RecipePrep
ECE1786
Final Report

Permission	Chenhong Qiu	Chujing Yu	Kassy Zeng	Yixuan Liu
Video	Yes	Yes	Yes	Yes
Final Report	Yes	Yes	Yes	Yes
Source Code	Yes	Yes	Yes	Yes

Chenhong Qiu, Chujing Yu, Kassy Zeng, Yixuan Liu

Word count: 1983 (excluding title and figure names)

Penalty: 0%

Introduction

The principles of a balanced diet are widely understood: increasing the consumption of fruits and vegetables while reducing processed foods. However, an estimated 9.5% of people face barriers to following the recommended Food Guide [1], due to the perception that recipes for a balanced diet are complex, as well as the additional difficulty of finding usable recipes online which are considered time-consuming [2].

To help overcome these barriers, this project developed RecipePrep, a GPT-4o powered personalized recipe generator. Users input ingredients and cooking requirements, and RecipePrep generates a balanced recipe with step-by-step instructions, offering clear guidance on how to use the ingredients they already have.

Illustration

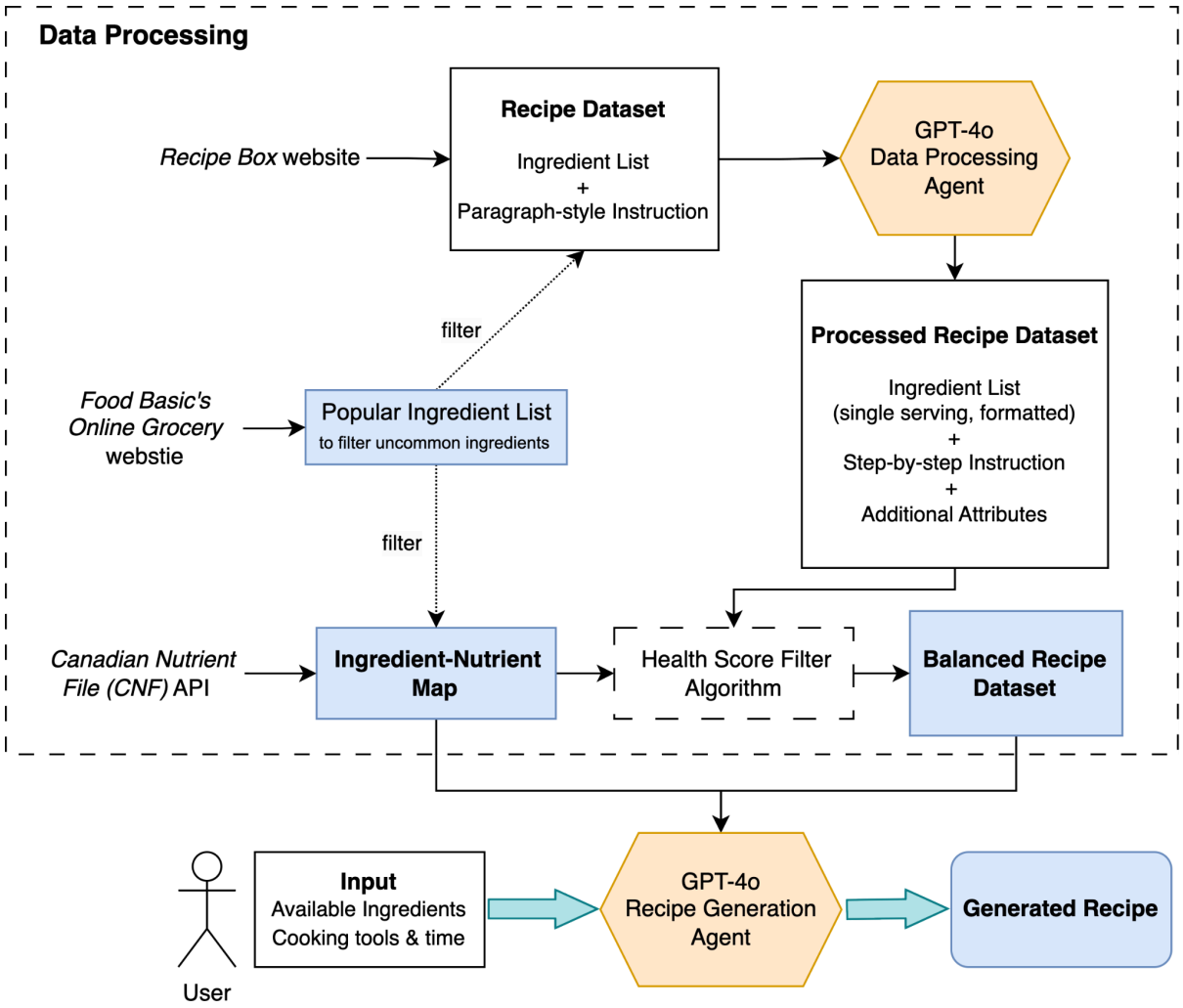

Figure 1: System Architecture Diagram

Background and Related Work

Machine learning has been developing in the culinary domain, especially in recipe generation using LLMs. Lee et al. [3] fine-tuned a GPT-2 model to generate instructions from recipe titles and ingredients, achieving a final F1 score of 0.77 when comparing generated ingredients to ground-truth ingredients. Aljbawi's model is a health-aware recipe generator that uses fine-tuned GPT-2 to create personalized and nutritious recipes [4]. Healthiness is defined by ingredient calorie content, and the model filters for low-calorie options to ensure balanced nutrition. Mean Squared Error (MSE) was used to measure ingredient quantity accuracy for calorie calculation and health goal alignment.

Data and Data Processing

Popular Ingredient List

The team visited the *Food Basics Online Grocery* website [5], sorted products by best-seller in categories (e.g. Fruits, Vegetables and Meat), and manually collected ingredients from the top 40 products per category. The finalized "Popular Ingredients" list includes 228 unique items across 10 categories.

Recipe Dataset

The primary recipe dataset, *Recipe Box Dataset* [6], contains numerous recipes with paragraph-style instructions, as shown in Figure 2. The dataset cleaning steps are summarized in Table 1.

Table 1: Recipe Box Dataset Cleaning Steps

Step	Initial Size	Process Detail	Final Size
1	85,362	Removed recipes with missing instructions, ingredients, or ingredients not in the Popular Ingredient List.	31,258
2	31,258	Randomly selected 600 recipes of various lengths.	600
3	600	Used a GPT-4o Data Processing Agent to convert paragraph-style instructions and sentence-style ingredient lists into structured labels (Figure 3), and adjust the recipe for single servings.	600
4	600	Applied the Health Score Filter Algorithm (explained in *Architecture* section), which uses the Ingredient Nutrient Map (detailed in the following section) to calculate each recipe's health score. Filtered out recipes that do not meet the health score requirements to create the final Balanced Recipe Dataset.	419

```
    "efQOFu8nwlZgli6eM3VRb34BGHZjkAe": {
        "instructions": "In a small bowl mash tuna with mayonnaise to taste to a smooth paste. Spread on
        bread and wrap securely in foil. Let sandwich sit 2 hours until mayonnaise has softened bread.
        Slice in half and serve with a glass of cold milk.;",
        "ingredients": [
            "1 small can solid white tuna packed in water, drained well",
            "3 to 4 tablespoons mayonnaise",
            "2 slices white bread"
        ],
        "title": "Tuna Sandwich",
        "picture_link": null
    },
```


Figure 2: Original Data from Recipe Box Dataset


```
"step_by_step_instructions": [
  "In a small bowl, mash 0.5 small can of tuna with 1.5 to 2 tablespoons of mayonnaise to a smooth paste.",
  "Spread the mixture on 1 slice of bread.",
  "Wrap the sandwich securely in foil.",
  "Let the sandwich sit for 2 hours until the mayonnaise has softened the bread.",
  "Slice the sandwich in half and serve with a glass of cold milk."
],
"processed_ingredients": [
  "0.50 can tuna",
  "1.50 tablespoon mayonnaise",
  "1 slice bread"
],
"pure_ingredients": [
  "tuna",
  "mayonnaise",
  "bread"
],
"cooking_time": "2 hours",
"required_tools": [
  "bowl",
  "foil",
  "knife"
],
```


Figure 3: Processed Recipe with Structured Labels

Ingredient Nutrient Map

The team combined the Popular Ingredient List with ingredient names from 600 randomly selected recipes and retrieved their nutrient information (per 100g) using the *Canadian Nutrient File API* [7]. This resulted in an Ingredient-Nutrient Map containing 641 items (example shown in Figure 4) for calculating health scores..

```
"ingredient_name": "tuna",
"nutrients": [
    {
        "value": 0.0,
        "nutrient_name": "Fibre, total dietary",
        "unit": "g"
    },
    {
        "value": 0.172,
        "nutrient_name": "Fatty acids, saturated, total",
        "unit": "g"
    },
    {
        "value": 0.49,
        "nutrient_name": "Total Fat",
        "unit": "g"
    },
    {
        "value": 24.4,
        "nutrient_name": "Protein",
        "unit": "g"
    },
    {
        "value": 45.0,
        "nutrient_name": "Sodium, Na",
        "unit": "mg"
    },
    {
        "value": 109.0,
        "nutrient_name": "Energy (kcal)",
        "unit": "kCal"
    },
```


Figure 4: Ingredient-Nutrient Map Example

Test Set

The Popular Ingredients List was used to generate 400 random ingredient combinations of varying lengths across categories, focusing on two cases:

1. Unbalanced: A list of ingredients from 1–2 categories.
2. Balanced: A list of ingredients from at least 3 categories.

Each case included short (1–5 ingredients) and long (5+ ingredients) lists, with 5 being the average ingredient count in the Recipe Box dataset.

After a manual review to identify combinations close to real-life scenarios, 10 were selected for tuning and 20 for testing. A detailed breakdown is shown in the table below.

Table 2: Tuning/Test Set Detail Breakdown

Case	Proportion	Sub-Type	Proportion	Tuning Set Number	Test Set Number
Unbalanced	40%	Long List	50%	2	4
		Short List	50%	2	4
Balanced	60%	Long List	50%	3	6
		Short List	50%	3	6
Total				**10**	**20**

Architecture and Software

The high-level system architecture diagram is shown in Figure 1.

Data Processing

The **Popular Ingredient List** filters the raw Recipe Box Dataset. The filtered dataset is then processed by the **GPT-4o Data Processing Agent**, which converts the recipes into structured labels. This results in the **Processed Recipe Dataset**, stored in a vector database where the key is the ingredient list and the value is the processed recipes.

The system uses the Canadian Nutrient File API to create an **Ingredient-Nutrient Map**, pairing each ingredient (key) with its nutrient content (value). The documents split from the map are embedded and stored in another vector database for the subsequent recipe filtering.

The **Health Score Filter Algorithm** filters each recipe in the Processed Recipe vector database by calculating its health score. To ensure accuracy, each recipe's ingredients are used as queries to retrieve their nutrient content from the Ingredient-Nutrient vector database. The retrieved data is aggregated and assessed using the health score metric and criteria, as detailed in the *Comparison* section. Recipes that do not meet the criteria are removed, with the remaining recipes forming the final **Balanced Recipe Dataset**.

Recipe Generation

The **Balanced Recipe Dataset** serves as the reference to guide the generator. To address GPT-4o's token limits and enhance reliability, the **Recipe Generation Agent** employs Retrieval-Augmented Generation (RAG) framework with three key components: Indexing, where recipes are split into chunks, embedded, and stored in a vector database; Retrieval, which fetches up to five similar recipes based on user-input ingredients using cosine similarity; and Generation, where retrieved recipes, user inputs, and predefined criteria are combined into a structured prompt for GPT-4o to generate recipes.

Baseline and Comparison

The generated recipes are evaluated using three key criteria: health score, relevance, and consistency. A recipe is considered a successful output if it meets all three criteria.

Health Score

Health score quantifies the recipe's nutrient balance by adapting the World Health Organization (WHO) Nutrient Intake Goals for three daily meals (Table 2). We focus on seven key macronutrients: proteins, carbohydrates, sugars, sodium, fats, saturated fats, and fibres. Recipes are scored from 0 to 7, where 0 indicates none of the macronutrients meet the recommended ranges (highly unbalanced) and 7 indicates all are satisfied (highly balanced). Since a meal typically consists of multiple dishes, recipes scoring 3 out of 7 are regarded as nutritionally acceptable, with higher scores preferred. A Python-based algorithm incorporating the above logic was implemented to evaluate the health score.

Table 3: WHO Ranges of Nutrient Intake Goals for a Single Meal [8]

Nutrient	**Range**
Proteins	10%-35% of total energy
Carbohydrates	45%-75% of total energy
Sugars	< 10% of total energy
Sodium	< 2.5g
Fats	15%-30% of total energy
Saturated Fats	<10% of total energy
Fibres	>12.5g

Relevance

Relevance assesses how well a recipe meets user requirements in terms of cooking tools, cooking times and ingredients alignment. Tools and times are hard constraints, recipes failing to meet these are rejected. Ingredients alignment measures the similarity between user-input ingredients and recipe ingredients, excluding seasonings and spices. To support substitutions, user input and recipe ingredients are mapped to root nouns using the nutrient map [3]. Root nouns are compared, and an overlap rate (similarity score) exceeding 70% is required for a recipe to be considered relevant.

Consistency

Consistency evaluates the quality of the generated recipe, focusing on text coherence and structure. It is assessed through three aspects: Instructional Clarity, where instructions must be clear, detailed, and free of ambiguity; Measurement Consistency, ensuring ingredient measurements are precise and consistent; and Logical Step Sequencing, requiring steps to follow a logical, complete order without missing or redundant actions. Recipes are assessed through human evaluation, and only those that meet all three aspects are considered to have successfully passed the consistency check.

Baseline: Single Prompt

We designed a single prompt that generates balanced recipes without referencing external data, serving as a baseline test. While the single prompt can produce coherent recipes, it often lacks precise ingredient descriptions (e.g., "salt and pepper to taste") and clear instructions (e.g., "cook until done"). Therefore, our baseline is that the system should exceed this performance by providing detailed ingredient specifications and clearer instructions.


```
"processed_ingredients": [
  "150g salmon fillet",
  "200g asparagus, trimmed",
  "1 tablespoon olive oil",
  "1/2 lemon, juice only",
  "Salt and pepper to taste"
 ],
 "instructions": [
  "1. Heat 1 tablespoon of olive oil in a skillet over medium-high heat.",
  "2. Season the salmon fillet with salt and pepper.",
  "3. Once the oil is hot, place the salmon fillet in the skillet, skin-side down. Cook for 4-5
minutes until the skin is crispy.",
  "4. Flip the salmon and cook for another 3-4 minutes until the salmon is cooked through.",
  "5. Remove the salmon from the skillet and set aside.",
  "6. In the same skillet, add the asparagus and a splash of water. Cover and steam for 4-5
minutes until the asparagus is tender-crisp.",
  "7. Squeeze the juice of 1/2 a lemon over the asparagus and season with salt and
pepper.",
  "8. Serve the salmon alongside the asparagus on a plate."
 ],
```


Figure 5: Single Prompt Output (“salmon, asparagus” as input)

Quantitative Results

Table 4: Success Test Cases among Three Criteria

Criteria	Health Score	Relevance	Consistency
Success cases	18/20	17/20	18/20
Cases passing all three criteria	14/20		

The recipe generator was tested on 20 examples (as mentioned in the Test Set section, with details provided in Appendix A). Among these, 18 recipes achieved an acceptable health score, 17 satisfied relevance requirements, and 18 met consistency standards. Overall, 14 recipes passed all three criteria, resulting in a success rate of 70%.

Both health scores and consistency achieved a 90% success rate, reflecting strong performance. Failed health score cases occurred with unbalanced inputs, like cheese and butter, which made it challenging to meet WHO nutritional guidelines. Consistency failures were due to missing ingredient measurements, often in recipes with numerous steps or many ingredients. Relevance scored 85%, slightly lower than the health score success rate, as the generator prioritizes health criteria. To fulfill nutritional goals, the system occasionally adds ingredients, reducing relevance in short-input test cases that cannot form balanced dishes.

The overall success rate (70%) effectively summarizes the system's ability to balance user requirements with health-focused adjustments while meeting all three criteria.

Qualitative Results

The following examples demonstrate both a successful and a failed recipe, illustrating the system's performance under different input conditions.

Successful Example

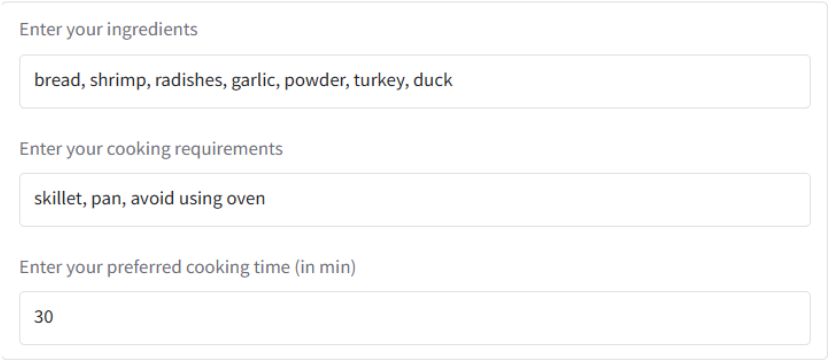

Figure 6: Input of a Successful Example

Title: Shrimp and Turkey Stir-Fry with Radishes

Ingredients:
- 0.25 lb shrimp, peeled and deveined
- 0.25 lb turkey breast, thinly sliced
- 0.5 cup radishes, sliced
- 1 slice whole-grain bread, cut into cubes
- 0.5 teaspoon garlic powder
- 1 tablespoon olive oil
- 0.25 teaspoon salt
- 0.25 teaspoon black pepper

Instructions:
1. Heat a skillet over medium-high heat and add 1 tablespoon of olive oil.
2. Add the shrimp and turkey breast to the skillet, seasoning with 0.25 teaspoon of salt, 0.25 teaspoon of black pepper, and 0.5 teaspoon of garlic powder.
3. Cook the shrimp and turkey for 5-6 minutes, stirring occasionally, until the shrimp is pink and the turkey is cooked through.
4. Remove the shrimp and turkey from the skillet and set aside.
5. In the same skillet, add the sliced radishes and sauté for 3 minutes until slightly tender.
6. Add the bread cubes to the skillet and sauté for an additional 2-3 minutes until they become crispy.
7. Return the shrimp and turkey to the skillet, stirring everything together for another 2 minutes to heat through.
8. Serve hot and enjoy your meal.

Cooking Time: 20 minutes

Cooking Tools:
- skillet
- pan

Suggestions:
Consider adding a handful of spinach or kale for additional fiber.
Incorporate a squeeze of lemon juice for enhanced flavor.

Figure 7: Output of a Successful Example

This recipe meets all three evaluation criteria: health score, relevance, and consistency. It achieved a health score of 4 by satisfying macronutrient requirements for protein, sugars, sodium, and saturated fats. However, the recipe shows a deficiency in fibre and carbohydrates, as it only includes 0.5 cups of radishes and 1 slice of whole-grain bread. To improve, the system suggests adding spinach or kale for a more balanced dish.

Regarding relevance, the cooking tools and time of the recipe align well with the user's requirements. Additionally, a similarity score of 75% indicates a reasonable substitution made by the system. The input ingredient “turkey” was replaced with “turkey breast” in the recipe. While not an exact match, this substitution is contextually appropriate due to the close relationship between these two ingredients.

The recipe also meets consistency criteria with coherent instructions and consistent measurements. This demonstrates the system’s ability to produce reliable and user-friendly outputs.

Failed Example

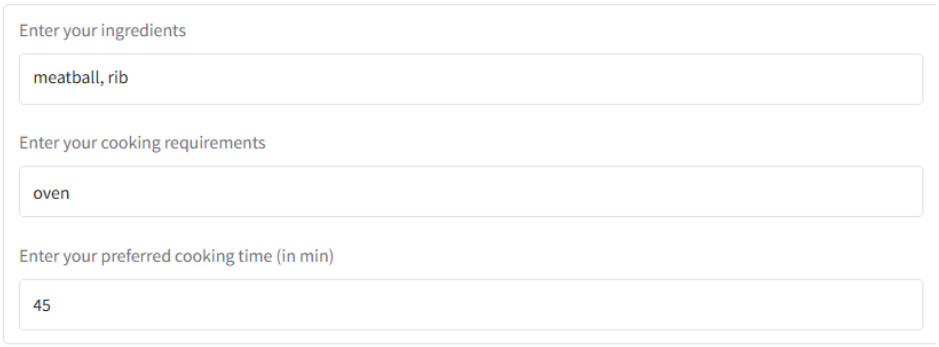

Figure 8: Input of a Failed Example

Title: Baked Meatball and Rib Medley

Ingredients:
- 3 meatballs
- 1 rib
- 1 tablespoon olive oil
- 0.5 teaspoon salt
- 0.25 teaspoon black pepper
- 1 cup mixed vegetables (e.g., bell peppers, zucchini, carrots)
- 2 tablespoons cooked brown rice

Instructions:
1. Preheat the oven to 375°F (190°C).
2. Lightly oil a baking dish with 1 tablespoon of olive oil.
3. Place 3 meatballs and 1 rib into the baking dish.
4. Season with 0.5 teaspoon salt and 0.25 teaspoon black pepper.
5. Add 1 cup of mixed vegetables to the dish, spreading them evenly around the meat.
6. Cover the baking dish with foil and bake in the preheated oven for 35 minutes.
7. Remove the foil and bake for an additional 10 minutes to let the meat brown.
8. While the meat and vegetables are baking, cook 2 tablespoons of brown rice according to package instructions.
9. Serve the baked meat and vegetables over the cooked brown rice.

Cooking Time: 45 minutes

Cooking Tools:
- oven
- baking dish
- foil

Suggestions:
Consider adding a source of plant-based protein, such as lentils or chickpeas, to increase fiber content.
Include a small amount of nuts or seeds (like almonds or sunflower seeds) for healthier fats.

Figure 9: Output of a Failed Example

In this failed example, the health score is 4, meeting most nutritional criteria. The recipe also satisfies the user's constraints for cooking tools and time. Additionally, it meets the consistency criteria with clear, logical, and easy-to-follow instructions. However, a similarity score of 50% falls below the acceptable score of 70%. This score is due to the limited input provided—only two ingredients, “meatball” and “rib,” both lacking carbohydrates and fibre.

To satisfy health score criteria, the system added additional ingredients such as mixed vegetables and brown rice to create a nutritionally balanced dish. While this adjustment improved the health score, it caused a difference in the recipe ingredients from the original input, resulting in a lower similarity score.

Therefore, the system performs well with balanced input but struggles with limited or unbalanced ingredients.

Compared to Single Prompt

Our system leverages ingredient maps and recipes to provide detailed measurements, logical step sequencing, and enhanced clarity (e.g., "sauté for 5-6 minutes"), delivering higher-quality and more usable results (Figure 10).


```
  "processed_ingredients": [
    "150 grams salmon",
    "200 grams asparagus",
    "1 tablespoon olive oil",
    "1 tablespoon lemon juice",
    "1 pinch salt",
    "1 pinch black pepper",
    "1 teaspoon sesame seeds"
  ],
  "instructions": [
    "1. Heat a skillet over medium-high heat and add 1 tablespoon of olive oil.",
    "2. Season 150 grams of salmon with a pinch of salt and black pepper.",
    "3. Add the salmon to the skillet and cook for 4-5 minutes on each side, or until golden brown and cooked through.",
    "4. Remove the salmon from the skillet and set aside.",
    "5. In the same skillet, add another drizzle of olive oil if needed and add 200 grams of asparagus.",
    "6. Sauté the asparagus for 5-6 minutes, or until tender-crisp.",
    "7. Add 1 tablespoon of lemon juice to the asparagus and toss to coat.",
    "8. Return the salmon to the skillet to warm through, about 1-2 minutes.",
    "9. Sprinkle 1 teaspoon of sesame seeds over the salmon and asparagus before serving."
  ],
```


Figure 10: Our System’s Output (“salmon, asparagus” as input)

Discussion and Learnings

The system demonstrates strong performance, with a success rate of 70%. This indicates that the model is effective in generating recipes that meet user requirements and evaluation criteria in most cases. However, some challenges limit its performance.

Challenges

1. Chain-of-Thought Prompting

Meeting multiple specific macronutrient requirements for health scores posed a challenge. While the chain-of-thought prompting significantly improved the GPT agent's reasoning during intermediate steps, this reasoning process cannot be explicitly included in the final recipe generation. Consequently, the system may miss critical intermediate steps, leading to outputs that do not fully meet expectations.

2. A Trade-off Between Health and Relevance Score

Balancing the health and relevance scores often involves trade-offs. For instance, when users provide limited ingredients, the system may add extra ingredients to improve the health score. While this adjustment enhances nutritional balance, it can lower the relevance score by deviating from the user's original input. This trade-off highlights the complexity of balancing multiple evaluation objectives in recipe generation.

Learning

A key lesson from this project is the importance of using GPT agents judiciously. While GPT agents can provide strong capabilities, their application should be purposeful and carefully integrated into the framework. Avoiding overuse, particularly in tasks requiring extensive calculations or involving ambiguous formulas, and ensuring clear alignment with project goals are essential for achieving better results.

Individual Contributions

Table 5: Team Members and Contributions

Chenhong Qiu	Chujing Yu	Kassy Zeng	Yixuan Liu
Collected popular ingredient list	Built RAG system and health score algorithm	Drafted Recipe Generation Agent	Conducted prompt engineering
Processed recipe dataset	Filtered recipes for nutrient balance	Tested system performance	Tested system performance and Chain of Thought
Developed ingredient nutrient map and Data Processing Agent	Implemented health score and relevance evaluation algorithm	Built user-interface with Gradio	Evaluated single prompt and system
Presented project proposal	Presented final results	Presented final results	Presented project proposal
Completed individual parts of all deliverables			

References

[1] S. Charlebois, “Canada’s Food Guide: Awareness, Understanding, Affordability, and Barriers to Adoption,” Canada’s Food Guide: Awareness, Understanding, Affordability, and Barriers to Adoption (Preliminary Results), https://cdn.dal.ca/content/dam/dalhousie/pdf/management/News/Canada%20Food%20Guide%20March%2014%20EN.pdf. Accessed: Dec. 9, 2024

[2] M. Solan, “Why is eating healthy so hard?,” Harvard Health, https://www.health.harvard.edu/nutrition/why-is-eating-healthy-so-hard. Accessed: Dec. 9, 2024

[3] H. Lee, H. Shu, K. Achananuparp, P. K. Prasetyo, Y. Liu, E. P. Lim, and L. R. Varshney, “RecipeGPT: Generative pre-training based cooking recipe generation and evaluation system,” in *Companion Proceedings of the Web Conference 2020*, Apr. 2020, pp. 181–184. [Online]. Available: https://doi.org/10.1145/3366424.3383536. Accessed: Dec. 9, 2024.

[4] Aljbawi, “Health-Aware Food Planner: A personalized recipe generation approach based on GPT-2,” Scholars Commons @ Laurier. Available: https://scholars.wlu.ca/etd/2311/. Accessed: Dec. 9, 2024.

[5] “Online grocery : Low prices every week | food basics,” Food Basics, https://www.foodbasics.ca/. Accessed: Dec. 9, 2024

[6] R. Lee, “Recipe Box,” *Eight Portions*. Available: https://eightportions.com/datasets/Recipes/. Accessed: Dec. 9, 2024.

[7] H. Canada and the Public Health Agency of Canada, “Canadian Nutrient File (CNF) API Guide,” *Canadian Nutrient File (CNF) API Guide*. Available: https://produits-sante.canada.ca/api/documentation/cnf-documentation-en.html. Accessed: Dec. 9, 2024.

[8] Joint WHO and FAO Expert Consultation, “Diet, nutrition and the prevention of chronic diseases,” World Health Organization Technical Report Series, no. 916, pp. i–viii, 2003. Available: http://health.euroafrica.org/books/dietnutritionwho.pdf. Accessed: Dec. 9, 2024

Appendix A: Test Cases and Results

Test Case	Health Score	Cooking Tool & Cooking Time Constraint	Ingredients Relevance	Consistency
Spring Rolls	5	pass	0%	failed
Egg, Spinach	6	pass	100%	pass
Meatball, Rib	4	pass	50%	pass
Salmon, Asparagus	3	pass	100%	pass
Bacon, Egg, Bread	5	pass	100%	pass
Rib, Pistachios, Green Beans	4	pass	100%	failed
Bagels, grapefruit, Cereal, Honey	4	pass	100%	pass
Pasta, rice, Peanut Butter, Tomato Sauce	4	pass	100%	pass
Egg, Lettuce, Tomato, Cucumber, Cheese	3	pass	100%	pass
Broccoli, Parsley, Swiss Chard, artichoke	5	pass	100%	pass
Butter, Cantaloupe, Salsa, Clams, Dill, Cookies	4	pass	100%	pass
Spring Rolls, Macaroni and Cheese, Fettuccine, Tofu, Raisins	2	pass	100%	pass
Beer, Grapes, duck, Pasta, Bell Peppers, Onion Powder, Crackers	4	pass	60%	pass
Mozzarella, Cider, Kale, bay leaf, Quinoa	5	pass	100%	pass
Waffles, Dairy, Cranberry Sauce, Cocktail Sauce, bay leaf, Vegetable	4	pass	75%	pass

Bread, Shrimp, Radishes, Garlic Powder, Turkey, Duck	4	pass	75%	pass
Eggplant, Coconut Milk, Tofu, Pepper, Thai Basil	4	pass	75%	pass
Beef, tomato, cabbage, celery, carrot, onion, garlic	4	pass	100%	pass
Cheese, Almond Beverage, Bacon, Pollock, Honey, Sunflower Seeds, Quinoa	2	pass	100%	pass
Pepper, Lemon, Squid, Shrimp, Mussel, Crab, Potato, Corn, Broccoli, Garlic	6	pass	100%	pass