essential for contextualizing the assignment and guiding the agents in responding. These input variables are the backbone of the application, because they contain the necessary information given by the users, who filled out the questionnaire. These values are the basis of the whole document creation process. The context sharing in the application enables agents to access variable values throughout the whole domain, therefore they can achieve consistency and coherence in referencing them through the whole document.

```
Business Name: {business_name}
Establishment Year: {start_year}
The reason for establishing the company: {business_reason}
Mission: {mission_vision}
Legal Structure: {legal_structure}
Financial Funding of the company: {financial_funding}
Business Sector the company operates in: {business_sector}
```

Source: Code from the implementation

The code snippet above shows how the input variables from the questionnaire are accessed and referenced in the task descriptions of the agents. They are accessed through string interpolation in the configuration files of the tasks. They are automatically inferred from the input variables that are passed to the crew during the execution of the application at runtime. The agents can access the values of these variables and use them as information about the company given and incorporate them into the generated sections.

The application employs a series of specialized tasks, each corresponding to a distinct section of the business plan. These include the creation of the company summary, product or service description, market analysis, marketing strategy, operating strategy, and financial plan sections. A defining feature of the task design in this application is the integration of prompt engineering techniques, specifically chain of thought prompting and few-shot learning. Chain of thought prompting is employed to encourage agents to reason through complex problems in a structured, multi-step manner. Rather than generating content in a single pass, the agent is guided to break down the assignment into logical components, consider the relationships between different pieces of information, and build a coherent whole that connects all relevant elements. This approach has been shown to significantly improve the quality of outputs in tasks that require deep reasoning and contextual understanding (*Wei et al., 2022*).

```
First, reason step by step to ensure no detail is missed. Follow these stages
while writing your final report:
1. Consider the values of the following variables provided by the user from a
questionnaire:
2. Break down the company summary into the following required components:
        (note: this part depends on which section it is for)
3. Plan the content for each section using bullet points.
4. Then, generate the full professional narrative based on the structured points
above. Do not include the bullet points in the final output. Write clearly,
concisely, and formally, and do not include the reasoning steps in the final
output.
```

Source: Code from the implementation

The code snippet above illustrates a general principle about the tasks of the 6 agents that are responsible for generating sections. This explicitly instructs the model to reason step-by-step when generating the final output, although these reasoning steps are not included in the final document, as they are not needed there. However, as mentioned in Wei et al., (2022), the chain-of-thought method relies on explicitly instructing the model to reason through the problem step-by-step to achieve a better result. The technique in the implementation follows specific stages, breaking down the complex problem of generating a section for the business plan into smaller chunks. It is in a structured format, explicitly stating the instructions in each step, providing more concrete guidance to the language model. Each 6 agent that generate the business plan follow these 4 general steps while creating their final outputs.

Few-shot prompting is another critical technique incorporated into the task prompts. This technique is implemented in a more detailed way. The few-shot technique is applied to guide the language model in generating a more coherent output from an array of information derived from the user-provided variables. The following example from the implementation illustrates the implemented technique:

```
Examples of good and bad usage:
BAD: "The company's value propositions are faster delivery, better quality, lower
prices"
GOOD: "The organization distinguishes itself through its commitment to efficient
delivery, maintaining high quality standards, and offering competitive pricing"

BAD: "The customer characteristics include tech-savvy, urban, young professionals"
GOOD: "The target market consists of young urban professionals who are comfortable
with technology and value convenience"
```

Source: Code from the implementation

One of the hardest challenges in the implementation was to accurately derive and contextualize information from the user inputs, especially when the information came from a pre-defined set of multiple-choice answers. By providing multiple good and bad examples the implementation aimed to achieve higher accuracy in generating a more coherent, better flowing text throughout the whole document.

The combination of chain of thought and few-shot prompting aims to enable the agents to produce outputs that are better aligned logically and coherently and are more interconnected throughout the entire generated business plan.

After generating the separate sections, the consolidator agent combines all the outputs into a single document and passes that to the evaluator agent. This agent has several criteria that must look at when evaluating the text. The code snippet below is taken directly from the prompt of the evaluator agent and shows what kind of evaluation steps it needs to take in reviewing the document:

```
Focus on the following aspects:
1. Coherence and Flow:
Check if the text flows naturally between sections
Verify that ideas are properly connected
Ensure there are no abrupt transitions
2.Professional Writing:
Look for any instances of bullet points or direct listings
Check for speculative language or assumptions
Verify that the tone is consistently professional
3. Variable Usage:
Ensure variables are properly integrated into the narrative
Check that list-type variables are transformed into flowing text
Verify that no variables are directly cited
4. Factual Accuracy:
Verify that all statements are based on provided data
Check for any made-up or assumed information
Ensure no speculative improvements are suggested
DO NOT:
Suggest additions or improvements that would require new information
Look for missing information or gaps in coverage
Recommend adding content that wasn't provided in the questionnaire
Make assumptions about what should be included
```

Source: Code from the implementation

The initial evaluation of the generated output is reviewed by these criteria. These criteria are based on feedback received from the human experts involved in the evaluation of the results, but the implementation of the evaluation method is explained in more detail in the Findings chapter.

4.2.3 Collaboration in a Crew

Collaboration is a defining feature of the agentic system implemented in this application. The CrewAI framework enables to create cooperative groups of agents, that are architected to mimic the dynamics of human teams, where specialized members work together to achieve complex goals.

In this implementation, the crew is composed of agents with defined roles. Each agent is responsible for generating a distinct segment of the business plan, leveraging its domain-specific knowledge and defined tasks to generate content. The collaborative process is created by managing the sequencing of tasks, the flow of information, and passing context between agents to share information through the whole system.

A key aspect of collaboration in this system is the use of shared context and variable inference. Agents do not operate on their own; instead, they have access to relevant data collected from the user and the outputs of other agents. This shared context enables agents to maintain consistency and coherence through separate parts of the business plan. For instance, the marketing expert agent can reference the market analysis produced by the market analyst, ensuring that the marketing strategy is aligned with identified market trends and customer needs.

5. Findings

This section examines the results of the systematic evaluation of the analysis. It examines and compares the evaluation results from the human expert side and from the LLM evaluation based on the metrics defined in the methodology section of this study. The results are presented in accordance with the iterative development and feedback cycle. This means that once the agentic system generated a set of report, the evaluation took place and feedback were given. The system was improved based on the feedback and new reports were generated.

5.1 Results of the evaluation

The evaluation of the results of the agentic system was conducted based on predefined evaluation metrics. Due to the fact that the evaluation of the results of a NLG task is subjective and, in most cases, requires human judgment for accurate evaluation results, a set of evaluation metrics had to be defined. Zheng et al., (2023) and Fabbri et al., (2021) mention in their research that human judgment of AI-generated texts still outperforms the LLM-as-a-Judge evaluation method, however, the latter provides a fast, and efficient method to quickly evaluate the initial results of a system. On the other hand, the external human judgment provides valuable, insightful feedback, although it takes significantly more time. The human review can pick up smaller mistakes, can understand the issues in coherence and overall logical flow of the whole document better, and humans are generally more critical in terms of evaluation than LLMs. This is especially observable when a language model tries to evaluate a generated text, which was produced by themselves. Zheng et al., (2023) examines this topic in detail and calls this phenomenon self-enhancement bias, referring to the practice of the model overestimating the quality or correctness of its own outputs.

As mentioned in the methodology in table 1, the results were measured across 5 metrics on a 1-5 scale, where the score of 1 meant a lower value and 5 meant a higher score. These metrics are connected to the linguistics, coherence, relevance, consistency and content of the generated business plans. Moreover, the structure and the clarity of the whole document had been measured and given a score. These metrics were derived from numerous previous academic research associated with evaluating NLG tasks. The research was also subjective in terms of the data available. The external evaluators provided data sources for fictive companies.

Moreover, they had access to real business plans and the research used one of them to simulate answers to the questionnaire using GPT-4.1. This synthetic data was later used to fill in the questionnaire and generate a new business plan purely based on the answers to the questions. It can be concluded that the quality of the data largely depends on the answers provided on the questionnaire form, because the basis of the agentic system is the information gathered from it.

The initial evaluation was conducted on two generated plans for two fictive companies, however, to gain an additional external validity evidence, the evaluation considered a third, different fictive company, whose generated business plan was evaluated only once, as a supplementary case to validate the influence of different data inputs on the generated plans.

5.1.2 Human expert evaluation

The generated reports undergone thorough examination and evaluation by the two external experts from Aalborg University, Peter Thomsen and Brian Balslev Andersen. They conduct research in the field of business planning for several years, therefore their expertise and knowledge provided valuable insights and feedback to the evaluation of the results of the analysis. They have been asked to evaluate the business plans based on the 5 metrics and give a rating for all aspects, additionally to provide comments for possible improvements. The first business plan was generated based on information provided by them in the form of predefined answers to the questionnaire about a fictive company called “Bake-Off DK”. The following table illustrates the evaluation conducted by them for this company.

Human evaluation of the generated plans for “Bake-Off DK”				
Aspects	1st set of plans		2nd set of plans	
	Score (1-5)	Comments	Score (1-5)	Comments
Relevance	5	Sections contain relevant information	5	Sections contain relevant information
Completeness	4	Includes all key elements, but misses a few details	5	Includes more smaller details than the previous version

Correctness	2	Output is too creative, contains explicit guesses	3	Less creative, still contains smaller hallucinations; mistakes
Consistency	3	Somewhat consistent, should be overall more coherent and capture the connections better	3	Smaller inconsistencies present, it may also be due to the data input
Structure & Clarity	3	Well written and easy to understand, but the structure is too mixed up back and forward between some sections	3	Too much repetition of the value proposition; redundancies in the Company Summary section; less explicit guessing; mentions of smaller, insignificant information; uses of too many synonyms for “the company”

Table 3: Human evaluation of the generated business plans for “Bake-Off DK” fictive company

Besides the comments seen in Table 3, additional, more detailed feedback were given for both versions of the plans. The feedback for the first version addressed the issue of too much creativity in the creation of the sections. Hallucinations, or made-up data were included throughout the document, that were not originated from the data provided. This contained explicit guesses and the use of expressions, such as “likely”, “may”, “suggests”. Furthermore, the model generated suggestions for improvements in the business plan, which were also not included in the input data. These issues pointed to the fact that the model is not deterministic enough, and considers adding too much additional information and suggesting improvements, rather than focusing on the facts and basing the section solely on the input information. After discussing for possible changes, it was concluded that the temperature hyperparameter (responsible for the level of creativity) of the model was set to a too high value of 0.7, it needed to be lowered, therefore it was set to 0.1 to be more deterministic and factual when generating the output (*Wang et al., 2023*). Moreover, the feedback addressed that the document contained the values of the variables directly in the generated text, contributing to a less coherent and less natural use of language. The feedback also mentioned that the coherence of the whole document was problematic at certain parts, because the agents were not able to connect information between separate sections well, that a human writer would have otherwise been capable of doing. The implemented changes for these issues were done in the prompts of the task definitions of the agents, because it is possible to direct the behaviour of the agents and the outcome of the generated text through experimenting with different prompting techniques. The Chain-of-Thought and Few-shot techniques were implemented after receiving this first, initial feedback to try to tackle the main issues and guide the model towards a more accurate representation of a coherent business plan. Additionally, the feedback contained guidance on the need for structural modifications in the document. It mentioned reorganizing certain headings to be more logically correct, and merging sub headers together. It also stated that one part of the business plan could be entirely removed, because they did not find it relevant to be included to provide a better flow of information in the text.

After the changes were implemented, a new set of business plans were generated and provided for evaluation. The results of this evaluation showed improvements overall in the generated plan for “Bake-Off DK”, however, there remained smaller mistakes spotted by the experts. The feedback stated that the plan improved in almost all the metrics for the second iteration. The creativity level decreased significantly; however they were able to spot parts in the resulted text, that contained information which was not originally in the dataset. The feedback detailed these instances, which makes it easier to find a solution for improvements. Noticeably less guessing occurred in the second iteration of plans, but still in some sections there were a few examples of the usage of suggestive phrases, such as “likely” and “suggests”. After this round of evaluation, the experts noticed few inconsistencies and contradictions in the data input for this company and advised us to try to validate the model’s capabilities with a third, new dataset that they have developed, which might contain data that is more consistent and contains more details. According to the opinion of experts, the reason behind this, might have been caused by the inconsistencies of data input of the “Bake-Off DK” dataset, that causes some of the mistakes in the language model’s generated outputs. Lastly, regarding the structure and clarity of the second batch of the “Bake-Off DK” plan, they observed two issues: an extensive use of synonyms for the phrase “the company”, which might be too exaggerated and needed to be modified, and the inclusion of unnecessary information in certain sections, that is not relevant to the operations of the observed company.

The other fictive company was called “The Enchanted Vineyard Bed and Breakfast”. The data used in this case was synthetic, generated by GPT-4.1. The LLM was prompted to act as the business owner of the specified company. It was provided with a the previously written business plan as a reference, and it had to provide answers for all the questions in the questionnaire based on this information. The agentic system utilised these answers in generating the business plan, which was subsequently reviewed by human evaluators.

Human evaluation of the generated plans for “The Enchanted Vineyard Bed and Breakfast”				
Aspects	1st set of plans		2nd set of plans	
	Score (1-5)	Comments	Score (1-5)	Comments
Relevance	4	Addresses the relevant topics aligned with the business plan framework and mirrors the overall intent of the original case effectively	5	Generally, it aligns well with the original case, contains all major sections required for a business plan for a Bed and Breakfast
Completeness	3,5	Presents the core elements, lacks specificity	4,5	More thorough in presenting operational and strategic details
Correctness	2	Several claims are based on assumptions or suggestions not supported by source data; implications for future strategies, which could be misleading	3	Mostly true to the original case, but introduces some misplaced elements in some sections
Consistency	2	Transitions are abrupt, and integration between concepts is weak	4	The narrative flow and internal structure are generally strong, but certain information are presented differently than in the original case

Structure & Clarity	3	Readable and professional style, clarity is weakened by mixed sequencing; overuse of bullet points; inconsistent use of terminology for key business concepts	4	Well-written, follows the format and tone of a "standard" business plan; has a clearer hierarchy and logical structure; but minor redundancies in value propositions are still present and lack personal tone in the narrative

Table 4: Human evaluation of the generated business plans for "The Enchanted Vineyard Bed and Breakfast" fictive company

The first iteration of this business plan received very similar feedback to the one generated for "Bake-Off DK". The evaluators commented that it had the same issues, and the output was overall too creative, and the suggestions made by the language model on how to possibly improve the business was not credible and potentially could be misleading in a professional business plan. However, it was also mentioned in their evaluation, that this business plan was in general noticeably higher quality than the "Bake-Off DK" plan, it is better written, more concise and is a clearer representation of a business plan. Their reasoning for this phenomenon was that the quality of the input information might have been better in this case, that flowed into the agentic system through the questionnaire. On the other hand, it was observed that this generated plan was also a bit too creative at times and needed similar modifications than the other one. The general observation was that seemingly it was close to a business plan, however it would need a firmer grounding in the source material, a more coherent structure and flow of text to increase its reliability and professional tone.

The second iteration of the generated plan for "The Enchanted Vineyard Bed and Breakfast" received improved scores from the human evaluation process in all aspects. Based on the feedback it had an improved structure, contained all major key components, and the content was relevant to each section. The feedback mentions that this version tries to reintroduce the financial context, however not in the same depth as the original business plan. However, as agreed before with the external experts during the coordination phase of the study, the financial aspect of business planning was purposefully omitted both from the questionnaire and from the generation tasks. Furthermore, this version's content is overall better backed up by source data, some misplaced elements were still observable, that were not explicitly stated in the original version. This implies that the language model suggested or recommended those elements. The narrative flow and the general structure of the second version overall improved from the first iteration, although some minor inconsistencies were still present, especially in the customer segmentation section. The evaluation mentions that the product description is more detailed than the marketing strategy, which counts as an inconsistency in the depth of these sections, however, it also mentions that this can partly be explained by the limitations of the questionnaire and the input data. Lastly, there was one exception, where the evaluation of experts emphasizes that the first version was better at incorporating a certain level of personal tone in the narrative, which was lacking in the second version.

The third round of evaluation in the iterative cycle was conducted on a new dataset, provided by the experts that contained information about a new fictive company called "DM Green Keeping". During the consultation and discussion with the evaluators, they suggested that it might be a good idea to try out the system on new data too, to test its capabilities. This dataset was available, and according to their analysis it contained more consistent information about the company than the "Bake-Off DK" case. The decision regarding the new case was that it can be a valuable addition to the study, since it gives a supplementary case besides the two other cases, where the continuous progress and iterative improvement cycle can be observed. This new case therefore provides insight into the system's handling of a new dataset. Table 5 illustrates the evaluation results for the generated business plan for this case.

Human evaluation of the generated plans for "DM Green Keeping"		
Aspects	Score	Comments
Relevance	5	No additional comments
Completeness	5	No additional comments
Correctness	4	2 explicit guesses in the "Market Analysis" section, therefore more concrete wording recommended; inaccurate textual representation of the self-service options compared to the original answer from the dataset
Consistency	5	"Market Positioning" could be deleted from the "Marketing Strategy" part, as it is already mentioned in "Market Analysis"

Structure & Clarity	4	Sub-headers are not necessary in the “Company Summary” section, but earlier redundancies are not present anymore; usage of a few uncommon words; mentions of all value propositions in the text, could be adjusted to group the ones similar in meaning; smaller mistakes regarding the financial context

Table 5: Human evaluation of the generated business plan for “DM Green Keeping” fictive company

The reason for evaluating a supplementary case for this research was that it gives additional, external validation to the system. The two other cases were followed through multiple development cycles, and the system incorporated the feedback and possible improvement steps deducted from the evaluation of the experts. This case provided a new perspective and new dataset for the system to prove its capabilities. The overall evaluation sentiment from the experts revealed that it contained much less redundancies and better structure as before. The generated plan contained every key aspect of the required format, with relevant information inside them. This indicated a better structure of the headings as well. The implemented improvements contained stricter rules and guidelines for the evaluator agent, therefore less suggestive language was used, and the creativity of the contained elements also decreased. From this case it can be concluded that the quality and level of detail in the input data from the questionnaire is crucial to generate better business plans. More detailed answers and consistent input information results a better text in terms of consistency, coherence and flow.

5.1.3 LLM-as-a-Judge evaluation

As mentioned previously in the methodology section, this method of evaluation provides an instant, cost-effective and time-saving way to get an initial evaluation for the generated business plans. The LLM used for evaluation is GPT-4.1, the OpenAI’s latest model. According to Zheng et al. (2023), OpenAI’s models demonstrated the highest alignment with human judgment in evaluating AI-generated text, even though the ratings and depth of analysis were still not at the same level. The following table illustrates the evaluation of the plans generated for the “Bake-Off DK” company.

LLM evaluation of the generated plans for “Bake-Off DK”				
Aspects	1st set of plans		2nd set of plans	
	Score (1-5)	Comments	Score (1-5)	Comments
Relevance	3	Generally, addresses the expected topics, but lacks specificity in certain areas	5	Each section addresses relevant topics
Completeness	2	Several elements are missing, like risk analysis, growth projections, market sizing	3	Lacks depth in certain areas
Correctness	3	Many claims are not supported by data or evidence	4	Most information is plausible, but some claims are not backed with data
Consistency	3	Notable inconsistencies, repeated value propositions	4	Observable inconsistencies and value propositions are repeated across sections
Structure & Clarity	2	The plan follows a logical structure, but often verbose, repetitive and lacks clarity; absence of clear transitions	4	Some transitions are abrupt and the text lacks flow in certain sections

Table 5: LLM-as-a-Judge evaluation of the generated plans for the “Bake-Off DK” fictive company

In the first iteration of the generated plans, in this case the LLM-as-a-Judge evaluation method gave a slightly lower score than the human evaluators did. The evaluation comments are aligned with the human feedback in certain aspects, however, the LLM evaluation was not able to detect the creativity factor, that the humans observed. It correctly identified the inconsistencies in the logical flow of the document, and the absence of clear transitions between sections. For the completeness metric, the evaluator LLM provided suggestions, that certain newly added sections could have been beneficial for the real-world application of the business plan, like growth analysis, market sizing, and SWOT analysis. However, these aspects are not explicitly covered in the questionnaire, therefore it was not the purpose of this study to be included in the generated plans.

The second round of business plans received a slightly higher rating, which was an improvement from the previous iteration, and this trend matches the human evaluators' feedback. The LLM evaluation identified repeated mentions of the value propositions across multiple sections throughout the document, that was also observed by the human experts. The LLM could correctly identify the logical gaps in transitions as well. For the correctness metric, it states that no data is backing up certain claims throughout the sections. This claim might originate from the nature of the input data, since no numerical data is provided in the questionnaire, neither any form of analysis, therefore the backing data is missing.

Overall, the LLM evaluation was able to correctly identify some aspects that the human judgment listed as well, however, it lacked the ability to spot the mistakes related to coherence and the level of creativity in the generated text.

LLM evaluation of the generated plans for "The Enchanted Vineyard Bed and Breakfast"				
Aspects	1st set of plans		2nd set of plans	
	Score (1-5)	Comments	Score (1-5)	Comments
Relevance	4	Generally, addresses the expected topics, but certain areas are underdeveloped	4	Each section addresses relevant topics, but certain sections contain too generic statements
Completeness	2	Several critical elements are missing, namely risk analysis, growth projections, market sizing	3	Same comment as for the first version
Correctness	3	Many claims are not supported by data or evidence, too optimistic at certain areas	4	Some assertions about guest loyalty, market position, and operational strengths are still made without concrete evidence or data

Consistency	3	Notable inconsistencies, repeated value propositions; abrupt shifts in the narrative	4	The narrative flows more clearly than before Minor inconsistencies, for example: ambitious positioning without clear differentiation
Structure & Clarity	3	The plan follows a logical structure, clear headings, but often verbose, repetitive and lacks clarity; absence of clear transitions	4	Improved transitions, clear headings and structure; writing is less repetitive, though some overlap remain

Table 6: LLM-as-a-Judge evaluation of the generated plans for the “The Enchanted Vineyard Bed and Breakfast” fictive company

The comments and evaluation scores are similar to the “Bake-Off DK” plan scores. The LLM correctly identified the absence of transitions and logical mistakes in the overall document. It also observed the phenomenon of the repeated mentions of value propositions through the sections. The mentions of adding critical elements, such as the risk analysis, growth projections or market sizing, and that many claims in the document are not backed up by data are also present in this evaluation as well. The reason for this is the same as for the company’s plans: the missing data from the user input, and the structure and intent of the questionnaire. The LLM states that the plans are too optimistic at certain areas, without supporting evidence. The second iteration of the generated plans also received higher scores for both companies, which aligns with the human judgment.

6. Discussion

This section presents a comparison of the research findings in relation to the two primary research questions that guided this study. The discussion combines the results from both the human evaluation and LLM-as-a-Judge assessment to provide information about the effectiveness and limitations of LLMs in business plan generation.

The first research question asks, "*How can Large Language Models be used to generate effective business plans?*".

The findings demonstrate that one approach of utilising LLMs to generate business plans can be implemented through a multi-agent agentic system.

The implementation of the multi-agent system using the CrewAI framework proved to be a proper approach for business plan generation. The division of the business plan creation process into six specialized sections, each managed by a dedicated agent, demonstrated better performance compared to a generalized approach, as the human evaluation revealed that all major sections of the plans contained relevant information and they were clearly distinguishable from each other. The task allocations and definitions were trying to follow the component structure mentioned in the literature review, namely through Georgetown University Law Center (2020), McKeever (2016), and Cambridge Judge Business School (2020). The findings align with the principle that AI agents perform better when assigned specific, focused tasks rather than general responsibilities (*Juang et al., 2024*).

The sequential workflow design, where each agent's output serves as context for subsequent agents, ensured that information passing between agents was implemented, improving coherence and flow throughout the generated documents.

The implementation of prompt engineering techniques, particularly Chain-of-Thought prompting and few-shot learning, introduced by Wei et al., (2022), and Brown et al., (2020), respectively, enhanced the quality of generated business plans. The evaluation results showed improvement between the first and second iterations, with human expert scores improving across multiple metrics after implementing these techniques. The Chain-of-Thought prompting enabled the models to break down complex business planning tasks into logical components, while few-shot prompting examples provided guideline for professional tone and appropriate detail levels in writing style.

The research revealed a critical dependency on input data quality. The comparison between the "Bake-Off DK" company, which input data was provided by the human experts through the questionnaire platform and "The Enchanted Vineyard Bed and Breakfast", which input data was synthetically generated from a real, reference business plan demonstrated that higher quality input data correlates with better output quality. Human evaluators noted that the second company's business plan was "noticeably higher quality", "better written", and "more concise" partly due to the smaller inconsistencies in the “Bake-Off DK” dataset. Furthermore, the third case which was evaluated, namely the “DM Green Keeping” business plan further reinforced this statement about the influence of the input data’s quality on the results of the system.

The second research question asks, “*What are the benefits and limitations of AI-generated business plans compared to traditional business planning methods?*”
The evaluation results provide insights into both the advantages and limitations of AI-generated business plans when compared to traditional methodologies.
The most significant advantage identified is the reduction in time and resource requirements for creating such a complex document. The system can generate business plans in significantly less time required for traditional methods, making business planning accessible to startups and small businesses that might otherwise overlook this essential step due to resource constraints. The evaluation framework demonstrated that AI-generated plans can be systematically improved through feedback incorporation. The second iteration showed improvements across all metrics for the companies, indicating that the system can be better adapted to the task based on expert feedback. The feedback incorporation into the system does not require the rewriting of the whole document, merely the modifications of the task prompts, or the system architecture. This could mean addition of new agents, tasks or employing external tools.

A significant limitation identified was the tendency for showing creativity and generating hallucinations, meaning made-up data, that were not explicitly present in the input dataset. Human evaluators consistently noted that the system generated "explicit guesses" and included information which was not present in the input data. The lower correctness scores across the iterations highlight this issue’s persistence in the research.
The evaluation revealed that while the system can organize and present information effectively, it lacks the deep business intuition and contextual understanding that experienced human business planners bring to the process. Human evaluators noted issues with coherence and the system's inability to "connect information between separate sections well, that a human writer would have otherwise been capable of doing".

The system's effectiveness is dependent on the quality and completeness of the input data obtained through the questionnaire. The LLM-as-a-Judge evaluation consistently noted the absence of supporting data for claims and the lack of critical elements, for instance, risk analysis, growth projections, and market sizing. These limitations partly originate from the questionnaire-based approach, and the quality of the input data, since some questions are implemented as open-ended, which means that the company, that fills out the questionnaire relies heavily on their creativity and own judgment about what information it provides through the application.

Both human and LLM evaluators identified problems with consistency and repetition. The system showed a tendency to repeat value propositions across sections and demonstrated inconsistencies in narrative flow. Consistency scores across evaluations indicate this as an area requiring significant improvement. In order to mitigate these issues, stricter rules and guidelines were defined in the prompts of the evaluator agents, specifically indicating the persistent issues. This targeted method proved to be successful as the third, supplementary evaluation case showed even better evaluation scores given by the experts.

The research confirms that human judgment remains superior in evaluating business plan quality. Human evaluators were more critical and better able to identify smaller issues such as creativity levels, coherence problems, and logical flow disruptions that the LLM-as-a-Judge method missed. This project supports the conclusion that human oversight remains essential in the business planning process.

The findings imply that AI-generated business plans are best positioned as complementary tools rather than complete replacements for traditional methods. The system is good at providing structured drafts and starting points that can save significant time and effort, but these outputs require human review, refinement, and validation to achieve professional standards.

The research reveals a very important trade-off between speed and quality. While the use of LLMs for content generation offers significant speed and cost-effectiveness advantages, according to this study, it currently cannot match the depth of analysis and contextual understanding that experienced human experts could provide in the field of business planning. This trade-off may be acceptable for early-stage startups seeking basic planning solutions but may be insufficient for more complex businesses.

7. Implications and Contributions

This research contributes to the field of AI-assisted business planning by showing both the potential and current limitations of LLMs in generating structured business plans. The findings reveal that LLMs can reduce the time and resource requirements that prevent many startups from investing in formal business planning, the system was not yet effective enough to replace human expertise entirely.

From a practical view, the research advises for entrepreneurs and startups to leverage AI-generated business plans as starting points rather than final products. The ability of the multi-agent system to produce structurally coherent documents with relevant content across all major sections shows that LLMs could provide access to professional-quality drafts for business plans. However, the issues with hallucination and the tendency to generate unsupported claims requires for human contribution and validation in any practical implementation. Revising these drafts by humans could raise costs for companies, however, the initial expense of creating the documents could be significantly reduced.

The evaluation methodology developed in this research contributes to the assessment frameworks for AI-generated content. The comparison between human expert evaluation and LLM-as-a-Judge approaches reveals differences in evaluation capabilities, with human evaluators showing higher ability to identify minor quality issues such as coherence problems and inappropriate creativity levels. The human evaluators still possess a better understanding of the overall document and the flow of text, which is crucial in identifying even minor issues with the generated texts. This finding has implications for quality checking processes in AI-powered content generation systems, conforming that human review remains essential for maintaining professional outputs.

The research also highlights the importance of input data quality in the generated outputs, a finding that is not limited to business planning tasks, because it has implications on all types of NLG applications, where user input is required to provide a good quality output using an LLM. This has practical implications for how researchers and developers design questionnaires that can collect as comprehensive and as precise information as possible from the users.

The iterative improvement through the feedback incorporation process shows that the agentic system could be improved through systematic evaluation and refinement cycles. However, certain issues, like hallucinations and consistency problems (despite targeted improvements), indicates that some limitations may be more present to current LLM architectures, are not easily addressable through prompt engineering or system design modifications.

The research also contributes to understanding the trade-offs that are appearing in AI-assisted professional services across several domains. The speed and cost advantages are important, the limitation in professional quality indicates that certain use cases may require different approaches. Early-stage startups seeking basic business planning could find AI-generated plans sufficient with minimal human review, while more complex businesses or companies applying for loans or grants require more extensive human refinement to achieve professional outputs.

The research indicates that the most effective approach may not be full automation using LLMs and an agentic system but rather intelligent use that leverages AI solutions while also employing human judgment, evaluation and refinement to reach optimal results.

8. Limitations and Future Research

This research provides insights into the application of LLMs for business plan generation; however, it contains several methodological and scope limitations. The evaluation was conducted using only a few fictional companies, which represents a limited sample size for drawing broader conclusions about the effectiveness of AI-generated business plans across more diverse business contexts.

The evaluation represents another limitation, as the assessment is only focused on the quality of the generated business plans without examining their practical usefulness or effectiveness in real world scenarios. The research does not address whether AI-generated business plans actually contribute to improved business outcomes, and successful funding acquisition. The reason for this limitation is the nature of fictional and synthetic data usage, which does not substitute data provided by real companies. This gap between evaluating the quality of the generated plans and their practical business impact represents an important area requiring further research to gain information about the true value of AI-assisted business planning tools. The results show that in the current state, the outcomes of the agentic system are not suitable to use for applications for loans at financial institutions or to be presented to apply for grants. One other aspect of limitations that is connected to this part, as the financial aspect of the business planning process is missing from the generated outputs, and the questionnaire does not cover it either. While conducting the research and developing the application, a discussion with the external experts revealed that this absence of the financial aspects is intentional, and the implementation of it to the questionnaire would provide complex and thoughtful planning, but it is a consideration for future improvement.

The generated outputs of the system are useful for templates and drafts in their current state, that require further refinement and evaluation by humans. This limits the practical usefulness of the application and requires further testing and improvements to the system.

Future research should prioritize studies to check the real-world performance of businesses using AI-generated plans compared to those employing traditional planning methods. Such comparisons would provide valuable insights into the practical effectiveness of AI-generated business plans and help identify the specific areas where these tools provide real value. A possible implementation for this would be to include real company data for input of the agentic system, which would enhance the ability to measure real-world implications.

Additionally, the inclusion of the financial part would be crucial in future research, to develop and extend the questionnaire, as well as the agentic system with more advanced, extensive coverage for financial aspects. This would greatly boost the real-world impact of the generated plans, as one of the most important deciding factors in applying for loans is financial performance and indicators. The future research could include presenting these extended business plans to financial institutions, to really capture the details of the financial aspects from industry experts and gain advice on improvement areas.

Another possibility for future improvement could be the diverse testing of different LLMs in content generation scenarios in the agentic system and comparing their results using the established evaluation metrics with the help of human expertise. This could give feedback about which LLMs perform better at these tasks.

Another important aspect of future research is the public deployment and visual improvements of the application and the user interface. This would include incorporating a new feature, where the users would be able to edit their generated documents in the application with the help of LLMs and AI (acting like an assistant), and modify the parts that they did not like, to their own tastes.

Finally, the ethical considerations surrounding AI-generated business plans, such as concerns about accountability, transparency, and bias in AI-generated business plans, require further examination. As these tools become more common, setting appropriate guidelines for their use, disclosure requirements, and data handling and storage will be crucial for maintaining trust and effectiveness in AI-powered business planning applications.

9. Conclusion

This research investigated the application of LLMs in business plan generation through the development and evaluation of a multi-agent system designed to help startups and small businesses with the process of creating business plans. The research stated two main research questions about how LLMs can be utilized for business plan generation and what benefits and limitations these AI-generated plans present compared to traditional planning methods.
The results show that LLMs can generate structurally coherent business plans through a specialized multi-agent architecture, with the CrewAI framework, which was proved to be effective for such tasks. The implementation of six specialized agents, each responsible for certain sections of the business plan was developed. Using different LLMs like Google's Gemini 2.5 Pro for content generation and Meta's Llama 3.3 70B versatile for assessment and refinement showed good properties in generating and evaluating reports. Overall, these could illustrate well the value of leveraging different AI tools within the agentic system.
However, the research also revealed some limitations that make it difficult to fully use AI-generated business plans currently in real-life situations. There were issues related to hallucinations, where systems generated made-up text, representing a challenge regarding credibility and real-life usage of the generated reports. The temperature hyperparameter of the LLMs were significantly reduced to prevent this phenomenon from happening, and stricter initial evaluation rules were implemented in the task definitions of the agents. These solutions proved to result in better quality business plans with less hallucinated information, however the issue has not vanished completely, only mitigated to a certain extent. Additionally, the systems had limited ability to naturally flow information across some sections and generate the deep business insights that experienced human planners could provide, although in the third round of assessment this aspect had also undergone improvement due to much stricter rules and guidelines in the prompt definitions.

The comparative analysis between AI-generated and traditional business planning methods reveals a trade-off between efficiency and quality. While AI systems offer advantages in terms of speed, cost-effectiveness, and accessibility, they currently lack the ability to capture the depth in analytics, understanding, and insight that human expertise could provide. This finding reveals that AI-generated business plans are best described as starting points rather than complete solutions and still require human review and refinement to achieve professional levels. An augmented approach, where the users could directly modify the input to the system after the initial draft has been generated for them might be a good solution and a good system architecture. After an initial business plan acting as a starting point, the users would have the opportunity to modify parts and information regarding sections in the plan that they find unpromising. Furthermore, this solution would be even more useful if a business planning expert would review the drafts, and modify the sections for the company, as they possess the expertise, and knowledge about every detail of business planning.

The research detailed the evaluation methodology in the Findings section, which is an additional contribution to the value of the research. Combining human expertise and LLM-as-a-Judge techniques for evaluating the generated documents proved to be effective, and resulted in improved versions through the iterations, although the development of the application could not entirely resolve the mentioned limitations. Another contribution to the field is the synthetic data generation using an LLM from the original business plan for the evaluation. This method proved to be effective, fast and resulted in high quality information extraction from the original business plan, which was later the basis for the agentic system as data input to generate the new, completely AI-generated plan for the company.

In summary, this research demonstrated that LLMs can contribute to business plan generation through a multi-agent agentic system approach, by providing useable drafts or starting points for business owners to create their final reports. The most promising path forward involves using this system as a helper in the creation process of a business plan rather than replacement of human expertise, leveraging AI capabilities while maintaining the expertise, knowledge, and critical thinking that experienced human business planners can provide.

10. References

Abdullah, R. (2020). *Importance and contents of business plan: A case-based approach. Jurnal Manajemen Indonesia, 20*(2), 161-173. https://scispace.com/pdf/importance-and-contents-of-business-plan-a-case-based-3szs1f75jb.pdf

Abrahamsson, P., Salo, O., Ronkainen, J., & Warsta, J. (2002). *Empirical studies of agile software development: A systematic review. Empirical Software Engineering, 13*(2), 197–234. https://www.researchgate.net/publication/222827396_Empirical_studies_of_agile_software_development_A_systematic_review

Barrow, C., Barrow, P., & Brown, R. (2012). *The Business Plan Workbook (7th ed.).* Kogan Page. https://students.aiu.edu/submissions/profiles/resources/onlineBook/d5a8R6_business%20plan%20workbook-2012.pdf

Bayat, S., & Isik, G. (2023, July 10–12). *Assessing the efficacy of LSTM, Transformer, and RNN architectures in text summarization*. 5th International Conference on Applied Engineering and Natural Sciences (ICAENS), Konya, Turkey. https://www.academia.edu/105718971/Assessing_the_Efficacy_of_LSTM_Transformer_and_RNN_Architectures_in_Text_Summarization

Brinckmann, J., Grichnik, D., & Kapsa, D. (2010). *Should entrepreneurs plan or just storm the castle? A meta-analysis on contextual factors impacting the business planning–performance relationship in small firms.* Journal of Business Venturing, 25(1), 24-40. https://www.sciencedirect.com/science/article/abs/pii/S0883902608001109

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). *Language Models are Few-Shot Learners.* Advances in Neural Information Processing Systems (NeurIPS), 33, 1877–1901. https://arxiv.org/abs/2005.14165

Cambridge Judge Business School. (2020). *How to write a business plan*. https://www.jbs.cam.ac.uk/wp-content/uploads/2020/08/how-to-write-a-business-plan.pdf

Cao, H., Driouich, I., Singh, R., Thomas, E. (2025). *Multi-Agent LLM Judge: automatic personalized LLM judge design for evaluating NLG applications.* https://arxiv.org/pdf/2504.02867

Delmar, F., & Shane, S. (2003). *Does business planning facilitate the development of new ventures?* Strategic Management Journal, 24(12), 1165–1185. https://sms.onlinelibrary.wiley.com/doi/10.1002/smj.349

Fabbri, A., R., Kryscinski, W., McCann, B., Xiong, C., Socher, R., Radev, D. (2021). *SummEval: Re-evaluating Summarization Evaluation.* Salesforce Research. https://arxiv.org/pdf/2007.12626

Georgetown University Law Center. (2020). *Elements of a business plan*. https://www.law.georgetown.edu/wp-content/uploads/2020/08/Elements-of-a-Business-Plan.pdf

Google Gemini API documentation. https://ai.google.dev/gemini-api/docs

Hiroko, N., & Tomoki, S. (2025). *Is Business Planning Useful for Entrepreneurs? A Review and Recommendations.* https://www.mdpi.com/2673-7116/5/1/10

Hu, Z., Wang, C., Shu, Y., Paik, H. Y., Zhu, L. (2024). *Prompt perturbation in retrieval-augmented generation based LLMs*. Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (pp. 1119-1130). https://arxiv.org/abs/2402.07179

Jain, N. (2023, September 8). *What is a research design? Definition, types, methods and examples*. https://ideascale.com/blog/what-is-research-design

Juang, S., Cao, H., Zhou, A., Liu, R., Zhang, N., Liu, E. (2024). *Breaking the mold: The challenge of large scale MARL specialization*. https://arxiv.org/pdf/2410.02128

Kerlinger, F N (1986). Foundations of Behavioural Research. New York: Holt Rinehart and Winston.

Kojima, T., Shane Gu, S., Reid, M., Matsuo, Y., Iwasawa, Y., (2023). *LLMs are Zero-Shot Reasoners.* https://arxiv.org/pdf/2205.11916

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., Zhu, C. (2023). *G-EVAL: NLG Evaluation using GPT-4 with Better Human Alignment.* Microsoft Cognitive Services Research. https://arxiv.org/pdf/2303.16634

Mahabub, S., MD Hossain, R., & Snigdha E. Z. (2025). *Data-Driven Decision-Making and Strategic Leadership: AI-Powered Business Operations for Competitive Advantage and Sustainable Growth. Journal of Computer Science and Technology Studies,* 326-336.

McKeever, M. (2016). *How to Write a Business Plan.* http://livre2.com/LIVREE/E1/E001026.pdf

Miller, J. K., & Tang, W. (2025, May 13). *Evaluating LLM metrics through real-world capabilities*. University of Sydney. https://arxiv.org/pdf/2505.08253

Naveed, H., Khan, A.U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., Mian, A. (2024). *A Comprehensive Overview of LLMs*. https://arxiv.org/abs/2307.06435

Official CrewAI documentation. https://docs.crewai.com/

Official FastAPI documentation. https://fastapi.tiangolo.com/

Official Streamlit documentation. https://docs.streamlit.io/

OpenAI. (2024). *GPT-4 Technical Report.* https://arxiv.org/abs/2303.08774

OpenStax. (2019). *Entrepreneurship*. OpenStax CNX. https://openstax.org/books/entrepreneurship/pages/11-4-the-business-plan

Osterwalder, A., & Pigneur, Y. (2010). *Business Model Generation: A Handbook for Visionaries, Game Changers, and Challengers.* *https://vace.uky.edu/sites/vace/files/downloads/9_business_model_generation.pdf*

Parsons, N. (2024, August 1). Do you need a business plan? Scientific research says yes. Bplans. https://www.bplans.com/business-planning/basics/research/

Romero Lauro, Q., Gautam, A., Kotturi, Y. (2025). *BizChat: Scaffolding AI-Powered Business Planning for Small Business Owners Across Digital Skill Levels.* https://arxiv.org/html/2505.08493v2

Russell, S. J., & Norvig, P. (2022). *Artificial Intelligence: A Modern Approach* (4th ed.). Pearson.

Scarborough, N. M., & Cornwall, J. R. (2018). *Essentials of entrepreneurship and small business management* (9th ed.). Pearson. https://students.aiu.edu/submissions/profiles/resources/onlineBook/W8z3L4_Essentials_of_Entrepreneurship_and_Small_Business_Management-_8.pdf

Topal, M. O., Bas, A., & van Heerden, I. (2021). *Exploring transformers in NLG: GPT, BERT, and XLNet*. Proceedings of the International Conference on Interdisciplinary Applications of Artificial Intelligence (ICIDAAI), 20–22. https://arxiv.org/abs/2102.08036

Tran, K.-T., Dao, D., Nguyen, M.-D., Pham, Q.-V., O'Sullivan, B., & Nguyen, H. D. (2025). *Multi-agent collaboration mechanisms: A survey of LLMs*. https://arxiv.org/html/2501.06322v1

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention is all you need*. Advances in Neural Information Processing Systems, 30. https://arxiv.org/abs/1706.03762

Wang, C., Liu, S., X., Awadallah, A., H. (2023). *Cost-Effective Hyperparameter Optimization for Large Language Model Generation Inference*. https://arxiv.org/pdf/2303.04673

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E.H., Le, Q.V., Zhou, D. (2022). *Chain-of-thought prompting elicits reasoning in LLMs*. 36th Conference on Neural Information Processing Systems (NeurIPS 2022). https://arxiv.org/abs/2201.11903

White, R. M. (2020). *The entrepreneur's manual: Business start-ups, spin-offs, and innovative management*. Echo Point Books & Media, LLC.

Wieringa, R. J. (2009). *Design Science as Nested Problem Solving. In Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology (pp. 1-12).* Article 10.1145/1555619.1555630 Association for Computing Machinery. https://research.utwente.nl/en/publications/design-science-as-nested-problem-solving

Xiao, T., Zhu, J. (2023). *Introduction to Transformers: an NLP Perspective.* https://arxiv.org/abs/2311.17633

Zheng, L., Chiang, W-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., P., Zhang, H., Gonzalez, J., E., Stoica, I. (2023). *Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena*. https://arxiv.org/pdf/2306.05685

Appendix A.

Detailed list of components of a business plan

1. Executive Summary
    - 1.1 Mission
    - 1.2 Objectives

1.3 Keys to success

2. Company Summary
    - 2.1 Company ownership
    - 2.2 Start-up summary
    - 2.3 Locations and facilities
3. Products & Services
    - 3.1 Description
    - 3.2 Competitive comparison
    - 3.3 Technology
    - 3.4 Product and service design and development
4. Market Analysis
    - 4.1 Market and industry trends
    - 4.2 Market segmentation
    - 4.3 Target market
    - 4.4 Competition

4.5 Buying patterns

5. Strategy and Implementation Summary
    - 5.1 Competitive analysis
    - 5.2 Marketing and Sales strategy
    - 5.3 Sales forecast
    - 5.4 Milestones
6. Operations & Management
    - 6.1 Management team
    - 6.2 Management team gaps
    - 6.3 Personnel plan
    - 6.4 Operational plan
7. Sustainability & Green Initiatives

7.1 Current sustainability efforts

7.2 Challenges in green transformation

7.3 Motivators for future green initiatives

8. Financial Plan

8.1 Important assumptions

8.2 Break-Even analysis

8.3 Projected Profit and Loss

8.4 Projected cash flow

8.5 Projected Balance Sheet

8.6 Business ratios

8.7 Exit strategy

9.Appendices

Questions presented in the original questionnaire provided by the external experts

Relevant Questions to company summary section:

- What is the name of your company?
- In what year was your company established?
- Kindly describe in maximum 500 characters why your company was established!
- Please state your company's long-term goal or vision!
- What type of business is your company?
- How is your company currently financed?
- Please describe the key people in your company, their positions, and core competencies.
- Which industrial sector does your company operate in?
- Please specify which country your company's primary market will be in the short-term (1-2 years).

Relevant Questions to products and services section:

- Please write a maximum of 500 characters about the products or services that the company offers to customers!
- Please specify what characterizes the product range of your company!
- Is product/service development centralized or decentralized?
- Please specify what characterizes the groups of end-consumers (private individuals, companies, etc.)!

- What type of business is your company? (e.g., raw materials, services, IT, etc.)

Relevant Questions to market analysis section:

- Name of your most relevant customer segment.
- Demographics of this customer segment (e.g., age, location, income level).
- Characteristics of this customer segment (e.g., needs, preferences, behaviours).
- How many customers does this segment have?
- Please briefly describe the problems or challenges that your company is trying to solve for the customer group.
- Please indicate and name the three biggest competitors in relation to your company's sales to this customer group.
- Please indicate the intensity of the competition in the market.
- How are the prices of your company's products/services compared to that of the competitors?
- Is the market best described as a niche market or a mass market?
- Please indicate this customer group's purchasing power.
- How easy is it for customers to switch to other providers of similar products/services?
- To what extent are online communities used to exchange information and solve the challenges of this customer group?
- To what extent is this customer group involved in the design or development process of products and services?
- How often does this customer group pay for after-sales services?
- How is the relation with this customer group in general?

Relevant Questions to strategy and implementation section:

- Which competitive parameters does your company excel at towards the customer group?
- What are the most important value propositions towards the private end-consumers.
- Which type of channels does the company use towards its customer group?
- How often is the customer group offered self-service and automated processes?
- What degree of personal assistance is offered?
- What is the price on the package solution compared to buying the individual products/services separately?
- To what extent are the prices for the customers negotiable?

- Please specify what determines the fixed/dynamic prices for the customers!
- How can your company's primary revenue from its customer group(s) be characterized?
- Please describe how you plan to use the requested funding (if applicable).

Relevant Questions to operations and management section:

- Please select the three most important material resources for your company to create/deliver value to customers!
- Please select the three most important intangible resources that your company can use to create/deliver value to customers!
- Please select the three most important activities for your company to create/deliver value to customers!
- Please select all the activities that are performed in-house!
- Please select all the activities that are outsourced!
- Please indicate if any of the following statements apply to your company (e.g., crowdfunding, white label, customer club partners)!
- Please select the three most important strategic partners of your company to create/deliver value to customers!
- What benefits does your company derive from cooperation with its three main partners?
- How dependent is your company on its collaboration with a specific company?
- Please select the three most cost-intensive components of your company!
- Please mark what technologies are actively used in your company!
- Please specify the other type of technology (if applicable)!
- Please rate the intensity of technological changes in the market!
- Please describe the key people in your company, their positions, and core competencies!

Relevant Questions to financial plan section:

- How is your company currently financed?
- If the business plan is used to apply for funding, please specify the amount that you apply for (in Danish Kroner).
- Please describe how you plan to use the requested funding!
- Please now select the three most cost-intensive components of your company!
- What is the price on the package solution compared to buying the individual products/services separately?
- To what extent are the prices for the customers negotiable?

- Please specify what determines the fixed/dynamic prices for the customers!
- How can your company's primary revenue from its customer group(s) be characterized?