Quantitative Evaluation of AI-generated Recipes

Abstract—**The rise of generative Artificial Intelligence (AI) has created the possibility of presenting novel recipes, i.e., recipes that do not exactly match any known recipe and this has led to the creation of AI-based recipe recommendation systems. AI-based recipe recommendation has the possibility of accommodating a variety of preferences – including a person's current health (e.g., diabetes), health goals (e.g., weight loss), taste preferences, cultural or ethical needs (e.g., vegan diet). However, unlike recipes recommended or created by a human dietitian, recipes created by generative AI do not guarantee accuracy, i.e., the generated recipe may not meet the requirements specified by the user. This work quantitatively evaluates how closely recipes generated by OpenAI's GPT4 large language models, created in response to specific prompts, match known recipes in a collection of human-curated recipes. The prompts also include requests for a health condition, diabetes. The recipes are from the largest online community of home cooks sharing recipes (www.allrecipes.com) and the Mayo Clinic's collection of diabetes meal plan recipes. Recipes from these sources are assumed to be authoritative and thus are used as ground truth for this evaluation. Quantitative evaluation using NLP techniques (Named Entity Recognition (NER) to extract each ingredient from the recipes and cosine similarity metrics) enable reporting the quality of the AI results along a continuum. The nutrient attributes of a recipe, such as its total calories, are also evaluated for accuracy. Each ingredient is looked up on USDA's FoodData Central API to retrieve its calories and other nutrient information to provide a true estimate. Our results show that the ingredients list in the AI-generated recipe matches 67-88% with the ingredients in the equivalent recipe in the ground truth database. The corresponding cooking directions match 64-86%. Ingredients in recipes generated by AI for diabetics match those in known recipes in our ground truth datasets at widely varying levels: between 26-83%.**

Index Terms—**GPT, recommender system, USDA, prompt engineering, NER, Mayo Clinic, diabetes, AI assistant**

I. Introduction

With the large amount of food and health-related sources on the Internet, it has become a challenge to identify the most relevant information for a specific person's situation. In particular, there are a large number of cooking recipes that are available but identifying a particular recipe that best matches a person's health, budget, and taste preferences can be a time-consuming process. Recipe recommendation systems have become a popular area of research [1]. The goal is to take into account individual preferences when recommending a recipe. In fact, data-driven approaches for exploring the vast amount of food-related information has given rise to the new field of "computational gastronomy" [2].

Many works use traditional recommender system technologies such as content-based and collaborative filtering [3]. More recently, deep learning and graph neural network-based approaches have been used [4]. However, all these works are restricted to selecting from a given set of recipes. The rise of generative Artificial Intelligence (AI) has now the possibility of presenting novel recipes, i.e., recipes that do not exactly match any recipe in a database, to the user. This has led to the creation of AI-based cooking recipe recommendation systems [5].

AI-based recipe recommendation has the possibility of accommodating a variety of preferences – including a person's current health (e.g., diabetes), health goals (e.g., weight loss), taste preferences, cultural or ethical needs (e.g., vegan diet). However, unlike recipes recommended or created by a human dietitian, recipes created by generative AI do not guarantee accuracy. It is quite possible that the generated recipe may not meet some of the requirements specified by the user.

In this work, we quantitatively evaluate how closely recipes generated by OpenAI's GPT-4 large language models, created in response to specific prompts, match known recipes in a collection of human-curated recipes. The prompts also include requests for a health condition, diabetes. The recipes are from the largest online community of home cooks sharing recipes (www.allrecipes.com) and the Mayo Clinic's collection of diabetes meal plan recipes. Recipes from these sources are assumed to be authoritative and thus are used as ground truth for this evaluation. Quantitative evaluation using string matching and cosine similarity metrics enable reporting the quality of the AI results along a continuum. In particular, we evaluate the ingredients list and the cooking directions parts of a recipe separately. We also evaluate the nutrient attributes of a recipe, such as its total calories, for accuracy. For this, we use Named Entity Recognition (NER) to extract each ingredient from the recipes. Each ingredient is looked up on USDA's FoodData Central API dataset to retrieve its calories and other nutrient information to provide a true estimate.

Our results show that the ingredients list in the AI-generated recipe matches 67-88% with the ingredients in the equivalent recipe in the ground truth database. The corresponding cooking directions match 64-86%. Ingredients in recipes generated by AI for diabetics match those in known recipes in our ground truth datasets at widely varying levels – between 26-83%.

The contributions of this work are: (1) An approach to quantitatively evaluate the quality of AI-generated recipes by comparing it with recipes in a ground truth collection using NLP techniques, and (2) quantitative evaluation of GPT4 models (gpt4-turbo and gpt4-1106) for generating recipes for diabetics.

II. Related Work

Most work on presenting recipes to a user based on specific preferences recommend recipes from a known dataset. Chen et al. [6] introduced an approach for food recommendation based on constrained question answering using a large-scale food knowledge base/graph (KBQA). Their work also integrates user-specific dietary needs and health guidelines into the recommendation process. The proposed KBQA-based framework demonstrated performance improvements over non-personalized methods. Their work was validated with a personalized QA-style dataset. Chen et al. [7] describe a framework designed to assist home cooks in finding recipes that match available ingredients while adhering to healthy eating guidelines. Their approach models ingredient interactions and proportion using an embedding-based predictor for ingredient relevance and a multi-layer perceptron for quantity prediction. This is used to generate a "pseudo-recipe" which is used to search from available recipe datasets. Chavan et al. [3] investigate the use of recommender systems in the nutrition domain to promote healthier dietary choices, leveraging Big Data analytics and machine learning. They focus on the development and evaluation of three recommendation models: content-based, collaborative filtering, and hybrid, each tailored to individual dietary preferences and restrictions. Wang et al. [8] describe a personalized health-aware food recommendation method that maps market ingredients to healthy home-cooked dishes. The method integrates three components: recipe retrieval from a dataset, user health profiles from social network data, and a category-aware hierarchical memory network for health-aware food recommendations. Tian et al. [4] introduce a heterogeneous graph learning model for recipe recommendation. They create user-recipe-ingredient graph to integrate relational structure information among users, recipes, and food items. The model enhances recommendation accuracy through a graph neural network with hierarchical attention and an ingredient set transformer, supported by a graph contrastive augmentation strategy for self-supervised learning. Khilji et al. [9] present a recipe recommendation system that utilizes a threshold parameter from the recommendation engine to ensure only relevant recipes are suggested in response to user queries. Their system integrates a question classification task alongside a question answering module.

As a list of ingredients is a common component across all recipes, researchers have also developed methods to identify ingredients. Goel et al. [10] explore named entity recognition (NER) in the context of recipe text. Their work evaluates different NER methods, including statistical analysis, deep learning model fine-tuning, and few-shot prompting on large language models. They found that the spaCy-transformer model fine-tuned for this task delivers the highest accuracy. Researchers have also used images instead of text as the basis for recipe recommendations. Morol et al. [11] describe a machine learning model using a convolutional neural network (CNN) to recognize food ingredients from images and recommend recipes based on these identifications. They evaluate their system on a custom dataset with 9,856 images across 32 different food ingredient classes.

Relatively few works have explored the possibility of generating new recipes instead of only recommending known recipes. Lee et al. [12] introduce a system, RecipeGPT, for the automatic generation and evaluation of cooking recipes, leveraging a GPT-2 model fine-tuned on a substantial dataset of online recipes. RecipeGPT features two text generation modes: generating cooking instructions from a recipe title and ingredients and generating ingredients from a recipe title and instructions. Additionally, a recipe evaluation module enables users to assess and store the quality of generated recipes for future reference. Our work extends the use of generative AI to output recipes based on specific health requirements. Specifically, we use GPT-3.5Turbo in our work and evaluate the quality of recommendation on datasets from the USDA.

III. Approach

We utilized the GPT-4 (GPT-4-1106 and gpt4-turbo) models from OpenAI using the OpenAI Assistant API to generate recipes. The prompt specifies that details including the recipe name, ingredients, directions, serving size, and total calories for each ingredient should be output. The basic prompt that we use in this work has the following structure:

```
You are a helpful recipe assistant. You
generate recipes in below format:
<recipe>
<recipe_name> {recipe_name} </recipe_name>
<ingredients> {ingredients} </ingredients>
<directions> {directions} </directions>
<nutrition> {nutrition} </nutrition>
</recipe>
Always use above format to give recipe.
```

An example of such an instruction is:

```
You are a helpful recipe assistant. You
generate recipe in below format:
<recipe>
<recipe_name> Apple-Cranberry Crostada
</recipe_name>
<ingredients> 3 tablespoons butter, 2 pounds
Granny Smith apples (or other firm, crisp
apples), peeled, quartered, cored and sliced
...
Optional: Ice cream or lightly sweetened
whipped cream </ingredients>
<directions> Heat butter in a large skillet
over medium-high heat. Add apples, ...
</directions>
<nutrition> Total Fat 18g 23%,
Saturated Fat 7g 34%,
Cholesterol 19mg 6%,
Sodium 128mg 6%,
Total Carbohydrate 60g 22%,
...
</recipe>
Always use above format to give recipe
```

The recipes are output in XML format based on the prompt instructions. The XML format recipes are then converted into JSON format.

After the generation of recipes, we quantitatively evaluate the results along different metrics. For this, we compare each generated recipe with the closest match in a known collection of recipes. These datasets are described next. The processing pipeline for evaluation is shown in Figure 1.

A. Datasets

1) Mayo Clinic diabetes meal plan recipes: The dataset was obtained through web scraping from the Mayo Clinic website[1] using the Beautiful Soup library for Python. This site is recognized for its assortment of healthy recipes. The recipes are sorted based on specific tags provided on the website, such as heart-healthy, low sodium, healthy carbohydrates, gluten-free, weight management, meatless, diabetic-friendly, and high-fiber, which aid in distinguishing between recipes. In this work, we used only the recipes recommended for diabetics. The dataset includes the following attributes: the recipe's name, ingredients, preparation instructions, nutritional analysis per serving, and calorie estimates. Figure 4 shows the top 100 ingredients in the recipes in this dataset.

2) Recipes from www.allrecipes.com: Allrecipes.com is the world's largest Internet-based community of home cooks. Cooks from around the world publish recipes and and share recipe photos and videos, and rate and review recipes. We used a subset of 961 unique recipes from www.allrecipes.com for evaluating the quality of AI-generated recipes. [2] Each URL in the dataset serves as a source link, tracing back to the origin of the recipe on www.allrecipes.com. The dataset includes the following attributes: recipe name, preparation time, cook time, total time, ingredients, directions, serving size, rating, URL, cuisine path, and nutritional information. Figure 5 shows the top 100 ingredients in the recipes in this dataset. It is notable that sugar is used in more than 600 of the 961 recipes.

3) USDA FoodData Central: The USDA FoodData Central API [3] is primarily designed for the users to integrate nutrient's data in their website or applications. The API offers a variety of data sources like Foundation Foods, SR(Standard Reference) Legacy, Surveys Foods (FNDDS, foods and nutrients database for dietary studies), Experimental Foods, and Branded Foods. For this project, we searched the item using the query keyword to efficiently access the food ingredients' weight and energy content.

4) Dataset for volume to weight conversion: We compiled a list of common unit conversions to accommodate the different serving sizes of ingredients with their respective weights in relation to the portions of these serving sizes. In particular, this enables the conversion from volume to weight as needed for all the ingredients in a recipe.

We use cosine similarity between the corresponding text strings from the AI generated recipe and the recipes in our dataset. The cosine similarity is computed separately for the text representing the ingredients and cooking directions.

We also evaluate the accuracy of the numerical attributes of a recipe, such as its total calories and nutrient (e.g., cholesterol) estimates. For this, we use Named Entity Recognition (NER) to extract each ingredient from the recipes. Each ingredient is looked up on USDA's FoodData Central API to retrieve its calories and other nutrient information. The calorie and nutrient information are scaled by the corresponding ingredient size and then summed to get the estimated total calories and nutrient profile for the complete recipe. We then compute and report the difference from the value included in the AI-generated recipe.

B. Closest match to known recipes

In order to quantitatively describe how similar are AI-generated recipes to known recipes, we performed the following series of steps:

1) Select a random recipe from the *allrecipes.com* dataset
2) Generate an AI recipe by prompting for a recipe with the same name as the randomly selected recipe
3) Use cosine similarity to compute the similarity between the ingredients in the selected and AI-generated recipe. Compute the cosine similarity of the directions separately.

C. Evaluating diabetic-friendly recipes

We next utilize the OpenAI Assistant API to randomly generate diabetic-friendly recipes for breakfast, lunch, and dinner. The model used is GPT-4-1106. We used the prompt shown earlier but with the additional sentences "You are helpful diabetic-friendly recipe assistant who give different recipe each time when user ask. Recipe can be for vegetarian, vegan, and non-vegetarian diets." and "Give total calories and calorie per serving."

The AI-generated recipe is matched with recipes in the Mayo Clinic dataset. We used string-matching to find similar recipes in the dataset based on titles (e.g., comparing different "chicken salad" recipes). We quantify the similarity of the best matched recipe by calculating the percentage of matching ingredients. Specifically, we use the following formula:

$$\frac{\text{number of matched ingredients in both recipes}}{\text{number of ingredients in AI-generated recipe}} \times 100$$

D. Evaluating accuracy of reported calories

Recipes also often report nutrient information, most commonly the total number of calories (per serving), carbohydrates, fats, and protein, and total sodium. We evaluated the accuracy of the reported number of calories in AI-generated recipes by comparing it with the estimate obtained from looking up authoritative sources (USDA dataset) for each ingredient in the recipe. The model used is gpt4-turbo. The following prompt was used:

```
You are helpful health-friendly recipe
assistant who give different recipe each time
when user ask. Recipe can be for vegetarian,
```

[1] https://www.mayoclinic.org/healthy-lifestyle/recipes/diabetes-meal-plan-recipes/rcs-20077150

[2] https://github.com/agm316/Food-Your-Way

[3] https://fdc.nal.usda.gov/

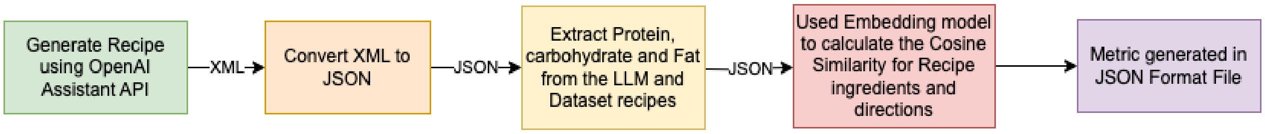
Generate Recipe using OpenAI Assistant API
XML
Convert XML to JSON
JSON
Extract Protein, carbohydrate and Fat from the LLM and Dataset recipes
JSON
Used Embedding model to calculate the Cosine Similarity for Recipe ingredients and directions
Metric generated in JSON Format File

Fig. 1. Processing pipeline for evaluating recipes.

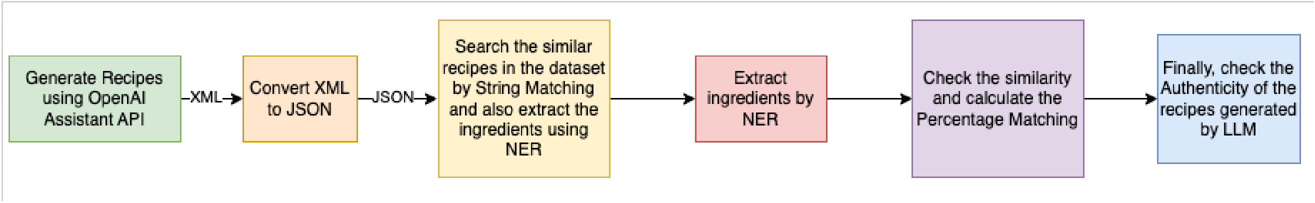
Generate Recipes using OpenAI Assistant API
XML
Convert XML to JSON
JSON
Search the similar recipes in the dataset by String Matching and also extract the ingredients using NER
Extract ingredients by NER
Check the similarity and calculate the Percentage Matching
Finally, check the Authenticity of the recipes generated by LLM

Fig. 2. Processing pipeline for evaluating recipes.

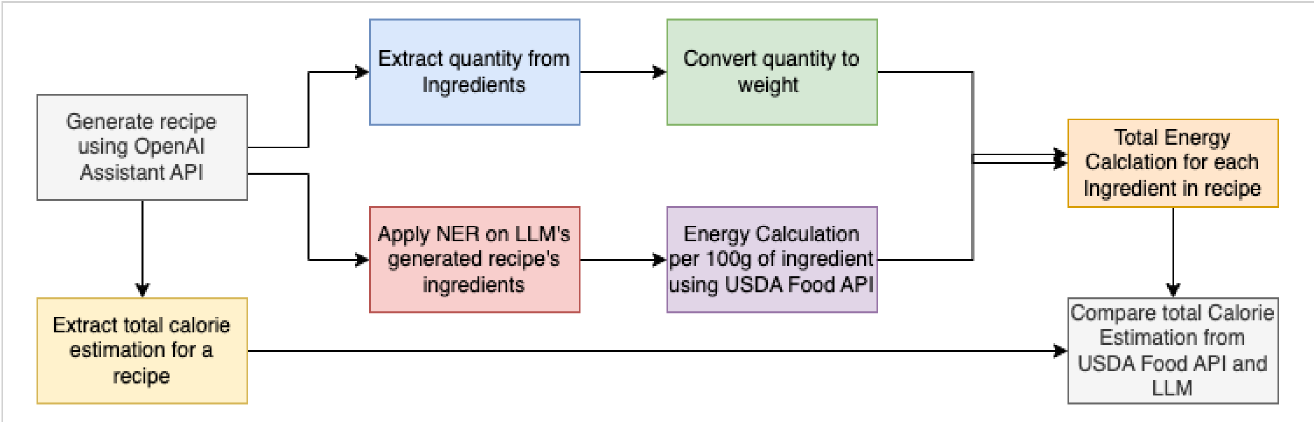
Generate recipe using OpenAI Assistant API
Extract quantity from Ingredients
Convert quantity to weight
Apply NER on LLM's generated recipe's ingredients
Energy Calculation per 100g of ingredient using USDA Food API
Total Energy Calclation for each Ingredient in recipe
Extract total calorie estimation for a recipe
Compare total Calorie Estimation from USDA Food API and LLM

Fig. 3. Processing pipeline for evaluating the accuracy of reported calories in an AI-generated recipe.

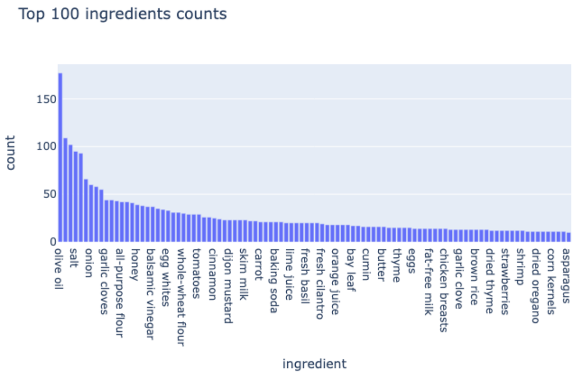

Fig. 4. Top 100 ingredients in the recipes in the Mayo Clinic dataset.

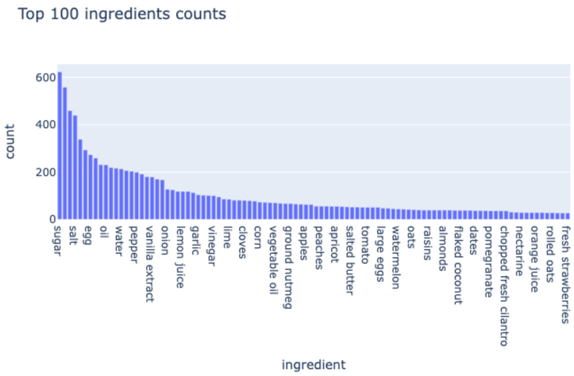

Fig. 5. Top 100 ingredients in the recipes in the www.allrecipes.com dataset.

```
vegan, and non-vegeterian diets. You generate
diabetic friendly recipe in below format:
<recipe>
  <recipe_name> {recipe_name} </recipe_name>
  <ingredients> {ingredients} </ingredients>
  <directions> {directions} </directions>
  <nutrition> {nutrients} </nutrition>
  <total_calories_estimation>
    {total_calorie_estimation}
  </total_calories_estimation>
</recipe>
Always use above format to give recipe.
Give total calories for recipe and refer
USDA Food API for calories.
```

We used a Named Entity Recognition (NER) algorithm on the ingredients of the AI-generated recipe to extract the names of the ingredients. We extracted the quantity of each ingredients by pattern matching using regular expressions. Energy calculations are performed for each ingredient per 100 grams using the USDA Food API. We converted any volume (e.g. ”cups of flour”) measurements to weight in grams. Results from the volume-to-weight conversion and per 100 grams calculations are combined to calculate the total energy for each ingredient in the recipe. We extract the total calories reported in the AI-generated recipe by using regular expressions. We then compare the total calories included in the AI-generated recipe and our calculated estimate obtained from the USDA Food APIs. This sequence of steps is shown in Figure 3.

IV. Results and Discussion

A. Closest match to known recipes

Figure 6 shows how closely does the ingredients list in the GPT4-generated recipe for a specific recipe name match the known recipe in the allrecipes.com dataset. Figure 7 shows the corresponding results for matching the directions.

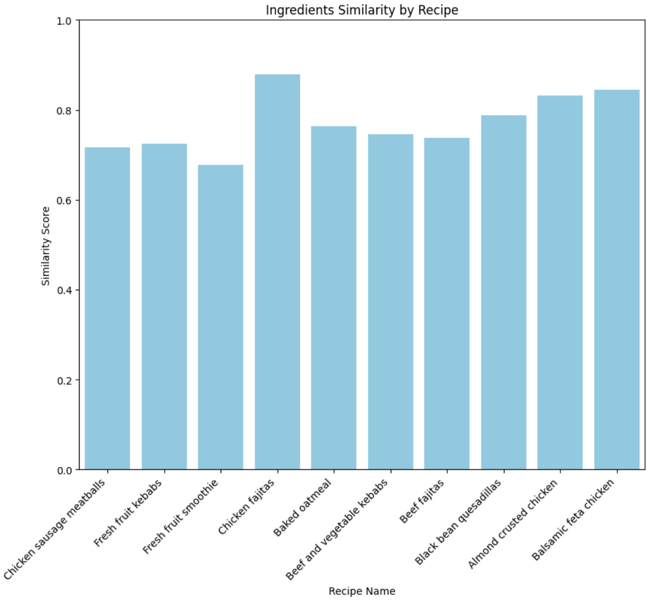

Fig. 6. Similarity score when matching the list of ingredients in a known recipe and an AI-generated recipe with the same name.

These results show that the ingredients list in the AI-generated recipe matches 67-88% with the ingredients in the equivalent recipe in the ground truth database, with an average of 77.1%. The corresponding cooking directions match 64-86% with an average of 74.3%.

B. Evaluating diabetic-friendly recipes

Table I shows the closest matched known recipe in the Mayo Clinic dataset when the GPT4 model is prompted for a diabetic-friendly recipe.

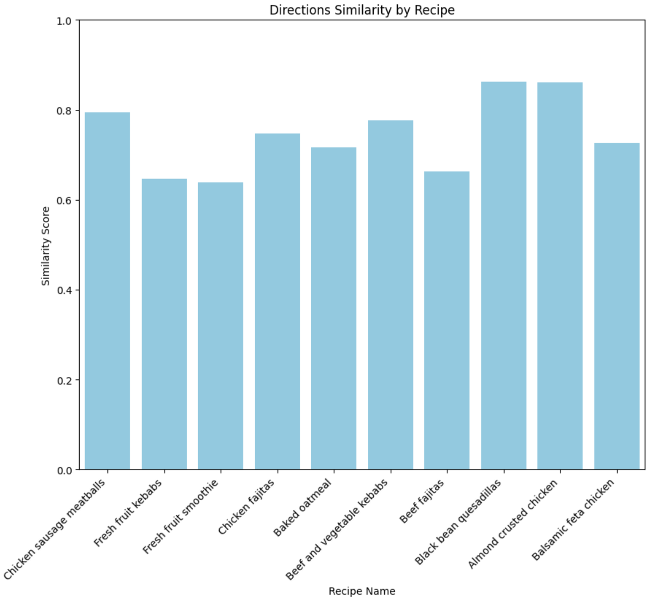

Fig. 7. Similarity score when matching the cooking directions in a known recipe and an AI-generated recipe with the same name.

We notice that the ingredients in recipes generated by AI for diabetics match those in known recipes in our ground truth datasets at widely varying levels – between 26-83%. For instance, the *Tuna and Chickpea salad* is matched with *Gazpacho with chickpea* which most likely does not have tuna. The *Almond and blueberry smoothie* is matched with one that is an orange smoothie, not blueberries. This indicates that AI-generated recipes can be improved by requiring a greater weight be assigned to the more important ingredients, either as part of the prompt, or in a post-processing step.

C. Evaluating accuracy of reported calories

Figure 8 shows the error in calories reported in AI-generated recipes when compared to the value calculated by looking up the calories in each ingredient according to the USDA FoodData Central API.

The mean absolute percentage error (MAPE) of the calories across the 14 recipes was 14%. Thus, we may be consider that the error in calories reported by GPT4-generated recipes is relatively small.

V. Conclusions and Future Work

We presented an approach to quantitatively evaluate the quality of AI-generated recipes by comparing it with a trusted collection of recipes. The data processing pipeline for this evaluation uses standard NLP techniques, including Named Entity Recognition (NER) to extract each ingredient from the recipes and cosine similarity to assign a match score on a continuum. We applied this data processing pipeline to evaluate the quality of GPT4 models for generating recipes for diabetics. The results show that the ingredients list in the AI-generated recipe matches 67-88% (average 77.1%) with the ingredients in the equivalent recipe in the ground truth

TABLE I
CLOSEST MATCH TO AI-GENERATED RECIPES

AI-generated recipe	Closest match to known recipe	Percentage match
Scrambled tofu with spinach and tomato	Vegetarian chili with tofu	44.4
Almond and blueberry smoothie	Orange dream smoothie	55.5
Almond flour pancakes	Whole-grain pumpkin pancakes	81.8
Spinach and mushroom egg Frittata	Southwestern frittata	83.3
Spinach and mushroom Frittata	Spinach and mushroom Frittata	75
Veggie-packed frittata	Smokey frittata	72.7
Chickpea and salad wraps	Chickpea polenta with olives	53.8
Mediterranean Chickpea salad	Chickpea polenta with olives	76.9
Quinoa Chickpea salad jars	Gazpacho with Chickpea	76.9
Tuna and Chickpea salad	Gazpacho with Chickpea	63.6
Grilled chicken salad with avocado dressing	Chicken salad with thai flavors	66.7
Grilled lemon herb chicken salad	Grilled chicken salad with olives and oranges	80.0
Stuffed bell peppers	Roasted red bell pepper pineapple salsa	29.4
Stuffed bell peppers with quinoa and black beans	Roasted red bell pepper pineapple salsa	26.7

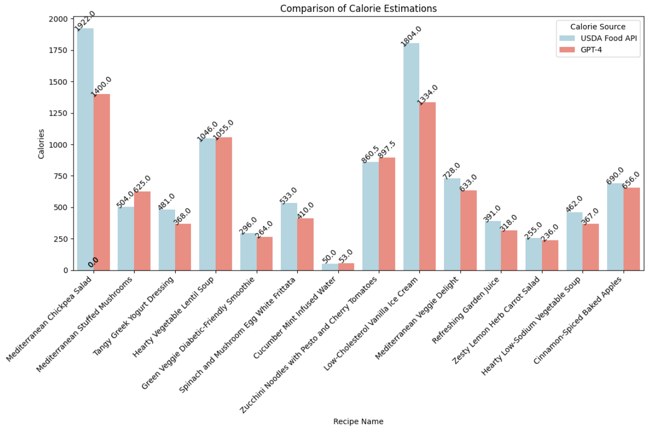

Fig. 8. Estimated difference in calories reported in AI-generated recipe from that computed using USDA FoodData API.

database. The corresponding cooking directions match 64-86% (average 74.3%). However, the ingredients in recipes generated by AI for diabetics match those in known recipes in our ground truth datasets at widely varying levels – between 26-83%. This indicates that AI-generated recipes can be improved by assigning a greater weight to the more important ingredients. On the other hand, the mean absolute percentage error in calories reported by GPT4-generated recipes is relatively small (0.14%).

For future work, we intend to develop more sophisticated prompts that account for the relative importance of ingredients in a recipe. We will also evaluate the other nutrient attributes of AI-generated recipes, such as the carbohydrate, fat, cholesterol, and sodium levels.

ACKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation under Grant No. 2125654.

REFERENCES

[1] M. Ge, F. Ricci, and D. Massimo, "Health-aware food recommender system," in *Proceedings of the 9th ACM Conference on Recommender Systems*, 2015, pp. 333–334.

[2] M. Goel and G. Bagler, "Computational gastronomy: A data science approach to food," *Journal of Biosciences*, vol. 47, no. 1, p. 12, 2022.

[3] P. Chavan, B. Thoms, and J. Isaacs, "A recommender system for healthy food choices: building a hybrid model for recipe recommendations using big data sets," in *Proceedings of the 54th Hawaii International Conference on System Sciences*, 2021.

[4] Y. Tian, C. Zhang, Z. Guo, C. Huang, R. Metoyer, and N. V. Chawla, "Reciperec: A heterogeneous graph learning model for recipe recommendation," *arXiv preprint arXiv:2205.14005*, 2022.

[5] R. Yera, A. A. Alzahrani, and L. Martínez, "Exploring post-hoc agnostic models for explainable cooking recipe recommendations," *Knowledge-Based Systems*, vol. 251, p. 109216, 2022.

[6] Y. Chen, A. Subburathinam, C.-H. Chen, and M. J. Zaki, "Personalized food recommendation as constrained question answering over a large-scale food knowledge graph," in *Proceedings of the 14th ACM international conference on web search and data mining*, 2021, pp. 544–552.

[7] M. Chen, X. Jia, E. Gorbonos, C. T. Hoang, X. Yu, and Y. Liu, "Eating healthier: Exploring nutrition information for healthier recipe recommendation," *Information Processing & Management*, vol. 57, no. 6, p. 102051, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S030645731930161X

[8] W. Wang, L.-Y. Duan, H. Jiang, P. Jing, X. Song, and L. Nie, "Market2dish: health-aware food recommendation," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 17, no. 1, pp. 1–19, 2021.

[9] A. F. U. R. Khilji, R. Manna, S. R. Laskar, P. Pakray, D. Das, S. Bandyopadhyay, and A. Gelbukh, "Cookingqa: answering questions and recommending recipes based on ingredients," *Arabian Journal for Science and Engineering*, vol. 46, no. 4, pp. 3701–3712, 2021.

[10] M. Goel, A. Agarwal, S. Agrawal, J. Kapuriya, A. V. Konam, R. Gupta, S. Rastogi, Niharika, and G. Bagler, "Deep learning based named entity recognition models for recipes," 2024.

[11] M. K. Morol, M. S. J. Rokon, I. B. Hasan, A. Saif, R. H. Khan, and S. S. Das, "Food recipe recommendation based on ingredients detection using deep learning," in *Proceedings of the 2nd International Conference on Computing Advancements*, 2022, pp. 191–198.

[12] H. H. Lee, K. Shu, P. 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*, 2020, pp. 181–184.