Crew-Driven Business Plan Generation using LLM-Based Multi-AI Agents

Catherine He
University of Southern California
hecy@usc.edu

Kevin Gerges
University of Southern California
kgerges@usc.edu

1 Introduction

The objective of this project is to develop a system that can autonomously generate high-quality, comprehensive business plans. A business plan is a written document that functions as a roadmap, describing the goals, strategies, business structure, milestones, and future plans of a company (Staff, 2024). By leveraging a multi-agent architecture using large language models (LLMs) that are specialized in different domains such as market research, competitive analysis, and financial projections, our goal is to create a structured, consensus-driven approach to business plan generation.

To achieve this, we introduce the Crew-Driven Business Plan Generator (CBPG), a collaborative, multi-agent system wherein each LLM agent provides domain-specific expertise. The output is a cohesive business plan that integrates the feedback from each agent, which can be evaluated against a custom Business Plan Rubric (BPR), a metric that considers dimensions such as thoroughness, feasibility, and quality. We hypothesize that by integrating diverse, domain-specific knowledge, CBPG will produce more comprehensive and realistic business plans compared to those generated by single-agent LLMs.

Given the scope and depth of domain expertise required, today's business plans are often manually created, which can be both time- and labor-intensive. This can be sub-optimal when opportunities are time-sensitive, and in competitive industries that move very fast. Hence, we are interested in *automatic* solutions for business plan generation. One trivial solution would be to use a single-agent approach, in which a single LLM is used to handle all aspects of the business strategy. Yet this leads to several problems: single agent LLMs often lack the depth needed for high-quality outputs in specialized fields, resulting in less coherent and poorer plans. Their generalist nature leads to broad, non-specific responses that fall short of the precision necessary for business plans, which may undergo rigorous human inspection and evaluation. Additionally, single-agent LLMs may hallucinate (Wang et al., 2023), producing unreliable information that is not grounded in strategic context and may be factually incorrect. Lastly, these models lack a structured feedback mechanism, a key limitation when addressing complex, multifaceted tasks.

CBPG introduces a novel multi-agent orchestration framework in which domain-specific LLM agents collaborate to produce comprehensive, structured business plans. Each agent functions as a specialist within a specific domain, enriching the plan with relevant, in-depth information. A central aggregator agent coordinates this process, delegating distinct tasks to each domain-specific agent and synthesizing their outputs into a cohesive, logically structured plan. We hypothesize that compared to a single-agent baseline, a multi-agent framework would be able to create business plans with higher BPR scores, generating outputs that are scored as more realistic, domain-relevant, and practical.

If CBPG outperforms the single-agent LLM approach, it has the potential to streamline and democratize the business planning process, particularly for organizations without specialized expertise or resources. CBPG could empower more informed ventures, especially among startups, by generating well-rounded plans that are aligned with market demands and regulatory guidelines. CBPG would provide value to startups, investors, and consultants by enabling the rapid development and evaluation of robust business strategies. Additionally, it could support established companies in shaping future initiatives and identifying growth opportunities.

2 Related Work

Our project draws upon multiple areas of research: multi-agent systems, LLM collaboration, and in-context learning.

2.1 Multi-Agent Systems and LLM Collaboration

Past work on multi-agent systems (MAS) has demonstrated that task complexity can be effectively managed through role specialization and agent collaboration. With the increasing capabilities of LLMs, recent approaches explore how agents in a MAS could integrate LLMs within its control loop to communicate and self-adapt using natural language. For example, Nascimento et al. (2023) introduced an innovative architecture that integrates GPT-4 into multi-agent systems, enabling agents to adapt to and execute complex tasks while exhibiting advanced communication capabilities. Forms of agent communication are also important to enhance the performance of LLMs in MAS. Zhuge et al. (2023) introduce the "Society of Mind" framework applied to modern LLMs, where multiple models execute tasks collaboratively using structured natural language interactions such as debate, feedback, and consensus. Further, Rasal and Hauer (2024) proposed a novel multi-agent communication framework that employs multiple LLM agents, each with a distinct persona, engaged in role-playing communication to individual subproblems that aggregate into an overarching complex problem.

This technique has proven effective in enhancing task quality and reliability, as structured interactions allow agents to refine output through iterative consensus building. This finding supports our adoption of the CrewAI architecture, which organizes agents as a coordinated 'crew' with specialized roles to optimize task performance. Masterman et al. (2024) further highlight that vertical multiagent architectures reduce cognitive load by delegating specific, complementary tasks to agents with distinct roles. High-quality task execution, they suggest, depends on the lead agent (in our case, the central aggregator) adhering to clear protocols for information sharing with domain-specific agents. Enforcing a structured hierarchy improves coherence and maintains the focus of each agent on its specialized area.

In addition, some studies have explored the use of LLMs in specific domains, such as manufacturing (Lim et al., 2024) and software development (Du et al., 2024). These studies show that LLMs can be used to construct specialized knowledge bases and tools, and to enable agents to adapt to and execute complex tasks in these domains. These findings inspire the decision to integrate domain-specific tools, or functions that agents use to execute their assigned tasks, including retrieval-based actions, web scraping, and more.

Overall, the integration of LLMs into MASs has the potential to revolutionize natural language processing and enable more sophisticated communication between agents, improved adaptability in dynamic environments, and more robust problem-solving capabilities.

2.2 In-Context Learning

Our approach also takes advantage of in-context learning. Brown et al. (2020) and Liu et al. (2021) demonstrate that massively over-parameterized language models, such as GPT-3, can perform effectively in few-shot settings, wherein by introducing a few examples in-context, the LM can be guided in a task-specific manner, without the need for additional training. Few-shot prompting enables agents to tailor their outputs using just a few examples, thereby enhancing task accuracy and relevance. Alternative approaches, such as domain-specific fine-tuning, are computationally intensive (requiring parameter weights to update), but few-shot performance has been found to match up to full fine-tuning with model scale (Brown et al., 2020). For tasks in business domains, where large labeled datasets are scarce or proprietary (and thus challenging to obtain), prompt-based learning offers a highly efficient alternative. This approach allows models to harness general world knowledge acquired from pre-training and quickly adapt to diverse tasks through carefully crafted prompts and minimal examples (Liu et al., 2021).

Specifically, in-context learning to assign *personas* is a rapidly growing area of research, with various approaches and techniques being explored to elicit and leverage diverse personas encoded in large language models (LLMs). Recent work has shown that LLMs can assume diverse personas or behaviors, spanning a wide spectrum of personality traits, political views, moral beliefs, and more (Ge et al., 2024). One approach to learning to specify personas in-context is via the Persona In-Context Learning (PICLe) framework, which elicits the target persona by few-shot prompting with carefully-selected demonstrative examples that maximize the likelihood of the persona (Choi and Li, 2024). This framework is shown to be effective in eliciting diverse behaviors and personas from LLMs such as

LlaMa-2, Vicuna, and GPT-J (Choi and Li, 2024). Another related work is the use of event schemas to augment an agent's persona and condition an LLM to generate narrative-like responses consistent with these schemas through in-context prompting (Kane and Schubert, 2023). This approach is effective in capturing habitual knowledge and eliciting personas that are consistent with the target persona.

Altogether, these approaches inform how we approach the design of the prompts for our agents to define personas, specifically as experts in key business domains.

2.3 Business Plan Quality

Nunn (2010) emphasizes the importance of a high-quality business plan, which ought to clearly communicate the research, strategy and planning behind a service or good, which serves as a critical step to winning over potential investors and ensuring the long-term success and viability of a business. The authors identify three forms of planning as key considerations: strategic, operational, financial budgeting; they emphasize that the Executive Summary, an overview of the key aspects of the plan, is the most important element.

We integrate these ideas into our formulation of the BPR, which will be expanded upon later.

3 Methods and Experiments

We hypothesize that a multi-agent large language model (LLM) system composed of multiple domain-specific agents will generate business plans of higher quality than a single-agent LLM system with zero-shot prompting conditions. This hypothesis holds true if business plans produced by the multi-agent system, augmented with few-shot prompting, achieve significantly higher scores on the Business Plan Rubric (BPR) and are preferred in a majority of blind comparative evaluations relative to plans generated by the single-agent model.

Testable Truth Condition: The multi-agent system achieves a statistically significant improvement in BPR scores ($p < 0.05$) and majority preference in blind evaluations.

Independent Variables include the **number of agents** in the system (multi-agent or single-agent) and the **prompting strategy** (few-shot or zero-shot).

Dependent Variables include the **business plan quality**, as measured via the Business Plan Rubric (BPR), which evaluates content-specific and style-specific attributes critical to effective business plans, as well as **blind comparative preference ratings** from human evaluators.

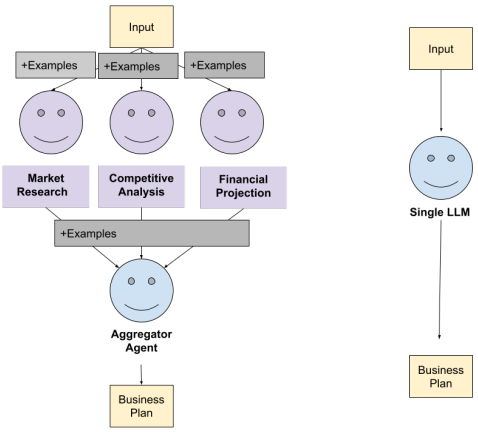

Figure 1: Visualization of Experiment A (CBPG) and Experiment B (baseline)

3.0.1 Definition of Business Plan Quality

Business Plan Quality is defined by the clarity, relevance, professionalism, and strategic depth of the plan. High-quality plans are data-driven, actionable, and tailored to the target audience. They demonstrate clear alignment between the user's stated goals and operational feasibility, and communicate a compelling vision and value proposition. Critical sections are identified to be the executive summary, market research, implementation plan, competitive analysis, and financial projection.

We design two experiments to test this hypothesis: **Experiment** A, which utilizes the CBPG approach incorporating a multi-agent LLM system with domain-specific agents, and **Experiment B**, which employs ChatGPT-4 as a single-agent LLM using zero-shot prompting. A simplified visualization of the two experimental setups is in Figure 1.

In the following sections, we provide detailed information about the implementation and methodology for each experiment.

3.1 Experiment A

Experiment A leverages the CBPG system, which consists of four domain-specific agents and one central aggregator agent. Each agent is assigned a specialized role: market research, competitive analysis, financial projection, and business plan generation. The aggregator agent synthesizes the outputs from these specialized agents into a coherent and actionable business plan. Domain-specific prompts tailored to each agent's task are utilized, supple-

mented with few-shot examples to enhance contextual understanding and task performance. Agents are further equipped with advanced tools, such as retrieval-augmented generation (RAG), to improve the quality and relevance of their outputs.

We implement CBPG using CrewiAI[1], a state-of-the-art framework designed to orchestrate multi-agent collaboration. Its architecture enables the creation of autonomous agents with clearly defined roles, goals, and customizable backstories, ensuring effective specialization for task execution. Additionally, the framework provides robust process management capabilities, supporting efficient coordination among agents. Throughout this paper, we utilize several key concepts from the CrewAI framework to guide our implementation, including Agent, Task, Tool, Crew, and Process. We chose CrewiAI over alternatives such as AutoGen, OpenAI Swarm, and LangChain because of its intuitive and simple framework, compatibility with open-source models, and role-based agent design.

Agent. These are autonomous LLM-based entities within the system, each designed to perform specific roles and achieve defined goals. Modeled to mimic real-world team members, agents possess distinct expertise and responsibilities. In CrewAI, agents are characterized by:

1. **Prompts**: Defined as Role, Goal, and Backstory to specify the agent's function, objectives, and optional context for enhanced task behavior.

2. **Tools**: Specialized functions available to the agent for executing its tasks effectively.

Tasks. Units of work assigned to agents. Each task is defined by:

1. **Prompts**: Descriptions and expected outputs that specify the task objectives and deliverables.

2. **Context**: Dependencies on outputs from other tasks to enable a collaborative workflow.

3. **Assigned Agent**: The specific agent responsible for executing the task.

A **Crew** represents the overall structured workflow of all Agents and their respective Tasks. The **Crew** runs a **Process**, which is the execution strategy, manages task assignment, and facilitates agent interactions.

[1]`https://www.crewai.com`

In summary, the **Crew** organizes the system's operation by coordinating domain-specific **Agents** to complete their specialized **Tasks** through collaborative **Process**, working towards the generation of a complete business plan.

3.1.1 System Architecture

Each agent in our system leverages `Meta-Llama-3-8B` (Grattafiori et al., 2024), a fully open-sourced LLM, to perform specialized tasks within their respective domains[2]. Sample prompts are included in Appendix A. Our **Domain-Specific Agents** are as follows:

1. **Market Research Agent**. This agent provides insights into market dynamics, including growth projections, segmentation, and market readiness. It identifies trends, opportunities, and risks using advanced analytical tools.

2. **Financial Planner Agent**. Responsible for developing financial projections, the Financial Planner Agent aims to perform cost analysis, revenue forecasting, break-even calculations, and sensitivity analysis. Expected outputs include detailed financial forecasts, funding strategies, and profitability metrics.

3. **Competitive Analysis Agent**. This agent evaluates competitor strategies, strengths, and market positioning through SWOT analysis and semantic parsing.

4. **Business Plan Aggregator Agent**. Integrates outputs from the domain-specific agents into a cohesive **business plan**. This includes crafting an **executive summary**, which highlights the main points of the plan, and an **implementation plan**, which provides a detailed strategy for achieving proposed objectives. Additionally, the Aggregator ensures document clarity, validates insights, and formats the business plan for readability.

To further boost the performance of the agents, we also use the **Retrieval-Augmented Generation (RAG) Tool**. We implemented the `DynamicRAGTool`, a real-time RAG-based function for the Market Research, Financial Planner, and Competitive Analysis agents. Unlike traditional RAG systems that rely on offline knowledge bases,

[2]Our codebase can be found at: https://github.com/kevingerges/Prodhunt

this tool integrates the the Serper search API for real-time web searches and web scraping to dynamically construct a domain-specific knowledge base.

1. **Workflow**: The tool generates search queries based on user-provided industry, competitive, and financial context. Text queries are encoded into dense vector embeddings using the `all-MiniLM-L6-v2` SentenceTransformer model, enabling similarity computations for relevance scoring. Retrieved content is semantically filtered with a cosine similarity threshold of 0.6 and categorized into structured insights across market trends, competitor analysis, and general data. This enables higher alignment with user's input, and the output also provides links and sources to substantiate claims, enhancing transparency and credibility.

2. **Fallback Mechanism**: When live retrieval fails, fallback options such as local embeddings or pre-defined default values ensure uninterrupted workflows.

We also leverage the **Market Research Analysis Tool**. The `MarketAnalysisPipeline` is designed to integrate sentiment analysis and semantic similarity matching for market evaluation, and executes four analytical processes: market sentiment evaluation, trend identification, opportunity detection, and risk assessment.

1. **Sentiment Analysis**. The system employs HuggingFace's `sentiment-analysis` pipeline to classify textual inputs into sentiment labels (e.g., POSITIVE or NEGATIVE). These results inform actionable insights, such as identifying promising market segments or potential challenges. For instance, strong positive sentiment could imply an area for entry.

2. **Semantic Search and Text Similarity**. Using `Sentence-BERT (paraphrase-MiniLM-L6-v2)`, the system encodes textual data into vector embeddings to calculate semantic similarity. Rule-based keyword matching combined with contextual sentence parsing detects market trends, opportunities, and risks. For instance, to identify **trends**, key words such as "growth" or "decline" are matched with semantically similar web search results to detect market patterns. To identify **opportunities and risks**, phrases such as "growing" (opportunity) or "challenge" (risk) are used. Semantically similar words are determined using a threshold of 0.6.

This approach ensures a structured, automated market analysis, allowing agents to extract reliable and context-rich information for generating actionable insights.

3.1.2 Post-Processing

Readability. To ensure coherence and accessibility, the system employs custom content processors to structure outputs effectively. These processors format the generated text into organized sections with headers, bullet points, and structured financial data, enhancing clarity and user readability.

Automated Quality Metrics The system automatically assesses the quality of its outputs using a combination of `spaCy` for linguistic analysis and the Transformers pipeline for sentiment evaluation. A validation process ensures the presence of critical sections: executive summary, market analysis, financial projections, competitive analysis, and implementation plan. Each section is evaluated against automated quality metrics across four dimensions: readability, business terminology, vocabulary diversity, and sentiment.

1. **Readability**: Assessed based on sentence structure, document flow, and formatting. Thresholds for readability include flagging overly complex sentences (> 25 words) and insufficient detail (< 10 words per sentence). These metrics ensure well-balanced and clear documentation.

2. **Business Terminology**: Evaluates the inclusion of key business-related terms such as "market," "revenue," and "strategy" to verify alignment with professional business standards.

3. **Vocabulary Diversity**: Measures the variety of words and phrases used, aiming to reduce repetition and promote a diverse lexicon.

4. **Sentiment Analysis**: Ensures an appropriate tone, prioritizing formal language while flagging excessive negativity, which could indicate a pessimistic outlook on the business plan's feasibility.

The system generates quality scores for each dimension, resulting in a composite score ranging from 0 to 1. These metrics serve as internal feedback for developers, guiding improvements and iterative refinements of the system's outputs.

3.2 Experiment B

Experiment B evaluates the performance of a single large language model (LLM) via non-iterative, zero-shot prompting. In this setup, the LLM assumes responsibility for all tasks, including market research, competitive analysis, financial planning, and business plan synthesis, simulating an end-to-end business plan generator.

We employ zero-shot prompting, relying solely on the model's pre-trained knowledge base without the addition of task-specific examples or contextual refinements. A singular prompt is crafted to include a high-level Role, Backstory, and Task description, but with no supplemental few-shot examples.

Experiment B utilizes OpenAI's ChatGPT, which is based on GPT-4o and is many times larger and generally more powerful than LlaMa 3.1 (Grattafiori et al., 2024). The absence of iterative processes or specialized domain agents reflects the distinction between this configuration and the multi-agent system in Experiment A.

3.3 Evaluation Methodology

To ensure robustness, each experiment is repeated ten times, generating a business plan for a unique evaluation set of ten input queries per iteration. The input queries vary across industries, scales, markets, and other factors to capture the systems' generalizability and adaptability under diverse conditions. In total, twenty business plans are generated across both experiments.

3.3.1 Business Plan Rubric

The Business Plan Rubric (BPR) serves as the primary evaluation framework, assessing the quality of the generated business plans. Developed based on prior research and expert recommendations, the rubric comprises six categories critical to business plan quality. Each category is scored on a scale of 0 to 10, yielding a maximum total of 60 points. Final scores are normalized to a range of 0 to 1. The rubric categories are:

- **Executive Summary**: Evaluates clarity, specificity, and alignment with business objectives.
- **Market Research**: Assesses the depth, relevance, and actionable insights provided.
- **Implementation Plan**: Measures the feasibility, clarity, and practicality of proposed steps.
- **Competitive Analysis**: Examines understanding of competitors and differentiation strategies.
- **Financial Projections**: Evaluates the logic and practicality of revenue, cost, and profitability estimates.
- **Styling & Presentation**: Focuses on organization, coherence, and visual appeal.

Rubric Formulation: The BPR was formulated using published research and expert advice in business strategy. The authors of this paper independently scored the twenty generated plans based on this rubric.

Statistical Analysis: A paired t-test is employed to compare the normalized scores of business plans generated in Experiment A (multi-agent system) and Experiment B (single-agent model). This statistical method is appropriate for determining whether the mean difference between two related groups is significant, particularly when the population standard deviation is unknown. Statistical significance is determined at $p < 0.05$.

1. Null Hypothesis (H_0): No significant difference in business plan quality exists between the multi-agent system and the single-agent model.
2. Alternative Hypothesis (H_a): The multi-agent system produces significantly higher-quality business plans compared to the single-agent model.

3.3.2 Blind Comparative Analysis

To minimize bias, the business plans generated by both setups are anonymized and randomly ordered (within each pair) in the blind comparative analysis setup. Human reviewers assess the plans without knowing the source system, ensuring an unbiased comparison. Three volunteer reviewers, all anonymous students from the USC Marshall School of Business, independently evaluated ten pairs of business plans. Instead of assigning scores, reviewers conducted a pairwise preference evaluation, selecting the superior plan in each pair. This method

increases the efficiency of evaluation while maintaining reliability.

Evaluation Protocol: Each reviewer was instructed to examine the two anonymized plans in a pair and select the higher-quality document based on clarity, relevance, professionalism, and strategic depth. The reviewers were asked to provide a brief 1–2 sentence rationale for their choice. The percentage of preferences for the multi-agent system over the single-agent model serves as an additional measure of relative performance.

We describe an end-to-end workflow in Appendix B.

4 Results and Findings

Refer to Appendix C for sampled generated outputs from Experiment A and Experiment B.

4.1 Business Plan Rubric Score Results

We evaluated the statistical significance of the Business Plan Rubric (BPR) scores for both systems; results are in Table 1.

4.2 t-Test Calculation

To determine whether the difference in mean BPR scores between the CBPG system and GPT-4 is statistically significant, we conduct a paired t-test. The t-test formula is:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where $\bar{X}_1, \bar{X}_2$ are sample means, s_1^2, s_2^2 are variances, and n_1, n_2 are the sample sizes (both equal to 10 here).

Recall that the null hypothesis states that there is no significant difference between the quality of business plans generated by the multi-agent system and the single-agent model. The t-test provided **4.31** for the **t-statistic**, and **0.00042** (with 18 degrees of freedom) for the **p-value**, thus leading to the rejection of the null hypothesis, and confirming that the difference is statistically significant. This result supports the hypothesis that the multi-agent CBPG system produces significantly higher-quality business plans compared to the single-agent GPT-4 system. The design of the multi-agent system, with its domain-specific expertise and few-shot prompting, appears to enhance the quality and relevance of the output.

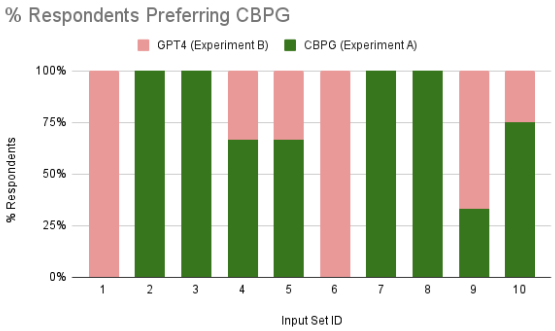

Figure 2: Percent of respondents preferring CBPG. Green indicates those who chose the CBPG approach as superior. Pink indicates percent of those who preferred the single-agent GPT4 approach.

4.3 Blind Comparative Analysis

To complement the quantitative results, a blind comparative analysis was conducted. The results, as shown in Figure 2 indicate that CBPG-generated plans were preferred in 20 out of 30 comparisons, or 66.67% of preferences. This preference is particularly strong in Input Sets 2, 3, 7, and 8, where CBPG dominates. However, GPT-4 outperforms CBPG for specific input sets (1 and 6).

Overall, CBPG's domain-specific approach and structured workflow seems to outperform the single-agent method, but these results indicate further room for improvement to better align with the end user.

Qualitative Analysis: Several key insights emerged from the BPR scoring process and from the rationale provided by reviewers in the blind comparative analysis:

1. Outputs showed a bias toward Western market contexts, particularly within the U.S. This is likely becuase LLMs are trained on disproportionately English data and thus inherit Western-centric biases (Durmus et al., 2024).

2. CBPG-generated plans occasionally include default agent phrases, such as "Hope this helps!", which disrupt the professional tone and coherence of the output.

3. Sections requiring advanced logical reasoning, particularly in financial projections, exhibit errors or inconsistencies in both experiments.

4. The single-agent GPT-4 system outputs display greater cohesion across sections; however, they lack detailed, data-driven insights,

	1	2	3	4	5	6	7	8	9	10	Mean
CBPG	0.85	0.87	0.83	0.85	0.87	0.85	0.87	0.92	0.85	0.92	0.867
GPT4	0.85	0.80	0.75	0.80	0.83	0.82	0.82	0.83	0.82	0.82	0.813

Table 1: BPR results for CBPG and GPT-4o. The raw scores are normalized to belong within the range [0, 1] (divided by 60).

especially when addressing recent trends or industry-specific data.

5. CBPG-generated plans excel in integrating detailed reasoning and data within individual sections, especially market research. They sometimes struggle with inter-section cohesion, leading to a fragmented narrative.

6. CBPG's performance appears to be sensitive to the complexity of input-specific tasks. For example, performance declines when handling intricate inputs, such as "a HIPAA-compliant platform designed to help healthcare professionals collaborate remotely."

4.4 Tool Integration

An unexpected but significant finding of this research was the critical role that **tools** play in enhancing the performance of multi-agent systems. While prompting techniques, including few-shot prompting, improve output to some extent, they do not consistently outperform single-LLM systems.

During the initial implementation phase, our **minimum viable product** (MVP) relied entirely on a multi-agent system setup that used well-crafted prompts with few-shot examples. However, the generated outputs exhibited similar shortcomings to those of the single-agent LLM, including ambiguity, lack of clarity, and insufficient depth.

External **Tools**—defined as functions or capabilities that agents can invoke to perform specific actions—were introduced to address these limitations. Tools provide agents with enhanced abilities, such as real-time web searches, web scraping, and high-complexity data analysis. By improving the accuracy and verifiability of outputs, tools address key concerns for potential adopters of this system.

The integration of tools into the framework yielded measurable improvements. Internal validation metrics demonstrated significant gains. For example, readability score increased by approximately 21%, reflecting improved sentence structure, clarity, and flow. Business terminology score also increased by approximately 14%, indicating better alignment with domain-specific language.

Qualitative observations further confirmed these quantitative improvements. Specifically, the market research agent's outputs showed enhanced relevance and depth, incorporating more detailed insights and contextually accurate information.

5 Future Work and Conclusion

This work demonstrates that a multi-agent LLM system, enhanced with tools and domain-specific roles, significantly outperforms single-agent LLM systems in generating high-quality business plans, as shown through quantitative metrics and blind comparative analysis.

It also indicates the potential of multi-agent systems to facilitate business plan generation, offering significant benefits for stakeholders such as entrepreneurs, strategists, and venture capitalists. This system can serve as an *initial research foundation*, by automating market research, financial projections, and competitive analysis, the system provides a strong initial direction for human experts, as well as a *decision-making aid*, in which key decision-makers can leverage the generated plans to prioritize promising ideas, enabling more efficient allocation of time and resources for in-depth analysis.

While CBPG currently performs well within a U.S.-centric context, extending its capabilities to international markets—considering variations in currency, regulation, and economic dynamics—remains an exciting avenue for future exploration. Beyond business applications, the results of this work highlight how well-designed multi-agent systems can address other multi-faceted problems that demand significant time and resources. Potential applications include legal domains, where automated document analysis could streamline case assessments, and social work, where these systems could assist in resource allocation and program evaluation.

6 Limitations

While promising, the performance and utility of this system depend significantly on its effectiveness for end users. Future work will need to involve extensive feedback through interviews with domain experts to align outputs with their expectations. Key areas for improvement include:

- Establishing a consensus on what constitutes a "good" business plan output with domain experts.
- Designing interfaces that allow experts to interact with, modify, and guide the system outputs in real time.
- Potential human-in-the-loop integration, by incorporating expert feedback directly into the system to iteratively refine outputs and ensure relevance.

It is also important to note that generative models remain susceptible to common limitations, including misaligned or unverifiable outputs. For instance, hallucination must be further addressed. While tools such as retrieval-augmented generation (RAG) improve data accuracy for market research, additional mechanisms are needed to verify outputs, particularly in financial projections. In addition, post-processing steps and industry-standard methodologies can be employed to filter out toxic results.

Finally, while this study concludes that the multi-agent system outperforms a single-agent LLM, further improvements are required to match the quality of human-generated business plans. Ideally, access to a **ground truth dataset**—consisting of high-quality business plans across diverse input configurations—would enable more rigorous benchmarking. However, the proprietary nature of business plans often limits public access to such datasets, presenting challenges for evaluation. In the absence of ground truth data, future work must focus on designing robust alternative evaluation methodologies to ensure fair, consistent, and reliable assessments of system performance.

References

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

Hyeong Kyu Choi and Yixuan Li. 2024. Picle: Eliciting diverse behaviors from large language models with persona in-context learning.

Zhuoyun Du, Chen Qian, Wei Liu, Zihao Xie, Yifei Wang, Yufan Dang, Weize Chen, and Cheng Yang. 2024. Multi-agent software development through cross-team collaboration.

Esin Durmus, Karina Nguyen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2024. Towards measuring the representation of subjective global opinions in language models.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal

Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun

Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models.

Benjamin Kane and Lenhart Schubert. 2023. We are what we repeatedly do: Inducing and deploying habitual schemas in persona-based responses.

Jonghan Lim, Birgit Vogel-Heuser, and Ilya Kovalenko. 2024. Large language model-enabled multi-agent manufacturing systems.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing.

Tula Masterman, Sandi Besen, Mason Sawtell, and Alex Chao. 2024. The landscape of emerging ai agent architectures for reasoning, planning, and tool calling: A survey.

Nathalia Nascimento, Paulo Alencar, and Donald Cowan. 2023. Self-adaptive large language model (llm)-based multiagent systems.

Leslie Nunn. 2010. The importance of a good business plan. *Journal of Business Economics Research (JBER)*, 8.

Sumedh Rasal and E. J. Hauer. 2024. Navigating complexity: Orchestrated problem solving with multi-agent llms.

Investopedia Staff. 2024. Business plan.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, Yidong Wang, Linyi Yang, Jindong Wang, Xing Xie, Zheng Zhang, and Yue Zhang. 2023. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity.

Mingchen Zhuge, Haozhe Liu, Francesco Faccio, Dylan R. Ashley, Róbert Csordás, Anand Gopalakrishnan, Abdullah Hamdi, Hasan Abed Al Kader Hammoud, Vincent Herrmann, Kazuki Irie, Louis Kirsch, Bing Li, Guohao Li, Shuming Liu, Jinjie Mai, Piotr Piękos, Aditya Ramesh, Imanol Schlag, Weimin Shi, Aleksandar Stanić, Wenyi Wang, Yuhui Wang, Mengmeng Xu, Deng-Ping Fan, Bernard Ghanem, and Jürgen Schmidhuber. 2023. Mindstorms in natural language-based societies of mind.

A Appendix A: Sample Prompts

We present the prompts used in our system below. They are slightly condensed for brevity.

A.1 Experiment A: CBPG Agent

- **Role**: Market Research Analyst
- **Goal**: Analyze market potential and provide actionable insights for `business_idea`. Use real-time market data and trend analysis for validation.
- **Backstory** You are an expert market researcher with access to advanced market analysis tools. You provide data-driven insights using real-time market intelligence.

A.2 Experiment A: CBPG Task

- **Description**: Conduct detailed MARKET RESEARCH for `business_idea` in the industry industry focusing on the `target_market` market:
- **Expected Output**:
  1. **Market Analysis:**
     - Total market size and growth potential
     - Market segmentation and trends
     - Geographic considerations for `scale` scale
     - Industry growth forecast
     - Market maturity assessment
  2. **Customer Analysis:**
     - Primary target customers and their needs
     - Customer behavior and preferences
     - Market readiness and adoption factors
     - Customer pain points
     - Buying patterns and decision factors
  3. **Market Environment:**
     - Industry regulations and requirements
     - Economic factors and market conditions
     - Technology and innovation impact
     - Political and legal considerations
     - Environmental factors
  4. **Opportunity Analysis:**
     - Market gaps and unmet needs

- Growth opportunities
- Potential market barriers
- Entry timing considerations
- Market accessibility

5. **Data-Driven Insights:**
   - Key market statistics
   - Growth rate projections
   - Market share potential
   - Consumer trend data
   - Industry benchmark data

Provide specific data points and justify all conclusions. Include market sizing calculations and growth projections. Use current industry data where available and note any assumptions made.

Expected Output Format:

- Use bullet points for key findings
- Include numerical data with sources where possible
- Organize insights by section
- Highlight critical market factors
- Include brief summaries for each major section

- **Example**:
  Context: "SaaS project management tool for small businesses"

  Analysis:

  Market Size: $10B globally, growing at a 15% YoY rate.

  Primary Segments:

  - Freelancers: 40% of the market
  - Small agencies: 35% of the market
  - Startups: 25% of the market

  Key Competitors:

  - Asana
  - Trello
  - Monday.com

  Main Differentiators:

  - Competitive price point
  - Ease of use
  - Seamless integration capabilities

  Analysis:

  - The market is experiencing rapid growth, especially in the remote work segment.
  - Target customers are highly price-sensitive.
  - Feature parity with established players is a key adoption barrier.
  - Integration capabilities are critical to attract small businesses using existing SaaS tools.

A.3 Experiment B: GPT-4 Agent

ROLE: Business Plan Generator. BACKGROUND:Seasoned business strategist who has helped launch diverse ventures. Expert at crafting actionable business plans across various industries. Skilled in identifying key success factors and risk mitigation. TASK: Create a comprehensive business plan for `business_idea` targeting scale operations in the target_market market with an initial investment of initial_investment and a timeline of timeline:

Ensure that your business plan:

- Maintains consistency across all sections
- Uses specific data points from expert analyses
- Provides clear rationale for recommendations
- Links strategies to market opportunities

Generate the business plan with the following sections.

1. Executive Summary:
   - Business concept overview
   - Market opportunity
   - Value proposition
   - Financial highlights
   - Implementation roadmap
2. Business Strategy:
   - Operating model
   - Revenue model
   - Growth strategy
   - Scaling approach
   - Core competencies
3. Go-to-Market Plan:
   - Market entry strategy

- Marketing approach
- Sales channels
- Customer acquisition strategy
- Partnership strategy

4. Implementation Plan:

- Key milestones
- Resource requirements
- Timeline and phases
- Operational setup
- Team structure

5. Risk Management:

- Key risks identification
- Mitigation strategies
- Success metrics
- Contingency plans
- Monitoring approach

Ensure that all recommendations are specific to industry industry standards. Integrate insights from MARKET RESEARCH, FINANCIAL ANALYSIS, and COMPETITIVE ANALYSIS.

A.4 Human Evaluation

The instruction provided to reviewers for pair-wise human preference evaluation is as follows:

"You are given several pairs of business plans. Each pair responds to the same set of input configurations. Between Document A (CBPG) and Document B (GPT4), select which document you believe is of higher quality. Provide a 1–2 sentence rationale for your decision."

B Appendix B: End-to-End Workflow

To better illustrate the functionality of CBPG, the following describes a high-level, end-to-end workflow:

1. **Input Collection**: The user provides the system with key inputs, including the business idea, industry, scale, target market, initial investment, and timeline.
2. **System Initialization**: The system initializes a Crew of Agents, each assigned specific Tasks. A Process is defined to guide the coordinated execution of these Tasks.
3. **Market Research Execution**: The Market Research Agent performs the Market Research Task, leveraging the `DynamicRAGTool` and the Market Research Analysis Tool.
4. **Financial Analysis Execution**: The Financial Planner Agent executes the Financial Analysis Task using the `DynamicRAGTool`.
5. **Competitive Analysis Execution**: The Competitive Analysis Agent conducts the Competitive Analysis Task with the assistance of the `DynamicRAGTool`.
6. **Implementation Plan Creation**: The Business Plan Aggregator Agent integrates the outputs from the Market Research, Financial Analysis, and Competitive Analysis Tasks to create a comprehensive Implementation Plan.
7. **Executive Summary Creation**: The Business Plan Aggregator Agent synthesizes an Executive Summary, contextualizing the outputs from the Market Research, Financial Analysis, Competitive Analysis, and Implementation Plan Tasks.
8. **Business Plan Generation**: The Business Plan Aggregator Agent compiles the outputs from all preceding Tasks into a cohesive, structured final business plan.
9. **Post-Processing**: The system performs post-processing using content processors and validators to ensure the final output is well-structured, readable, and meets quality metrics. These metrics provide transparency and actionable feedback.

C Appendix C: Outputs

Pasted outputs from both experiments. Note that certain structural elements, such as bullet points, were omitted to condense the information more efficiently. Here are the sample inputs we used:

- **Business Idea:** Greek Yogurt Ice Cream
- **Industry:** Food
- **Scale:** National

- **Target Market:** Health Adults
- **Initial Investment:** $500,000
- **Timeline:** 1 year

C.1 Experiment A: Full Output

Comprehensive Business Plan

Table of Contents

Executive Summary

The Greek Yogurt Ice Cream business combines the growing health-conscious trend with the indulgence of premium frozen desserts. Positioned in the Healthy Adults segment, this business leverages the increasing demand for natural, high-protein, and low-sugar frozen desserts.

Vision: To lead the Greek Yogurt Ice Cream market with innovative, health-focused products that prioritize quality, taste, and sustainability. Mission: To offer high-protein, probiotic-enriched ice cream with unique flavors tailored to health-conscious adults, ensuring convenience and taste.

Key Highlights:

- Industry Opportunity: The global ice cream market is projected to reach $68.4 billion by 2025, with Greek Yogurt Ice Cream accounting for 10% of this market.
- Target Market: Health-conscious adults aged 25–45, particularly women prioritizing wellness and fitness.
- Initial Investment: $500,000 allocated for equipment, marketing, employee salaries, and working capital.
- Projected Revenue: By Year 3, forecasted revenue is $380,000 with net income of $197,500, supported by a gross margin of 65%.
- This plan addresses market gaps, leverages trends such as the prioritization of gut health and sustainability, and ensures operational scalability to establish a strong foothold in a growing market.

Market Analysis

Total Market Size and Growth Potential

The global ice cream market is projected to grow from $55.1 billion in 2020 to $68.4 billion by 2025, at a CAGR of 4.2% [Grand View Research]. Greek Yogurt Ice Cream accounts for approximately 10% of the total market, with significant growth potential driven by health trends. The US ice cream market is expected to grow at a CAGR of 3.2%, while the Greek Yogurt Ice Cream segment is projected to grow at 5.1% [Statista].

Market Segmentation and Target Audience

Primary Target Customers: Health-conscious adults aged 25–45, with a focus on women who prioritize wellness and fitness. Key Drivers: Increasing demand for natural ingredients, high-protein content, and lower sugar levels in frozen desserts.

Behavioral Insights:

70% of consumers prioritize sustainability when purchasing frozen desserts [Nielsen]. Online shopping, social media influence, and product reviews are key decision factors.

Trends and Drivers

- Health and Wellness: Rising interest in gut health and probiotics is driving demand for functional frozen desserts.
- Sustainability: 64% of consumers prioritize natural, eco-friendly ingredients in frozen desserts [Nielsen].
- Convenience: Growth in e-commerce platforms and online delivery services is reshaping purchasing behavior.
- Opportunity Analysis
- Market Gaps and Unmet Needs: Limited options for high-protein, low-sugar frozen desserts that align with health and fitness goals.
- Growth Opportunities: Expansion into new geographic markets, strategic partnerships with fitness and wellness brands, and the development of innovative flavors.

Financial Analysis

Setup and Operating Costs

$200,000 for equipment and infrastructure.
$150,000 for marketing and advertising.
$50,000 for employee salaries.
$100,000 for working capital.

Revenue and Profitability

Online sales through e-commerce platforms.

Retail partnerships with specialty grocery stores and fitness-focused outlets.

Wholesale distribution to cafes, restaurants, and wellness events.

Refer to Table 2 to see the generated financial summary table.

Key Performance Indicators (KPIs)

Revenue Growth Rate: 10% CAGR over three years.

Gross Margin: 65%.

Operating Expense Ratio: 12% of revenue.

Net Income Margin: 55%.

Competitive Analysis *Competitor Landscape*

Established Players: Chobani, FAGE, and niche frozen dessert brands dominate the segment.

Market Positioning: Greek Yogurt Ice Cream, with its health-focused appeal, occupies a unique niche but faces competition from other high-protein and functional dessert options.

Differentiation Strategy

USP: Natural, probiotic-enriched ingredients. Unique flavor profiles (e.g., matcha, lavender honey, pistachio). Eco-friendly and sustainable packaging.

Customer Loyalty: Leverage influencer marketing, loyalty programs, and community engagement to build a strong brand.

Growth Potential

Market Share Opportunity: With a focused strategy, the business can capture 10–15% of the Greek Yogurt Ice Cream segment.

Implementation Plan

Actionable Steps

- Months 1–3:

  Finalize recipes and secure ingredient suppliers. Acquire commercial kitchen space and equipment. Develop an e-commerce platform and establish social media presence.

- Months 4–6:

  Launch digital marketing campaigns, leveraging influencer partnerships and SEO. Form initial retail partnerships with health food stores in key regions (California, New York, Texas, Florida).

- Year 1:

  Expand into wholesale distribution for cafes and restaurants.

  Introduce seasonal and limited-edition flavors to drive consumer interest.

- Year 2–3: Scale operations nationally.

  Expand product lines with functional ingredients such as probiotics and adaptogens.

Resource Allocation

- Marketing and Advertising 25%
- Employee Salaries 30%
- Equipment Maintenance 20%
- Working Capital 25%

Risk Mitigation

- Supply Chain: Establish multiple suppliers to mitigate disruptions.
- Market Risks: Monitor trends and adjust product offerings to meet consumer preferences.
- Regulatory Compliance: Ensure adherence to food safety and labeling requirements.

Conclusion

The Greek Yogurt Ice Cream business is well-positioned to capitalize on health and wellness trends while addressing unmet needs in the frozen dessert market. By prioritizing sustainability, innovation, and convenience, the business can achieve strong financial growth and long-term market relevance.

Strategic Priorities:

- Develop premium product lines with innovative flavors and natural ingredients.
- Build a strong digital presence through e-commerce and social media.
- Expand into retail and wholesale channels with a focus on health-conscious consumers.

This data-driven, comprehensive approach ensures a sustainable and profitable future for the business.

C.2 Experiment B: Full Output

Business Plan: Greek Yogurt Ice Cream

Executive Summary

The business aims to introduce a Greek yogurt ice cream that merges indulgence with wellness, catering to the growing demand for healthier dessert options. With a focus on quality, natural ingredients, and appealing flavors, the product is positioned to capture the attention of health-conscious

Year	Revenue ($)	COGS ($)	Gross Profit ($)	Operating Expenses ($)	Net Income ($)
1	340,000	120,000	220,000	34,500	185,500
2	360,000	130,000	230,000	38,500	191,500
3	380,000	140,000	240,000	42,500	197,500

Table 2: Financial Summary Table displayed in Experiment A Business Plan under Financial Analysis section

adults across the nation. The venture will leverage a $500,000 initial investment to develop the product, establish a strong market presence, and scale operations within one year.

Market Research

The frozen dessert market is experiencing significant growth, with a notable shift toward health-oriented products. Consumers increasingly seek desserts that are high in protein, low in sugar, and made from natural ingredients. Greek yogurt ice cream uniquely fulfills this demand, offering a balance of taste, nutrition, and innovation. Competitor success in adjacent niches, such as protein-enriched ice creams and Greek yogurt products, indicates strong market potential for this offering.

Financial Analysis

The $500,000 initial investment will be allocated strategically:

- Product Development and Manufacturing: 40%
- Marketing and Branding: 30%
- Distribution and Logistics: 20%
- Operational Reserves: 10%

Revenue projections anticipate strong early sales, with potential break-even within 18–24 months. The financial plan includes a phased scaling strategy to minimize risk while maximizing market reach.

Competitive Analysis

The competitive landscape features established brands in the frozen dessert market, including Halo Top and Yasso. These competitors have demonstrated the viability of health-focused ice creams but leave room for differentiation through premium Greek yogurt formulations and innovative flavor profiles. The proposed product aims to fill this gap by offering superior nutritional benefits and catering to niche health-conscious audiences.

Implementation Plan

The one-year timeline includes key phases:

- Product Development (Months 1–3): Finalizing recipes, securing suppliers, and developing packaging.
- Market Testing (Months 4–6): Conducting trials in select regions and gathering consumer feedback.
- Brand Launch (Months 7–9): Building awareness through social media campaigns, influencer partnerships, and sampling events.
- National Rollout (Months 10–12): Expanding distribution through retail partnerships and e-commerce platforms.

Conclusion This business leverages the growing demand for healthier dessert options to introduce a distinctive Greek yogurt ice cream. With a strategic focus on innovation, quality, and consumer engagement, the venture is positioned for long-term success in a competitive yet evolving market.