Universiteit Leiden
The Netherlands

Opleiding Informatica

AI-Driven Creativity:

The Power of Dual-Agent Systems

Mai Nehoray

Supervisors:
Dr. A.H. Zohrehvand & Dr. G.J. Wijnholds

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)

1 Abstract

This thesis explores the potential of the FunSearch algorithm, introduced by Google DeepMind, to enhance creativity in domains without exact solutions, specifically in recipe generation. Building on foundational work in mathematical problem-solving, FunSearch integrates a pre-trained Large Language Model (LLM) with an evaluator to iteratively produce innovative and practical outputs. The study adapts FunSearch to emulate human collective creativity through a structured, iterative process, addressing the dual criteria of originality and effectiveness.

The methodology involves modifying FunSearch's components—database, samplers, and evaluators—to suit the subjective nature of recipe generation. The generated recipes are evaluated for creativity using both qualitative and quantitative methods, including principal component analysis and cosine similarity. Results indicate that AI-generated recipes exhibit greater ingredient diversity and shorter cooking times than human-created ones, yet they lack the nuanced creativity found in traditional culinary practices. This research demonstrates that FunSearch can extend its utility to creative and open-ended domains, offering a framework for AI-driven innovation across various fields. Future directions include refining evaluation methods and balancing exploration and exploitation to enhance both the novelty and quality of AI-generated outputs.

Contents

2 Introduction

As advancements in Artificial Intelligence (AI) continue to accelerate, its applications are increasingly recognized for their capacity to enhance creativity across various sectors. This transformative potential has made AI an attractive tool for innovation, promising to redefine traditional processes and generate novel solutions. However, while AI has shown remarkable capabilities in domains requiring exact solutions, such as mathematics and computer science, its potential to foster creativity in more subjective and open-ended domains remains less explored. This thesis explores the integration of AI in creative fields, specifically focusing on recipe generation, to understand how AI can enhance creativity and produce practical, innovative outputs.

Research into creativity and AI has demonstrated that systems can generate outputs that are both novel and effective. For instance, Jia et al. (2023) explored how AI can enhance employee creativity by automating repetitive tasks, thereby allowing more cognitive resources to be allocated to higher-level problem-solving and innovative thinking [11]. Additionally, Leahey et al. (2023) used computational text analysis to categorize scientific papers based on the type of novelty they presented, distinguishing between disruptive and consolidating influences [13]. These studies collectively highlight AI's capacity to contribute to creative processes by optimizing routine tasks and categorizing the impact of different types of innovations. Furthermore, studies by Kenett (2019) and Greenacre (2022) introduced quantitative measures such as semantic distance, alongside traditional qualitative assessments, to robustly evaluate the creativity of AI-generated outputs [3, 6]. These advancements underscore AI's potential to innovate within defined parameters and generate outputs that are both novel and effective.

While existing AI systems have shown success in generating novel solutions for problems with well-defined criteria, their application in domains where solutions are inherently subjective and multifaceted is less understood. This gap is particularly evident in creative fields, where the evaluation of outcomes is not straightforward and often relies on human judgment. The current research lacks comprehensive methodologies to leverage AI in generating creative outputs that are both original and effective. Addressing this gap is crucial for advancing AI's role in fostering innovation in subjective fields.

This thesis aims to bridge this gap by adapting the FunSearch algorithm, traditionally used for problems with exact solutions, to the subjective field of recipe generation. By combining an evolving algorithm with a large language model (LLM), this study seeks to enhance the creativity of AI-generated recipes. The research will evaluate the effectiveness of this dual-agent system in producing creative and practical solutions, demonstrating that AI can innovate in creative domains and providing a framework for AI-driven innovation across various fields. To guide this investigation, the question that will be answered in this research is:

> **Does applying the FunSearch algorithm in a context without exact answers still provide creative results?**

3 Theoretical Background

3.1 Creativity

Creativity is a subjective and open-ended domain, characterised by two essential criteria: originality and effectiveness [2]. An idea is considered creative when it is both novel and useful. Originality alone is insufficient; an idea must also have practical utility to be deemed creative [2]. For instance, a random process might generate something original, but without utility, it would not be considered creative. Thus, an idea must meet both criteria — originality and effectiveness — to be recognized as creative.

While individual creativity is crucial, it often thrives within a collaborative context where diverse perspectives can lead to more innovative solutions [8]. Collective creativity, which significantly contributes to innovation and problem-solving, leverages the strengths of multiple individuals working together. Acar et al. identify three types of collectives that foster creativity: attention-based, divergence-based, and convergence-based [8].

Attention-based collectives focus on the initial generation of diverse ideas, leveraging the unique insights of individual members. Divergence-based collectives enhance solution quality through iterative development and refinement, facilitated by collaborative interaction and frequent feedback. Convergence-based collectives aim to synthesize diverse ideas into a cohesive and practical solution, utilising joint decision-making and integrating the best elements from individual contributions.

Understanding the types of novelty is crucial for evaluating the impact of creative outputs, particularly in scientific and technological contexts. Novelty can manifest as new methods, new theories, or new results, each playing a distinct role in the progression of knowledge These different types of novelty have varying impacts on scientific influence due to their inherent nature. Leahey et al. [13] found that new methods tend to be more disruptive than new theories or new results. New methods often introduce fundamentally different ways of approaching problems, leading to shifts in research practices and priorities. Conversely, new theories typically complement existing research, providing deeper understanding without completely overhauling foundational principles. New results, while adding to the existing knowledge base, do not necessarily change foundational approaches or theories, making their impact more incremental.

3.2 Artificial Intelligence

The concept of creativity is not inherently human; Artificial Intelligence (AI) can replicate creativity by producing outcomes that are perceived as both novel and useful [5, 9]. AI aims to develop advanced computers and machines that possess intelligence equal to humankind's [1]. AI can mimic human creativity by generating unique combinations of familiar ideas, creating new works based on the attributes of previous creations, and offering innovative ideas that combine attributes in ways humans might not have considered [12].

Research has shown various ways AI can enhance creativity. For instance, Jia et al. investigate how AI can enhance employee creativity by taking over repetitive and mundane tasks, therby allowing employees to allocate more cognitive resources to higher-level problem-solving and innovative thinking [11]. By freeing up employees from routine tasks, AI can significantly improve overall creative output in the workplace.

Additionally, Leahey et al. [13] used AI to categorise scientific papers based on the type of novelty (new results, new theories, new methods) and their impact on the scientific community. By distinguishing between disruptive and consolidating influences, they provided insights into the dynamics of scientific progress.
However, there are challenges to integrating AI in creative processes. Doshi et al. [16] observed a reduction in collective diversity when using AI for creative writing, likely due to the anchoring effect of the provided story ideas. Writers tend to stick closely to these initial AI-generated prompts, resulting in higher cosine similarity scores between stories within the AI conditions compared to the human-only condition. This suggests that while AI can introduce novel elements, it might also limit creative diversity by anchoring human creators to specific ideas.

Seeming collusion is another challenge described by Abada and Lambin [7]. This phenomenon arises from the incomplete exploration of possible states and actions inherent to Q-learning algorithms, a type of reinforcement learning algorithm. Q-learning algorithms strive to balance exploration (seeking out new possibilities) with exploitation (optimizing known good solutions). However, as these algorithms learn, the rate of exploration diminishes, leading to a preference for previously visited states, even if these states represent suboptimal outcomes. This presents a significant challenge in the application of AI to creative tasks, where continuous exploration is essential to achieving innovative solutions.

To address these challenges, it is useful to draw parallels with human problem-solving methods. In human problem-solving, tasks are often approached through iterative modification and evaluation of existing designs. This process continues until a satisfactory performance level is achieved or until it becomes apparent that no better solutions can be found, resulting in a restart with a different initial solution. This method resembles a hill-climbing process. While effective for certain tasks, this method has limitations in creative domains due to the vast, high-dimensional solution space with multiple peaks and valleys [4]. Traditional computer methods struggle with these complex, non-linear landscapes.

According to Miikkulainen, Evolutionary Computation offers a promising alternative for these complex tasks where traditional methods fall short [4]. Evolutionary computation is a subset of AI that mimics the process of natural evolution. It uses mechanisms inspired by biological evolution, such as reproduction, mutation, recombination, and selection, to evolve solutions to problems over successive generations. This approach conducts parallel searches across different areas of the solution space, which allows for comprehensive exploration and prevents the search from getting stuck in local optima. By maintaining multiple candidate solutions simultaneously, evolutionary algorithms can explore a diverse range of possibilities. When a promising solution is found in one area, it can be shared with other parallel searches, thereby enhancing the overall search process and avoiding premature convergence on suboptimal solutions. This process mirrors the iterative improvement and diverse idea generation seen in divergence-based collectives, described by Acar et al. [8]. By continuously evolving and refining solutions, evolutionary computation fosters creativity and innovation, making it a powerful approach for addressing complex and high-dimensional problems.

Large Language Models (LLMs) have emerged as powerful tools in AI-driven creativity, significantly enhancing the capability of AI systems to produce creative and contextually appropriate outputs. LLMs are advanced neural network models with billions of parameters, designed to understand and generate human-like text. These models are trained on vast datasets using self-supervised learning approaches, where the model learns to predict parts of the data from other parts without the need for manually labeled examples [15].
The training process involves converting text into tokens, which are basic units of language the model can process. LLMs use attention mechanisms, which allow the model to focus on different parts of the input text dynamically, establishing context and understanding the relationships between words in a way that mimics human comprehension. This mechanism enables LLMs to generate coherent and contextually relevant text based on the input they receive [17].
The ability of LLMs to produce creative outputs lies in their training on diverse and extensive datasets, which imbues them with a broad understanding of language and concepts. They can generate novel text by recombining learned patterns in innovative ways, making them valuable tools for tasks that require a high degree of creativity, such as writing and storytelling. By leveraging their vast parameter space, attention mechanisms, and sophisticated learning processes, LLMs have demonstrated their indispensable role in various creative applications. This capability is crucial for this research, as it highlights the potential of AI to generate human-like, contextually appropriate, and innovative content, thereby pushing the boundaries of what AI can achieve in creative domains.

3.3 FunSearch as an AI Tool for Creativity

Google DeepMind introduced the FunSearch algorithm in December 2023, marking a significant advancement in AI's ability to discover novel and valuable solutions [14]. Designed to search for new solutions in mathematical and computer science problems, FunSearch combines a pre-trained Large Language Model (LLM) and an evaluator. The LLM enhances creativity by generating novel ideas, while the evaluator ensures the validity and practicality of these ideas, guarding against potential hallucinations and faults. This iterative process mirrors how the human brain continuously evaluates and refines its thoughts, similar to divergence-based collectives in human creativity and the iterative improvement seen in evolutionary computation [4, 8].

Unlike traditional AI models, which often focus on either generating ideas or evaluating them, FunSearch integrates both processes iteratively, ensuring that the creative outputs are both innovative and practical. By integrating these iterative processes, FunSearch marks the first time an LLM has been used to solve long-standing scientific puzzles, producing verifiable new information that did not previously exist [10]. This capability underscores its potential as a powerful tool for creativity.

The FunSearch algorithm stands out from other notable AI systems developed by DeepMind, like AlphaTensor and AlphaDev, for four distinct reasons [14].
To begin with, the FunSearch algorithm leverages existing knowledge by starting with a foundation of common knowledge about the problem. This could be a basic, even non-functional, program, or it might already incorporate some problem-specific insights. However, it's expected to be suboptimal, serving as a starting point for the evolutionary process. Starting from here enables the algorithm to concentrate on finding the most critical parts of the solution, thereby avoiding the need to reinvent already established knowledge. By building on what is known, FunSearch can more efficiently explore new solutions and improve upon existing ones.
Furthermore, FunSearch generates output in the form of a program that details how the solution is derived. This is similar to scientific explanations, where researchers detail the methods used to make discoveries. Unlike traditional computer search techniques that only produce the solution, often as a list of vectors, FunSearch provides a program capable of generating this solution. This transparency enables researchers to trace the solution's origin and understand the reasoning behind it, making the algorithm's output more trustworthy and interpretable.
Even more, FunSearch enhances solution diversity through an island-based evolutionary model, which prevents stagnation in local optima. The algorithm divides the program population into several independent islands that evolve autonomously. Periodically, the programs in the worst-performing half of the islands are discarded and replaced by clones of the best programs from the surviving islands. This strategy promotes the exploration of diverse solutions and ensures that the search process does not become trapped in suboptimal regions, thereby increasing the likelihood of discovering innovative solutions.
Finally, FunSearch utilizes parallel processing to enhance efficiency. By running each island in parallel, the algorithm can explore a larger solution space within a shorter time frame. This parallelisation allows FunSearch to simultaneously evaluate multiple candidate solutions, significantly speeding up the search process and enabling the discovery of high-quality solutions more quickly.

Romera-Paredes et al. applied the FunSearch algorithm to several well-defined mathematical problems, providing concrete examples of its capabilities and effectiveness. Two notable applications include the cap set problem in extremal combinatorics and the online bin packing problem in combinatorial optimization, both of which are detailed in the main paper.

The cap set problem involves finding the largest possible set of vectors within a specific mathematical space, where no three vectors sum to zero. This problem is challenging due to the significant gap between the known upper and lower bounds. FunSearch discovered new constructions of large cap sets, surpassing the best-known solutions and improving the lower bound on cap set capacity. Additionally, by analyzing the generated programs, researchers identified new symmetries in the admissible set problem, providing deeper insights and further enhancing the results.
The online bin packing problem involves sequentially packing items of varying sizes into a

limited number of fixed-size bins, aiming to minimize the number of bins used. This problem is highly relevant to real-world scenarios where future item sizes are unknown. FunSearch discovered new heuristics that outperform traditional ones like first fit and best fit. These heuristics are defined as programs that prioritize bins based on the item and an array of bins as input, outputting a priority score for each bin. The evolutionary process starts from the best fit heuristic and evolves to find even better heuristics.

For both the Cap Set Problem and the Online Bin Packing Problem, clear criteria exist for evaluating solutions. Romera-Paredes et al. identified several key characteristics that enable FunSearch to work effectively for mathematical problems:

- The availability of an efficient evaluator
- A rich scoring feedback quantifying the improvements
- The ability to provide a skeleton with an isolated part to be evolved

Unlike mathematics, many real-world problems do not have exact solutions, and their evaluation criteria are more subjective. In theory, FunSearch can be used to find solutions to a wide range of problems because it produces code rather than the solution itself [10]. However, FunSearch's application in domains without exact solutions remains unexplored. This presents an opportunity to extend FunSearch's utility to more subjective and open-ended problems. By understanding these characteristics and the successes in these specific problems, we can appreciate how FunSearch leverages its iterative nature to produce innovative and practical solutions. This foundation sets the stage for exploring FunSearch's application in more subjective and creative domains.

4 Methodology

Building upon the foundational work of FunSearch in mathematical problem-solving, this thesis can extend FunSearch to more subjective and open-ended problems. Building upon the insights from Jia et al. on enhancing creativity by automating repetitive tasks, this thesis explores whether AI can directly participate in creative tasks.

The recipe generation domain has been selected for this research to explore the concept of creativity, which emphasizes both originality and effectiveness. Recipes should not only be innovative in their combinations of ingredients and techniques, but also practical and appealing to users. This dual focus on originality and effectiveness aligns with the broader definition of creativity [2].

4.1 Implementation of FunSearch

Buidling on the insights from Acar et al. regarding human collective creativity, this thesis proposes that the FunSearch algorithm can emulate the different collectives found in human groups. The mechanisms underlying human creativity can be mirrored in technical agents, facilitating the generation of creative outputs through the structured, iterative processes of the FunSearch algorithm.

The algorithm encompasses three phases that correspond to the types of collectives in human creativity: attention-based, divergence-based, and convergence-based. Initially, independent islands work on generating diverse solutions, similar to individual members in an attention-based collective. Iterative feedback refines these solutions, similar to the divergence-based collective. Finally the convergence-based phase integrates the best elements to produce the final output.

To adapt the FunSearch algorithm for the creative task of recipe generation, several modifications were made to its three main components: the database, the samplers, and the evaluators. These adaptations are essential to address the subjective and open-ended nature of creativity in this domain. For a technical explanation of the FunSearch notebook, refer to Appendix A.

4.1.1 The database

Solutions are stored in the database as a dictionary, where the keys are scores and the values are solution IDs. Initially, the database consists of a user-generated input called a skeleton, containing boilerplate code (sections of code that are repeated in multiple places with little to no alteration), and previous knowledge of the problem. This setup allows the algorithm to focus on evolving only the critical parts of the problem, ensuring efficient computational resources. After each iteration, the new program's output is stored in the database.

For recipe generation, the database retains the structure of the original FunSearch implementation, storing solutions as a dictionary with scores as keys and recipe IDs as values. This structure enables efficient retrieval and updating of recipes based on their evaluated scores.

To mitigate the anchoring effect observed by Doshi et al. [16], the initial skeleton for recipe generation is designed to be intentionally basic. This basic skeleton provides minimal constraints, allowing the LLM to explore a wide range of creative possibilities without being anchored to specific starting points. The details of this skeleton can be found in Appendix D.

4.1.2 The samplers

The samplers generate prompts, to feed to the LLM, by combining programs from the database with a standard instruction. These prompts resemble the skeleton, but can incorporate new programs from the first iteration onwards. For each iteration, different programs are sampled for the prompt based on their evaluation by the evaluators.

The recipes are sampled randomly, including both high and low-scoring examples. The focus is on exploring the full range of creative possibilities rather than solely generating the best recipe. This approach encourages the creation of novel combinations and innovative techniques, thereby fostering creativity.

The use of random sampling from the database for each iteration, is specifically designed to address the challenges identified by Abada and Lambin [7] regarding the balance between exploration and exploitation for mitigating seeming collusion.
This approach contrasts with traditional Q-learning algorithms, which may prematurely reduce exploration and become trapped in suboptimal states.

The prompt design includes three recipes from the database, accompanied by the standard text that can be found in Appendix C.

4.1.3 The evaluators

All outputs generated by the LLM are scored by the evaluators. This evaluation is done using an aggregation function, typically a mean of inputs. The evaluator creates a tuple containing the program score and the program itself. Each tuple is stored in the database for future use by the samplers.
For the recipe adaptation, a predictive model is developed to predict user ratings for each recipe, ensuring that generated outputs are not only novel but also appealing and effective. The predictive model is trained on a combination of two Kaggle datasets of 230.000 human recipes, each complemented with a total of 1.000.000 user interactions. It contains approximately 13.000 unique ingredients. The data is gathered in the period of 2000 until 2018, sourced from *Food.com*. The dataset features are categorised into numeric features and text features. The target variable, y, is the user rating. The numeric features are cooking time, number of steps and number of ingredients. They are scaled using a standard scaler. The text features are recipe ingredients and recipe steps. They are vectorised using the TF-IDF method, with a maximum of 100 features. The reason for choosing this method can be found

in section 4.3.

Various algorithms, including Random Forest, XGBoost, Linear Regression, Support Vector Machine, were evaluated, using Grid Search and automatic fine-tuning packages, to identify the most effective model. Performance metrics such as Mean Squared Error (MSE) and R-squared were used for model selection. Ultimately, a neural network was chosen as the final prediction model. The detailed process of creating this prediction model can be found in Appendix B.
Each generated recipe is evaluated by the predictive model, which assigns a score rounded to two decimal places.

4.2 Evaluating creativity

Evaluating creativity in AI-generated outputs is crucial to understanding the effectiveness of the models and the novelty of their outputs. Creativity can be assessed using both qualitative and quantitative methods.

Qualitative assessment relies on human judgments to evaluate the creativity of AI-generated outputs. This method involves experts or general users providing subjective opinions on whether the outputs are original and effective. While this approach captures the nuanced, context-dependent aspects of creativity that are difficult to quantify, it can be subjective and vary significantly between evaluators, emphasizing the need for complementary quantitative measures.

Quantitative assessment methods offer objective metrics to evaluate creativity, providing a necessary complement to qualitative judgments. One common quantitative measure is semantic distance, which quantifies the novelty of an idea by calculating the distance between concepts in a semantic space. Recent studies have used semantic distance to assess the originality of AI-generated ideas, helping to validate subjective assessments [3].

Building upon techniques used by Leahey et al. to measure the novelty and impact of scientific contributions [13], Term Frequency - Inverse Document Frequency (TF-IDF) will be employed in this study. It is a powerful tool for converting textual data into numerical features, reflecting the importance of each term within the context of the entire dataset. TF-IDF will be used in two different contexts.

Firstly, TF-IDF will be utilized to convert the textual data of ingredients and cooking steps into numerical vectors. To manage the high dimensionality of these vectors, Principal Component Analysis (PCA) will be applied. PCA reduces the large set of correlated variables into a smaller set of uncorrelated principal components. These principal components are linear combinations of the original variables that maximize the explained variance [6]. This dimensionality reduction will help identify the most significant variations in the recipes, and these variations will be visualized using a scatterplot. The scatterplot will display the distribution of human-created and AI-generated recipes along the principal components, facilitating a clear comparison of their creative differences and similarities.

Secondly, the TF-IDF vectors will be used to measure the collective diversity between recipes through cosine similarity. This method, referred to by Doshi et al. [16], calculates the cosine of the angle between two vectors, providing an objective measure of their similarity. The cosine similarity between pairs of recipes A and B will be calculated as follows:

$$\text{Cosine Similarity} = \frac{A \cdot B}{\|A\|\|B\|}$$

where $A \cdot B$ is the dot product of the vectors, and $\|A\|$ and $\|B\|$ are the magnitudes (norms) of the vectors.
There are three different pairs to measure: within the human-created recipes, within the AI-generated recipes, and between the human-created and AI-generated recipes.

5 Results

Before delving into the results, it is important to note that these findings are based on a limited number of observations and should be viewed as preliminary and exploratory. The PCA results, in particular, are subject to variability due to the small sample size. Therefore, while these initial insights are promising, they should be interpreted with cautious optimism.

5.1 Data overview

The analysis was conducted using two final datasets: one containing 20 human-created recipes and another containing 20 AI-generated recipes. Table 1 provides an overview of statistics for both types of recipes. The table includes the average number of ingredients, average cooking time, average number of steps, standard deviation of cooking time, standard deviation of the number of ingredients, and average rating for both human-created and AI-generated recipes.

The human-created recipes were sourced from *Kaggle*. There were two datasets available: the recipes themselves, and the interactions between users and recipes. These two datasets were merged on recipe ID. This created a total of 912 complete rated recipes. After removing 601 duplicate recipes, 311 recipes remained. For each recipe, the columns: minutes, number of steps, number of ingredients, steps, ingredients and rating were kept. A random sample of 20 recipes was taken for the analysis.

The AI-generated recipes were produced using the FunSearch algorithm. 109 recipes were generated in total. 48 of those had missing values for the number of steps, number of ingredients or cooking time. These were left out of the analysis. From the 61 remaining recipes, a random sample of 20 is taken for the analysis.

Table 1: Basic Statistics of Human-created and AI-generated Recipes

Metric	**Human-created Recipes**	**AI-generated Recipes**
Number of Recipes	20	20
Average Number of Ingredients	7.8	8.6
Average Cooking Time (minutes)	59.5	36.0
Average Number of Steps	8.1	6.5
S.D. of Cooking Time	92.862377	8.675434
S.D. of Number of Ingredients	2.44088	4.10904
Average Rating	4.6500	4.5895

To illustrate the capabilities of the AI in generating creative and practical recipes, two examples are showcased in figure 1.

Recipe 1

- **Minutes:** 45
- **Number of Steps:** 9.0
- **Steps:**
  1. Preheat the oven to 400°F (200°C).
  2. In a small bowl, mix together pepper, cumin, coriander, cinnamon, allspice, and nutmeg for the spice blend. Set aside.
  3. Season the lamb chops with salt and pepper.
  4. Heat a large oven-safe skillet over medium-high heat. Add the lamb chops and sear until browned for 2-3 minutes on each side. Remove from heat and set aside.
  5. Add the butternut squash to the same skillet and cook until slightly softened for 3-4 minutes.
  6. Add the spice blend to the skillet and toast for 1-2 minutes, stirring constantly.
  7. In a separate bowl, mix together the honey, maple syrup, apple cider vinegar, and Dijon mustard for the glaze. Pour the glaze over the lamb chops and butternut squash in the skillet.
  8. Transfer the skillet to the oven and bake for 15-20 minutes or until the lamb chops reach the desired level of doneness and the butternut squash is tender.
  9. Remove from the oven and let rest for 5 minutes before serving.
- **Ingredients:**
  - 4 lamb chops
  - 1 large butternut squash, peeled and cubed
  - 1 tsp pepper
  - 1 tsp cumin
  - 1 tsp coriander
  - 1 tsp cinnamon
  - 1 tsp allspice
  - 1 tsp nutmeg
  - 1/4 cup honey
  - 1/4 cup maple syrup
  - 1/4 cup apple cider vinegar
  - 1 tbsp Dijon mustard
  - Salt and pepper, to taste
  - Optional: fresh rosemary sprigs, chopped (for garnish)
- **Number of Ingredients:** 12
- **Rating:** 4.55

Recipe 2

- **Minutes:** 30
- **Number of Steps:** 10.0
- **Steps:**
  1. Preheat the oven to 400°F (200°C).
  2. Sous vide the quail eggs for 12 minutes at 145°F (63°C) to achieve a runny yolk.
  3. While the eggs are cooking, mix 1/4 cup of saffron threads with 1/4 cup of olive oil in a small saucepan over low heat.
  4. Remove the eggs from the sous vide bath and carefully crack them into a bowl. Whisk the eggs gently and add a pinch of salt and pepper.
  5. In a separate bowl, mix 1/2 cup of smoked paprika with 1/4 cup of olive oil.
  6. Dip each egg in the smoked paprika mixture, coating them evenly.
  7. Place the eggs on a baking sheet lined with parchment paper and bake for 10-12 minutes or until the whites are set and the yolks are still runny.
  8. While the eggs are baking, toast 1/2 cup of breadcrumbs in a pan with 1 tablespoon of olive oil until golden brown.
  9. Top each egg with a spoonful of the saffron-infused olive oil and a sprinkle of the toasted breadcrumbs.
  10. Serve immediately and enjoy!
- **Ingredients:**
  - 6 quail eggs
  - 1/4 cup saffron threads
  - 1/4 cup olive oil
  - 1/2 cup smoked paprika
  - 1/4 cup breadcrumbs
  - 1 tablespoon olive oil
  - Salt and pepper to taste
- **Number of Ingredients:** 6
- **Rating:** 4.58

Figure 1: Examples of FunSearch-generated recipes

5.2 Principal Component Analysis

PCA was applied to the combined dataset to reduce its dimensionality to two principal components. This reduction facilitates the visualization of the distribution of recipes in a two-dimensional space, making it easier to compare the human-created and AI-generated recipes.

Figure 2 displays the PCA visualization of human-created and AI-generated recipes. The plot shows the distribution of recipes along the two principal components. Human-created recipes are marked in blue, while AI-generated recipes are marked in orange. Key observations from the plot include a noticeable overlap between human-created and AI-generated recipes. Both types of recipes cluster in certain regions. Additionally, some recipes, especially human-created ones, appear as outliers.

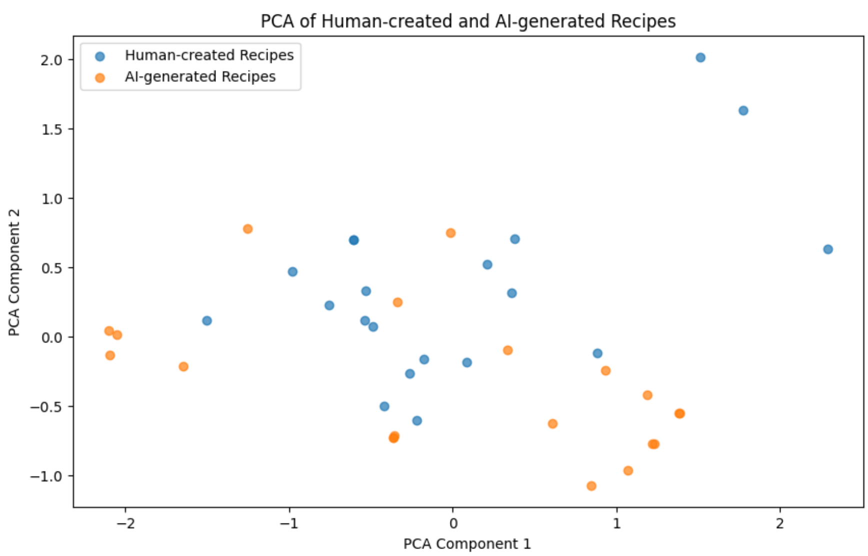

Figure 2: PCA of Human-created and AI-generated Recipes

5.3 Cosine Similarity

Cosine similarity was employed to measure the similarity between the human-created and AI-generated recipes based on their ingredient and step features. The cosine similarity values range from -1 to 1, where 1 indicates identical vectors, 0 indicates orthogonal vectors (no similarity), and -1 indicates diametrically opposed vectors. Table 2 summarizes the average cosine similarity within human-created recipes, within AI-generated recipes, and between human-created and AI-generated recipes.
The analysis of the cosine similarity scores reveals several important observations. The average cosine similarity within human-created recipes is 0.1155, indicating a moderate level of similarity among these recipes. In contrast, the average cosine similarity within AI-generated recipes is higher at 0.2217. When examining the similarity between human-created and AI-generated recipes, the average cosine similarity is 0.1078, which is lower than the within-group similarities.

Table 2: Average Cosine Similarity Statistics

Similarity Measure	Human-created Recipes	AI-generated Recipes
0.1155	0.2217	0.1078

6 Discussion

6.1 Interpretation of Results

6.1.1 Recipe analysis

Statistical analysis The table presenting the basic statistics of human-created and AI-generated recipes offers several insights into the characteristics and tendencies of each type of recipe.

Firstly, AI-generated recipes have a higher average number of ingredients compared to human-created recipes. This suggests that the AI tends to include more components in its recipes, possibly to enhance flavour complexity, or to experiment with novel combinations. One potential reason for this could be the LLMs access to a large dataset of ingredients, allowing it to explore a broader range of ingredients than a human chef might consider.

Furthermore, the average cooking time for AI-generated recipes is significantly lower than for human-created recipe. This difference might reflect the AI's tendency to optimise for efficiency, creating quicker recipes that could be more practical for everyday cooking. This efficiency could make AI-generated recipes more appealing to busy people looking for quick meals without compromising on flavour. In contrast, human chefs might invest more time in complex preparation to achieve a broader range of flavours, contributing to the higher average cooking time observed in human-created recipes.

Additionally, human-created recipes have a higher average number of steps compared to the AI-generated recipes. This could indicate that human chefs often engage in more complex preparation, which might contribute to the broader range of creativity observed in human-generated recipes. The complexity of these steps might involve advanced techniques that AI, despite its data-driven approach, might simplify for practicality and efficiency.

The standard deviation of cooking time is much higher for human-created recipes compared to AI-generated ones, suggesting greater variability in the time humans are willing to invest in cooking. This variability reflects the diverse culinary traditions, personal preferences and different approaches of human chefs. On the other hand, the standard deviation in the number of ingredients is higher for AI-generated recipes, indicating more experimentation by the AI in ingredient selection. This might be due to the AI's algorithmic approach to maximising flavour and novelty by combining a wide range of ingredients.

Finally, the average rating difference is minimal, with human-created recipes slightly higher. This suggests that human-created recipes may still be perceuved as slightly more successful overall, possibly due to the human touch in the traditional recipes. Despite the AI's efficiency and innovation, the nuances and experiental aspects of human cooking might contribute to slightly higher ratings for human-created recipes.

These differences help better understand the strengths and limitations of AI in culinary creativity. The AI excels in efficiency and ingredient diversity, making it a valuable tool for generating quick and innovative recipes. However, it may still lack the depth of creativity and cultural context that human chefs bring to their culinary creations.

Qualitative analysis The qualitative analysis of AI-generated recipes reveals the AI's ability to create dishes with innovative pairings of ingredients and the use of creative cooking techniques. For instance, consider the first recipe, which showcases a creative blend of spices and sweet elements by using honey and maple syrup with spices like cumin and nutmeg. This combination not only enhances the flavor complexity but also demonstrates the AI's ability to experiment with both common and uncommon pairings, resulting in a dish that is both novel and palatable.
In another example, the second recipe highlights advanced techniques such as sous vide cooking and the use of saffron-infused olive oil. Sous vide cooking, a method typically used in professional kitchens, shows the AI's capability to incorporate complex methods into its recipes. The use of saffron-infused olive oil makes it more luxerious, reflecting the AI's understanding of high-end culinary techniques.
To highlight the unique aspects of AI-generated recipes, it is essential to compare them directly to human-created ones. Human chefs often rely on cultural traditions, personal experiences, and taste evaluations, for adding a distinct touch of creativity and authenticity. For example, a classic Italian "Risotto alla Milanese" by a human chef relies on simple ingredients and the personal touch in preparation to achieve perfect texture and flavor balance. In contrast, an AI-generated risotto might include additional ingredients like truffle oil or a unique spice blend to enhance flavor complexity. While this introduces novelty, it might lack the cultural nuances and experiential aspects of a human-crafted dish.
Moreover, human-created recipes involve complex steps and intuitive adjustments based on taste, something AI-generated recipes might simplify for efficiency. For instance, a human chef might adjust seasoning based after tasting the recipe, whereas an AI-generated recipe follows a set of predetermined instructions. This comparison underscores the AI's innovative potential while highlighting the depth and authenticity that human chefs bring to their creations.

6.1.2 PCA

The PCA visualisation indicates a noticable overlap between human-created and AI-generated recipes, suggesting that some AI-generated recipes closely resemble those created by humans. This overlap demonstrates that the FunSearch algorithm can generate recipes that are not only novel but also align with human creativity. Several observations and considerations are worth discussing in detail.

The clustering of both human-created and AI-generated recipes in certain regions of the PCA plot suggests that there are common patterns followed by both types. This indicates that the AI is capable of learning and replicating the typical structures and combinations that are familiar to human chefs. However, the noticeable spread and outliers in the human-created recipes highlight the broader range of creativity and uniqueness that humans can achieve, which AI has yet to fully replicate.

Outliers in the PCA plot, particularly among human-created recipes, represent unique or innovative dishes that stand out from conventional patterns. For example, a human-created recipe with an unconventional regional ingredient might be an outlier, demonstrating cultural diversity and personal creativity. In contrast, the absence of such distinct outliers in AI-generated recipes suggests that while AI can mimic typical recipes well, it may not yet achieve the same level of radical innovation that human creativity can produce.
Looking at it from another perspective, an outlier in the human recipes could also be a chef that just started his cooking experience, making mistakes and learning from these.

The wider spread of human-created recipes along the second principal component axis indicates greater diversity in the approaches and techniques used by human chefs. For instance, a human recipe that incorporates a unique preparation method or a rare ingredient reflects the broader range of creativity seen in human chefs. AI-generated recipes tend to cluster more around the center, balancing familiarity and innovation but lacking the extreme variability observed in human-created recipes. This clustering suggests that AI tends to stay within safer bounds, potentially limiting its creativity.

6.1.3 Cosine Similarity

The analysis of cosine similarity scores provides additional insights into the characteristics and diversity of human-created and AI-generated recipes. The average cosine similarity within human-created recipes is 0.1155, indicating a moderate level of similarity among these recipes. This suggests that human chefs tend to create diverse recipes with varying ingredients and steps, reflecting individual creativity and traditional culinary practices.
In contrast, the average cosine similarity within AI-generated recipes is higher at 0.2217. This higher similarity score implies that the AI-generated recipes are more homogeneous, possibly due to the AI's reliance on learned patterns and optimization for ingredient combinations and cooking steps. The AI's approach may focus on generating recipes that are efficient and practical, leading to less variation compared to human-created recipes.
When examining the similarity between human-created and AI-generated recipes, the average cosine similarity is 0.1078, which is lower than the within-group similarities. This lower score indicates a noticeable difference between the two sets of recipes, highlighting the distinct approaches taken by human chefs and the AI. While the AI can produce recipes that resemble human creations to some extent, it still lacks the depth and variability that human creativity brings to culinary arts.
Overall, the AI excels in generating consistent and efficient recipes but may not yet fully replicate the diversity and innovative flair found in human-created recipes.

6.2 Contributions

This research extends the findings of Jia et al., who demonstrated that AI could enhance employee creativity by automating routine tasks [11]. While their work focused on the benefits of AI in handling repetitive tasks to free up cognitive resources for higher-level problem-solving, this study takes a step further. By leveraging Large Language Models (LLMs), it showcases AI's capability to directly generate creative outputs, thereby bridging the gap between routine task management and active participation in creative processes. This dual-agent system demonstrates the comprehensive potential of AI in enhancing creativity, not just by offloading routine tasks but by contributing to the creative process itself.

The study demonstrates that the FunSearch algorithm can function as creative collectives, similar to human groups. By mirroring the mechanisms of human collective creativity, such as diversity, interaction, and integration, the algorithm fostered significant creativity in its outputs. This suggests that Acar et al.'s framework [8] can be effectively extended to technical agents. The diversity of solutions generated by the independent islands, the iterative refinement process, and the final integration of the best solutions all contribute to the collective creativity of the FunSearch algorithm. This finding underscores the potential of AI systems to emulate human-like creative processes, thereby enhancing their capability to generate innovative solutions.

Moreover, the FunSearch algorithm exemplifies the introduction of a new method, aligning with Leahey et al.'s categorization of disruptive innovations. According to their research, new methods tend to be more disruptive than new theories or results because they introduce fundamentally different approaches to problem-solving [13]. FunSearch changes how AI systems explore and generate solutions by providing a more efficient and effective methodology. This study demonstrates its broad applicability, having been tested in a setting with no exact solutions, such as enhancing creativity in the culinary arts. This success indicates its potential for implementation in critical sectors, further showcasing its disruptive capability across various fields.

In addition, the study addresses the challenges of seeming collusion, as identified by Abada and Lambin, which arise from the balance between exploration and exploitation in Q-learning algorithms [7]. Traditional Q-learning approaches may prematurely reduce exploration, becoming trapped in suboptimal states. In contrast, this study employs a random sampling approach from the database for each iteration, ensuring a high exploration rate and mitigating the risk of collusion. This method allowed for a diverse exploration of possible solutions, leading to a significant degree of innovation in the outputs.
However, despite this high level of creativity, the overall scores of AI-generated recipes did not significantly surpass those of human-created recipes. This outcome can be attributed to the algorithm's emphasis on exploration over exploitation. By prioritising the generation of novel combinations rather than optimising for known successful outcomes, the algorithm produced creative but sometimes less refined solutions. This trade-off highlights the importance of balancing exploration and exploitation to achieve both innovation and high performance, as [7] suggests. By drawing on Abada and Lambin's insights, the necessity of maintaining a balance between exploration and exploitation becomes clear. This balance prevents the system from favoring suboptimal solutions while still fostering creativity. Future research should explore strategies to fine-tune this balance, ensuring that AI systems can generate both innovative and high-scoring outputs.

Finally, the analysis showed that AI-generated recipes had lower average cosine similarity scores compared to human-created recipes, suggesting greater diversity. Specifically, the average cosine similarity score for AI-generated recipes was significantly lower, indicating a wider range of creative possibilities. These findings contrast with Doshi et al.'s results, where higher similarity scores in GenAI-generated stories suggested reduced diversity due to the anchoring effect of provided prompts [16]. In this study, the use of a basic initial skeleton mitigated this effect, promoting greater diversity in generated recipes. This highlights the importance of initial design in AI systems for creative tasks. By avoiding the anchoring effect, the AI was able to explore a broader spectrum of creative solutions, enhancing collective diversity and demonstrating a more effective approach to fostering creativity in AI-generated outputs.

6.3 Considerations, Limitations and Future Directions

One critical aspect to consider is the evaluation method used to rate the generated recipes. The LLM-based approach inherently avoids generating 'bad' recipes, unlike humans who may produce a wider range of quality, including unsuccessful experiments. This raises questions about the appropriateness of the evaluation method, as it might bias the AI towards generating only acceptable or good recipes, potentially limiting its exploratory creativity. Developing a more nuanced evaluation framework that allows for the generation and assessment of a broader range of outputs, including less successful experiments, could provide a better understanding of AI's creative capabilities. Incorporating human expert reviews and feedback could also enhance the evaluation process.

Another important consideration is the potential for cultural bias in the generated recipes. Since the recipes in this study were generated and analyzed in English, there is a risk that non-English-speaking cuisines and culinary techniques are underrepresented. This limitation could restrict the diversity and creativity of the AI-generated recipes. To address this, future research should incorporate a broader range of languages and cultural influences. By sourcing recipes and techniques from diverse cultures and languages, the training data can become more representative, thereby expanding the creative potential of AI-generated outputs.

The random sampling approach used in this study prioritizes exploration over exploitation. While this fosters creativity and diversity, it might also result in less refined solutions. Future work should focus on finding an optimal balance between exploration and exploitation to enhance both the novelty and quality of AI-generated outputs. By utilizing the exploitation phase more, high-scoring recipes will be evolved, potentially improving the ratings of newly generated recipes. Further research should explore strategies to fine-tune this balance, which could involve adaptive techniques that adjust the exploration rate based on the performance of generated outputs. By optimizing this balance, AI systems can achieve both high levels of creativity and practical, high-quality solutions.

The analysis in this study was conducted on a relatively small sample size, which might limit the generalizability of the results. The AI-generated dataset consisted of 20 recipes out of an initial 109, and the human-created dataset also contained 20 recipes out of an initial 912.

This small sample size makes the results preliminary and exploratory. Conducting analyses on larger datasets in the future would provide more robust conclusions.

While PCA is a powerful tool for dimensionality reduction, it has its limitations. The method relies on linear combinations of variables and might not capture complex, non-linear relationships in the data. Additionally, PCA results can be sensitive to outliers and the scaling of variables. The PCA findings in this study should be interpreted with caution, and further validation using different techniques is recommended.

Future research could expand the application of FunSearch to both creative domains and critical areas where innovation is essential. By applying FunSearch in various fields such as artistic creation, music composition, product design, and critical sectors like healthcare, transportation, and legal practices, the potential to generate innovative and effective solutions in subjective and open-ended domains can be explored.

6.4 Conclusion

The exploration of the FunSearch algorithm, traditionally applied to exact solutions in mathematical problems, has shown promising results when adapted to the subjective domain of recipe generation. This thesis highlights the potential of AI, specifically through the integration of large language models (LLMs), to enhance creativity in more open-ended fields. The dual-agent system of FunSearch, combining creativity from an LLM with the evaluative rigor of a traditional algorithm, successfully generated novel and practical recipes.

The statistical analysis revealed that AI-generated recipes tend to use a greater variety of ingredients and have shorter cooking times compared to human-created recipes. This suggests that the AI prioritizes efficiency and complexity in its ingredient combinations. However, while the AI recipes were diverse and innovative, they often lacked the nuanced touch of human-created dishes, which benefit from cultural context and personal experience. Principal Component Analysis (PCA) and cosine similarity measurements showed that AI-generated recipes closely mimic human creativity, though with less variability and radical innovation.

The research extends existing knowledge on AI's role in enhancing creativity, showing that AI can participate in creative tasks beyond automating repetitive ones. It demonstrated that the FunSearch algorithm could emulate the collective creativity mechanisms found in human groups, suggesting a framework for technical agents to foster innovation.

Future research should focus on addressing the limitations observed, such as the need for more nuanced evaluation methods that allow for a broader range of creative outputs, and mitigating cultural biases inherent in the training data. Additionally, finding an optimal balance between exploration and exploitation in the algorithm could enhance both the novelty and quality of AI-generated outputs.

The findings underscore the potential of AI to innovate across various fields, suggesting that with further refinement, AI systems could significantly contribute to creative processes in domains traditionally dominated by human expertise.

7 Appendix

7.1 Appendix A

The evolve notebook contains all code that contributes to the workings of the FunSearch algorithm. The following detailed steps outline the framework employed within the notebook.

The prediction model is loaded into the notebook, as well as the vectorisers that were created for the prediction model data. Python libraries and tools are imported, setting up the environment for the task. These libraries include pandas, numpy, vectorising packages from sklearn and replicate. This setup phase also involves initialising the connection with the LLM API (Llama 2 is chosen for their performance in NLP tasks) and initiating a foundational skeleton of recipes, which provides a template from which the model can learn.

The process begins with the retrieval of the initial skeleton, which serves as a baseline structure for generating new programmatic content. This skeleton is retrieved only once at the start to initialise the loop. The while loop works until the time passed reaches 1800 seconds, 30 minutes. First, parent programs are chosen from the dictionary (which only contains the skeleton, at the beginning). This is done by a random sampling method. These parent programs are used as input to the 'evolve program' function, which calls the LLM using the prompt. The output from the LLM is cleaned and structured to a dataframe, using the 'clean output' function to ensure usability in predicting the rating. Each generated program is assigned a unique ID and is saved with its identifier for future use.

The new program is then evaluated using the prediction model, and its score is recorded. The score is added to the dictionary using a mechanism that checks for existing scores in the dictionary to group programs by their performance. If a program's score is already present, its ID is added under this score category. If not, a new entry is created in the dictionary.

7.2 Appendix B

The prediction model is developed in a Jupyter Notebook. This notebook outlines the process of developing a neural network model to predict recipe ratings based on features derived from recipe data. The following steps outline the procedures used in this notebook.

Python libraries for data manipulation, machine learning and neural network development are imported. These include pandas, numpy and various packages from sklearn and tensorflow. Data is loaded from two datasets, one containing recipe details and the other documenting interactions between users and recipes. These datasets include 100,000 rows which are then merged based on recipe IDs.

The features: cooking time, number of steps, number of ingredients, textual descriptions of the steps and ingredients, are specified. The target variable is the user-provided rating. Empty fields within the data are filled with an empty string to maintain consistency in data format.

Using Term Frequency-Inverse Document Frequency (TF-IDF), textual data from the steps and ingredients are vectorized considering the top 100 most relevant categories. This transforms the text into a numerical format that can be processed by machine learning models. The numeric features (number of steps, number of ingredients) are scaled to address the skewness in the data distribution.

The dataset is divided into training and test sets. A neural network is constructed using grid search to identify the optimal parameters. The model's performance is assessed using the mean squared error and R-squared metrics to quantify prediction accuracy and proportion of the variance explained by the model.

The final model architecture consists of an input layer with 48 neurons using ReLU activation and L2 regularization, followed by a dropout layer with a rate of 0.3 to prevent overfitting. It includes a second hidden layer with 64 neurons and ReLU activation, followed by another dropout layer with a rate of 0.1. The output layer contains a single neuron, suitable for regression tasks. The model is compiled with the Adam optimizer, using Mean Squared Error (MSE) as the loss function and Mean Absolute Error (MAE) and MSE as metrics. Early stopping is employed to monitor the validation loss, with a patience of 10 epochs and restoring the best weights observed during training. The model is trained on the train set for up to 100 epochs with a validation split of 20 percent, using early stopping to ensure the best model performance.

After training and evaluation, the final model is saved in an HDF5 format, ensuring that it can be loaded and used without the need to retrain.

7.3 Appendix C

"""**Minutes:** [Specify total cooking time in minutes here, using integer values]

Steps: List each cooking step clearly. Follow the numbering format below for clarity and consistency.

- 1. [Step 1]
- 2. [Step 2]
- 3. [Step 3]
- ...

Ingredients: Clearly list each ingredient with exact quantities and units, e.g., '1 cup of sugar', '2 tablespoons of olive oil'. Use the bullet point format to list items distinctly.

- * [Ingredient 1]
- * [Ingredient 2]
- * [Ingredient 3]

- ...

n_steps: [Automatically count the number of steps listed above]

n_ingredients: [Automatically count the number of distinct ingredients listed above]"""

7.4 Appendix D

Recipe 1:

- **Minutes:** 30
- **Steps:**
  1. Lightly sauté the main ingredient to enhance flavor.
  2. Gently fold in a blend of aromatic herbs and a dash of exotic seasoning.
  3. Wrap in parchment and bake to perfection.
- **Ingredients:**
  - 1 large zucchini, sliced lengthwise
  - 1 tsp Himalayan pink salt
- **Number of Steps:** 3
- **Number of Ingredients:** 2

Recipe 2:

- **Minutes:** 40
- **Steps:**
  1. Marinate the main ingredients in a fusion of citrus and herbs.
  2. Layer in a cast iron skillet and drizzle with infused oil.
  3. Slow roast to meld the flavors, finishing under the broiler for a crispy top.
- **Ingredients:**
  - 2 chicken breasts, cubed
  - 2 tsp herb de Provence
- **Number of Steps:** 3
- **Number of Ingredients:** 2

Recipe 3:

- **Minutes:** 15
- **Steps:**
  1. Toss the fresh ingredients in a homemade citrus dressing.
  2. Garnish with edible flowers and serve chilled.
- **Ingredients:**
  - 1 cup mixed berries
  - 1 tbsp honey-lime dressing
- **Number of Steps:** 2
- **Number of Ingredients:** 2

Recipe 4:

- **Minutes:** 50
- **Steps:**
  1. Brown the protein to lock in juices.
  2. Add layers of colorful vegetables and grains to a rich broth.
  3. Let simmer slowly, allowing the spices to infuse.
- **Ingredients:**
  - 1 cup beef chunks
  - 2 cups root vegetables, cubed
  - 1 cup barley
  - 1 tbsp smoked paprika
- **Number of Steps:** 3
- **Number of Ingredients:** 4

Recipe 5:

- **Minutes:** 20
- **Steps:**
  1. Blend the base ingredient with the sweetener and an exotic spice mix.
  2. Spread evenly in a silicone mold and bake until golden.
- **Ingredients:**
  - 1 cup almond flour
  - 1/2 cup agave syrup
  - 1 tsp cardamom
- **Number of Steps:** 2
- **Number of Ingredients:** 3

Figure 3: Initial Skeleton of Recipes

References

1. Wang. Moving towards complex intelligence? https://ieeexplore.ieee.org/abstract/document/5172882 (Aug. 1, 2009).
2. Runco, M. A. & Jaeger, G. J. The standard definition of creativity. *Creativity research journal* **24,** 92–96. https://doi.org/10.1080/10400419.2012.650092 (Jan. 1, 2012).
3. Kenett, Y. N. What can quantitative measures of semantic distance tell us about creativity? *Current opinion in behavioral sciences* **27,** 11–16. https://doi.org/10.1016/j.cobeha.2018.08.010 (June 1, 2019).
4. Miikkulainen, R. Creative AI through Evolutionary Computation: Principles and Examples. *SN Computer Science/SN computer science* **2.** https://doi.org/10.1007/s42979-021-00540-9 (Mar. 23, 2021).
5. Cropley, D. H., Medeiros, K. E. & Damadzic, A. *The intersection of human and artificial creativity* 19–34. https://doi.org/10.1007/978-3-031-14549-0_2 (Jan. 1, 2022).
6. Greenacre, M. *et al.* Principal component analysis. *Nature reviews methods primers* **2.** https://www.nature.com/articles/s43586-022-00184-w (Dec. 22, 2022).
7. Abada, I. & Lambin, X. Artificial intelligence: Can seemingly collusive outcomes be avoided? *Management science* **69,** 5042–5065. https://doi.org/10.1287/mnsc.2022.4623 (Sept. 1, 2023).
8. Acar, O. A., Tuncdogan, A., Van Knippenberg, D. & Lakhani, K. R. Collective Creativity and Innovation: An interdisciplinary review, integration, and research agenda. *Journal of management.* https://doi.org/10.1177/01492063231212416 (Nov. 14, 2023).
9. Haase, J. & Hanel, P. H. Artificial muses: Generative artificial intelligence chatbots have risen to human-level creativity. *Journal of creativity* **33,** 100066. https://www.sciencedirect.com/science/article/pii/S2713374523000250#bib0047 (Dec. 1, 2023).
10. Heaven, W. D. Google DeepMind used a large language model to solve an unsolved math problem. https://www.technologyreview.com/2023/12/14/1085318/google-deepmind-large-language-model-solve-unsolvable-math-problem-cap-set/ (Dec. 18, 2023).
11. Jia, N., Luo, X., Fang, Z. & Liao, C. When and how Artificial intelligence augments employee creativity. *Academy of Management journal/The Academy of Management journal.* https://doi.org/10.5465/amj.2022.0426 (Mar. 28, 2023).
12. Kirkpatrick, K. Can AI demonstrate creativity? *Communications of the ACM* **66,** 21–23. https://doi.org/10.1145/3575665 (Jan. 20, 2023).
13. Leahey, E., Lee, J. & Funk, R. J. What types of novelty are most disruptive? *American sociological review* **88,** 562–597. https://doi.org/10.1177/00031224231168074 (May 12, 2023).
14. Romera-Paredes, B. *et al.* Mathematical discoveries from program search with large language models. *Nature.* https://www.nature.com/articles/s41586-023-06924-6 (Dec. 14, 2023).

15. Shen, Y. *et al.* ChatGPT and other large language models are double-edged swords. *Radiology* **307.** https://pubmed.ncbi.nlm.nih.gov/36700838/ (Apr. 1, 2023).

16. Doshi, A. R. & Hauser, O. P. Generative AI enhances individual creativity but reduces the collective diversity of novel content. *UCL School of Management, University of Exeter.* Available at https://link-to-the-paper.com (2024).

17. Raiaan, M. A. K. *et al.* A review on large language models: architectures, applications, taxonomies, open issues and challenges. *IEEE access,* 1. https://doi.org/10.1109/access.2024.3365742 (Jan. 1, 2024).