©Copyright 2026
Liam Bakar

Enabling Personal Nutrition Tracking with AI-Supported Food Logging

Liam Bakar

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science in Electrical and Computer Engineering

University of Washington
2026

Committee:
Shwetak Patel
Vikram Iyer

Program Authorized to Offer Degree:
Electrical and Computer Engineering

University of Washington

Abstract

Enabling Personal Nutrition Tracking with AI-Supported Food Logging

Liam Bakar

Chair of the Supervisory Committee:
Shwetak Patel
Department of Electrical and Computer Engineering

Food logging, both self-directed and prescribed, plays a critical role in uncovering correlations between diet, medical, fitness, and health outcomes. Through conversations with nutritional experts and individuals who practice dietary tracking, we find current logging methods, such as handwritten and app-based journaling, are inflexible and result in low adherence and potentially inaccurate nutritional summaries. These findings, corroborated by prior literature, emphasize the urgent need for improved food logging methods. In response, we propose SnappyMeal, an AI-powered dietary tracking system that leverages multimodal inputs to enable users to more flexibly log their food intake. SnappyMeal introduces goal-dependent follow-up questions to intelligently seek missing context from the user and information retrieval from user grocery receipts and nutritional databases to improve accuracy. We evaluate SnappyMeal through publicly available nutrition benchmarks and a multi-user, 3-week, in-the-wild deployment capturing over 500 logged food instances. Users strongly praised the multiple available input methods and reported a strong perceived accuracy. These insights suggest that multimodal AI systems can be leveraged to significantly improve dietary tracking flexibility and context-awareness, laying the groundwork for a new class of intelligent self-tracking applications.

Acknowledgements

I would like to express my gratitude to my advisor, Prof. Shwetak Patel. I am truly grateful for the opportunity to have learned under his guidance. I am equally indebted to Prof. Vikram Iyer, whose expertise and patience were instrumental in shaping this work.

I owe a special thanks to PhD candidates Zachary Englhardt, Vidya Srinivas, and Girish Narayanswamy. Our late afternoons spent brainstorming and writing made the challenging moments of this journey far more manageable. I am also grateful for the emotional support and perspective they provided. On a similar note, I am grateful to Dr. Richard Li, who always grounded me with sage advice and candid humor.

Beyond academics, I am grateful to the UW men's club volleyball team. Stepping onto the court provided a much-needed mental reset and a sense of community. I would also like to thank Haein Joy Kim. Her persistent encouragement to leave the confines of my room and seek out the city's third spaces was vital to my well-being. Without her, I certainly would not have secured the necessary vitamin D to survive the Seattle winters. To my best friend Henry in New York, thank you for the countless calls. Despite the three-hour time difference and the miles between us, your perspectives keep me sane. Finally, I want to thank my parents. Their belief in my potential has been my primary source of motivation, keeping my spirits high through every setback and success.

Contents

List of Figures

List of Tables

1 Introduction

It is well studied that proper nutrition and eating habits are correlated with healthier living and reduced risk for a number of conditions [25]. However, what constitutes a healthy diet and dietary patterns varies greatly from individual to individual, and is influenced by factors such as cultural and demographic background, geographic location, pre-existing conditions, and food intolerance/sensitivities. The definition of healthy diet is further convoluted by factors such as self-image, societal pressures, socioeconomic status, and personal goals. As such there is not a "one-size-fits-all" solution to individual eating habits. To this end, food logging, both self-directed and prescribed, plays a critical role in uncovering correlations between diet, medical, fitness, and health outcomes for individuals [15, 49].

Nutritional tracking tools have emerged as a way to bring more awareness and structure to eating habits. These tools enable the 69% of U.S. adults keeping track of at least one health indicator (weight, diet, exercise regimen, or symptom) [20] to better monitor intake, set goals, and reflect on patterns using smartphone and wearable-based logging. However, unlike fitness trackers (e.g., smartwatches) that can automatically measure health metrics (e.g., step counts, heart rate), modern nutritional tracking tools, both handwritten and app-based, largely depend on manual inputs.

To better understand the current landscape of nutrition tracking apps, we conduct a formative study with nine participants spanning both professional dietitians and food journalers (those who log diet). Our results show that usability and convenience, psychological impact, long-term sustainability, and the risk of tracking fatigue (a common phenomenon in which users disengage due to the effort and mental load of constant self-monitoring [12]) all contribute to low adherence and tracking accuracy [46, 21]. Notably, users seek intuitive and flexible systems that align better with their personalized goals and lifestyle. Unlike physical activity tracking, which has benefited from the transition to passive, ubiquitous sensing (e.g., accelerometry in wearables), nutritional tracking remains one of the most significant tasks in the quantified-self movement. This friction creates a cognitive tax where the user must not only remember to log but also decompose complex, multi-ingredient meals into searchable database entries, a process that often leads to data abandonment or estimation bias.

Designing a single tracking solution that will seamlessly fit into diverse users' lifestyles is challenging, as each individual has varying dietary needs and nutrition tracking objectives. A person aiming for muscle gain has different nutritional needs (e.g., higher protein intake) than someone trying to manage cholesterol or lose weight [47]. Dietary guidelines provided by organizations such as the FDA are not specific or tailored enough to provide nutritional guidance for each individual [1, 37]. While many apps exist for self-tracking (MyFitnessPal, LoseIt, etc), our formative study finds they fall short when estimating portion sizes and matching food items [9, 21]. Many tools do not log important context about food preparation or seasonings, and struggle on the challenging problem of estimating specific nutrients. The fundamental challenge in automated nutrition tracking is the gap between a visual representation of food and its biochemical composition. A single photograph of a meal is an under-determined signal; it cannot capture hidden caloric density from cooking oils, the specific sodium content of seasonings, or the exact volume of ingredients obscured beneath the surface.

Recent advances in artificial intelligence (AI) open up new opportunities to reimagine how nutrition tracking systems can be designed and experienced. In particular, multimodal foundation models can jointly process text, images, and audio, allowing users to describe meals in natural language, upload photos of food or receipts, or record brief voice notes. These capabilities enable more flexible and personalized methods of logging that may better fit diverse lifestyles, contexts, and abilities. At first glance, such technology appears to offer a straightforward solution to the long-standing challenges of nutritional tracking. However, deploying AI-powered systems reveals that many core issues remain unsolved. Even with multimodal inputs, models often lack the contextual awareness needed to make accurate inferences: photos may be taken at poor angles, ingredients may be occluded, or voice notes may omit crucial details. The challenge, then, is how to fill in this missing context without increasing user burden or disrupting the tracking experience. While previous iterations of AI food logging relied on closed-set image classification models, the emergence of Large Multimodal Models (LMMs) allows for a more nuanced interpretation of dietary data. However, these models are prone to hallucination regarding nutritional values. By integrating Retrieval-Augmented Generation (RAG), we can anchor these models in verified nutritional databases and the user's own historical data, transforming the AI from a simple classifier into a personalized nutritional auditor.

To explore this opportunity and its associated challenges, we develop *SnappyMeal*, an AI-powered nutrition tracking application that integrates multimodal inputs, retrieval-augmented context, and adaptive interactivity. *SnappyMeal* surfaces the tensions at the intersection of AI, HCI, and system design, balancing automation with user flexibility and context-seeking, and highlights that simply "adding AI to the loop" does not automatically solve the difficult problems of nutrition tracking. Instead, meaningful progress depends on understanding what users value, how they wish to engage with AI assistance, and how systems can adapt to individual goals and contexts without imposing additional cognitive or interactional load. We introduce three complementary strategies for context augmentation in AI-powered nutrition tracking: (1) leveraging RAG to retrieve not only structured nutritional data but also visually and semantically similar food images, enhancing contextual understanding, (2) incorporating receipts to provide additional context about what the user actually bought, reducing reliance on potentially incomplete images or descriptions, and (3) selectively generating follow-up questions that directly elicit missing information from users in a goal-directed manner. Together, these strategies aim to improve flexibility and personalization of food logging while mitigating user effort and cognitive load.

We summarize our contributions below:

- We conduct a formative study identifying both the key gaps in current leading food tracking technologies and opportunities for AI systems to close these gaps.
- We develop an end-to-end mobile and cloud system supporting multimodal inputs—images, natural language text, and speech—to capture meal context, and **implement a Retrieval-Augmented Generation (RAG) pipeline** that interactively queries users to fill in missing details.
- We investigate these AI-powered features through a real-world, three-week field pilot study yielding over 500 food logs revealing both progress and key challenges in using AI to create flexible, context-aware nutrition tracking systems.

2 Related Work

2.1 The Quantified Self

The practice of nutritional tracking is a central pillar of the "Quantified Self" movement, a cultural and technological phenomenon focused on "self-knowledge through numbers" [18]. This movement has evolved from early, niche experiments with wearable computers in the 1970s to a pervasive global market where health and fitness apps represent one of the largest categories in digital storefronts [19]. The proliferation of smartwatches and fitness trackers has made self-tracking accessible to a large segment of the population, yet the "little data" generated by individuals often remains siloed from clinical care [19]. As self-tracking has become mainstream, it has faced criticism regarding "data fetishism," where the satisfaction of numerical achievement can lead to a reductionist view of health. Critics argue that predetermined ideals of health—often embedded in the algorithms of these apps—can distance users from their own bodily sensations and intuition, offering a normative framework that may not account for the complexity of individual lived experiences. Despite these concerns, the potential of self-tracking to provide a "mirror" for personal insight continues to drive innovation in the field, with researchers exploring how these datasets can be used for more nuanced, longitudinal health management.

2.2 Nutrition Tracking for Health

Dietary tracking can help individuals with chronic conditions better understand their health, treatments, and how these correlate with dietary intake. For example, systems like DIETOS [2], a recommendation system developed to deliver nutritional information to improve the quality of life of healthy subjects and patients with diet-related chronic diseases, promotes dietary mindfulness and through this demonstrate how personalized solutions may enhance quality of life (QoL). Similarly, dietary tracking may better inform the treatment of such chronic conditions. Several studies have shown the utility of dietary tracking in enabling diabetics to better dose insulin based on their logged carbohydrate intake [16, 48]. In the same vein, another study, Misra and James [33] find that participants with type two diabetes who consistently tracked their diet improved dietary self-efficacy and intake over 6 months. Such tracking solutions have also shown utility for those in periods of convalescence, such as cancer survivors. Specifically, Wang et al. [51] find that dietary tracking drives significant positive behavioral change and improved QoL for 11 out of 18 individuals. Finally, SenthilKumar et al. [40] attest that dietary tracking is more than a tool for observation. Their study demonstrates that personalized dietary counseling and reminders result in higher dietary adherence, increased QoL, and improved metabolic health.

2.3 Technology in Nutrition Tracking

Emergent technologies are rapidly transforming the landscape of nutrition tracking, allowing users to move beyond the traditional self-report methods to incorporate automated, continuous, and objective data streams. These tracking approaches can be largely categorized as those which derive additional insights from either body-worn sensors (e.g., wearables) or from ubiquitous mobile devices.

2.3.1 Wearables

The first wave of ADM focused on physical indicators of consumption. Wearable microphones have been used to classify food types based on chewing sounds, while neck-worn sensors can detect the motion of swallowing [14]. More recent advancements have utilized wrist-worn devices to identify hand-to-mouth gestures as a proxy for eating occasions [14]. However, these methods primarily track when someone eats, rather than the specific nutritional content of what they eat.

Several studies have explored the use of wearable devices for ADM. Amft et al. [4] utilized wearable microphones to classify food type and quantity based on chewing sounds. Similarly, other researchers have developed specialized wearable sensors for a similar purpose, such as the neckband created by Cheng et al. [10] to detect swallowing motions. More recently, Reščič et al. [38] demonstrated the potential of gesture recognition from wrist-worn devices to quantify food intake. Furthermore, Mirtchouk et al. [32] developed a multi-sensor approach, combining in-ear audio sensors with head and wrist motion detectors to classify consumption. More recently, the commoditization of continuous glucose monitors (CGM) has enabled the tracking of the metabolic responses directly from the bloodstream [3, 44]. While promising, these methods largely rely on specialized hardware. In contrast, our work focuses on leveraging ubiquitous mobile phones and their sensing capabilities to provide a more convenient and unobtrusive method for nutritional tracking.

2.3.2 Mobile Device Images

Image-based tracking provides an unobtrusive method of logging that can be facilitated through modern mobile devices. It utilizes the high-resolution cameras already present in smartphones. Image-based tracking provides a convenient, "point-and-shoot" method that can be facilitated through both manual uploads and automated analysis. The Remote Food Photography Method (RFPM) has been validated as an effective tool, particularly in pediatric research, where it minimizes the recall errors inherent in self-report methods [5]. Towards this, Wang et al. [50] proposed a model capable of estimating nutritional content from images of foods. However, images alone are often not enough to gain a comprehensive understanding of nutrition. Biel et al. [6] validate that combining mobile phone image capture with contextual metadata provides a convenient and non-intrusive method for nutritional tracking.

The Foodprint study [13] demonstrates that photo-based dietary diaries act as crucial "boundary negotiating artifacts," structuring data exchange to enable health experts to more quickly focus on the patient's context and specific goals. Recent research has focused on the use of Vision-Language Models (VLMs) to estimate nutritional content directly from meal images. These models aim to automate the detection, segmentation, and classification of food items [39]. However, fine-grained recognition—such as distinguishing between different cooking styles (e.g., grilled vs. fried) or identifying subtle differences in ingredients—remains a significant challenge [39].

A study conducted by Shahabi et al. [41] demonstrates the power of combining passive visual data with in-the-moment psychological and contextual data. This combination of data can be used to predict overeating episodes and identify distinct psychological phenotypes, underscoring the importance of collecting contextual data alongside primary food logs.

2.4 Contextual AI

Context provides the essential information needed to resolve ambiguity and validate the significance of a measurement. A primary limitation of image-only AI systems is "context vulnerability." Models often hallucinate portion sizes or misidentify region-specific dishes when crucial cues such as meal time, geolocation, or ingredient lists are absent [14]. To address this, researchers are integrating Retrieval-Augmented Generation (RAG) to ground AI responses in external, authoritative knowledge sources [14]. RAG frameworks improve the factual accuracy of Large Multimodal Models (LMMs) by retrieving relevant information before generating a response. This process significantly reduces hallucination rates, which is critical in the medical and nutritional domains where misinformation can lead to adverse health outcomes [31, 52]. The retrieval process typically operates in two modes: (1) Sparse Retrieval, which uses lexical overlap to match keywords between the query and documents and (2) Dense Retrieval, which maps queries and documents into a shared embedding space using neural encoders [35]. Models for dense retrieval often utilize Euclidean distance to find the nearest neighbor in a vector space:

$$\arg\min_i ||q - d_i||_2$$

where q is the query vector and d_i represents the document vectors.

Contextual data, acquired automatically or through prompting in AI systems, transforms food log entries into actionable, accurate records. Grocery receipt scanning provides a detailed "ground truth" for what a user actually purchased, which can be used to resolve ambiguities in meal photos. Follow-up questions can bolster contextual human-led inputs or conversations [53]. Dynamically generated follow-up questions represent a largely unsolved problem in the human-centered computing space [34]. Zhang et al. [54]'s findings demonstrate the feasibility and effectiveness of integrating AI-generated follow-up questions into real-time, semi-structured interviews. Kuric et al. [26] validates the ability of models like GPT-4 to generate follow-up questions in usability testing contexts. In general, information elicitation tasks, supported by conversational agents, greatly benefit from follow-up question generation [22, 30]. These techniques can be extended to other domains such as clinical [27, 17] and nutrition [42, 11] settings.

2.5 Evaluating Logging Apps

Mobile food logging, though proven to have many benefits, remains tedious and difficult, and is a significant focus in recent literature. Some logging methods, such as selecting a meal from a large food database, can present usability challenges due to the vast amount of information crowded onto a small screen. Jung et al. [24] addressed this issue through the design and evaluation of the EaT app, analyzing the timeliness of logging and identifying the causes of search failures, including an analysis of 1,163 user-created entries.

In a similar vein, Griffiths et al. [21] conducted a study to assess the precision of five popular free apps (including MyFitnessPal, Lose It!, and Fitbit) by comparing their nutrient intake estimates against calculations from the research-grade Nutrition Data System for Research (NDSR). From this study, it is clear that the demonstrated accuracy of automated technologies must be balanced with user compliance to ensure utility in real-world settings. The respective benefits and drawbacks of manual food journaling (high detail, high burden) and automated dietary monitoring (ADM) (low burden, lower context/detail) suggest the value of semi-automated journaling systems that combine both approaches.

Lu et al. [29] address this gap by examining how people anticipate and accept these hybrid systems. Their findings establish critical design trade-offs: User satisfaction is contingent on the quality of intervention. Participants showed more positive anticipation for prompts that contained information relevant to their journaling goals, aided recall of specific foods, and did not provide too much logging burden. This validates the need for semi-automated systems to produce high-value, food-specific prompts, even if this task is "more challenging to produce than manual reminders." This work suggests that the true measure of a nutrition-tracking app lies not just in its technical accuracy, but in its ability to balance sensing performance with user-anticipated burdens, reinforcing the need to select tracking approaches based on individual and practitioner journaling needs. We extend and complement this body of work studying food logging apps by specifically investigating how multimodal and conversational AI features can be incorporated into nutrition tracking apps.

3 Formative Study

We performed a series of semi-structured interviews to identify strengths and shortcomings of current dietary self-tracking techniques from the perspective of both dietitians and individuals who participate in self-tracking (food journalers). Building off the insights of the literature discussed in Sec. 2, we compiled a list of interview questions to better understand how these challenges manifest and identify potential opportunities for improvement. We focused on the nutritional tracking process itself, including frequency and consistency of tracking, methods individuals use or prescribe, how individuals interpret and use nutrition data, and the specific goals motivating dietary self-tracking. After conducting a pilot interview, we determined a semi-structured interview format would be most suitable to elicit detailed responses.

One crucial goal of our formative study was to identify areas in which perspectives on desired improvements in tracking aids vary between dietitians and food journalers, especially relating to data accuracy and behavior change. By synthesizing these unique viewpoints, we developed a nuanced understanding of the current limitations and future opportunities for improving nutrition tracking applications and synthesized insights to guide the subsequent design of SnappyMeal.

3.1 Methods

3.1.1 Participants

We recruited 4 professional dietitians (Table 2) and 5 food journalers (Table 1) to interview on Zoom video conferencing. Dietitians were recruited through emails to School of Public Health of a major research institution, as well as other smaller departments from the research team's professional connections. Members of the general public that participate in self-tracking were recruited through word of mouth or digital and physical flyers. To aid in recruitment, participants were offered an electronic gift card, with a value of \$50 USD for dietitians and \$20 USD for members of the general public. Prior to recruiting participants, our study protocols for each population group were submitted to and approved by the IRB at the host institution for this study.

3.1.2 Interview

We conducted a semi-structured interview using our prior knowledge of nutrition tracking as a baseline for questions, asking tailored follow-up questions and allowing participants to dive deep

Participant ID	Age Range (Years)	Sex	Tracking Frequency	Tracking History	Occupation
L1	18-24	M	Several times a week	6 months to 1 year	Student
L2	18-24	M	More than once a day	1+ years	Student
L3	18-24	M	More than once a day	1-6 months	Sports Operations
L4	18-24	F	More than once a day	1-6 months	Engineer
L5	18-24	F	More than once a day	6 months to 1 year	Student

Table 1: Characteristics of Formative Study Participants (Food Journalers)

Participant ID	Years of Experience	Clinic Size (# of Patients)	Focus
D1	6	10-20	Weight loss & metabolic health
D2	6	>100	Geriatrics
D3	33	15-20	Eating Disorders
D4	21	>1000	Weight management

Table 2: Characteristics of Formative Study Participants (Dietitians)

into their experience assisting food journalers or their experience journaling themselves. These interviews lasted approximately one hour and were recorded and transcribed asynchronously.

3.1.3 Analysis

Audio recordings of all interviews were transcribed using Zoom teleconferencing software. These transcripts were then subjected to an open-coding analysis. The raw audio recordings were retained as a fallback to resolve occasional transcription errors and to ensure the fidelity of the data. Subsequently, the research team performed a thematic analysis [7] to identify key insights. This process yielded 9 initial codes, which were grouped into the four thematic categories described in Section 3.3. Our analysis focused on two distinct research topics: understanding dietitians' perspectives on how self-tracking can be improved to facilitate the promotion of healthier eating habits, and identifying self-tracking individuals' perspectives on how logging applications can be enhanced to better support self-tracking.

3.2 Findings

3.2.1 Dietitians

Interviews with licensed dietitians revealed significant challenges with current dietary tracking, both manual and digital. Dietitians reported that patients, particularly older adults with limited tech skills, struggle with consistency and accuracy. As D2 stated, *"Think about the least amount of work patients need to do."* Manual logs are often incomplete, missing details like portion sizes, seasonings (especially sodium), drinks, meal timing, and eating speed. Digital apps, such as MyFitnessPal and LoseIt, frequently contain inaccurate food labels, struggle to differentiate food types (e.g., steak vs. roast), and make tracking specific nutrients like fiber, sodium, calcium, and potassium difficult for many.

Photos, while sometimes helpful, still present difficulties in accurately estimating quantities and lack context regarding preparation methods or processed versus fresh status. D2 noted, *"The challenge with photos is I can't tell how big the plate is in the image and I can't tell that there are seasonings."* Patient engagement varies, often dropping off without clear feedback or visualization of how tracking impacts their condition or relates to personal goals (e.g., energy to play with grandkids, bone health). D4 also highlighted this challenge, saying, *"Patients don't track enough and people only remember 30% of the stuff that they track."*

Based on these challenges, dietitians expressed a clear need for improved tools. They desire simplified, low-effort tracking methods, perhaps using pre-printed templates or involving family members. Enhanced granularity is needed, capturing not just food but also seasonings, portion sizes, timing, eating speed, and associated GI symptoms or stress levels, particularly for conditions like IBS. Better data integration and visualization are crucial, moving beyond "eyeballing" trends to clear graphs showing changes in weight, calories, and nutrient ratios over time. D2 explained, *"I eyeball the trends. It's helpful to see these numbers in a graphical form. Calorie differences and trends week by week—see how much they increased."*

Dietitians emphasized the need for patient-centered, adaptive consultations that track motivation drivers and use self-tracking data for gradual changes. They also highlighted the importance of a mindful approach, such as using a *"hunger and fullness scale (1-10 scale) tracking on a per-meal basis"* [D1] and looking at emotional hunger. D3 stated, *"It would be nice to see a graph of energy as they were eating... We can compare it to normal eating habit graphs."* This shows a desire for tools that provide insights beyond basic caloric intake.

Furthermore, dietitians highlighted the importance of understanding the connection between food intake and the patient's subsequent emotional and physical state. There is a need for tools that allow patients to easily log not just what they ate, but also how they felt afterward—both emotionally (e.g., stressed, satisfied, guilty) and physically (e.g., energy levels, specific GI symptoms like bloating or pain). D1 noted that they look at the *"emotional hunger and symptoms related to medication."* Capturing this information alongside dietary data could provide valuable insights into food triggers, sensitivities, and the complex relationship between diet, mood, and physical well-being.

3.2.2 Food Journalers

The interviews revealed that nutritional tracking is a complex and highly personal process for individuals driven by goals related to health and wellness. A key shared theme was the motivation to gain control and awareness over one's diet, often spurred by a desire to optimize physical performance or simply feel better. As L1 stated, *"Once you start cooking for yourself, you have more control over how healthy you eat."*

The data highlights a reliance on technology, with participants frequently using apps like MyFitnessPal and Lose It. However, this reliance is met with significant challenges, primarily related to inaccurate data and the tedious nature of manual logging. As L2 noted, *"I don't think it's very accurate because the apps miscalculate on protein and calorie intake."* Participants struggle with the time-consuming process of inputting information, especially for home-cooked meals or when dining out, leading to tracking fatigue. L3 explained, *"[I get deterred] When I'm hungry, it takes time and energy and thought. It's sometimes tedious to find the exact product that I am eating."* Busy days and snacks are especially difficult to track accurately, and L1 pointed out that *"Snacks are difficult to remember."*

A clear desire for more personalized and effortless tracking was evident, with suggestions for features like photo-based food analysis and integration with other devices. L3 wished to *"upload it directly from my scale or if I could just take a pic and it could tell me."* While some apps offer photo features, journalers find them inaccurate, with L2 calling MyFitnessPal's feature a *"scam kind of - it didn't really know the accurate measurement."* The challenge of estimating portion sizes was also a common theme. L1 expressed frustration with the difficulty of determining food weights, noting, *"You can't bring a scale with you everywhere."*

Furthermore, the findings show a strong preference for visual feedback and actionable insights over raw numbers. L1 stated, *"Numbers are good, but visualizations are easier to understand."* Participants also desire personalized progress indicators, such as *"Progress pictures or indicators on the app that I've reached my goal (protein goal, fiber goal, calorie goal, etc.)"* [L4].

The ultimate goal for many was not just to log data, but to feel a sense of mental ease and accomplishment without the obsession that can sometimes accompany meticulous tracking. L2 explained, *"If I think I'm feeling obsessive and compulsive... I'll take a break. It's bad to focus too much on the numbers."* L5 echoed this sentiment, stating, *"Nutritional tracking can be very harmful, I would want to take away the obsessive manner of tracking."*

3.3 Implications for Design of Self-Tracking Tools

The findings from our interviews with both dietitians and food journalers reveal a clear set of design implications for future dietary self-tracking tools. The current landscape of tools, both manual and digital, fails to meet the core needs of accuracy, ease of use, and a holistic approach to health. They lack the flexibility to adapt to a user's changing needs and the context awareness to understand the "why" behind the "what," resulting in a rigid, high-friction experience.

3.3.1 Effortless and Accurate Data Capture

The most significant barrier to consistent tracking is the high cognitive and physical effort required. Both dietitians and journalers highlighted this, with D2's advice to *"Think about the least amount of work patients need to do"* and L3's complaint that when hungry, tracking *"takes time and energy and thought."* Future tools must tackle this by leveraging AI to provide a flexible logging process. Language models for multimodal interaction allow users to log meals conversationally through text or audio, offering input flexibility and eliminating the tedious and often inaccurate process of manual entry. As L2 noted, current photo features are a *"scam,"* and L1 complained about the difficulty of weighing food. By using a pre-trained model to semantically search a robust database of food images and their nutritional data, models can provide a more reliable and accurate way for users to *"just take a pic and it could tell me"* [L3], moving beyond simple visuals to provide a precise nutritional breakdown without the need for a scale.

3.3.2 Actionable Visualizations and Personalized Feedback

Simply logging data is not enough; users seek meaningful insights that connect their habits to their goals. L1's desire to *"check if I ate too much of a certain food"* and L4's preference for *"Progress pictures or indicators on the app that I've reached my goal"* highlight a strong need for data visualization. Dietitians echoed this, with D2 stating, *"I eyeball the trends. It's helpful to see these numbers in a graphical form."* By storing structured data on calories, protein, fat, and other metrics, systems can generate clear, intuitive graphs that show trends over time. Systems can move from simple loggers to providing the type of visual feedback that adapts to the motivational context of a user and helps them feel a sense of mental ease and accomplishment.

3.3.3 Integration of Holistic, Contextual Data

The interviews consistently revealed that food intake is just one part of the health equation. Dietitians emphasized the importance of tracking a wider range of contextual data, from emotional states to physical symptoms. D1's focus on a *"hunger and fullness scale"* and "emotional hunger" demonstrates the value of capturing this qualitative data. A reliance on conversational AI interfaces can capture this holistic information. Systems can be designed to ask about a user's emotional state or energy levels in the natural flow of conversation. This comprehensive approach, combined with the ability to store a user's goals, allows for highly personalized prompts that directly address a user's unique health motivations.

3.3.4 Focus on a Positive and Non-Obsessive Approach

A critical finding, particularly from the journalers, is the risk of tracking leading to an *"obsessive and compulsive"* mindset [L2]. L5's warning that *"Nutritional tracking can be very harmful"* is a powerful design constraint. By using conversational, non-numeric approaches, systems can reduce the focus on meticulous, number-driven logging, which can be a source of anxiety. Additionally, adding personalization to the prompts of these conversations allows systems to tailor their feedback to focus on a user's broader goals, such as feeling *"in control"* [L1] or simply building a healthy lifestyle, rather than being a *"slave to the app"* [L2]. By leveraging user goals and sentimental tracking [L5], the system can promote a healthy relationship with food, shifting the focus from perfect data to sustainable, positive habits.

4 System Design

The SnappyMeal system employs a microservices-oriented architecture that separates the frontend, backend, and specialized AI services into modular components. This structure allows for independent development, deployment, and evaluation of major system components. An overview of key system modules and how they interact can be found in Figure 1.

4.1 Mobile App

We design our app with 5 screens: a pantry screen where users can view their uploaded receipt items, a log (2c) screen where users can view their generated food logs, a dashboard screen (2b) where users can see general progress charts as well as efficiently add new food logs or receipt uploads, a trend screen (2a) where users can see trends and visualizations about their nutritional data, and a profile screen (2d) where users can view and edit their personal information and goals. When users first create an account, they input their numeric and personal nutrition goals. This information forms the initial context which is used to tailor prompts, ask relevant follow-up questions, and provide support to help a user achieve these goals. When users log food, they are offered input flexibility, allowing them to upload an image, text description, or audio description of their meal. Subsequently, an LLM (Gemini), examines the uploaded media to determine whether enough information is present. If not, the model generates a follow-up question to gain a better understanding of the media. Finally,

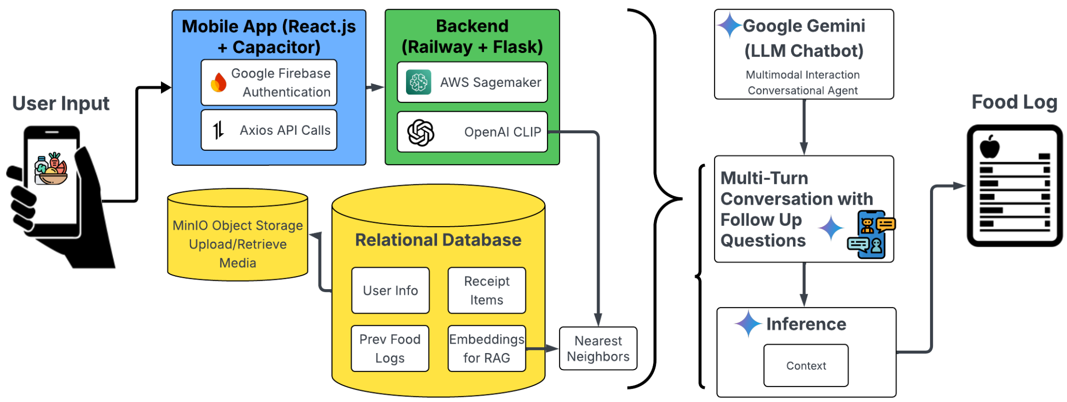


Figure 1: System overview: Users input multimodal food logs which are processed along with relevant context to extract nutritional information. This data is fed into an LLM to validate the nutritional information and determine if there is any missing information. Finally, all the information is sent to Gemini to generate food log data. The resulting data is displayed as individual food logs that can be examined and aggregate graphical visualizations.

the original media, user's nutritional goals, receipt context, and clarifying conversation history are sent to Gemini to generate a comprehensive nutrition log.

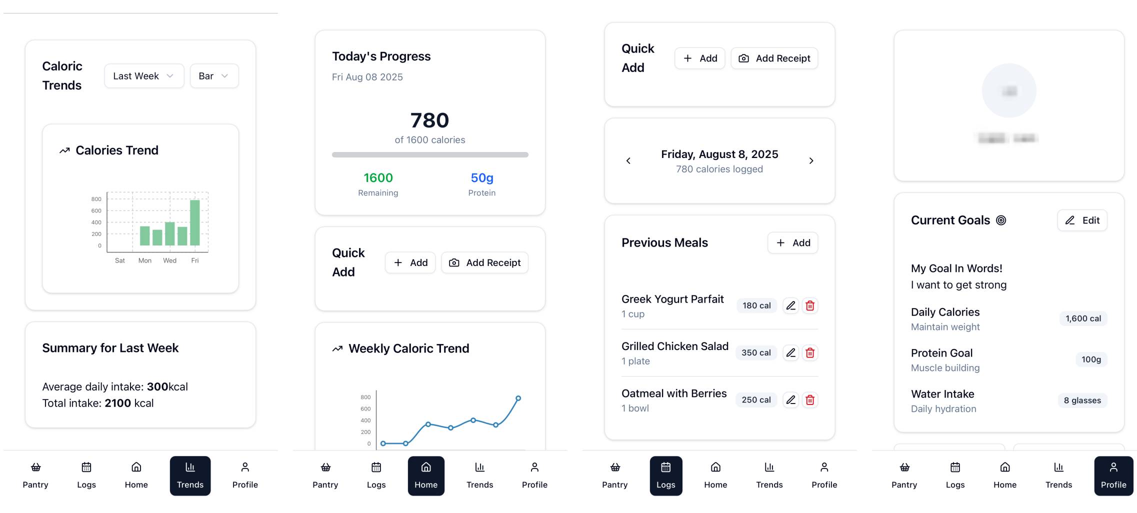


Figure 2: Overview of the application's main interface screens.

4.2 Data Storage

A relational database is used to store structured user data, meal logs, and conversation history. This choice provides strong data integrity and transactional consistency. The provided schema (see Appendix A) outlines the relationships between key entities like *users*, *food_logs*, and *receipt_items*.

- ***users*:** Stores core user information, including demographics (*age*, *height*, *weight*), health goals (*target_calories*, *target_protein*, *target_water*), and a free-form *text_goals* field for personalized prompts. The *user_id* serves as the primary key.
- ***food_logs*:** This is the central table for meal tracking. Each entry records a single meal, including the nutritional breakdown (*calories*, *protein*, *fat*, etc.), the meal's name, and a link to the *chat_history* and any associated *media_url* for multimodal logging.
- ***receipt_items*:** This table stores data extracted from user-uploaded grocery receipts. It helps the system understand the user's general dietary habits and food preferences by linking items purchased at a store to a user's ID.
- ***conversations*:** This table stores the history of user interactions with the Gemini API. It records each conversational turn, including the original prompt and the API's response, allowing for context-aware, follow-up interactions.
- ***food_embeddings*:** This is a crucial table for the RAG functionality. It stores pre-computed vector embeddings for a large corpus of food items. These embeddings, generated by the CLIP model from the Nutrition5k[45] dataset, allow the system to efficiently find the most semantically similar food items based on a user-uploaded image embedding. The table also includes metadata such as *food_label*, *estimated_calories*, and other nutritional data sourced directly from the dataset.
- ***personalized_prompts*:** This table stores the user's personalized prompt derived from their goals during signup. These prompts are subject to change if user goals are edited, providing flexibility and ensuring the system's context awareness is up-to-date.

A separate object storage system, MinIO, is used to store unstructured data such as user-uploaded images and audio files. This offloads large binary data from the primary database, improving database performance and allowing for efficient retrieval and processing by the AI models.

4.3 Multimodal Logging

To achieve effortless, accurate, and contextually-aware data capture—the primary goals identified in our formative study—the SnappyMeal system is designed to accept and process food logs across multiple modalities: image, text, and audio. Our approach fundamentally relies on multimodality, which allows the system to process and relate different data types.

4.3.1 Image inputs

The OpenAI CLIP (Contrastive Language–Image Pre-training)[36] model, hosted on AWS Sagemaker, is used for Retrieval-Augmented Generation (RAG). CLIP's fundamental contribution to downstream applications stems from its ability to map text and images into a shared embedding space where their vector representations are aligned based on semantic content. This allows us to use the same model for text and image modalities. When a user uploads an image, the CLIP model generates an image embedding. This embedding is then used to perform a semantic search using cosine similarity against the *food_ embeddings* table to find the most relevant food items in the database. For this purpose, we used the Nutrition5k[45] dataset, a large-scale, pre-annotated dataset of food images and their corresponding nutritional information, to create the foundational *food_ embeddings* for our database. This dataset serves as the knowledge base for the RAG system, significantly improving the accuracy of food identification. Using a nearest neighbors search, the system identifies the most semantically similar foods in the database and presents its verified nutritional information as context in the aggregate prompt.

4.3.2 Text inputs

Similarly, when a user uploads a text description of their meal, the CLIP model generates an embedding. Due to CLIP's contrastive nature and its training on text-image pairs, we are able to match text embeddings to image embeddings. Again, using a nearest neighbors search, the system identifies foods in the database that are most semantically similar to the user's input and presents their verified nutritional information as context in the aggregate prompt. Unlike with images, there are no food-text embeddings in the vector database, so we rely on the multimodality of CLIP.

4.3.3 Audio inputs

When a user uploads an audio description, the audio file is directly sent to the aggregate Gemini prompt input with the user context since CLIP does not support audio inputs. The prompt includes instructions explaining that this file is user uploaded media. While we could have included speech-to-text preprocessing to utilize CLIP, the additional step would have increased latency, making it challenging to provide users real-time interaction.

4.4 Interactive prompting

The Gemini API is the core of the app's multimodal interaction and personalized prompting. It processes natural language queries (text and audio) and analyzes uploaded images to identify food items and provide conversational feedback. For this task, we use Gemini 2.5-Flash due to its balance of strong multi-modal support. While Gemini 2.5-Pro is more powerful for many tasks, it would have introduced more latency and cost.

4.4.1 Follow-up Questions

By default, every new food log receives a follow-up question with the intention of clarifying any missing information from the raw uploaded media. This process occurs through a Gemini multi-turn conversation stemming from the prompt in Appendix B.3. Contrary to single-turn LLM calls, the multi-turn conversation has a sense of conversation history that allows the model to lead goal-oriented "conversations." After every answer, the model is asked if it has enough information to generate a comprehensive food log; if not, it is asked to generate another follow-up question. This conversation history is then included in the aggregate prompt to provide the model with additional context and improve the accuracy.

4.4.2 Receipt Context

Users are asked to upload their receipts into the system to improve their generated food logs. Receipts are parsed using Gemini to extract name, quantity, and source (where they were bought). Utilizing the publicly available nutrition information of many popular grocery store products, we use Gemini to generate nutritional summaries of each of the purchased food ingredients. These ingredient data are then added as context in the aggregate prompt to improve accuracy. Sometimes the follow-up questions can clarify if the meal being analyzed was cooked using any pantry ingredients.

5 System Evaluation

To assess the performance and reliability of the nutrition tracking software, we carried out a systematic technical evaluation focused on quantifying its accuracy in food recognition and nutritional estimation. This assessment used the publicly available Nutrition5k dataset, a benchmark repository comprising 3,490 food images meticulously paired with ground-truth nutritional information. The following section details the specific experimental setup, the definition of the performance metrics employed, and the results derived from challenging the software against this diverse labeled data corpus.

5.1 Experiment Setup

The technical evaluation was conducted utilizing the Gemini Batch API to facilitate efficient inference across the entire Nutrition5k dataset. A standardized, zero-shot prompt was designed to instruct the model to identify food items and output the corresponding nutritional breakdown in a structured JSON format.

To determine the impact of individual architectural components on overall performance, we performed a controlled ablation study. This involved comparing the baseline model's performance against configurations where key features were selectively introduced:

5.1.1 Baseline Model:

To test the core visual recognition and vanilla LLM inference, we sent in the same prompt (see Appendix B.1) for each food image instance.

5.1.2 Ingredient Addition:

The model was augmented with the capacity to infer and explicitly list constituent ingredients beyond the main food item, such as those present in a receipt. The Nutrition5k dataset provides ground-truth ingredients for each image. To mitigate bias toward the visible ingredients, we deliberately introduced negative sampling. Specifically, we began by establishing a repository containing all unique ingredients present across the dataset, denoted as $\{Ing\}_{\text{all}}$. For an image i containing k true ingredients ($|\{Ing\}_i| = k$), an equivalent set of k negative samples was randomly drawn from the set of all ingredients not present in the image, $\{Ing\}_{\text{all}} - \{Ing\}_i$. The ingredients and their nutritional values were included in the original prompt.

5.1.3 Retrieval-Augmented Generation (RAG):

Since we had used the Nutrition5k dataset as our vector database, we performed RAG by omitting the image being evaluated and finding the nearest images based on OpenAI's clip model. When evaluating image i, we performed cosine similarity nearest neighbor search on images indexed $0, 1, \ldots, i-1, i+1, \ldots, n$. The top 5 closest matches and their nutritional information were added into the prompt as additional context for the model.

5.1.4 Follow-Up Questions:

Due to the large size of the dataset, we could not answer follow-up questions for every image. Instead, we randomly sampled 100 images. Four members of the research team split these 100 samples, and manually answered the generated follow-up questions. This ensured a diversity in the answers of the questions to further generalize our evaluation. Upon completing this, the vanilla prompt was sent with the answered question for evaluation in the following format.

```
Here is a clarifying question and answer that can help you better understand the food:
{question}
{answer}
```

5.2 Performance Metrics

Nutritional performance was quantified using the Mean Absolute Error (MAE) between the estimated calories (kcal), protein (g), fat (g), and carbohydrates (g) generated by our experiments and the ground-truth values from the Nutrition5k dataset. We also evaluate the Root Mean Squared Error (RMSE) which is more sensitive to outliers (see Appendix C.1 for equations).

To assess the statistical significance of our results, we constructed 95% confidence intervals using the percentile bootstrap method. This technique allows us to estimate the uncertainty of the MAE and RMSE given the varying size of our evaluation sets. The procedure involved generating B bootstrap samples (in our case $B = 1000$) by sampling with replacement from the original n data points. For each of these B samples, we re-calculated our metric, which yielded a distribution of B metric estimates. The 95% confidence interval was then derived directly from this distribution by taking the 2.5th and 97.5th percentiles as the lower and upper bounds, respectively. A formal description of this method is provided in Appendix C.2.

5.3 Results

The evaluation results of the full dataset can be found in Table 3 where the point metrics and their confidence intervals are reported. Due to follow-up involving human input, we only evaluated the feature on 100 samples. The results of the ablation study can be found in Table 4 where we evaluated every feature and some combinations on the same 100 food image samples. We highlight some specific examples where follow-up questions help and hurt in Table 5.

5.4 Discussion

In summary, the RAG model performed the best for protein, carbohydrate, and fat estimations in Table 3. However, the confidence intervals mostly overlapped, so we cannot confidently conclude that RAG is the best model. Additionally, we noticed that the error values increased when

Table 3: Evaluation Metrics and Confidence Intervals of Individual Model Performance by Nutritional Value ($n = 3466$).

Nutritional Value	Model	MAE (95% CI)	RMSE (95% CI)
Calories (kcal)	vanilla	120.38 (115.54, 124.92)	**188.58** (171.16, 213.45)
	receipt	121.57 (116.47, 127.20)	192.87 (175.24, 220.50)
	RAG	**120.21** (115.40, 125.23)	188.94 (171.73, 214.79)
Protein (g)	vanilla	7.72 (7.39, 8.10)	12.86 (12.17, 13.50)
	receipt	7.67 (7.32, 8.03)	12.73 (12.11, 13.32)
	RAG	**7.40** (7.05, 7.75)	**12.32** (11.67, 12.97)
Carbohydrates (g)	vanilla	12.33 (11.69, 13.04)	23.75 (18.10, 31.42)
	receipt	12.46 (11.79, 13.20)	24.21 (18.68, 32.18)
	RAG	**12.19** (11.56, 12.97)	**23.48** (18.00, 31.34)
Fat (g)	vanilla	7.98 (7.66, 8.27)	12.36 (11.82, 12.83)
	receipt	8.05 (7.72, 8.38)	12.50 (11.94, 13.07)
	RAG	**7.81** (7.50, 8.13)	**11.90** (11.43, 12.38)

evaluating the same model on the $n = 3466$ dataset and the $n = 100$ dataset. This suggests the 100 images evaluated in the smaller dataset can be some of the more unclear images for the model. In Table 4, the overlap for RAG and receipt is not as pronounced. We noticed it has the best performance of all models for each nutritional category except carbohydrates. This is likely because those two features combined provide the most numeric nutritional information without introducing extra textual noise. Unlike receipts alone, the RAG data in "RAG and receipts" helps standardize the input by providing a grounding estimation.

Contrary to expectations, we observed no conclusive evidence that the follow-up questions directly improved nutrition estimation. This outcome is likely tied to two major factors: the model's small evaluation sample size and the observed behavior of the LLM. The confidence interval analysis suggested that the model may be second-guessing itself; generating a question, receiving an imperfect or ambiguous answer from the user, and then allowing that conflicting information to degrade the final estimation rather than refine it. A clear benefit of the follow-up questions is observed in the first two examples (rows 1 and 2) of Table 5, where the follow-up question successfully disambiguated key nutritional factors that are difficult or impossible to determine from visual data alone. In the first case, the baseline model's estimation for a meat dish was significantly improved after the user clarified the food type as "Beef." The initial "vanilla MAE" was high, suggesting the model may have defaulted to a generic "meat" profile or an incorrect specific type (e.g., chicken). The user's textual input allowed the model to apply a more accurate nutritional profile, resulting in a uniform improvement across all four measured metrics (Calories, Protein, Carbohydrates, and Fat). Similarly, the query "How were the vegetables prepared?" provided critical, non-visual context. The user's response, "Fried eggplants and steamed cauliflower," resolved ambiguity about preparation methods that significantly impact nutritional content. The baseline model cannot visually distinguish "steamed" from "fried," "boiled," or "roasted." The clarification allowed for a major correction, particularly in fat and calorie estimations, and again resulted in a uniform improvement.

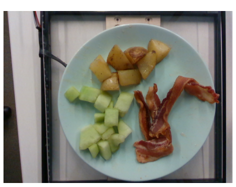
Conversely, the final two examples (rows 3 and 4) of Table 5 illustrate moments where the model's performance degraded despite receiving correct information from the user. The third case presents

Table 4: Ablation Evaluation Metrics and Confidence Intervals of Model Performance by Nutritional Value ($n = 100$).

Nutritional Value	Model	MAE (95% CI)	RMSE (95% CI)
Calories (kcal)	vanilla	148.88 (122.51, 177.29)	208.00 (173.08, 243.00)
	receipt	148.42 (117.40, 180.84)	219.67 (176.78, 260.31)
	RAG	145.45 (114.11, 180.28)	223.14 (173.50, 270.91)
	follow-up	161.82 (123.14, 209.24)	274.48 (197.67, 358.78)
	RAG + follow-up	168.87 (128.05, 218.25)	288.50 (199.23, 377.33)
	receipt + follow-up	144.79 (109.99, 187.02)	242.07 (174.63, 323.50)
	RAG + receipt	**123.96** (97.16, 153.03)	**191.92** (149.13, 232.17)
	RAG + receipt + follow-up	153.00 (111.04, 199.29)	273.19 (181.77, 359.27)
Protein (g)	vanilla	10.06 (7.92, 12.22)	14.72 (11.88, 17.39)
	receipt	9.38 (7.37, 11.84)	14.72 (11.57, 17.96)
	RAG	9.05 (7.06, 11.11)	13.98 (10.77, 16.99)
	follow-up	10.27 (8.03, 12.57)	15.64 (12.01, 18.94)
	RAG + follow-up	10.60 (8.24, 13.23)	16.92 (13.08, 20.59)
	receipt + follow-up	9.02 (7.01, 11.12)	14.10 (10.58, 17.22)
	RAG + receipt	**8.75** (6.87, 10.88)	**13.66** (10.77, 16.57)
	RAG + receipt + follow-up	9.08 (6.82, 11.60)	15.06 (11.16, 18.64)
Carbohydrates (g)	vanilla	15.30 (12.44, 18.76)	22.77 (18.17, 27.82)
	receipt	15.81 (12.50, 19.27)	23.76 (18.29, 29.62)
	RAG	**13.80** (10.43, 17.29)	**22.06** (16.64, 26.81)
	follow-up	18.00 (13.31, 23.49)	31.99 (20.92, 43.00)
	RAG + follow-up	17.21 (12.56, 22.87)	31.75 (20.58, 42.51)
	receipt + follow-up	16.43 (12.49, 21.42)	28.27 (19.53, 37.60)
	RAG + receipt	13.81 (10.36, 17.53)	22.43 (16.63, 27.92)
	RAG + receipt + follow-up	16.56 (11.79, 21.72)	31.00 (19.40, 40.69)
Fat (g)	vanilla	8.05 (6.53, 9.86)	11.73 (9.74, 13.76)
	receipt	8.56 (6.59, 10.53)	12.95 (10.23, 15.50)
	RAG	8.88 (6.89, 10.93)	13.61 (10.62, 16.37)
	follow-up	9.58 (7.35, 12.15)	15.31 (11.93, 18.93)
	RAG + follow-up	9.42 (7.20, 11.91)	15.26 (12.13, 18.38)
	receipt + follow-up	8.23 (6.25, 10.20)	13.25 (9.94, 16.19)
	RAG + receipt	**7.17** (5.50, 8.98)	**11.34** (8.59, 14.10)
	RAG + receipt + follow-up	8.45 (6.24, 10.77)	14.63 (10.84, 18.00)

a mixed result. When the user identified the food as "Egg whites," the model correctly updated its fat estimation, resulting in an improvement. However, the estimations for Calories and Protein both declined. This suggests that while the model correctly associated "egg whites" with near-zero fat, its internal profile or subsequent quantity re-estimation for "egg whites" was less accurate than its baseline assumption (perhaps "whole eggs"). The new information, therefore, improved one metric while introducing significant error in others. The most severe friction is observed in the fourth case. The user's answer, "Three" (strips of bacon), prompted a uniform decline across all metrics, making the final estimate significantly worse than the "vanilla" visual-only guess. Here, the

Table 5: Food Image Samples from the Nutrition5k Dataset where follow-up questions caused benefit or friction. In the Δ MAE column, a ↓ means improvement (MAE decreased), a ↑ means decline (MAE increased), and a ◯ means no change.

Food Image	Follow-up Q&A	Metric	vanilla MAE	follow-up MAE	MAE Impr
	SnappyMeal: *What kind of meat is that?* User: *Beef.*	Calories (kcal) Protein (g) Carbohydrates (g) Fat (g)	240.46 19.46 6.20 16.07	112.46 17.56 2.90 4.27	↓ ↓ ↓ ↓
	SnappyMeal: *How were the vegetables prepared?* User: *Fried eggplants and steamed cauliflower.*	Calories (kcal) Protein (g) Carbohydrates (g) Fat (g)	389.83 6.57 58.88 17.45	234.83 0.57 20.88 14.45	↓ ↓ ↓ ↓
	SnappyMeal: *Are these egg whites or whole eggs?* User: *Egg whites.*	Calories (kcal) Protein (g) Carbohydrates (g) Fat (g)	45.81 24.25 7.96 4.79	138.89 25.25 7.96 0.21	↑ ↑ ◯ ↓
	SnappyMeal: *How many strips of bacon did you eat?* User: *Three.*	Calories (kcal) Protein (g) Carbohydrates (g) Fat (g)	101.06 2.09 18.046 2.89	173.06 5.09 35.046 6.89	↑ ↑ ↑ ↑

model's internal database entry for "three strips of bacon" may be highly erroneous, and applying this flawed data point "poisoned" the entire meal calculation.

A major limitation of this study was the necessary reliance on a small sample size for the follow-up question models due to the requirement for human input. Future research must prioritize gathering a significantly larger dataset of human-labeled follow-up-question responses to definitively determine the true potential of interactive estimation models and achieve tighter confidence intervals for all models. Ultimately, while the follow-up questioning technique showed promise for specific, clear-cut cases (Table 5 rows 1 and 2), the most reliable immediate path to improved nutrition estimation lies in the synergistic combination of visual RAG context and explicit external data, as demonstrated by the robust performance of the RAG and receipt approach.

6 Field Pilot Study

While the preceding technical evaluation established SnappyMeal's strong foundation, validating its computational feasibility, these results are based on isolated performance metrics. Crucially, they do not account for the human factors critical to sustained dietary tracking, such as motivation, adherence, and the potential for technological fatigue over time. Therefore, to holistically assess how the system's primary design principles of input flexibility and deep context-awareness translate into real-world usability and impact, we conducted a 3-week field pilot study. This study transitioned our focus from the technical capabilities of the back-end architecture to the long-term changes in user behavior, adherence rates, and the evolving relationship between the user and the system's personalized conversational interface.

6.1 Methods

6.2 Participants and Recruitment

Recruitment was conducted via word-of-mouth within a university setting and a formal advertisement announcement posted in university departments' official Slack channel and mailing lists. To encourage completion and mitigate attrition, participants were offered a compensation of one $20 USD electronic gift card for each full week of study completion, totaling $60 for the entire 3-week period. A total of 12 eligible participants were initially recruited for this field study. Of the eligible participants, 8 individuals downloaded the app, and the final sample consisted of 6 participants who completed the full 3-week study period. The participant pool primarily comprised individuals in the 18-24 age range, consistent with the recruitment strategy. 2 of the 6 participants reported prior experience logging food.

6.2.1 Inclusion/Exclusion Criteria

Inclusion criteria for participation included smartphone ownership and a commitment to consistently track dietary intake throughout the study. Exclusion criteria were applied to individuals with a self-reported history of disordered eating or diagnosed eating disorders, as the nature of the study could pose a potential health risk.

Table 6: Characteristics of 3-Week Study Participants

Participant ID	Age Range (Years)	Sex	Tracking History	Occupation
P1	18-24	M	6 months to 1 year	Student
P2	18-24	F	1-6 months	Engineer
P3	18-24	M	Less than 1 month	Software Engineer
P4	18-24	F	1-6 months	Student
P5	25-34	F	1-6 months	Student
P6	18-24	M	6 months to 1 year	Student

6.2.2 Ethics

A detailed study protocol was submitted to and approved by the IRB at the host institution for this study prior to recruiting participants.

6.3 Experiment Setup

The study was designed as a 3-week longitudinal study to evaluate the performance and user experience of the novel nutrition tracking app. The experiment was conducted in a real-world, naturalistic setting, with participants using the app in their daily lives.

Participants were onboarded one at a time so we could ensure they correctly downloaded the app and set up an account. This ensured a smooth data collection afterward.

6.4 Study Procedure

The study was not divided into distinct phases but rather operated as a continuous, 3-week active tracking period for each participant.

6.4.1 Onboarding

Participants created an account, with basic demographic information and initial nutrition goals.

6.4.2 Active Tracking

For the 3-week duration, participants were instructed to use the app to log all meals, snacks, and beverages. They were encouraged to utilize the app's multimodal features, including text, image, and audio inputs, as the primary method for logging.

6.4.3 Data Collection

The system automatically extracted and saved all raw user data. This included:

- **Conversational Data:** All interactions with the Gemini API were stored in the conversations table.
- **Food Tracking Data:** The raw media files (images, audio) and their corresponding estimated nutritional logs were stored in the food_logs table, with media files offloaded to the MinIO file bucket.

6.5 Challenges and Data Consistency

A key challenge observed during the study was the low consistency in participants' food tracking habits. This resulted in a dataset with significant variability in the number of logged meals per day and the thoroughness of the logs. This finding will be addressed in the discussion section, as it highlights a common hurdle in longitudinal nutrition tracking studies and provides a realistic context for the app's performance. The study focused on extracting and analyzing the available raw data to understand user behavior and system performance under real-world usage conditions, even with inconsistent input.

6.6 Results

6.6.1 Survey results

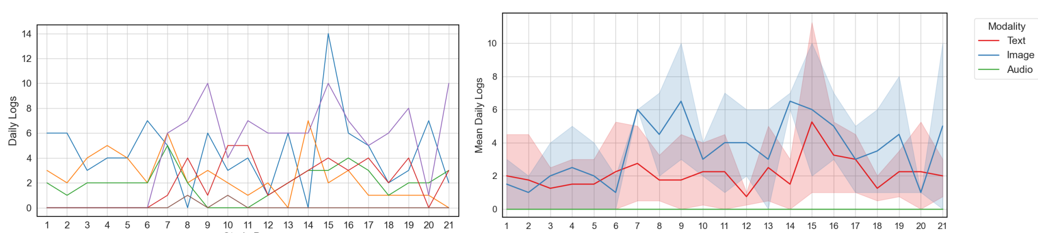
An exit interview form was sent to the 6 participants. From these interviews, we found high engagement for most users, suggesting the process was manageable for many, despite some frustration.

A section of the form was dedicated to rating their agreement with several statements with a scale of 1 to 5 with 1 meaning "Strongly Disagree" and 5 meaning "Strongly Agree." Participants reported that the app made them feel more aware of their eating habits ($\bar{x} = 3.67$). Additionally, participants found the follow-up questions relevant to the food they were uploading ($\bar{x} = 3.83$) and relevant to the goals they were seeking to achieve ($\bar{x} = 3.83$). When it came to the system's usability, participants appreciated the ability to edit their logs ($\bar{x} = 4.33$).

6.6.2 User engagement

Engagement with the application was generally high during the three-week trial, with four of six participants reporting daily logging (7 days per week), one logging 5–6 days per week, and one logging 1–2 days per week. However, the estimated time required for an individual food log showed high variance, ranging from under one minute ($n = 2$) to over five minutes ($n = 1$). Outside of technical failures, the primary self-reported reasons for missed logging were that the process felt too time-consuming or cumbersome ($n = 3$) or forgetting to log ($n = 3$).

All users completed the 21 days of data collection. Figure 3a shows the daily number of log entries per user over the course of the study. With the exception of one user who only entered two logs, users generally logged their food multiple times per day. Participants generally had a high variability in logging frequency, with some users forgetting to log on some days and compensating by logging extra on other days.



(a) Frequency of user logs over time. (b) Mean frequency of modalities logged over time.

Figure 3: User logs over time.

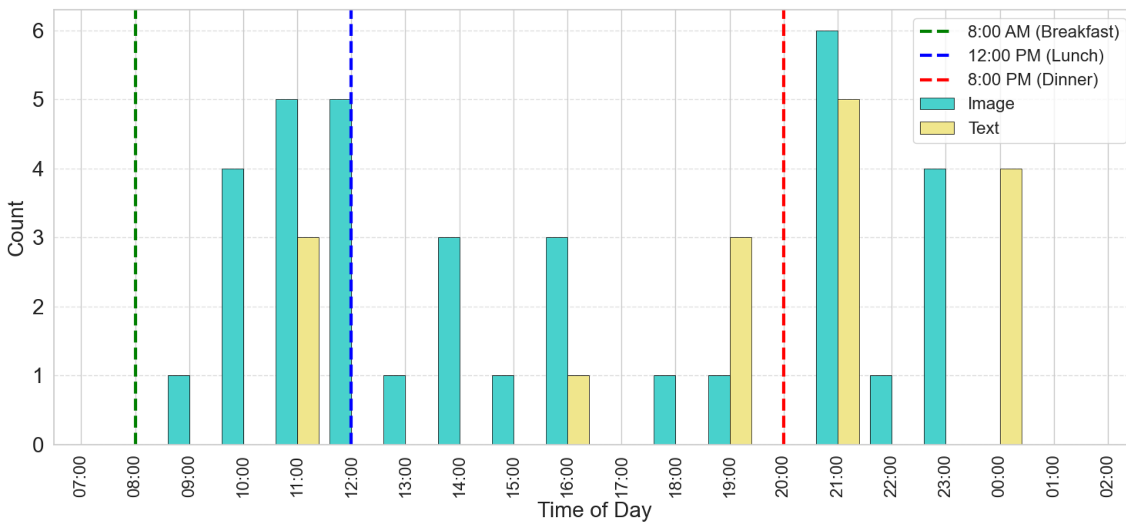


Figure 4: Timeline of meal logging modalities for a representative participant (P1) over a 24-hour period

Fig 3b shows the number of logs by modality over the course of the study. There was a robust mix of modalities, with many users preferring images. Interestingly, we note that no users chose to use audio. This could be because public settings such as restaurants are not conducive to recording audio, or they were eating while logging. We note that our participants represent a younger demographic. Older users less familiar with technology may be more inclined to use audio inputs.

Preference for logging method was split, with three participants preferring Image Logging and three preferring Text Logging for efficiency. However, the image recognition component was a source of friction. One participant rated the accuracy as "Sometimes accurate (25-50% of the time)." The results highlight a significant disconnect between the perceived benefit of the application—increased awareness and data visualization—and the core interaction cost associated with the logging process.

In addition to these high-level trends, we also observe that users generally prefer a mix of modalities. Fig 4 shows a timeline representative of an average day for one user. This user seems to prefer logging with image around actual meal times while preferring to log with text for snack times. These figures show that user preferences for how they enter data varies and our application accommodates this to reduce the mental load of tracking.

6.6.3 Interactive AI follow-up

Follow-up questions were a generally new concept in food logging for the participants. Fig 5 reveals participants generally believed the follow-up questions were relevant to their food and personal goals as well as helped clarify details their initial uploads did not cover. Table 7 exhibits the distribution of follow-up question categories with most of the questions being about quantity & portion size and food type & detail, questions that were important to clarify in Section 3.2.

While the follow-up questions were rated as relevant to the food, the quantitative results show

Table 7: Gemini Classification of follow-up questions classified. Definitions for categories and the prompt used to classify can be found in Appendix B.4.

Category	Percentage (%)	Count (n)
Quantity & Portion Size	34.3	610
Food Type & Detail	31.8	566
Preparation & Source	20.1	357
Consumption Ratio	12.4	221
Other	1.4	25

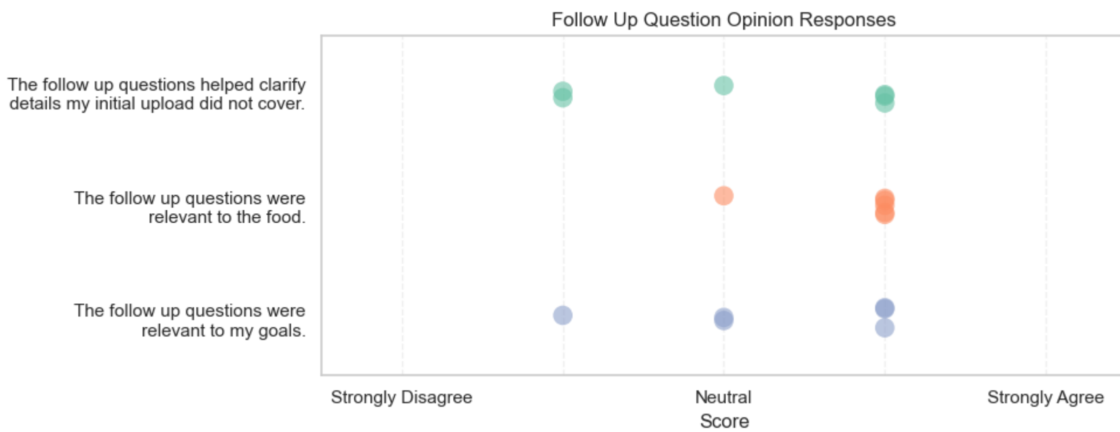


Figure 5: Responses to "follow-up"-focused questions in exit survey.

they did not make the process easier than traditional methods. Fig 6 reveals there was no clear correlation between the number of follow-up questions received and users' opinions reflecting the follow-up question methodology. This suggests that the survey results were not significantly biased by the number of follow-up questions administered. However, Fig 7a reveals more experienced food loggers (at least 6 months of experience) generally received more follow-up questions. This is likely because more experienced loggers tended to log more in general, as illustrated in Fig 7b.

The follow-up questions were intended to improve data quality and self-awareness when logging food, but they inadvertently became a source of user friction. While they were rated as relevant to the food, the quantitative results show follow-up questions did not make the process easier than traditional methods. The feedback suggested that the follow-up questions failed to adapt to input context. For instance, a user employing Text Logging reported receiving follow-up questions "phrased for photo input," and another complained of generic, frustrating defaults: "...it would default to 'how much chicken did you bake' which was frustrating to edit." Furthermore, users requested the ability to skip follow-up questions for simple, single-ingredient foods (e.g., a banana), emphasizing the need for greater efficiency.

Beyond the exit interviews, analysis of the application's usage data provided quantitative metrics on logging accuracy and behavior. Due to the app's configuration, we tracked the frequency of log edits and deletions. Out of 502 total food logs, 104 (20.7%) were edited and 29 (5.8%) were deleted.

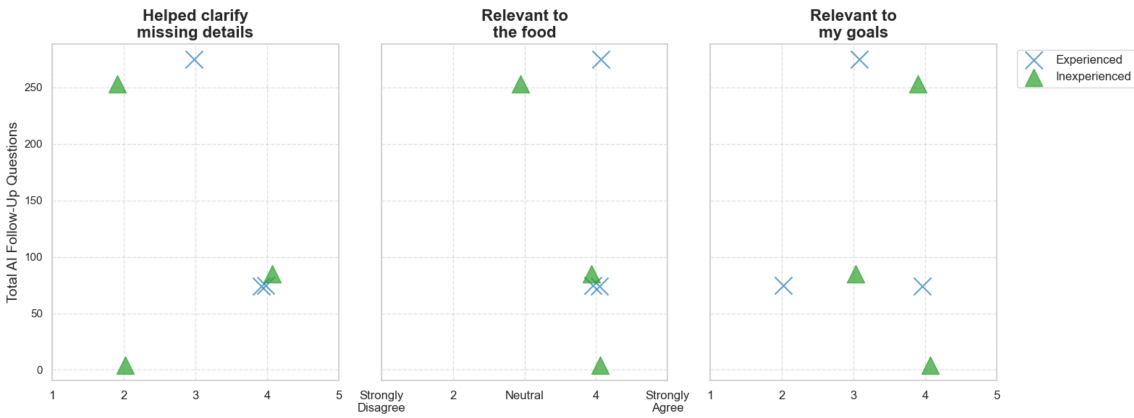
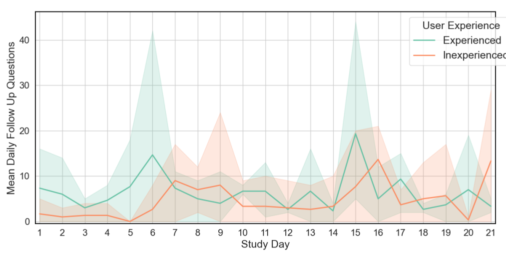
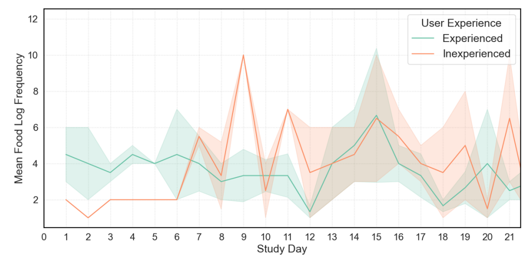


Figure 6: User perception vs. interaction volume categorized by experience.



(a) Mean daily follow-up questions over time.



(b) Mean daily user food logs over time.

Figure 7: Food logging trends by user experience (more than 6 months of experience with nutrition tracking).

6.7 Field Pilot Study Discussion

The 21-day deployment successfully evaluated the novel nutrition tracking application in a real-world setting, providing critical insights into the trade-off between minimizing tracking fatigue and maintaining data accuracy. The study demonstrates two significant positive outcomes: (1) The system successfully fostered high user engagement and adherence over the 21-day trial, and (2) it achieved its primary goal of increasing users' self-awareness of their eating habits.

Engagement was notably high, with all six participants completing the entire 21-day study and four of six logging daily. This high adherence, despite self-reported frustrations with the time required ($n = 3$), suggests that the perceived benefits of the system—namely, feeling more aware of their eating habits ($\bar{x} = 3.67$)—were compelling enough to overcome usability hurdles. This finding is a key success, as "logging fatigue" and user drop-off are primary challenges for longitudinal health apps [12].

Despite these successes, our findings also reveal a critical tension: the very AI features intended to enhance the logging experience also became new sources of friction. The discussion section below will unpack this by examining (1) the success and challenges of the AI-driven follow-ups, (2) the value and friction of the multimodal design, and (3) the implications for designing future context-aware food logging systems.

A key success of our system was the content and relevance of the AI-driven follow-ups. Participants strongly agreed that the questions were relevant to their food ($\bar{x} = 3.83$) and their personal goals ($\bar{x} = 3.83$). The system demonstrated its utility by helping users "clarify details their initial uploads did not cover" (Fig 5). Our quantitative analysis confirms this, showing the AI correctly targeted the most critical missing information: "Quantity & Portion Size" (34.3%) and "Food Type & Detail" (31.8%) (Table 7). This demonstrates the potential of an interactive AI to intelligently guide users toward higher-quality data collection.

6.7.1 Limitations

Several limitations must be considered when interpreting these findings. First, while our three-week deployment yielded high-resolution qualitative insights, the sample size ($N = 6$) is small and consists of a tech-forward cohort of students and engineers. This study serves as a feasibility pilot rather than a large-scale generalizable evaluation; the high engagement observed may be partially attributable to the participants' technological aptitude. Future work should evaluate SnappyMeal with more diverse populations, including those with varying levels of digital literacy.

Second, we observed instances of compensatory logging, where participants logged multiple meals in a single session to catch up on missed days. This introduces potential recall bias, a well-documented challenge in dietary self-tracking [8]. While SnappyMeal's interactive follow-up questions were designed to prompt for missing details, the accuracy of recall-based entries (especially text-only logs without a reference photo) may be lower than real-time entries. Future iterations could investigate proactive notifications to reduce the need for compensatory logging.

The 21-day study duration was sufficient to observe sustained use without significant drop-off (Fig 3), but a longer longitudinal study would be needed to assess the long-term effects of "logging fatigue." Finally, our metrics for "logging burden" were based on self-report; future studies could benefit from more objective measures, such as timestamped interaction logs to precisely quantify the time spent on each log and edit.

In summary, the field study yielded positive indicators of high user engagement and perceived awareness benefits among a tech-savvy cohort. The critical and specific feedback regarding contextual failure in the follow-up questions and friction in image recognition offers a powerful set of instructions for future iterations, allowing us to focus on dramatically lowering the interaction cost to further maximize compliance and data quality. Ultimately, the goal is to make the system proactively adaptive rather than merely reactive.

7 Discussion

Our formative study revealed that dietitians and food journalers prioritize flexible, contextually-aware data capture. While many of the model features occurred in the backend, follow-up questions allowed users to focus on the experience of eating. Though this feature resulted in a decreased accuracy during our system evaluation (Section 5), the field study (Section 6) demonstrated its significant value to real-world users, who appreciated the added awareness it provided.

7.1 AI Follow-ups

While challenging to quantify, the follow-up questions improved user engagement. Users reported feeling more involved in the food logging process. The act of being asked served as a behavioral intervention, driving the perceived value in the field study. Despite this, the follow-ups resulted in user effort that led to statistically noisier data, sometimes due to the poor nutrition annotations of users. The negative effect of textual follow-ups on nutrition value generation was expected as the resolution and timing of follow-up-question answers significantly impact the quality of generation, especially when the user themselves does not know the exact answer the model is searching for [43]. The observed negative impact of the follow-up questions is likely compounded by the inherent noise in the human response data. The specific nature of the follow-up questions was non-trivial to annotate, requiring detailed subjective estimations not readily available from the image and on foods the study team did not cook or eat. Consequently, the system evaluation annotators frequently struggled to answer these questions with real confidence, introducing a significant source of error or conflicting data into the prompt. Given the ambiguity of the required input, we believe this particular follow-up question evaluation on this constrained dataset is not truly representative of a real-world deployment where a user might be present at the time of eating.

However, this success in *what* to ask was disconnected from *how* and *when* to ask. This "context-insensitivity" was a primary source of user frustration. The main complaints included: lack of modality awareness, generic and inefficient defaults, and lack of adaptive friction. This finding has significant implications for HCI: an "intelligent" system is not just one that identifies correct information, but one that knows when not to intervene. The lack of correlation between the number of follow-ups and user opinion (Fig 6) reinforces this; it was not the quantity of questions that mattered, but the quality and context of each interruption.

Future systems should be trained to actively default to user convenience and only introduce friction when the benefit of the improved estimation significantly outweighs the burden on the user. This necessitates a new set of system intervention thresholds based on user-centric criteria, ultimately leading to more usable and trustworthy automated logging tools.

7.2 User-Driven Flexibility

The multimodal input design was a clear success in providing user-driven flexibility. Participants organically developed logging strategies to fit their daily context. As visualized in Fig 4, a user might prefer images for complex meals (e.g., at mealtimes) and text for simple snacks. This accommodation for varying user preferences and situations is a positive outcome, demonstrating that the system successfully reduced the mental load of tracking by allowing users to choose the most efficient path. This benefit was, however, hampered by the failure of image recognition. While there was a preference for image logging ($n = 3$), it was directly undermined by the poor performance of the image recognition model ("Sometimes accurate (25-50%)"). The high "edit" rate (20.7% of all logs) and the corresponding high appreciation for the edit feature ($\bar{x} = 4.33$) are symptoms of this system making frequent mistakes.

In the future, systems should support more modalities such as inferences from wearable metrics and more personalized meta-data. Time of day, location, and other meta-data can assist the inference of models. For example, if the user's location is at a restaurant a system could automatically retrieve an online menu to supplement its context. With the wide access to user data [23], systems can now perform extrapolations that would improve the accuracy of nutrition estimates and infer diet choice consequences for each user.

7.3 Moving from "Interactive" to "Context-Aware"

A core lesson from this study is the need to shift from an interactive system to a context-aware one. The current system places the burden of context-switching and error correction on the user. Based on our findings, we propose two key design implications for future work:

1. **Pipe Modality Context to the AI:** The AI model (Gemini) must be explicitly prompted with the input modality. This would prevent basic errors like asking photo-phrased questions for text logs and represents a straightforward, high-impact fix.

2. **Implement Adaptive Friction:** The system should dynamically adjust its level of intervention based on the confidence of the initial log. For a high-confidence text log ("banana") or a high-accuracy image recognition, the system should default to no follow-up question to maximize efficiency.

System intelligence must practice restraint. Future work should focus on building context engines that manage the *when* and *how* of user intervention, allowing users to concentrate on their primary tasks rather than correcting an overzealous digital assistant. This paradigm shift will lead to logging tools that are not only accurate but truly supportive of real-world behavior change.

7.4 Implications for Future Work

In summary, the design and evaluation of SnappyMeal revealed a fundamental tension in the development of automated logging tools: the conflict between algorithmic accuracy and user-perceived value. Our system evaluation demonstrated that follow-up questions, while intending to gather crucial missing data, introduced noise and cognitive load that reduced the final accuracy of the nutrition estimates. However, the subsequent field user study demonstrated that this very friction was valuable, acting as a prompt for user self-contextualization and reflection—the true goal of many behavior change applications.

This outcome necessitates a paradigm shift from building merely "interactive" systems to engineering truly "context-aware" systems that practice restraint. Our key design implication is the implementation of a system's decision to intervene being governed by the input modality, the confidence of the initial log, and the real-world cost of user interruption. This system intelligence must be rooted not only in what information is missing but in when and how to ask for it.

The challenges we identified extend beyond nutrition directly into other high-stakes, data-intensive domains. Any task where an LLM needs to make a high-stakes decision based on incomplete user input will require it to actively probe for necessary details rather than relying solely on the initially logged information. This extends to clinical settings as well. Englhardt et al. [17] find that adding more data to their prompt, even if the data is accurate, does not necessarily improve LLM reasoning performance on extrapolating depression and anxiety from activity, sleep and social interaction data. Their findings suggest that improvement does not come simply from logging more information, but from designing systems that can contextualize, interpret, and communicate logged data in clinically useful ways.

Similarly, Li et al. [27] emphasize the importance of follow-up questions in clinical pre-consultation systems and argue that such systems must be able to ask targeted questions based on fine-grained patient attributes. In this framing, the value of follow-up questions is not that they make the interaction more conversational, but that they improve the quality and completeness of the collected information. By eliciting relevant missing context at the point of logging, these systems may reduce the need for extensive clarification later during provider communication. Additionally, Li et al. [28] demonstrate the importance of comprehensive summaries for both clinicians and users in pre-consultation settings, further highlighting the need for systems that can transform fragmented user input into structured, interpretable, and actionable records.

Together, these works and the findings from this thesis point toward the design requirements for comprehensive holistic health logging systems. Rather than treating logging as the passive accumulation of more data, future systems should treat logging as an adaptive process of contextualization. Such systems should estimate their confidence in generating a complete and useful log before deciding whether to intervene. They should also reason across multiple forms of context, including physiological signals, geospatial patterns, behavioral routines, temporal trends, and user-provided reflections. For example, a holistic health logging system might connect a user's meal, sleep, activity, mood, symptoms, and location patterns to help explain why a particular health event occurred or why a symptom may have intensified. However, this expansion of context must be balanced against the risk of overburdening users with excessive prompts or producing misleading interpretations from loosely related data.

This broader vision motivates HoloLog, an ongoing direction for future work that extends beyond single-domain tracking toward the integrated capture and interpretation of everyday health experiences. HoloLog explores how fragmented forms of health context might be organized into structured longitudinal records. HoloLog represents a natural extension of the design implications surfaced by SnappyMeal: logging systems should not simply ask users for more information, but should help them structure, connect, and make sense of the information that matters across domains. The tension identified in SnappyMeal between useful friction and cognitive burden becomes even more important in holistic health logging, where the amount of potentially relevant context is much larger and the risk of overwhelming users is greater.

The long-term opportunity is therefore not simply to build systems that collect more health data, but to build systems that know how to ask for, organize, and interpret the right data at the right time. SnappyMeal demonstrates this opportunity within the domain of nutrition tracking: even when follow-up questions do not directly improve estimation accuracy, they can still help users reflect on their behaviors and provide richer contextual logs. HoloLog builds on this same insight at a broader scale by imagining how LLM-powered systems might support sensemaking across the many interconnected experiences that shape everyday health. Extending these insights to holistic health logging suggests a future in which automated logging tools support retrospective recall as well as deeper reflection, contextual understanding, and communication across all aspects of personal health.

8 Conclusion

While dietary tracking is critical for understanding health outcomes, current methods like app-based journaling are inflexible, resulting in poor user adherence and imprecise nutritional summaries. SnappyMeal, our proposed AI system, enhances data quality by intelligently posing goal-dependent follow-up questions to acquire missing context and by utilizing information retrieval from grocery receipts and nutritional databases. We validated SnappyMeal through public benchmarks and a 3-week, in-the-wild deployment. Participants reported high satisfaction with the multiple input methods and strong perceived accuracy, a sentiment supported by objective benchmark performance. These findings suggest that multimodal AI systems can substantially improve adherence and accuracy, heralding a new class of intelligent self-tracking applications. Ultimately, SnappyMeal demonstrates the need for restrained intelligence. The most usable and trustworthy logging tools will be those that prioritize user convenience by defaulting to maximum efficiency and only intervening when the expected benefit to the estimate significantly outweighs the cognitive cost to the human.

References

[1] Sean H Adams, Joshua C Anthony, Ricardo Carvajal, Lee Chae, Chor San H Khoo, Marie E Latulippe, Nathan V Matusheski, Holly L McClung, Mary Rozga, Christopher H Schmid, Suzan Wopereis, and William Yan. Perspective: Guiding principles for the implementation of personalized nutrition approaches that benefit health and function. *Advances in Nutrition*, 11(1):25–34, 2020. ISSN 2161-8313. doi: https://doi.org/10.1093/advances/nmz086. URL https://www.sciencedirect.com/science/article/pii/S2161831322002320.

[2] Giuseppe Agapito, Mariadelina Simeoni, Barbara Calabrese, Ilaria Caré, Theodora Lamprinoudi, Pietro H. Guzzi, Arturo Pujia, Giorgio Fuiano, and Mario Cannataro. Dietos: A dietary recommender system for chronic diseases monitoring and management. *Computer Methods and Programs in Biomedicine*, 153:93–104, 2018. ISSN 0169-2607. doi: https://doi.org/10.1016/j.cmpb.2017.10.014. URL https://www.sciencedirect.com/science/article/pii/S0169260716306927.

[3] Sameera Sh. Mohammed Ameen, Khalid M. Omer, Fotouh R. Mansour, Alaa Bedair, and Mahmoud Hamed. Non-invasive wearable electrochemical sensors for continuous glucose monitoring. *Electrochemistry Communications*, 173:107894, 2025. ISSN 1388-2481. doi: https://doi.org/10.1016/j.elecom.2025.107894. URL https://www.sciencedirect.com/science/article/pii/S1388248125000335.

[4] Oliver Amft, Mathias Stäger, Paul Lukowicz, and Gerhard Tröster. Analysis of chewing sounds for dietary monitoring. In *Proceedings of the 7th International Conference on Ubiquitous Computing*, UbiComp'05, page 56–72, Berlin, Heidelberg, 2005. Springer-Verlag. ISBN 9783540287605. doi: 10.1007/11551201_4. URL https://doi.org/10.1007/11551201_4.

[5] Traci A Bekelman, Corby K Martin, Susan L Johnson, Deborah H Glueck, Katherine A Sauder, Kylie K Harrall, Rachel I Steinberg, Daniel S Hsia, Dana Dabelea, and program collaborators for Environmental influences on Child Health Outcomes (ECHO). A comparison of the remote food photography method and the automated self-administered 24-h dietary assessment tool for measuring full-day dietary intake among school-age children. *Br. J. Nutr.*, 127(8):1269–1278, April 2022.

[6] Joan-Isaac Biel, Nathalie Martin, David Labbe, and Daniel Gatica-Perez. Bites'n'bits: Inferring eating behavior from contextual mobile data. *Interact. Mob. Wearable Ubiquitous Technol.*, 1(4), January 2018. doi: 10.1145/3161161. URL https://doi.org/10.1145/3161161.

[7] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006. doi: 10.1191/1478088706qp063oa. URL https://doi.org/10.1191/1478088706qp063oa.

[8] Tracy L Burrows, Rebecca J Martin, and Clare E Collins. A systematic review of the validity of dietary assessment methods in children when compared with the method of doubly labeled water. *J. Am. Diet. Assoc.*, 110(10):1501–1510, October 2010.

[9] Juliana Chen, William Berkman, Manal Bardouh, Ching Yan Kammy Ng, and Margaret Allman-Farinelli. The use of a food logging app in the naturalistic setting fails to provide accurate measurements of nutrients and poses usability challenges. *Nutrition*, 57:208–216, 2019. ISSN 0899-9007. doi: https://doi.org/10.1016/j.nut.2018.05.003. URL https://www.sciencedirect.com/science/article/pii/S0899900718303678.

[10] Jingyuan Cheng, Bo Zhou, Kai Kunze, Carl Christian Rheinländer, Sebastian Wille, Norbert Wehn, Jens Weppner, and Paul Lukowicz. Activity recognition and nutrition monitoring in every day situations with a textile capacitive neckband. In *Proceedings of the 2013 ACM Conference on Pervasive and Ubiquitous Computing Adjunct Publication*, UbiComp '13 Adjunct, page 155–158, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450322157. doi: 10.1145/2494091.2494143. URL https://doi-org.offcampus.lib.washington.edu/10.1145/2494091.2494143.

[11] Han Shi Jocelyn Chew. The use of artificial intelligence–based conversational agents (chatbots) for weight loss: Scoping review and practical recommendations. *JMIR Med Inform*, 10(4): e32578, Apr 2022. ISSN 2291-9694. doi: 10.2196/32578. URL http://www.ncbi.nlm.nih.gov/pubmed/35416791.

[12] Eun Kyoung Choe, Nicole B. Lee, Bongshin Lee, Wanda Pratt, and Julie A. Kientz. Understanding quantified-selfers' practices in collecting and exploring personal data. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, page 1143–1152, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450324731. doi: 10.1145/2556288.2557372. URL https://doi.org/10.1145/2556288.2557372.

[13] Chia-Fang Chung, Qiaosi Wang, Jessica Schroeder, Allison Cole, Jasmine Zia, James Fogarty, and Sean A. Munson. Identifying and planning for individualized change: Patient-provider collaboration using lightweight food diaries in healthy eating and irritable bowel syndrome. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 3(1), March 2019. doi: 10.1145/3314394. URL https://doi.org/10.1145/3314394.

[14] Bruce Coburn, Jiangpeng He, Megan E. Rollo, Satvinder S. Dhaliwal, Deborah A. Kerr, and Fengqing Zhu. Comprehensive evaluation of large multimodal models for nutrition analysis: A new benchmark enriched with contextual metadata, 2025. URL https://arxiv.org/abs/2507.07048.

[15] Ayush Dahake, Roshan Adhau, Manthan Ghosh, and Laxman Thakre. A holistic approach to health management and integrated fitness mobile application for nutritionists, dietitians, and patients. In *2025 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)*, volume 3, pages 1–7, 2025. doi: 10.1109/IATMSI64286.2025.10984725.

[16] Alaina Darby, Matthew W. Strum, Erin Holmes, and Justin Gatwood. A review of nutritional tracking mobile applications for diabetes patient use. *Diabetes Technology & Therapeutics*, 18

(3):200–212, 2016. doi: 10.1089/dia.2015.0299. URL https://doi.org/10.1089/dia.2015.0299. PMID: 26950679.

[17] Zachary Englhardt, Chengqian Ma, Margaret E. Morris, Chun-Cheng Chang, Xuhai "Orson" Xu, Lianhui Qin, Daniel McDuff, Xin Liu, Shwetak Patel, and Vikram Iyer. From classification to clinical insights: Towards analyzing and reasoning about mobile and behavioral health data with large language models. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 8(2), May 2024. doi: 10.1145/3659604. URL https://doi.org/10.1145/3659604.

[18] Shan Feng, Matti Mäntymäki, Amandeep Dhir, and Hannu Salmela. How self-tracking and the quantified self promote health and well-being: Systematic review. *J. Med. Internet Res.*, 23(9):e25171, September 2021.

[19] Brent I Fox and Bill G Felkey. The quantified self. *Hosp. Pharm.*, 51(2):189–190, February 2016.

[20] Susannah Fox and Maeve Duggan. Tracking for health. *Pew Research Center*, 2013. URL https://www.pewresearch.org/internet/2013/01/28/tracking-for-health/.

[21] Carly Griffiths, Lisa Harnack, and Mark A Pereira. Assessment of the accuracy of nutrient calculations of five popular nutrition tracking applications. *Public Health Nutrition*, 21(8): 1495–1502, 2018. doi: 10.1017/S1368980018000393.

[22] Jiaxiong Hu, Jingya Guo, Ningjing Tang, Xiaojuan Ma, Yuan Yao, Changyuan Yang, and Yingqing Xu. Designing the conversational agent: Asking follow-up questions for information elicitation. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW1), April 2024. doi: 10.1145/3637320. URL https://doi.org/10.1145/3637320.

[23] Robert S.H. Istepanian and Turki Al-Anzi. m-health 2.0: New perspectives on mobile health, machine learning and big data analytics. *Methods*, 151:34–40, 2018. ISSN 1046-2023. doi: https://doi.org/10.1016/j.ymeth.2018.05.015. URL https://www.sciencedirect.com/science/article/pii/S1046202318300860. Health Informatics and Translational Data Analytics.

[24] Jisu Jung, Lyndal Wellard-Cole, Colin Cai, Irena Koprinska, Kalina Yacef, Margaret Allman-Farinelli, and Judy Kay. Foundations for systematic evaluation and benchmarking of a mobile food logger in a large-scale nutrition study. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 4(2), June 2020. doi: 10.1145/3397327. URL https://doi.org/10.1145/3397327.

[25] Sean Kandel. An evidence-based look at the effects of diet on health. *Cureus*, 11(5):e4715, May 2019.

[26] Eduard Kuric, Peter Demcak, and Matus Krajcovic. Unmoderated usability studies evolved: Can gpt ask useful follow-up questions? *International Journal of Human–Computer Interaction*, 41(15):9752–9769, 2025. doi: 10.1080/10447318.2024.2427978. URL https://doi.org/10.1080/10447318.2024.2427978.

[27] Brenna Li, Ofek Gross, Noah Crampton, Mamta Kapoor, Saba Tauseef, Mohit Jain, Khai N. Truong, and Alex Mariakakis. Beyond the waiting room: Patient's perspectives on the conversational nuances of pre-consultation chatbots. In *Proceedings of the 2024 CHI Conference*

on Human Factors in Computing Systems, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703300. doi: 10.1145/3613904.3641913. URL https://doi.org/10.1145/3613904.3641913.

[28] Brenna Li, Liam of Bakar, Anna Kirik, Jiaqi Guo, Alex Mariakakis, and Khai N. Truong. Digitizing the pre-consultation experience: Impacts and design recommendations. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400722783. doi: 10.1145/3772318.3791354. URL https://doi.org/10.1145/3772318.3791354.

[29] Xi Lu, Edison Thomaz, and Daniel A. Epstein. Understanding people's perceptions of approaches to semi-automated dietary monitoring. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 6(3), September 2022. doi: 10.1145/3550288. URL https://doi.org/10.1145/3550288.

[30] Yan Meng, Liangming Pan, Yixin Cao, and Min-Yen Kan. Followupqg: Towards information-seeking follow-up question generation, 2023. URL https://arxiv.org/abs/2309.05007.

[31] Yiqun Miao, Yuhan Zhao, Yuan Luo, Huiying Wang, and Ying Wu. Improving large language model applications in the medical and nursing domains with retrieval-augmented generation: Scoping review. *J Med Internet Res*, 27:e80557, Oct 2025. ISSN 1438-8871. doi: 10.2196/80557. URL https://doi.org/10.2196/80557.

[32] Mark Mirtchouk, Christopher Merck, and Samantha Kleinberg. Automated estimation of food type and amount consumed from body-worn audio and motion sensors. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '16, page 451–462, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450344616. doi: 10.1145/2971648.2971677. URL https://doi-org.offcampus.lib.washington.edu/10.1145/2971648.2971677.

[33] Ranjita Misra and Delores James. The role of dietary tracking on changes in dietary behaviour in a community-based diabetes prevention and management intervention. *Public Health Nutrition*, 28(1):e73, 2025. doi: 10.1017/S1368980025000436.

[34] J.D. Moore and V.O. Mittal. Dynamically generated follow-up questions. *Computer*, 29(7): 75–86, 1996. doi: 10.1109/2.511971.

[35] Fnu Neha, Deepshikha Bhati, and Deepak Kumar Shukla. Retrieval-augmented generation (rag) in healthcare: A comprehensive review. *AI*, 6(9), 2025. ISSN 2673-2688. doi: 10.3390/ai6090226. URL https://www.mdpi.com/2673-2688/6/9/226.

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021. URL https://arxiv.org/abs/2103.00020.

[37] Britta Renner, Anette E. Buyken, Kurt Gedrich, Stefan Lorkowski, Bernhard Watzl, Jakob Linseisen, Hannelore Daniel, Johanna Conrad, Paola G. Ferrario, Christina Holzapfel, Michael Leitzmann, Margrit Richter, Marie-Christine Simon, Christian Sina, and Jan Wirsam.

Perspective: A conceptual framework for adaptive personalized nutrition advice systems (apnass). *Advances in Nutrition*, 14(5):983–994, 2023. ISSN 2161-8313. doi: https://doi.org/10.1016/j.advnut.2023.06.009. URL https://www.sciencedirect.com/science/article/pii/S2161831323013364.

[38] Nina Reščič, Eva Valenčič, Enej Mlinarič, Barbara Koroušić Seljak, and Mitja Luštrek. Mobile nutrition monitoring for well-being. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*, UbiComp/ISWC '19 Adjunct, page 1194–1197, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450368698. doi: 10.1145/3341162.3347076. URL https://doi-org.offcampus.lib.washington.edu/10.1145/3341162.3347076.

[39] Sergio Romero-Tapiador, Ruben Tolosana, Blanca Lacruz-Pleguezuelos, Laura Judith Marcos Zambrano, Guadalupe X. Bazán, Isabel Espinosa-Salinas, Julian Fierrez, Javier Ortega-Garcia, Enrique Carrillo de Santa Pau, and Aythami Morales. Are vision-language models ready for dietary assessment? exploring the next frontier in ai-powered food image recognition, 2025. URL https://arxiv.org/abs/2504.06925.

[40] Gopika SenthilKumar, Aronne M. Schottstaedt, Lindsay L. Peterson, Lauren N. Pedersen, Christopher R. Chitambar, Alexis Vistocky, Anjishnu Banerjee, John M. Longo, Tracy Kelly, Adam Currey, Melinda R. Stolley, and Carmen Bergom. Stay on track: A pilot randomized control trial on the feasibility of a diet and exercise intervention in patients with breast cancer receiving radiotherapy. *Cancer Research Communications*, 4(5):1211–1226, 05 2024. ISSN 2767-9764. doi: 10.1158/2767-9764.CRC-23-0148. URL https://doi.org/10.1158/2767-9764.CRC-23-0148.

[41] Farzad Shahabi, Boyang Wei, Christopher Romano, Rowan McCloskey, Annie W. Lin, Mahdi Pedram, Jacob M. Schauer, Tammy Stump, and Nabil Alshurafa. Unveiling overeating patterns within digital longitudinal data on eating behaviors and contexts. *npj Digital Medicine*, 8(1):567, 2025. doi: 10.1038/s41746-025-01698-9. URL https://doi.org/10.1038/s41746-025-01698-9.

[42] Andrea Sosa-Holwerda, Oak-Hee Park, Kembra Albracht-Schulte, Surya Niraula, Leslie Thompson, and Wilna Oldewage-Theron. The role of artificial intelligence in nutrition research: A scoping review. *Nutrients*, 16(13), 2024. ISSN 2072-6643. doi: 10.3390/nu16132066. URL https://www.mdpi.com/2072-6643/16/13/2066.

[43] Vidya Srinivas, Xuhai Xu, Xin Liu, Kumar Ayush, Isaac Galatzer-Levy, Shwetak Patel, Daniel McDuff, and Tim Althoff. Substance over style: Evaluating proactive conversational coaching agents, 2025. URL https://arxiv.org/abs/2503.19328.

[44] Frédérique N. Sunstrum, Jawairia Umar Khan, Nga-Wun Li, and Alec W. Welsh. Wearable textile sensors for continuous glucose monitoring. *Biosensors and Bioelectronics*, 273:117133, 2025. ISSN 0956-5663. doi: https://doi.org/10.1016/j.bios.2025.117133. URL https://www.sciencedirect.com/science/article/pii/S0956566325000077.

[45] Quin Thames, Arjun Karpur, Wade Norris, Fangting Xia, Liviu Panait, Tobias Weyand, and Jack Sim. Nutrition5k: Towards automatic nutritional understanding of generic food. *CoRR*, abs/2103.03375, 2021. URL https://arxiv.org/abs/2103.03375.

[46] KS Todd, M Hudes, and DH Calloway. Food intake measurement: problems and approaches. *The American Journal of Clinical Nutrition*, 37(1):139–146, 1983. ISSN 0002-9165. doi: https://doi.org/10.1093/ajcn/37.1.139. URL https://www.sciencedirect.com/science/article/pii/S0002916523157455.

[47] Ricardo Uauy. Defining and addressing the nutritional needs of populations. *Public Health Nutrition*, 8(6a):773–780, 2005. doi: 10.1079/PHN2005774.

[48] Muhammad Usman, Kashif Ahmad, Amir Sohail, and Marwa Qaraqe. The diabetic buddy: A diet regulator and tracking system for diabetics. In *2021 International Conference on Content-Based Multimedia Indexing (CBMI)*, pages 1–4, 2021. doi: 10.1109/CBMI50038.2021.9461897.

[49] Meghna Verma, Raquel Hontecillas, Nuria Tubau-Juni, Vida Abedi, and Josep Bassaganya-Riera. Challenges in personalized nutrition and health. *Front. Nutr.*, 5:117, November 2018.

[50] Liangyu Wang, Yoko Yamakata, Ryoma Maeda, and Kiyoharu Aizawa. Measure and improve your food: Ingredient estimation based nutrition calculator. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 11273–11275, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706868. doi: 10.1145/3664647.3684997. URL https://doi-org.offcampus.lib.washington.edu/10.1145/3664647.3684997.

[51] Lufan Wang, Crystal S. Langlais, Stacey A. Kenfield, June M. Chan, Rebecca E. Graff, Isabel E. Allen, Chloe E. Atreya, and Erin L. Van Blarigan. mhealth interventions to promote a healthy diet and physical activity among cancer survivors: A systematic review of randomized controlled trials. *Cancers*, 14(15), 2022. ISSN 2072-6694. doi: 10.3390/cancers14153816. URL https://www.mdpi.com/2072-6694/14/15/3816.

[52] Shan Xu, Zhaokun Yan, Chengxiao Dai, and Fan Wu. Mega-rag: a retrieval-augmented generation framework with multi-evidence guided answer refinement for mitigating hallucinations of llms in public health. *Frontiers in Public Health*, Volume 13 - 2025, 2025. ISSN 2296-2565. doi: 10.3389/fpubh.2025.1635381. URL https://www.frontiersin.org/journals/public-health/articles/10.3389/fpubh.2025.1635381.

[53] Michael Yeomans, Alison Wood Brooks, Karen Huang, Julia Minson, and Francesca Gino. It helps to ask: The cumulative benefits of asking follow-up questions. *J. Pers. Soc. Psychol.*, 117(6):1139–1144, December 2019.

[54] He Zhang, Yueyan Liu, Xin Guan, Jie Cai, and John M. Carroll. Harnessing the power of ai in qualitative research: Role assignment, engagement, and user perceptions of ai-generated follow-up questions in semi-structured interviews, 2025. URL https://arxiv.org/abs/2509.12709.

A Database Schema

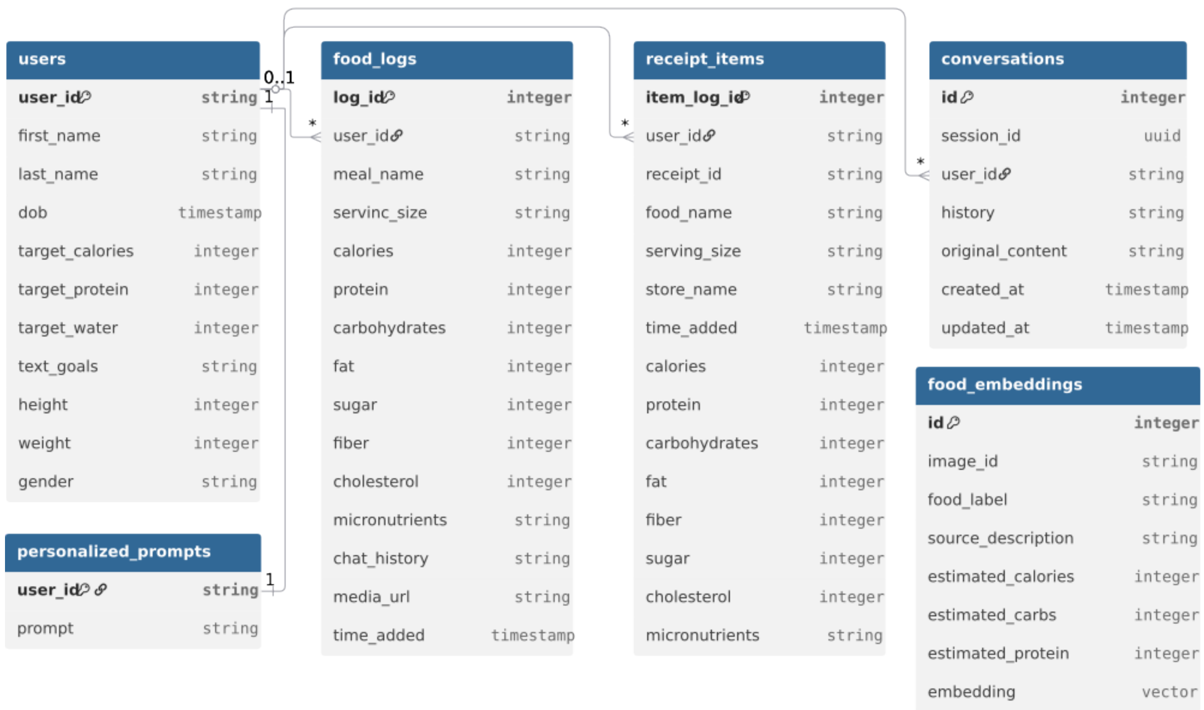


Figure 8: Relational Database Schema

B Prompts

B.1 Vanilla Food Log Generation Prompt

```
Analyze the nutritional content of the food items in this piece of media. The media
can be image, text, or audio.
You are also provided with a chat history that clarifies potentially missing information.
Please identify the main ingredients and estimate the approximate macronutrient breakdown
(calories, protein, carbohydrates, fat) and highlight any significant micronutrients
that are visually discernible.
Output the results in a structured JSON format with the following keys: "ingredients",
"macronutrients", and "micronutrients".
Under "meal_name", come up with a name that summarizes the food in as few words as
possible.
Under "ingredients", list all the ingredients that could affect the nutritional value.
Under "serving_size", estimate how much food there is in customary units (oz, cups,
tbsp, etc.).
Under "calories", estimate how many calories are in the food. Don't provide the units,
use the most standard unit.
```

```
Under "protein", estimate how much protein is in the food.  Don’t provide the units,
use the most standard unit.
Under "carbohydrates", estimate how many carbohydrates are in the food.  Don’t provide
the units, use the most standard unit.
Under "fat", estimate how much fat is in the food.  Don’t provide the units, use the
most standard unit.
Under "fiber", estimate how much fiber is in the food.  Don’t provide the units, use
the most standard unit.
Under "sugar", estimate how much sugar is in the food.  Don’t provide the units, use
the most standard unit.
Under "cholesterol", estimate how much cholesterol is in the food.  Don’t provide the
units, use the most standard unit.
Under "micronutrients", list any significant vitamins or minerals that are visible
in the image and their amounts.  Only include one number and nothing else.  If you
think it’s between some range, use the average.  Don’t include units.
Reminder that JSON formatting requires double quotes.
Example output:
{example_output}
Do not output anything except the JSON.
```

B.2 Example Generated Food Log JSON


```json
{
    "meal_name": "Peanut butter and celery",
    "ingredients": ["peanut butter", "celery"],
    "serving_size": "1 large celery stalk with 2 tablespoons creamy peanut butter",
    "meal_type": "snack",
    "date": "2025-05-07T10:13:27Z",
    "calories": 280,
    "protein": 11,
    "carbohydrates": 16,
    "fat": 20,
    "fiber": 4,
    "sugar": 7,
    "saturated_fat": 4,
    "cholesterol": 0,
    "micronutrients":
    {
        "vitamin_k_mcg": 30,
        "vitamin_a_iu": 500,
        "folate_mcg": 40,
        "potassium_mg": 450,
        "magnesium_mg": 60,
        "phosphorus_mg": 120,
        "vitamin_e_mg": 2,
        "niacin_mg": 3,
        "zinc_mg": 1
    }
}
```

B.3 Follow-Up Question Generation Prompt

You are an AI assistant for a food logging service. Your main goal is to gather detailed information about a user’s meal to create a complete food log entry. Users will provide media (image, text, or audio) of their food.
Task Workflow
1. **Analyze Input:** Examine the user’s media to identify the food and any information already provided.
2. **Check for Missing Information:** Compare the extracted information against all required fields for a food log entry: ‘serving_size‘, ‘calories‘, ‘protein‘, ‘carbohydrates‘, ‘fat‘, ‘sugar‘, ‘fiber‘, ‘cholesterol‘, and ‘micronutrients‘.
3. **Formulate a Single Clarifying Question:** If any crucial information is missing or ambiguous, you must formulate **one** concise, direct follow-up question.
* Prioritize questions that fill the most critical gaps first (e.g., what the food is, then the quantity, then preparation method).
* The question must be relevant to the user’s input. For example, do not ask about an image if one was not provided.
* For homemade meals, if possible, refer to recent receipts to ask a targeted question about ingredients.
* Avoid greetings, conversational filler, or explanations. The output should be **only** the question.
* The question should directly help gather information for the food log fields. Do not ask about nutritional measurements (e.g., "How many calories?")
* **Portion Size:** Try to understand how much food the person ate (e.g. "How many slices of toast were there?").
4. **Determine Question Type:** The question must be one of the following types:
* **text:** Answered with free-form text.
* **select:** Answered by selecting one option from a list. This is the preferred type and should have no more than 3 options.
5. **Output Format:** If a question is needed, output only the question, its type, and any options (if applicable). These three pieces of information must be separated by a semicolon (;).
* Example: ‘How many slices of pizza did you eat?;select;[1,2,3,4,5]‘
* Example: ‘Are there any unseen ingredients in the lasagna?;select;[yes,no]‘
* Example: ‘What is inside your burrito?;text;[]‘
Constraints
* **One Question Only:** Generate a single question at a time.
* **Relevance:** All questions must be directly relevant to completing the food log fields.
* **No Timestamps:** You are forbidden from asking about the time the meal was eaten.
* **No Extraneous Text:** Do not output any text other than the formatted question.
* **Clarity on Units:** For ‘select‘ questions, explicitly state the units (e.g., "How many glasses of milk?"). Do not ask vague questions like "What is the serving size?" and do not ask about measurable units like "How many grams of protein?". Remember, it’s hard for individuals to guess portion sizes.
* **Receipts:** Only ask questions about receipts if the meal seems to contain something from the receipt list and the question is relevant.

```
* **Necessity:** Do not ask unnecessary questions.  If you can determine a value or
fact, don't ask about it.  If the value or fact is hard for a human to estimate, don't
ask about it.
Example output:
What percentage of the food did you consume?;text;[]
Example output:
What is inside of your burrito?;text;[]
Example output:
Are there any unseen ingredients in the lasagna?;select;[yes,no]
Example output:
Is this curry homemade?;select;[yes,no]
Example output:
How was the chicken cooked?;select;[roasted,fried,other]
Example output:
How were the vegetables prepared?;select;[stir fried,steamed,raw,other]
Example output:
How many slices of pizza did you eat?;select;[1,2,3,4,5,6,7,8,9,10]
Example output:
Was this protein shake store bought or homemade?;select;[store bought,homemade]
Example output:
Where did you buy your protein powder from?;text;[]
Example output:
Is this the Chobani yogurt you bought from Safeway?;select;[yes,no]
```

B.4 Follow-Up Question Classification Prompt

```
You have been tasked with classifying questions surrounding nutrition tracking.
There are four categories:
1.  Preparation & Source:  This category focuses on how the food was prepared or where
it came from.
2.  Food Type & Detail:  These questions seek specific detail about a particular ingredient
or food item, useful for precise nutrient tracking.
3.  Quantity & Portion Size:  This is for tracking the absolute amount or portion of
a specific food item consumed.
4.  Consumption Ratio:  This category is for questions that gauge the user's portion
of a larger, visible meal.
You must output a JSON containing the fields "question" and "category".  Only output
this JSON with no other formatting.
If you think the question does not belong to any of the aforementioned categories,
you can set the "category" to "None".
Here is the question:
```

C Equations

C.1 Error Formulae

$\hat{y}$ denotes an estimated value while y denotes an observed or ground-truth value. N represents the number of samples being evaluated.

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} |\hat{y}_i - y_i| \tag{1}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (\hat{y}_i - y_i)^2} \tag{2}$$

C.2 Percentile Bootstrap Formulation

Let the original evaluation dataset be $D = \{(y_{\text{true},i}, y_{\text{pred},i}) \mid i = 1, \ldots, n\}$, where n is the total number of evaluation samples.

We generate B bootstrap samples. Each sample D_b for $b = 1, \ldots, B$ is created by drawing a set of n indices I_b randomly with replacement from the original indices $\{1, \ldots, n\}$. This results in the bootstrap sample:

$$D_b = \{(y_{\text{true},i}, y_{\text{pred},i}) \mid i \in I_b\}$$

For each bootstrap sample D_b, our evaluation metric θ is calculated, yielding a distribution of B bootstrap estimates:

$$\hat{\theta}_b = \text{Metric}(D_b) \quad \text{for } b = 1, \ldots, B$$

The set of all estimates is $\{\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_B\}$. To construct a $(1-\alpha)$ confidence interval, we first define the significance level. For our 95% CI, $\alpha = 0.05$. The lower and upper percentile bounds, P_1 and P_2, are calculated as:

$$P_1 = \left(\frac{\alpha}{2}\right) \times 100 = \left(\frac{0.05}{2}\right) \times 100 = 2.5$$

$$P_2 = \left(1 - \frac{\alpha}{2}\right) \times 100 = \left(1 - \frac{0.05}{2}\right) \times 100 = 97.5$$

The final $(1-\alpha)$ percentile confidence interval for the metric θ is:

$$\text{CI}_{1-\alpha}(\theta) = \left[\hat{\theta}^{(P_1)}, \hat{\theta}^{(P_2)}\right]$$

where $\hat{\theta}^{(p)}$ denotes the p-th percentile of the sorted list of bootstrap estimates $\{\hat{\theta}_b\}_{b=1}^{B}$.