PersonaPlate: Contextual Language Modeling for Individualised Diets

Ajjkumar Patel **Mansi Nanavati** **Shweta Pardeshi**
University of California, San Diego
{ajp011, mnanavati, spardeshi}@ucsd.edu

Abstract

The increasing prevalence of personalized health and wellness trends necessitates advanced approaches for tailored dietary recommendations. This paper introduces PersonaPlate: Contextual Language Modeling for Individualised Diets, a novel framework designed to generate personalized meal plans tailored to individual dietary needs and preferences. We present and evaluate three distinct methods for achieving this: a multi-agent collaborative system, a Retrieval Augmented Generation (RAG)-based approach integrating external nutritional information, and a QLoRA fine-tuned Llama-3.2-1B Large Language Model. Our evaluation methodology focuses on quantitatively analyzing the macro and micro-nutrient intake of generated meal plans against the average required nutrients for various user profiles. Results indicate that the multi-agent collaborative system significantly outperforms other methods, followed by the RAG-based system, while the QLoRA fine-tuned Llama-3.2-1B model exhibited limited coherence in its outputs. Link to the Github repo for this project: `https://github.com/glitched-shadeslayer/PersonaPlate_ai_nutritionist`

1 Introduction

In recent years, the advent of Large Language Models (LLMs)[1] has shifted the focus of artificial intelligence research toward their integration into domains that demand comprehensive data understanding and expert-level decision-making. Among these, healthcare has emerged as a prominent application area, where LLM-based systems are increasingly being developed to interpret complex medical data, support clinical reasoning, and assist in diagnostic and therapeutic processes. Within the broader scope of healthcare, nutrition represents a critical yet underexplored niche.

Despite the growing interest in personalized medicine, there remains a lack of accessible AI-driven nutritionists capable of generating individualized diet plans grounded in a user's medical history and lifestyle data. Existing AI nutritionists typically offer generalized recommendations without adequately accounting for individuals' comprehensive medical histories, genetic predispositions, and specific health conditions [2]. Most existing solutions focus on basic calorie counting or adherence to predetermined diet patterns, failing to integrate critical personalized medical data that could significantly impact nutritional needs [2].

In this work, we address this gap by introducing an advanced AI nutritionist that leverages the capabilities of LLMs to provide personalized, context-aware dietary recommendations. Specifically, our research addresses the following key objectives: (1) integrating comprehensive medical histories; (2) generating personalized dietary recommendations tailored to individual health profiles; (3) developing predictive models for future health parameters (e.g., blood glucose, vitamin, and hormone levels) based on dietary adherence; and (4) incorporating the impact of human behavioral variability, such as missed or skipped meals[2], [3].

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

2 Motivation

The development of a personalized AI nutritionist addresses several critical healthcare challenges.

Health Impact and Chronic Disease Management. Chronic diseases, including diabetes and hypertension, represent major global health challenges with significant economic and social costs [4]. A personalized AI nutritionist could dramatically improve health outcomes by providing tailored dietary guidance based on an individual's medical history.

Accessibility to Nutritional Expertise. Professional nutritionists and dietitians remain inaccessible to many due to cost and geographical constraints [2]. An AI nutritionist provides access to personalized nutrition advice without the limitations of traditional healthcare models [5].

Integration of Medical Data. Diet recommendations that fail to consider an individual's complete medical profile can be ineffective [4]. Our proposed solution addresses the challenge of integrating medical data to develop truly personalized nutritional guidance, representing a significant advancement in precision nutrition [3].

3 Literature Review

3.1 LLM in dietary recommendation

Recent years have seen a surge of interest in applying artificial intelligence to improve dietary guidance. While early tools focused on calorie tracking or rigid meal plans, newer systems increasingly offer flexible, personalized recommendations. For example, MedDietAgent[5] leverages reinforcement learning and computer vision to promote adherence to the Mediterranean diet, demonstrating strong alignment with population-level dietary targets. However, its recommendations are not tailored to individual medical conditions or clinical markers.

Other research has explored preference-based models, such as collaborative filtering and variational autoencoders, to suggest meals users are likely to enjoy [2]. Although these approaches offer personalization based on user tastes, they typically ignore the physiological effects of specific foods. Amiri et al.[4] proposed a more adaptive framework using fuzzy logic and multicriteria decision-making to accommodate dietary restrictions. While effective for managing constraints, such rule-based systems lack the flexibility to respond dynamically to evolving health data or behavior patterns.

Large language models have shown promise in healthcare tasks such as clinical summarization and decision support. Alkhalaf et al.[3] fine-tuned transformer-based models for nutrition decision-making in residential aged care using structured medical records, but their approach was limited to static datasets and predefined prompts.

3.2 LLM-based Multi-Agent Systems

With recent advances in Large Language Models (LLMs), Agentic AI [6], [7] has become phenomenal in real-world applications, moving toward multiple LLM-based agents to perceive, learn, reason, and act collaboratively. These LLM-based Multi-Agent Systems (MASs) enable groups of intelligent agents to coordinate and solve complex tasks collectively at scale, transitioning from isolated models to collaboration-centric approaches [8]. These systems are structured around components such as agent profiles, perception, self-action, interaction, and evolution, allowing agents to dynamically adapt and operate within various environments [9]. MedAgents [10], proposed by Tang et. al., leverages LLM-based agents in a role-playing setting that participate in a collaborative multi-round discussion, thereby enhancing LLM proficiency and reasoning capabilities. We will take inspiration from these papers to implement our multi-agent system.

3.3 Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) has recently gained traction as a way to enhance LLM outputs by incorporating retrieved context from external corpora at inference time. Originally introduced by Lewis et al. [11], RAG systems combine dense neural retrieval with text generation, improving factual grounding without retraining. Despite its success in biomedical question answering, multi-document summarization, and evidence-grounded dialogue systems [12], the application of RAG to personalized dietary planning remains underexplored. Most existing RAG pipelines are optimized for static question answering or factual retrieval, rather than dynamic synthesis of structured meal plans based on individual medical profiles.

3.4 LoRA Fine-Tuning

Low-Rank Adaptation, or LoRA,[13] freezes the pre-trained model weights and injects trainable rank decomposition matrices into each layer of the Transformer architecture, greatly reducing the number of trainable parameters for downstream tasks. LoRA is a fine-tuning technique that reduces the number of trainable parameters in large language models. In this project, we explore QLoRA fine-tuning technique. Quantized Low-Rank Adapter (QLoRA)[14] is a parameter-efficient fine-tuning method that enables large language models to be trained using 4-bit quantization, significantly reducing memory usage without sacrificing performance. It combines low-rank adaptation (LoRA) with quantized model weights, allowing full-model fine-tuning on consumer hardware like a single GPU. Such an initialization can be useful to improve generalization as well as computation usage with the help of quantization.

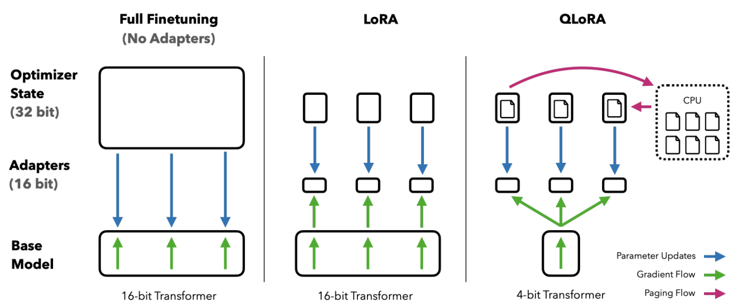

Figure 1: Different finetuning methods and their memory requirements. QLORA improves over LoRA by quantizing the transformer model to 4-bit precision and using paged optimizers to handle memory spikes.

4 Methodology

We aim to build a modular AI system that uses pretrained large language models (LLMs) to generate personalized dietary plans based on a user's medical and lifestyle data. Instead of training new models, we'll rely on open-source LLMs capable of interpreting both structured inputs (like lab results) and unstructured data (like meal logs or symptoms). As part of determining which AI system serves our purpose best, we are following three different techniques: (i) Multi-agent collaboration, (ii) RAG Based search and generation, and (iii) LoRA fine-tuned small language model.

4.1 Multi-Agent Collaboration

In this system, there exist four domain expert agents, each responsible for a specific subtask in the meal planning pipeline.

- Nutrition Agent ($agent_n$): Computes nutritional requirements based on an initial user report R_o, which includes health goals, dietary preferences, and restrictions.
- Planner Agent ($agent_r$): Generate a meal plan consisting of recipes or meals for a given time period.
- Optimizer Agent ($agent_o$): Refines the meal plan to minimize deviations from nutritional targets while optimizing secondary objectives.
- Feedback Agent ($agent_f$): Evaluates the plan's quality against report R_o, and provides feedback to guide further refinements.

The agents collaborate iteratively, sharing information to progressively improve the meal plan until it meets predefined criteria or exhausts a maximum number of attempts. The collaborative meal planning process is formalized in Algorithm 1 .

Algorithm 1: Collaborative Meal-Planning

Input: Domain experts agents $A = \{agent_1, \ldots, agent_4\}$, initial report R_o, maximum attempts t, prompts $\{p_{\text{nutrition}}, p_{\text{planner}}, p_{\text{optimizer}}, p_{\text{feedback}}\}$
Output: Final plan P_f

```
//Initialize variables
```
$optimized \leftarrow False, \ \ n_{try} \leftarrow 0$
$nutri_{req} \leftarrow agent_n(p_{nutrition}, R_o)$
```
//Iterative planning
```
while $optimized\ is\ False$ ***and*** $n_{try} < t$ **do**
- $n_{\text{try}} \leftarrow n_{\text{try}} + 1$;
- $recipe_{plan} \leftarrow agent_r(p_{planner}, nutri_{req})$;
- $optimized_{plan} \leftarrow agent_o(p_{optimizer}, recipe_{plan}, nutri_{req})$;
- $feedback \leftarrow agent_f(p_{feedback}, R_o, optimized_{plan})$;
- $optimized \leftarrow feedback$;
- **if** $optimized\ is$ no **then**
  - //propose modifications
  - $recipe_{plan} \leftarrow agent_r(p_{planner}, nutri_{req}, feedback)$;
  - $optimized_{plan} \leftarrow agent_o(p_{optimizer}, recipe_{plan}, nutri_{req}, feedback)$;
  - $feedback \leftarrow agent_f(p_{feedback}, R_o, optimized_{plan})$;
  - $optimized \leftarrow feedback$;
- **else**
  - $optimized \leftarrow True$

return $P_f \leftarrow optimized_{\text{plan}}$

4.1.1 Mathematical Formulation

We model meal planning as a constrained optimization problem. Let $P = m_1, m_2, ..., m_k$ represent a meal plan, where each m_i is a meal with associated nutritional values. Define $N(P)$ as a function mapping P to a nutritional vector $N(P) = [calories, fat, carbs, protein, \ldots]$ and R as the target nutritional vector derived from $nutri_{req}$. The objective is to minimize the deviation between $N(P)$ and R:

$$J(P) = ||N(P) \ - \ R||_2^2$$

where $||.||_2$ is the L2 norm, capturing the squared difference across all nutritional dimensions. The optimization is subject to constraints:

- **Nutritional Bounds:** $N_{min} \leq N(P) \leq N_{max}$, ensuring each nutrient falls within acceptable ranges.
- **Dietary Restrictions:** Constraints such as allergen avoidance or vegetarian preferences, encoded in R_o.

Convergence: The iterative loop converges to an optimized plan if the Feedback Agent accurately evaluates P and the Optimizer Agent reduces $J(P)$ in each iteration. Given that our system employs a LLM in the backend, convergence to an optimized plan is not always guaranteed, as LLMs may introduce stochasticity in agent outputs. However, the collaborative multi-agent framework ensures that the quality of the meal plan is likely to improve compared to the initial $recipe_{plan}$, as each iteration refines the plan based on nutritional targets and feedback. To prevent excessive recursive computation, we limit the number of iterations to a maximum of t attempts, balancing computational efficiency with quality improvement.

4.2 RAG-Based Search and Generation

We develop a retrieval-augmented generation (RAG) pipeline to produce personalized daily meal plans grounded in structured health profiles. The pipeline consists of the following key components:

- **Knowledge Base Construction:** We combine two nutrition-focused datasets: (1) the USDA SuperTracker food database, which includes structured nutrient values (e.g., calories, fat, sodium), and (2) a curated set of real-world recipes annotated with ingredients. Each record x_i is converted into a natural language sentence (e.g., "Spinach: 2.9g protein, 0g fat, 23 kcal") and stored as part of a corpus $X = \{x_1, x_2, \ldots, x_n\}$.
- **Semantic Retrieval:** Each document $x_i \in X$ is embedded using a sentence transformer $f(\cdot)$, yielding vector representations $v_i = f(x_i) \in \mathbb{R}^d$. At inference time, a user profile q—comprising medical history, lab results, BMI, and symptom descriptions—is embedded as $v_q = f(q)$. We retrieve the top-k documents from X using cosine similarity:

$$\text{Sim}(v_q, v_i) = \frac{v_q \cdot v_i}{\|v_q\| \, \|v_i\|}, \quad x_i \in \text{TopK}(X, v_q).$$

- **Meal Plan Generation:** The retrieved set $R = \{x_{i_1}, \ldots, x_{i_k}\}$ is concatenated with the user profile to form a structured prompt $P = [q; R]$, where $[\,\cdot\,;\,\cdot\,]$ denotes text concatenation. This prompt is passed to an open-weight LLM (e.g., Mistral-7B-Instruct), which generates a meal plan structured by meal type (breakfast, lunch, snacks, dinner). The model is instructed to prioritize foods from the retrieved context and to adhere to clinical constraints (e.g., low sodium for hypertension).

4.3 QLoRA Fine-Tuned Small Language Model

LLMs are known to have billions of parameters. Considering the computational limitations of Datahub (1 GPU, 16GB RAM), we are applying fine-tuning techniques to smaller language models to facilitate strong performance despite fewer parameters. Models such as Llama-3.2[15], SmolLM2[16], and Gemma3-1b[17] were some of the candidates for such a language model. After experimenting with all three models, the *"meta-llama/Llama-3.2-1B"* gave more coherent answers.

NutriBench is a publicly available benchmark of 11,857 human-verified meal descriptions annotated with macronutrient labels, designed to support accurate nutrition estimation from natural language. For fine-tuning, each NutriBench example is formatted as a dialogue-style prompt where the user provides a meal description and the assistant responds with a structured nutritional breakdown. The text is tokenized with a maximum sequence length of 512, and the input IDs are duplicated as labels to enable supervised learning with causal language modeling. This approach helps the model learn to generate accurate nutrient estimates directly from natural language inputs.

In our pipeline, the base model is Llama-3.2-1B. We are configuring LoRA parameters, specifically setting `lora_r` to 64 for rank approximation and `lora_alpha` to 16 for scaling. For enhanced memory efficiency, the base model was loaded with **4-bit precision** utilizing $nf4$ quantization. Training was done with PeftModel and SFTTrainer modules, and an `AdamW` optimizer. It was fine-tuned for 2 epochs. This integrated configuration facilitated robust supervised fine-tuning, optimizing for performance while maintaining resource efficiency on a single GPU.

5 Experiments

We compare three approaches for meal plan generation: (1) a baseline "Normal" planner using a single-agent LLM without iterative refinement, (2) a multi-agent system that decomposes the task into nutrient-to-ingredient and ingredient-to-meal sub-agents, (3) a retrieval-augmented generation (RAG) system that grounds outputs in structured nutritional and recipe knowledge retrieved at inference time, and (4) QLoRA fine-tuned Llama-3.2-1B model's output. Table 1 summarizes the meal plans, and Table 2 provides the total daily nutrient breakdown for user profile with hypertension and vegetarian dietary preference. Evaluation considers both macronutrients (calories, fat, carbohydrates, protein) and key micronutrients (fiber, sodium, potassium, calcium, vitamin D, omega-3, saturated fat, sugar).

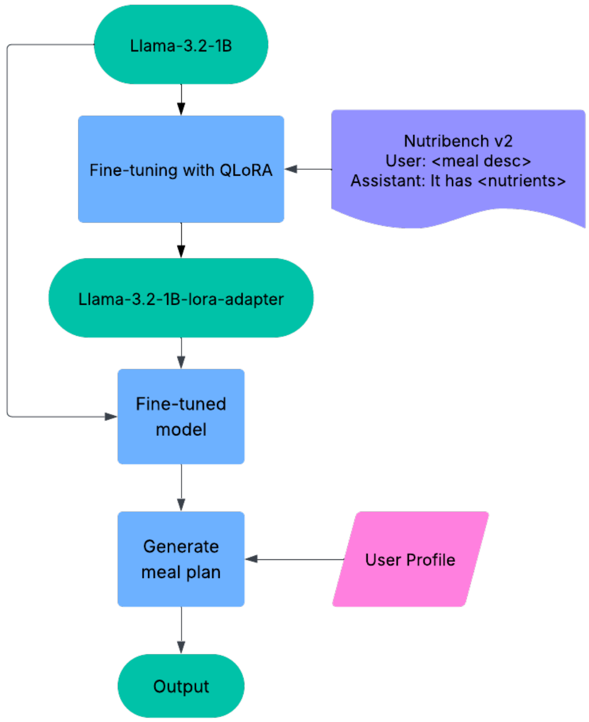

Figure 2: Pipeline for fine-tuning Llama-3.2-1B over Nutribench dataset using QLoRA technique

5.1 Meal Planning Analysis

The multi-agent system achieved a moderate increase in caloric coverage (+3%) and showed consistent improvement across most micronutrients. Notably, it increased fiber and omega-3 by 40%, and reduced sodium intake by 46.67% compared to the baseline [18]. Additionally, it successfully generated an entirely vegetarian meal plan, whereas the baseline system failed to meet that constraint. This highlights the benefits of iterative planning and role specialization in the multi-agent approach.

The RAG-based system produced a more conservative plan with significantly lower total calories (1377 kcal) but achieved the highest protein density (102.32g) and the lowest sodium (600mg), sugar (10g), and saturated fat (6% of kcal). These results indicate RAG's strong alignment with condition-specific nutritional priorities—especially for cardiovascular health. However, its output lacked caloric sufficiency, suggesting that retrieval-guided generation may benefit from explicit constraints on energy balance or retrieval-coverage objectives. In addition, the system faced challenges with output consistency. Meal plans occasionally omitted sections (e.g., snacks or dinner), repeated input context, or deviated from the expected structure—requiring post-processing to extract usable outputs. Prompt refinement and response filtering were necessary to achieve reliable formatting in most cases.

5.2 Analysis Macro Nutrients

Figure 3 (left) compares sodium prescriptions for 50 hypertensive profiles. The single-agent baseline routinely overshoots the recommended *1500 mg* ceiling, exceeding it in ~60 % of the cases and occasionally crossing even the *2300mg* general-adult limit. In contrast, the proposed multi-agent scheduler (MAS, black) keeps majority of the plans at or below the hypertension guideline with mean value of *1469mg*. MAS also succeeds on three profiles for which the baseline fails to return any plan (filled markers), underscoring its improved robustness. Figure 4 reports results for an obesity cohort, with calories (left) and total fat (right). Across all 50 profiles MAS reduces caloric prescriptions by an average of *210 kcal (-12 %)* and positions all of the plans inside the *2000 kcal* reference window, compared with only 82% for the baseline. A similar trend holds for fat: MAS keeps all of meals

Meal	Meal Name	Plan Type	Cal.	Fat (g)	Carbs (g)	Prot. (g)
Breakfast	Oatmeal with Banana, Almond, and Honey	Normal	400	7	60	10
	Whole-grain Oatmeal+Almond Butter, Banana, Walnuts	Multi-Agent	650	11	95	18
	Oatmeal with Cooked Heart, Palm Hearts, Mint	RAG-Based	379	4.9	35	31.9
	Breaded Chicken with Banana	QLoRA-finetuned	573	19	50.5	24.2
Lunch	Grilled Chicken and Vegetable Wrap	Normal	500	22	110	40
	Grilled Vegetable and Avocado Wrap	Multi-Agent	880	20	60	25
	Grilled Chicken with Greens and Vinaigrette	RAG-Based	340	3.5	20	30
	Steamed Rice with Sweet Potato and Mixed Sausages	QLoRA-finetuned	201	29.75	33.33	6.3
Dinner	Lentil and Vegetable Curry	Normal	550	12	70	20
	Quinoa+Black Bean Bowl with Roasted Vegetables	Multi-Agent	655	10	90	22
	Grilled Salmon with Quinoa and Broccoli	RAG-Based	330	10	20	24
	Chicken, Ground Beef, and Cooked Potatoes Dinner	QLoRA-finetuned	700	41.24	29.32	51.94
Snacks	Apple Slices with Almond Butter, Pear	Normal	210	18	35	4
	Apple Slices with Almond Butter, Pear, Walnuts	Multi-Agent	390	20	39	8
	Almonds + Energy Drink, Apple + Yogurt	RAG-Based	328	16.5	30	16.42
	Cooked White Rice Snack with Sugar	QLoRA-finetuned	220	1.07	47.29	4.94
Total Daily		Normal	2500	59	275	74
		Multi-Agent	2575	61	284	73
		RAG-Based	1377	34.9	105	102.32
		QLoRA-finetuned	1694	91.06	160.44	87.38

Table 1: Meal plan comparison across Normal, Multi-Agent, and RAG-based generation for a user with hypertension.

Plan Type	Fib. (g)	Sod. (mg)	Pot. (mg)	Cal. (mg)	Vit. D (IU)	O-3 (mg)	Sat. Fat (% kcal)	Sug. (g)
Normal	25	1500	3800	1000	600	250	10	20
Multi-Agent	35	800	4500	1100	700	350	10	15
RAG-Based	22	600	3700	900	500	300	6	10
QLoRA-finetuned	4.52	4140	N/A	0.05	N/A	N/A	N/A	0.0
Diff (Multi-Agent vs Normal)	**+40%**	**-46.67%**	**+18.42%**	**+10%**	**+16.67%**	**+40%**	**0%**	**0%**
Diff (RAG vs Normal)	**-12%**	**-60%**	**-2.63%**	**-10%**	**-16.67%**	**+20%**	**-40%**	**-50%**
Diff (QLoRA-finetuned vs Normal)	**-81.92%**	**+176%**	**N/A**	**-99.995%**	**N/A**	**N/A**	**N/A**	**-100%**

Table 2: Total Daily Nutrient Breakdown: Normal vs Multi-Agent vs RAG-based vs QLoRA-finetuned LLM Meal Plans.

inside the Dietary Reference Intake band of 45-78g and majority towards lower end, whereas the baseline fluctuate a lot. Notably, MAS never exceeds the upper fat bound.

Figure 3 (right) illustrates sodium intake across 50 hypertensive profiles generated using the RAG-based planner. The RAG system demonstrates consistent alignment with dietary constraints, with *94% of profiles* below the 1500 mg hypertension threshold. No plan exceeds the general adult upper limit of 2300 mg. The average intake is approximately *1380 mg*, with controlled fluctuation and no outliers. These results highlight RAG's ability to generate clinically relevant recommendations while maintaining stability across diverse medical inputs.

5.3 QLoRA Limitations

The QLoRA fine-tuned Llama-3.2-1B was unable to generate as coherent a response compared to the two prior methods. The model is inconsistent with the output format. It injects "User" and "Assistant" in the output, which could be due to the data format it was fed from Nutribench while fine-tuning. This implies that the model is memorizing, making it unfit for meal plan generation across diverse user profiles. The response from the model is also not correctly able to calculate the nutrients such as proteins, sodium, fats, carbohydrates, and calories (refer to Table 2 for comparison). Due to the inconsistent and incompetent generation from the model, we did not generate over multiple profiles since most responses did not mention the nutrient intake from the meal. To further improve the fine-tuning process, we recommend using a mid-sized LLM (for example: Llama-3.2-3B or 5B) as the base model and a higher quantization value for improved precision.

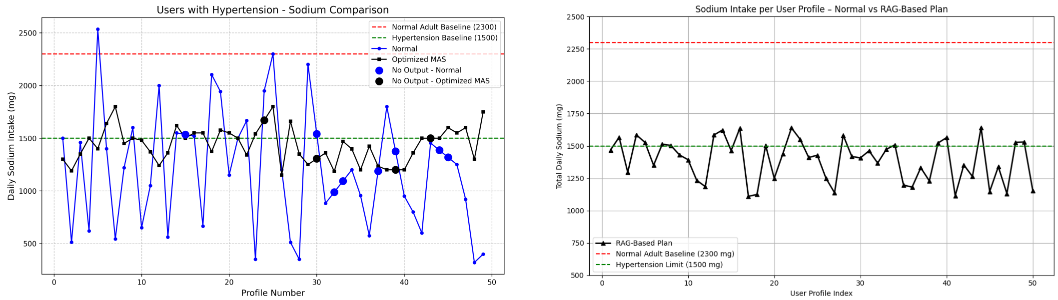

Figure 3: (a) Daily dietary sodium prescribed for 50 hypertensive user profiles. Blue lines show the "Normal" (single-agent) meal plans currently produced by the system; black lines show the plans produced by our optimized multi-agent system (MAS). The horizontal dashed lines mark guideline–based upper limits for adults without hypertension (red, 2300 mg per day) and with hypertension (green, 1500 mg per day). Large filled markers indicate profiles for which the respective planner failed to return a meal plan ("No Output"). (b) Sodium intake distributions generated using the RAG-based planner for the same 50 profiles. The RAG system successfully maintains sodium levels below the 1500 mg hypertension threshold in 94% of cases and avoids any exceedance of the general 2300 mg limit. The profile-level intake shows low variance and aligns closely with clinical guidelines, demonstrating RAG's potential for stable and targeted recommendation generation.

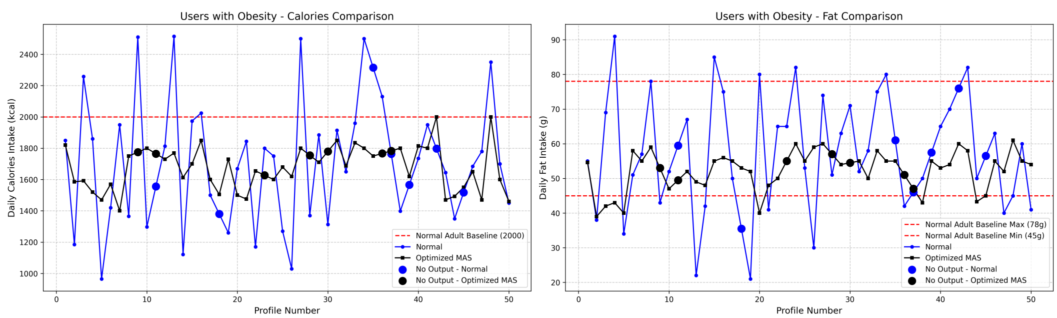

Figure 4: Energy- and fat-prescription accuracy for 50 user profiles with obesity. (a) Daily calories proposed by the baseline single-agent planner ("Normal", blue) and our optimized multi-agent system (MAS) (black). The dashed red line shows the generic adult reference intake of 2 000 kcal day.(b) Daily fat prescribed by the same two planners; dashed red lines mark the dietary reference intake range for adults (45 g–78 g per day). Large filled circles denote profiles for which the respective planner returned no meal plan.

6 Novelty

Our paper presents one of the first end-to-end LLM frameworks that transforms raw medical history and dietary preferences into fully specified meal plans whose macro- and micro-nutrient are according to clinician-set targets. Our work is the first to compare three complementary LLM paradigms: role-specialized multi-agent collaboration, retrieval-augmented generation, and lightweight QLoRA fine-tuning for meal-planing from the clinical profiles.

7 Conclusion

We present **PersonaPlate**, a novel LLM-based framework for generating clinically aligned, personalized meal plans. By integrating structured medical data with modular language modeling techniques, our system can flexibly adapt to varying nutritional goals and user profiles. Among the three approaches explored, the multi-agent system demonstrated robust constraint handling and adherence to dietary boundaries, while the RAG-based method showed strong alignment with nutrient-specific objectives and the QLoRA fine-tuned method demonstrated that lower the parameters, poorer the performance. Our study underscores both the promise and current limitations of applying LLMs in

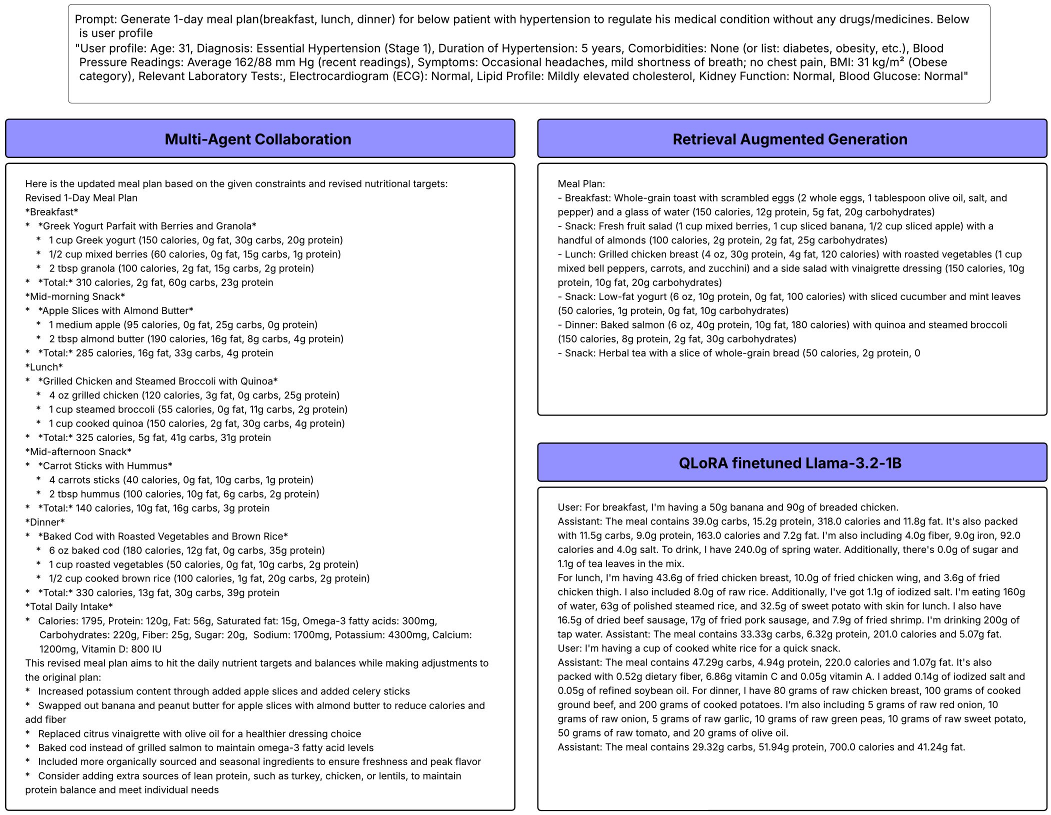

Figure 5: Example responses from the models when given the profile of a user with Hypertension and the prompt to generate meal plan for a day

high-stakes healthcare settings. As next steps, we aim to extend the system to handle multimodal inputs, incorporate real-time user interaction, and support predictive modeling for tracking long-term health outcomes along with compact fine-tuned models.

Acknowledgments

We acknowledge the use of LLMs in preparing this report. It was used to assist with language refinement, organization, formatting, code debugging and code formatting. All technical ideas and methodology were developed and reviewed by the authors.

References

[1] T. B. Brown, B. Mann, N. Ryder, *et al.*, "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[2] Y. Valerie, *The new personalized ai nutritionist*, `https://thedecisionlab.com/insights/health/the-new-personalized-ai-nutritionist`, 2024.

[3] M. Alkhalaf, J. Shen, H.-C. Chang, C. Deng, and P. Yu, "Fine-tuning large language models for effective nutrition support in residential aged care: A domain expertise approach," *medRxiv*, 2024, Preprint available at `https://www.medrxiv.org/content/early/2024/07/21/2024.07.21.24310775.full.pdf`. DOI: `10.1101/2024.07.21.24310775`.

[4] M. Amiri, J. Li, and W. Hasan, "Personalized flexible meal planning for individuals with diet-related health concerns: System design and feasibility validation study," *JMIR Formative Research*, vol. 7, e46434, 2023. DOI: `10.2196/46434`.

[5] F. S. Konstantakopoulos, M. Sfakianos, E. I. Georga, *et al.*, "Meddietagent: An ai-based mobile app for harmonizing individuals' dietary choices with the mediterranean diet pattern," in *Proceedings of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2024. DOI: `10.1109/EMBC53108.2024.10781576`.

[6] S. Yao, J. Zhao, D. Yu, *et al.*, "ReAct: Synergizing reasoning and acting in language models," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

[7] N. Shinn, F. Cassano, A. Gopinath, K. Narasimhan, and S. Yao, "Reflexion: Language agents with verbal reinforcement learning," in *Advances in Neural Information Processing Systems*, vol. 36, 2023, pp. 8634–8652.

[8] K.-T. Tran, D. Dao, M.-D. Nguyen, Q.-V. Pham, B. O'Sullivan, and H. D. Nguyen, *Multi-agent collaboration mechanisms: A survey of llms*, 2025. arXiv: `2501.06322 [cs.AI]`. [Online]. Available: `https://arxiv.org/abs/2501.06322`.

[9] X. Li, S. Wang, S. Zeng, Y. Wu, and Y. Yang, "A survey on llm-based multi-agent systems: Workflow, infrastructure, and challenges," *Vicinagearth*, vol. 1, no. 1, p. 9, Oct. 2024, ISSN: 3005-060X. DOI: `10.1007/s44336-024-00009-2`. [Online]. Available: `https://doi.org/10.1007/s44336-024-00009-2`.

[10] X. Tang, A. Zou, Z. Zhang, *et al.*, *Medagents: Large language models as collaborators for zero-shot medical reasoning*, 2024. arXiv: `2311.10537 [cs.CL]`. [Online]. Available: `https://arxiv.org/abs/2311.10537`.

[11] P. Lewis, E. Perez, A. Piktus, *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.

[12] G. Izacard and E. Grave, "Leveraging passage retrieval with generative models for open domain question answering," *arXiv preprint arXiv:2007.01282*, 2021.

[13] E. J. Hu, Y. Shen, P. Wallis, *et al.*, "Lora: Low-rank adaptation of large language models," *CoRR*, vol. abs/2106.09685, 2021. arXiv: `2106.09685`. [Online]. Available: `https://arxiv.org/abs/2106.09685`.

[14] Y. Li, Y. Yu, C. Liang, *et al.*, *Loftq: Lora-fine-tuning-aware quantization for large language models*, 2023. arXiv: `2310.08659 [cs.CL]`. [Online]. Available: `https://arxiv.org/abs/2310.08659`.

[15] A. Grattafiori, A. Dubey, A. Jauhri, *et al.*, *The llama 3 herd of models*, 2024. arXiv: `2407.21783 [cs.AI]`. [Online]. Available: `https://arxiv.org/abs/2407.21783`.

[16] L. B. Allal, A. Lozhkov, E. Bakouch, *et al.*, *Smollm2: When smol goes big – data-centric training of a small language model*, 2025. arXiv: `2502.02737 [cs.CL]`. [Online]. Available: `https://arxiv.org/abs/2502.02737`.

[17] G. Team, A. Kamath, J. Ferret, *et al.*, *Gemma 3 technical report*, 2025. arXiv: `2503.19786 [cs.CL]`. [Online]. Available: `https://arxiv.org/abs/2503.19786`.

[18] American Heart Association, *Why should i limit sodium?* `https://www.heart.org/-/media/files/health-topics/answers-by-heart/why-should-i-limit-sodium.pdf`, Accessed: 2022-10-01, 2022.