



# HHS Public Access

Author manuscript

*Proc SIGCHI Conf Hum Factor Comput Syst.* Author manuscript; available in PMC 2022 November 29.

Published in final edited form as:

*Proc SIGCHI Conf Hum Factor Comput Syst.* 2022 April ; 2022: . doi:10.1145/3491102.3501886.

## Examining AI Methods for Micro-Coaching Dialogs

**Elliot G Mitchell,**

Columbia University, Department of Biomedical Informatics, New York, New York,

Geisinger, Steele Institute for Health Innovation, Danville, Pennsylvania

**Noémie Elhadad,**

Columbia University, Department of Biomedical Informatics, New York, New York

**Lena Mamykina**

Columbia University, Department of Biomedical Informatics, New York, New York

### Abstract

Conversational interaction, for example through chatbots, is well-suited to enable automated health coaching tools to support self-management and prevention of chronic diseases. However, chatbots in health are predominantly scripted or rule-based, which can result in a stagnant and repetitive user experience in contrast with more dynamic, data-driven chatbots in other domains. Consequently, little is known about the tradeoffs of pursuing data-driven approaches for health chatbots. We examined multiple artificial intelligence (AI) approaches to enable *micro-coaching* dialogs in nutrition — brief coaching conversations related to specific meals, to support achievement of nutrition goals — and compared, reinforcement learning (RL), rule-based, and scripted approaches for dialog management. While the data-driven RL chatbot succeeded in shorter, more efficient dialogs, surprisingly the simplest, scripted chatbot was rated as higher quality, despite not fulfilling its task as consistently. These results highlight tensions between scripted and more complex, data-driven approaches for chatbots in health.

### Keywords

Health coaching; chatbots; conversational agents; self-management; reinforcement learning

## 1 INTRODUCTION

*Health coaching* is a promising approach to support self-management for the millions living with chronic conditions [18], but there are not enough coaching practitioners to provide care to those in need, let alone provide preventative support. Technology-based approaches can make coaching accessible to larger and more diverse populations [34]. Mobile health

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

[egm2143@cumc.columbia.edu](mailto:egm2143@cumc.columbia.edu) .



applications can scaffold health goal setting, and self-monitoring enables individuals to collect data about their behaviors and health state, which hold potential to personalize coaching support to each individual.

One particularly promising approach for delivering coaching support is through conversational interaction, for example chatbots. Automated conversational approaches like chatbots may be better able to follow clinical communication best practices than human practitioners [3], and are a natural fit with the conversational interaction style at the heart of health coaching [73]. However, some have argued that effective health coaching relies on uniquely human abilities, like empathy and problem solving, and researchers have debated whether automated approaches are suited to health coaching [48, 54]. These arguments are justified, at least in part, by the relatively simple approaches to chatbot design that are typical for health applications. Until recently, most chatbots in health have been scripted or rule-based [38], and while these approaches have shown promising results, they can also be perceived as stagnant and repetitive, which can hamper long-term engagement [38, 48]. In other domains, massive, openly available data sets enable more dynamic, flexible, and passingly intelligent chatbots [2, 59]; incorporating these newer methods could alleviate some of the weaknesses of health chatbots and provide more intelligent, dynamic support, with the potential to increase their positive impact on individuals' health. However, there is a lack of such corpora for health applications, at least partly due to data privacy and confidentiality concerns — consequently there is a lack of knowledge about the potential benefits and limitations of applying data-driven dialog modeling approaches in a health context [38].

In this paper, we investigate one particular dimension of automated health coaching that could benefit from a more intelligent, data-driven approach. In particular, prior work has shown that in the context of goal-oriented health coaching individuals are often unsure about whether they are achieving their health goals, and are eager for feedback and contextually tailored suggestions [47, 48, 53]. This is particularly the case for nutrition-related goals, as opposed to other components of self-management like physical activity, sleep, or medication adherence. Because of the degree of nutrition knowledge and literacy required to categorize and estimate the amounts of foods in one's meal in relation to one's goal, arriving at accurate assessment may present challenges for individuals with varying levels of nutritional literacy [9]. Therefore, goal achievement feedback is an important component of coaching tools. Generating such feedback computationally requires digital records of individuals' health behaviors, for example meal logs, that include a sufficient level of detail to enable comparison between captured meals and nutritional goals. However, comprehensive self-tracking can be burdensome and impede long term engagement [15, 56]. Furthermore, offering feedback on already captured meals has a limited impact on individuals' ability to change their choices. In a coaching context, if an experienced health coach were available when an individual was planning their meals, they might try to help by asking a few carefully selected questions to assess the individual's plan and offer insightful feedback and suggestions. We refer to such brief, health coaching conversations related to specific behaviors as *micro-coaching* and argue that effectively implementing automated micro-coaching dialogs may benefit from more sophisticated Artificial intelligence (AI)



approaches, in particular the need to understand natural language descriptions of meals, and to select the most informative follow-up questions.

In this research we explore multiple approaches to implementing micro-coaching dialogs, including a fully scripted approach and several approaches that incorporate AI with the following research questions: *What are the comparative benefits of these different types of dialog management approaches for coaching chatbots? How do these chatbots compare in their ability to reach their end goal as quickly as possible, and their perceived quality and user experience?*

To answer these questions, we implemented four micro-coaching chatbots; each chatbot could generate a set of dialogs to automatically assess whether a planned meal is consistent with a given nutrition goal. First, a *fully-scripted* chatbot was designed through a user-centered approach with input from diabetes educators. Then, we developed two additional chatbots that incorporated common types of AI: rule-based and data-driven. Applying AI to assess the attainment of nutritional goals requires representing meals in a computer-understandable form. To that end, we developed a knowledge-based system for natural language understanding (NLU) and generation—FoodNLU—to parse and interpret utterances describing an individual’s meal. The *rule-based* chatbot chooses follow-up questions with pre-defined logic based on information extracted from an initial description of a planned meal using FoodNLU. The data-driven chatbot builds on the same FoodNLU system, but instead of a rule-based system, a Reinforcement Learning (RL) model learns how to efficiently manage dialogs from many examples. Finally, as a fourth control condition, we implemented a chatbot with the FoodNLU system and a *random* policy for choosing the next question.

To evaluate these approaches, we conducted a study with crowd workers recruited on Amazon Mechanical Turk. In this study, crowd workers were presented with records of meals and responded to prompts from one of the four chatbots; we used the resulting dialogs to create a crowdsourced corpus for training and evaluation. We compared the resultant dialogs on their length, and also enlisted additional crowd workers to assess the *quality* of dialogs, as well as the perceived *user experience*. To assess *quality*, crowd workers were presented with pairwise comparisons of two dialogs side-by-side, and asked to rate which was superior in terms of the coach’s *question-asking strategy*, as well as the *coherence* and *naturalness* of messages from the coach. To assess perceived *user experience*, crowd workers reviewed multiple dialogs from the same chatbot, and completed the subjective assessment of speech systems questionnaire (SASSI; [28])

In the evaluation study, we found that the RL chatbot succeeded in generating conversations that were significantly shorter than those generated with the other chatbots. Considering the perceived quality of the chatbots, we expected the AI-based chatbots to out-perform the scripted chatbot, but found mixed results, particularly on the user ratings of the chatbot coach’s strategy. Raters assessed longer conversations as more strategic, not shorter, pointing to a tension between the efficiency of dialogs and dimensions of their perceived quality. However, despite the higher quality ratings, the fully-scripted chatbot was able to collect necessary information about users’ meals only 65% of the time. In contrast, the data-driven



(RL) and rule-based chatbots always collected the necessary information. Performance differed across the three nutrition goals considered as case studies in the evaluation; for one of the quantitative goals, the RL-based chatbot was rated as more natural and coherent than the others. There were no differences between the four chatbots in perceived user experience.

These results have implications for the application of data-driven approaches to health chatbots. While the data-driven, RL-based approach was most successful in the intended outcome of brief conversations, crowd workers did not always rate it as high quality. This may have been due to the counterintuitive question-asking strategy of the RL approach that did not aim to preserve a logical flow of questions, as opposed to scripted approach, which was more transparent in why each question was being asked. This work also contributes a dialog corpus that can be used by other researchers to further refine data-driven dialog management approaches. Overall, these results suggest the plausibility and potential benefits of more complex computational approaches to chatbots in health, while highlighting the value of simple and scripted approaches, which have shown promising results in many health domains.

## 2 RELATED WORK

### 2.1 Health coaching

Self-management for many chronic conditions requires changes to daily lifestyle behaviors, like diet, exercise, and sleep [5]. *Health coaching* has been shown to be an effective intervention to support healthy lifestyle changes [18]. Coaching centers on collaborative goal setting, where the coach and client work together in a supportive, longitudinal relationship, aiming to increase motivation and help the client learn self-management strategies and skills [52, 73].

Some have argued that the human element is essential to effective health coaching, and placed particular emphasis on building interpersonal relationships and flexible, contextual thinking [54]. However, human-powered coaching interventions are limited in their ability to reach everyone in need of support because there are not enough coaching practitioners, resources are not always available in underserved communities, reimbursement varies, and there are barriers for individuals in accessing coaching support, including access to transportation or time off from work.

### 2.2 Conversation agents

Conversational agents — sometimes referred to as chatbots, or intelligent assistants — are a class of applications driven by the exchange of natural language between a user and the system. One of the first conversational agents, ELIZA, was developed in the 1960's [72]. The first use case for ELIZA was to emulate a Rogerian psychotherapist with rule-based responses. Decades later, in the present day, more sophisticated conversational agents are nearly ubiquitous, from Siri in smartphones to Alexa in smart speakers [12].

Chatbots can be categorized into three categories based on their functionality: finite state-based, frame-based, and agent-based [44]. Finite state-based systems follow a deterministic,



structured dialog tree or use rule-based language processing to respond to user input. Frame-based systems are useful for task-based applications, where the designer can specify the types of tasks and pieces of information necessary to complete the task, or slots. Frame-based approaches then utilize natural language processing (NLP) to classify the user's intent, and fill the necessary slots in the frame to execute the task. Finally, agent-based or AI systems come closest to replicating human-human dialog, and rely on more complex logic to determine responses, often through data-driven dialog models trained with machine learning (ML). Many of the early examples of chatbots were finite state-based [4, 72], and rule-based agents continue to be developed [23].

Recent approaches to designing more dynamic conversational AI agents rely on training statistical dialog models, for example deep neural networks, that learn from large corpora of example conversations [25]. Through thousands of examples, these models learn a mapping from input messages to output responses. This approach is made possible by many large, openly available corpora of dialogs in many domains like IT support and restaurant searching [43, 59], as well as advances in computational power to train ever-larger language models. In particular, transformer-based models like BERT, and GPT, have shown astounding success in many natural language understanding tasks [6, 17, 69].

Applying these advances to other areas like health coaching presents challenges due to the nature of the task and the need for domain knowledge to accomplish it. Certain types of conversational interaction are well studied, including task-based agents, open domain chit-chat, and question answering systems [1, 2, 25, 42]. However, there may be challenges applying these approaches to different types of interaction like health coaching, which are not primarily user-initiated task-based or question-answering system, nor are they fully unstructured open domain chit chat conversations. Second, while language models can be trained on large corpora and transferred to other, smaller data sets for fine-tuning, healthcare and health coaching dialogs include domain-specific knowledge, for example about nutrition and health state. Pre-trained language models have been developed for biomedical literature [39], but existing models may not necessarily transfer to health coaching, particularly because of the nutrition terminology and informal language used by coaches and clients. In either case, the creation and open availability of corpora would be necessary to train or tune data-driven dialog models for health coaching.

As shown in Figure 1, a common architecture for chatbots separates interpreting user utterances (understanding), from *what* an agent should say in response to user input (dialog management) from *how* to say it (dialog generation) [25]. Importantly, data-driven approaches can be applied to a subset of these components, requiring less data than end-to-end models [25].

### 2.3 Reinforcement learning for conversational agents

A common approach to improve the efficiency of dialog management is to apply Reinforcement Learning (RL) [25, 63]. RL is a distinct machine learning approach, and is separate from supervised learning — where the task is to predict a label or classification for instances in a data set — and unsupervised learning — where the task is to find hidden structures or relationships within a data set. The task for RL is to learn a policy



for what actions to take in a given environment with a certain state. RL agents learn from trial-and-error, collecting rewards as they move through the environment, and keeping track of which actions in which situations bring about the highest long-term reward [63]. In recent years, RL has shown strong performance playing many different games without any expert knowledge about the game's rules or strategy [61].

For conversational agents, RL has been applied to improve an agent's dialog management. For example, RL can help task-based agents accomplish the user's aim with fewer back-and-forth turns, or in a manner that is perceived as more natural, depending on how the rewards are defined [40, 42, 62]. In the case of open domain chat-chat bots, reinforcement learning has been used to choose responses that result in longer engagement with the agent and positive user sentiment during dialogs [58]. RL refinement of dialog management can be accomplished as a part of end-to-end models, or as a separate dialog management component [25].

A key distinction in RL is between online and offline learning. With online learning, the agent is able to interact with a simulated or actual environment to directly observe the impact of their actions on the state and reward collected by the agent [63]. This is partly attributable to the high-profile success of RL in playing many common games, where thousands upon thousands of iterations of the game can be simulated. In the case of dialog agents, this is not always possible, and learning in real time with actual users would take too long when the RL model is in the early stages of training and makes many errors. Many RL algorithms are able to learn offline from data generated by some other process. Offline learning with an existing data set can be used to train or pre-train RL models before deploying them into an actual environment, where they can continue learning over time [63, 65]. When the generating policy for a data set is not known, it can introduce methodological challenges for offline learning, which is an open research area [31, 65]. If the generating process *is* known, it can simplify the RL approach substantially [65].

## 2.4 Conversational agents in health

Conversational agents in healthcare are often applied in mental health settings, and are predominantly finite-state based or frame-based [38, 49], as opposed to AI-based or data-driven. The continued focus on rule-based and scripted agents is partly because of a low tolerance for error in the health domain [16, 35]. With scripted agents, the designer knows exactly how the agent will respond to a given input from the user. With more dynamic, data-driven approaches, the models are probabilistic, and because the responses can be more variable the designer has much less control over what the agent might say to a user, which is not a desirable risk if delivering health-related advice. In addition, because of HIPAA and other data privacy protections, health-related data sets are rarely made openly available for researcher use, and there is therefore a lack of publicly available dialog corpora in health domains [38, 59].

Advances in data-driven and ML approaches have been used primarily for Natural Language Understanding (NLU) in the health domain. Some examples take advantage of commercial services and platforms (e.g. Google Home) [10] or preexisting NLU services [24, 68], while others develop their own ML models [30, 51]. These systems use ML to parse user input,



classify the user's intent, and identify key entities, but do not use ML to manage or generate responses, and are typically classified as frame-based systems. Other researchers have incorporated ML into chatbots for personalization and clustering users [22, 36] to analyze personal health data for insights or trends [37, 48, 77], or to analyze user engagement [29]. Another related body of research seeks to analyze speech and verbal patterns for diagnostic purposes, for example detecting dementia [64]. The use of data-driven ML approaches for *dialog management* has been less thoroughly explored in a health context. In one notable example, Yasavur and colleagues applied reinforcement learning for dialog management for brief virtual counseling interventions [76], however they did not compare their system with other dialog management approaches.

### 3 MICRO-COACHING

Prior work has shown that self-monitoring is burdensome which can adversely impact engagement [15, 56]. In addition, for chatbots, brief dialogs are preferable for sustained engagement [37, 48]. In the context of health coaching and goal setting, it is well documented that individuals pursuing health goals may not always be sure of their progress in achieving these goals and are eager for feedback and for personalized and contextually appropriate suggestions to help meet their goals [47, 53].

These findings helped us to formulate several design needs for micro-coaching dialog systems (Figure 2). First, the system needs to be able to **automatically assess whether the user is on track to achieve their goal** with a planned meal. The assessment must be automatic in order to provide timely, in-the-moment support. Second, the system must **offer feedback** to the user based on the goal assessment. This could be positive reinforcement if the user is on track, or an acknowledgement and explanation if they are not. Third, if the user is not on track, the system must **offer suggestions** for how to modify their plan to better align with the goal. These suggestions should be personalized to an individual's preferences, and to alternate options available to them. Throughout all three phases of support, an overarching design need is for conversations to be as **brief and targeted** as possible.

The three design needs can be met with three distinct phases of a micro-coaching conversation; implementing each incurs their own potential complexities and nuances. The remainder of this paper focuses on the first need — the ability to automatically assess whether an individual's planned meal is likely consistent with their nutrition goals. This step is a prerequisite to enable the subsequent steps of offering feedback and suggestions, and itself presents considerable complexity.

### 4 CHARACTERIZING EXPERT APPROACHES TO MICRO-COACHING DIALOGS

To explore how expert coaches approach asking follow-up questions about meals, we conducted a qualitative study with health coaches. In particular, we wanted to know *types of questions* coaches would ask their clients about specific meals in order to assess whether a meal is consistent with a nutrition goal.



## 4.1 Methods

Health coaches, who were Certified Diabetes Care and Education Specialists (CDCESs), were recruited from professional networks to participate in the study.

First, CDCESs joined for an interview where we asked them how they would interact with clients in a hypothetical scenario when they were always available in real time to discuss their clients' planned meals. In addition, each coach completed a survey that prompted them to list the questions they would ask a hypothetical client about their meal if the only information the coach had available was a brief text description of the meal, and the nutrition goal the client is working on. In each survey, the prompt was repeated for 10 meals across 5 nutrition goals, and coaches were asked to list 3 to 5 questions per meal. We inductively categorized the yielded set of questions listed for each meal/goal pair in the survey to find patterns and groupings.

After completing the survey on their own time, participants returned for a second interview to discuss their specific responses, as well as to member-check findings.

## 4.2 Results

Two CDCESs participated, completing surveys for a total of 20 meals covering 10 distinct nutrition goals and generating 60 questions.

We found that there was a very limited set of question types across all of the meal-goal pairs. At the highest level of distinction, some questions sought to *search* by asking individuals to list additional food items not already mentioned, while other questions sought to *drill-down* on the details of food items that had already been mentioned. As shown in Table 1, the four main question types were “what else?”, “what kind?”, “how much?”, and “how was it prepared?”.

Within the question types, there were some variations. Some questions apply generically to the entire meal (e.g., “What else will you have with your meal?”) while other question reference specific components of the meal (e.g., “What else will you put in your *burrito*?”). In addition, meal-specific questions sometimes referenced *sub-components* of a meal that were not explicitly stated in the meal description, for example asking about the amount of bread in “a ham sandwich.”

Considering which questions were applicable to which goals, we found that *search* questions were applicable across all goals. In contrast, *drill-down* questions were applicable to some goals and not others. For example, “How much?” questions were applicable to quantitative goals, while “What kind?” questions were more applicable to qualitative goals. In addition, some of the questions took different forms in the context of different goals. For example, “What kind?” questions might be asking about the fat content of yogurt (e.g., 0%, 2% or full fat) for a goal about lean proteins, while asking if the yogurt is plain or flavored for a goal about added sugars.

Using the results of this study as a guide, we designed four different chatbots that utilize different approaches to NLU and dialog management. The first is a fully scripted, while the



other three utilized an AI system for NLU, called FoodNLU. The AI-based chatbots utilize different approaches for dialog management: rule-based, data-driven with Reinforcement Learning, and random as a control condition. In the following sections, we describe the design of these different chatbots and results of intrinsic evaluation studies for the underlying components of the chatbots.

## 5 A SCRIPTED, FINITE STATE-BASED CHATBOT

The first chatbot condition was a fully-scripted, finite state-based chatbot. Using the question types identified from the qualitative study with CDCESs, the scripted chatbot always asks the same questions in the same order for a given goal; the questions differed between nutritional goals but not between meals. The scripted chatbot did not include any NLP components, and always asked the same questions regardless of how the user responded, or the details of their meal, so conversations would always be exactly the same length. The complete scripted dialogs are included in Table 10.

## 6 AI-BASED CHATBOTS

In order more intelligently respond to user's descriptions of their meals with follow-up questions, a chatbot first needs to be able to “understand” the nutrition content of the meal being described. In the next section, we describe a system for Natural Language Understanding (NLU) of meal-related dialogs, called FoodNLU, based on the findings from the study with CDCESs. This system performs NLU of user utterances, and also generates a set of goal- and meal-dependent follow-up questions. We then describe chatbots with multiple dialog management approaches to choosing between these possible responses: rule-based, data-driven, and random.

### 6.1 Knowledge-based natural language understanding (FoodNLU)

In this section, we describe the design and evaluation of the natural language understanding system, called FoodNLU. A visual overview of the pipeline is presented in Figure 3

**6.1.1 Overview of FoodNLU.**—First, to parse food items from natural language descriptions of meals, we utilized *Nutritionix*, a commercial solution for named entity recognition (NER) of food items [78]. Nutritionix has been used as a component of other natural language food projects [50], and can handle common misspellings as well as brand names. For many multi-component foods, Nutritionix includes a sub-recipe of component ingredients. For example, “ham sandwich” has the components “ham,” and “bread,” which enables asking questions about implicit sub-components of the meal. In addition, we extended existing open source code to identify quantities [27].

In order to both determine whether food items were consistent with a given goal, as well as to determine which questions would be applicable to which food items, we incorporated food types and categories from an existing and widely used food ontology, FoodOn [19]. For example, considering the question “What else will you put in your *<food\_item>*?”, some foods are likely to be containers for other foods, like sandwiches or burritos. In FoodOn,



these types of foods are listed as “multi-component food items,” which can be used as a heuristic to determine which food items the question is applicable to.

These attributes also enable the system determine when a meal is or is not consistent with a goal. For example, for the goal “Choose lean proteins,” attributes indicating which foods are proteins, and which proteins are lean or fatty, can be used to determine when all proteins have been clarified to be either fatty or lean, and the stop criteria are met.

In the last step, the system considers the question types relevant to the goal and the attributes present in the user’s meal description to generate a set of possible follow questions.

**6.1.2 Implementation of FoodNLU with a sample of nutrition goals.**—Because the applicable question types vary for different goals, it was important to consider multiple different nutrition goals when designing and evaluating the system. Specific nutrition goals can vary for different individuals, but there are many themes and similarities across them. We chose three nutrition goals to examine as case studies. Candidate goals were compiled from an existing knowledge base of diabetes-focused health goals [14].

Goals were chosen to give coverage across key attributes, and were intended to be a reasonable level of difficulty for most individuals with diabetes, both in terms of how often individuals achieve each goal, as well as how accurate individuals are in self-assessing goal attainment. From an analysis of goal achievement in a prior data set with over 3,000 meals, we selected a set of 3 goals, presented in Table 2.

For each of the three goals, we determined the set of potentially relevant follow-up questions based on the results of Study 1. See Table 3 for a summary of which question types apply to each goal.

In addition, we wrote a set of rule-based stop criteria based on the logic underlying each of the three goals and the attributes of foods in the meal. The stop criteria indicate when there is enough information to say whether a given meal is likely consistent with a nutrition goal. The initial version of all stop criteria are presented in Appendix Table 11

**6.1.3 Intrinsic evaluation of FoodNLU.**—To evaluate FoodNLU, we focus specifically on evaluating the stop criteria. Because these criteria are the cumulation of the steps in the system (Figure 3), if the stop criteria perform well, it suggests that the upstream components are reasonably performant as well. In order to evaluate the stop criteria we used crowdsourcing to create a set of dialogs regarding a diverse set of meals.

**Crowdsourcing dialogs.:** Each dialog started by presenting crowd workers with a photo of a “seed” meal to describe. Meal images were drawn from prior self-tracking studies. We selected 10 meals at random, balanced on the user, the type of meal (e.g., breakfast, lunch, or dinner) and the length of the user-entered meal description. Each image was manually reviewed to ensure the food item(s) were clearly visible. Based on the image, and user-entered description, a member of the research team wrote an ingredient list, plainly listing the names of the food items in the photo.



To create the dialogs, we posted human intelligence tasks (HITs) to Amazon's Mechanical Turk (mTurk) platform. Each HIT included the seed meal image and ingredients, which were rendered as a photo to prevent copy-pasting verbatim, and a text-message conversation history between a fictitious health coach and their client (Figure 12). Each crowd worker was asked to review the conversation history and the meal image/ingredients, and answer the question posed by the health coach. Each meal was used as a seed for 3 dialogs per goal, for a total of 90 dialogs (30 per goal). Each dialog continued until it was clear that there was enough information to determine whether the described meal achieved the goal, according to the stop criteria defined in Table 11

**Registered Dietitian surveys.:** For each of the 3 goals, we selected 5 dialogs where the stop criteria were met, indicating that there *was* enough information and the conversation could end, as well as 5 dialogs where the stop criteria had not been met (Figure 4). The dialogs were balanced on the number of turns to prevent any potential confounding effects of conversation length.

In a Qualtrics survey, Registered Dietitians (RDs;  $n=2$ ) assessed whether they thought there was enough information to determine whether the meal the individual was describing would likely meet their nutrition goal, or not for each of the 30 dialogs. If there was enough information, RDs also labeled whether the goal was met or not, and if there was not enough information, indicated what necessary information was missing.

We calculated inter-rater agreement with Cohen's Kappa statistic. After adding their initial labels, disagreeing items were discussed, and RDs had the option to change their labels. We calculated both the inter-rater agreement and accuracy of the FoodNLU's determinations with those of the RDs.

In addition to inter-rater reliability and accuracy, we performed a qualitative *error analysis* to better understand the cause of situations where the system's predictions were incorrect. For each of the dialogs where one of the RDs disagreed with the prediction of the rule-based stop criteria, we categorized the reason for the disagreement and tabulated the frequency of each type of error.

**Results.:** Interrater agreement between the two RDs was initially only moderate ( $\kappa = 0.46$ ). Most of the disagreements were due to differing definitions of lean proteins between the two RDs. After clarifying the rubric for the 3 goals, RDs adjusted some of their initial labels, resulting in a substantially improved inter-rater agreement score ( $\kappa = 0.87$ ).

Considering the agreement between FoodNLU and the RDs, the average inter-rater agreement score indicated substantial agreement about whether there was enough information in the dialog to determine if the goal would be achieved ( $\kappa = 0.67$ ). Considering the overall accuracy of predictions from FoodNLU, the terminal states were accurate 83% of the time, and accuracy decreased as the goals increased in complexity (Table 4).

When there was enough information for the system to make a prediction about whether the meal was consistent with the goal, those labels were 81.8% accurate with RD labels. These evaluation results suggest that the rule-based FoodNLU system is reasonably performant.



Results of the error analysis are presented in Appendix Table 12. The most common reason for error was that the dialog did not include a drill-down question asking about a food that likely contained a large quantity of other food items, like a smoothie. A handful of additional errors were due to disagreements about food item attributes with the labels from the FoodOn ontology, or errors with the Nutritionix named entity recognition system. The results suggest that there was not a single point of failure responsible for all of the errors.

## 6.2 Rule-based dialog management

While FoodNLU represents what a user has said, and the possible follow-up questions the chatbot coach could ask, it does not include logic for *which* question to ask next. The food items and attributes identified by FoodNLU, though, can be used as a straightforward set of features for rule-based dialog management. For example, for the lean proteins goal, if one of the food items identified is either lean or fatty protein (e.g., skim or whole for milk) then the system should ask “what kind” in an attempt to disambiguate whether the protein is lean or fatty.

---

### Algorithm 1 Rule-based logic for dialog management

---

#### Repeat

If there is a goal-related food item to ask a drill-down question about, then ask that question (e.g., if there is an ambiguously fatty protein for the lean proteins goal, then ask “what kind of <ambiguously fatty protein>?” or if there is a carbohydrate for the carbohydrate portions goal, then ask “how much <carbohydrate>?”)

Else if there is a “container food”, then ask “what else in <container food>?”

Else ask “what else?”

until stop criteria is met and at least one search question has been asked

---

Informed by the qualitative findings from health coaches, we built on FoodNLU to design a simple, rule-based algorithm to choose the next action (Algorithm 1). To prevent premature closure of conversations, the rule-based system had a constraint to always ask one search question before the dialog was considered complete. For instance, for a goal about carbohydrate portions, if two high-carb food items were eaten, but only one was mentioned in the initial description, the conversation might end prematurely without searching for unmentioned food items. This constraint ensures there are at least some amount of search questions in each dialog.

## 6.3 Data-driven dialog management with Reinforcement Learning (RL)

The same food items and attributes identified by FoodNLU system can also be used as features for an ML-based dialog management system. Reinforcement Learning (RL) is a machine learning approach that is well suited to the task of learning to choose the best action in a given circumstance [63]. However, data-driven approaches like RL require a corpus of examples to learn from.

Without an existing data set for micro-coaching dialogs, we used *crowdsourcing* to create a corpus of meal-related dialogs. Because of the potentially high costs and wasted resources



of paying crowd workers for multiple iterations of *online* learning, we created a corpus of dialogs for *offline* learning.

In this section, we introduce the RL algorithm used in this analysis, q-learning, followed by a description of the state space and rewards. Then, we present two validation studies, first with simulated data, and then with a new, crowdsourced data set of meal dialogs.

**6.3.1 RL algorithm: Q-learning.**—Q-learning [63, 70, 71] is an off-policy algorithm that aims to learn the action-value function  $Q(s, a)$ , which estimates the value of taking a particular action  $a \in A$  while in a discrete state  $s \in S$ . The value is the reward  $r \in R$  gained from moving to the next state  $s'$  plus the sum of rewards that could be accumulated from  $s'$  onwards, reduced by a discount factor  $\gamma \in [0, 1]$ . By observing the reward when moving from  $s$  to  $s'$ , the q values are updated iteratively following a temporal distance learning algorithm (Algorithm 2). Through these iterations, the learned action-value function  $Q$  approximates  $q^*$ , which is the optimal action-value function.

---

**Algorithm 2** Q-learning. Adapted from Sutton & Barto [63]

---

**Initialize**  $Q(s, a) = 0$  for all  $s \in S, a \in A$

**Repeat** (for each dialog)

**Initialize**  $S$  from the initial meal description

**Repeat** (for each dialog turn)

        Choose  $A$  following a policy (e.g. random or  $\epsilon$  greedy)

        Take action  $A$ , observe  $R$  and  $S'$

$Q(S, A) \leftarrow Q(S, A) + \alpha[R + \gamma \max_a Q(S', a) - Q(S, A)]$

$S \leftarrow S'$

$\alpha \leftarrow \alpha - \omega$

**until**  $S$  is terminal

---

The hyperparameters for q-learning are the learning rate  $\alpha \in [0, 1]$ , which controls the step size of each q-value update, the learning rate decay  $\omega \in [0, 0.01]$ , which gradually decreases the learning rate  $\alpha$  over the course of training, and the discount-rate  $\gamma \in [0, 1]$ , which discounts the value of future rewards thereby increasing the influence of immediate rewards.

Once q-values have been learned offline from an existing data set, the algorithm can be applied to prospectively collected dialogs, following a policy based on the pretrained Q-values. In a given state  $S$ , the best action according to the Q-values can be attained from  $\max_a Q(S, a)$ . However, always greedily following the best action can pigeonhole the algorithm to following a particular path, and will not be able to continue learning about other paths. Therefore, the greedy algorithm can be modified so that at each turn a random action is taken with probability  $\epsilon \in [0, 1]$ .

**6.3.2 State space and reward function.**—Two key considerations in applying q-learning to micro-coaching dialogs are representing the state space  $\mathcal{S}$  and the reward function.

Considering  $\mathcal{S}$ , the larger the state space is, the more observations that are necessary for the algorithm to converge. Therefore, smaller state spaces are desirable for a proof of concept. The same attributes used for the rule-based system can be used as features to represent the



state of the conversation. For instance, the number of food items identified by FoodNLU may be informative, as would the presence or absence of certain types of food items. For example, for the goal to “Choose lean proteins” the presence or absence of any proteins in the meal would be a relevant feature to determine which questions would be informative. Based on these features, we designed minimalist, discrete state spaces for each of the three nutrition goals used as case studies in this analysis (Table 5).

Considering the reward function, it was of primary importance to reward reaching a terminal state, meaning a state where the stop criteria are fulfilled, with as few conversational turns as possible. The highest reward ( $r = 10$ ) was given for reaching a terminal state. To reward questions that resulted in additional information, for example, finding additional food items or identifying a goal-relevant food item, a smaller reward was given ( $r = 3$ ). To incentivize short conversations, a small penalty ( $r = -1$ ) was given for questions that resulted in no changes to the state representation, suggesting that they were non-informative.

**6.3.3 Creating a corpus for offline learning.**—In order to train a model with actual data, we need data to learn from. To create a corpus for training the RL model, we used crowdsourcing following a similar process to the intrinsic evaluation of FoodNLU. For the crowdsourced corpus, the dialog management was handled by a *random* policy — the coaches’ follow-up question was chosen at random from the possible question type.

We selected 25 meal images and ingredient lists to serve as seed meals for crowdsourced dialogs. Each meal was the seed for 4 dialogs per goal, for a total of 300 dialogs (100 per goal). Each dialog continued for a total of 10 turns. The resulting corpus included 300 dialogs and 3,000 total conversational turns. The corpus is available for other researchers to use on GitHub in a JSON format similar to other open dialog data sets [60]. Descriptive statistics of the corpus are presented in Appendix B.

**6.3.4 Validation experiments with simulated data.**—To validate this q-learning approach as a proof of concept, we first conducted an experiment with simulated data. The simulation was also designed so that certain questions would be more informative in certain states. For example, with the lean protein goal, asking “what kind of *<ambiguously fatty protein>*?” would find a non-ambiguous lean or fatty protein and receive a high reward with an 80% probability if there were proteins present in the meal.

**Methods.:** We iteratively trained the q-learning model with the simulated data and a random policy for hundreds of episodes (one training episode corresponded to a complete dialog from beginning to end). We tuned the hyperparameters by examining the changes in q value for convergence and the consistency in performance across multiple rounds of training.

To examine the performance of trained q-values, we then simulated the prospective, *online* collection of new dialogs between two policies: 1) a policy that greedily follows the action with the highest q-value, and 2) the same random policy that was used for training. We compared the average length of dialogs between the greedy-*q* and random policies, as well as the average reward attained per episode.



**Results.:** As shown in Table 6, the tuned q-learning algorithm was able to learn a policy that resulted in shorter conversations, compared to a random policy.

Examining the change in q-values over the course of training suggested that the algorithm was correctly unpacking the signal in the simulated data, and finding different actions to be more valuable in different states. A side-by-side comparison of the q-value history for two different states and the “lean protein” goal is presented in Figure 5, which shows that the most valuable actions (the actions with the highest q-values) were correctly identified;. In addition, the q-value histories show that the q-values begin to find signal and converge after 25 to 50 episodes, suggesting that a corpus of 100 dialogs should be sufficient for training.

**6.3.5 Validation experiments with crowdsourced data.**—After validating the q-learning approach with simulated data, we trained a q-learning agent for each of the three goals using the crowdsourced dialogs.

**Methods.:** The training data set was the corpus of 100 dialogs per goal. Following similar methods to the simulated data, we trained 3 separate q-learning models, one for each of the 3 goals, for 150–200 episodes each. In addition, we examined the dataset’s coverage of the state space; sufficient coverage of the state space is an assumption needed for q-learning to converge [63, 70].

**Results.:** Examining the changes in q-values over the course of training demonstrated similar patterns to those found with simulated data. As seen in Figure 6, for the goal “Choose lean proteins,” a policy based on q-values correctly learned to ask “what kind” questions when ambiguously fatty proteins are present and a number of other foods had been identified, but instead asked “what else” to continue searching if no proteins have been mentioned.

Importantly, the RL agent did not always favor exploiting possible drill-down questions: If there were few food items present (1 or 2) the agent would continue to *search* by valuing “what else” or “what else in” questions (Figure 7).

Results for the other two goals followed a similar pattern, and are included in Appendix C, along with the results of the state-space coverage evaluation.

## 6.4 Random question dialog management

As a control condition to compare rule-based and data-driven dialog management strategies against, the next question can be chosen randomly from the possible responses generated by FoodNLU. This random policy is how the dialogs for training RL were generated, and represents a minimally intelligent dialog management policy that the other two AI-based chatbots should be able to outperform.



## 7 COMPARISON OF MULTIPLE CHATBOT APPROACHES FOR MICRO-COACHING

In the prior section, we described four chatbots for micro-coaching dialogs. One was simple and fully scripted, while the other three incorporated some elements of AI. The three AI-based chatbots used FoodNLU to parse user input and generate possible follow-up questions, but employed different dialog management approaches: rule-based, data-driven with RL, and random as a control. In this section, we describe an evaluation study comparing the four chatbots in dialog length, and perceived quality and user experience. An example of dialogs generated using each of the 4 chatbots is presented in Figure 8

### 7.1 Methods

**7.1.1 Crowdsourced dialog test set.**—With 10 meal images that were not a part of the training set for RL, we crowdsourced 2 dialogs per meal, per goal for each of the 4 conditions, resulting in a total of 240 evaluation dialogs.

**7.1.2 Dialog length.**—To examine the length of conversations, we compared the average number of dialog turns across each of the conditions, and tested for significance with pairwise Wilcoxon tests between the RL condition and three comparators, using a Bonferroni correction for multiple hypothesis tests.

**7.1.3 Perceived dialog quality.**—For each of the four chatbots, we solicited crowd worker feedback on the quality of coaching dialogs with is design. Pairwise comparison a *pairwise comparison* commonly used to compare multiple entities on some subjective property, for example in preference elicitation and decision-making research [26, 55]. Pairwise comparisons have the advantage of avoiding floor and ceiling effects [11] that can occur when using Likert scales to measure subjective attributes, and therefore ensure the ability to understand which chatbots were preferred to others. Crowd workers were asked to consider the overall quality of the coach’s *question-asking strategy* (following [41]), as well as the *naturalness* and *coherence* of messages from the coach (following [42]).

With 10 dialogs per goal and 4 conditions, there were a total of 60 unique comparisons per goal. Crowd workers completed surveys on mTurk with 30 randomly-selected comparisons. For each comparison, participants were shown two dialogs, and asked which of the two was superior in each of the 3 quality constructs (Figure 9).

Participants were recruited from mTurk, and needed to be United States residents with a 90% approval rate to be eligible. Participants were compensated \$5 for completing the survey.

We calculated how often each of the four chatbots was rated as better than another in pairwise comparison, for each of the three quality constructs, resulting in an overall “win percentage” for each chatbot. We considered the win percentage for each of the 3 quality constructs (strategy, naturalness, coherence) individually, as well as a composite quality score from averaging all three together.



In addition, to compare quality assessment based on the length of dialogs, we examined how often the winning dialog was longer (more turns), or shorter (fewer turns), or deemed it a tie if the dialogs were the same length.

**7.1.4 Perceived user experience.**—A separate set of participants was recruited to evaluate the perceived user experience of interacting with the coach using the Subjective Assessment of Speech System Interfaces measure (SASSI; [28]), with a *between subjects* design. Participants reviewed 10 dialogs from the same chatbot, related to the same goal, and then were asked to consider the experience of the user and complete the full SASSI questionnaire. Each participant was compensated with \$8 for completing the survey through the mTurk platform.

To test for differences in survey responses, scores were compared between the four chatbot conditions. Because survey measures are ordinal, values between the conditions were compared with the Kruskal-Wallis test, a non-parametric version of a one-way ANOVA.

**7.1.5 Stop criteria for the scripted chatbot.**—Unlike the other 3 chatbots, which continued until reaching the stop criteria, the scripted chatbot always asked the same questions, regardless of the responses offered by the user. This meant that the scripted dialogs may not contain sufficient information to determine if the described meal is consistent with a goal. To quantify this discrepancy, I applied the same stop criteria to the scripted dialogs, to examine how often the scripted dialogs reach the stop criteria. If a dialog does not reach the stop criteria, there is likely insufficient information to determine if the meal is likely consistent with the goal.

## 7.2 Results

**7.2.1 Dialog length.**—As shown in Figure 10, conversations with the RL chatbot were consistently shorter to meet their stop criteria. Conversations were an average of 3.56 turns long in the RL condition, compared with 4.18 turns in the rule-based condition, and 5.75 turns in the random condition. Scripted conversations were predictably an average of 4.33 turns long. A breakdown of conversation length across the 3 goals is presented in Table 7. The more complex goal “Make ½ my meal fruits and/or non-starchy vegetables” generally had much longer conversations on average than the other two goals. RL showed the most improvement over the random baseline for the goal “Eat no more than 2 portions of carbs (30g)”

**7.2.2 Perceived dialog quality.**—15 participants completed the pairwise quality comparison survey. The win percentage results are presented in Table 8. The higher quality condition varied by goal. The scripted condition won most often in head-to-head quality comparisons, especially for goal #1, “choose lean proteins,” and goal #3 “1/2 fruits and non-starchy vegetables. For goal #2, “no more than 2 portions carbs,” the RL chatbot was the most natural and coherent, while the rule-based chatbot had the better question-asking strategy.

Considering the length of conversations (Table 9), shorter dialogs were considered natural more often, while longer dialogs were considered to have a better question-asking strategy.



**7.2.3 Perceived user experience.**—When examining differences in perceived user experience through the SASSI, which has a minimum score of 1 and a maximum of 5, 36 individuals completed the survey, and no statistically significant differences were detected (Figure 11)

**7.2.4 Stop criteria for the scripted chatbot.**—Dialogs from the scripted chatbot reached the stop criteria only 65% percent of the time. Dialogs for the other chatbots reached their stop criteria 100% of the time.

## 8 DISCUSSION

In this research, we examined multiple approaches to design a conversational coaching intervention. Informed by prior research, we proposed a set of design needs for *micro-coaching* dialogs — brief conversations to provide support for planning specific meals. Enabling such an approach required the ability to automatically determine whether an individual is likely to achieve their chosen goal, based on the description of their meal, which was the focus area of this research.

Specifically, we designed and evaluated a knowledge-based system that processes user utterances describing their meals and generates a set of possible follow-up questions. In addition, we compared multiple approaches to dialog management, including a simple fully-scripted approach, approaches that utilize different types of AI, rule-based, data-driven, and an approach that selected questions at random. Below we discuss the main findings of this work and their implication for future research in micro-coaching chatbots.

### 8.1 Comparative advantages of rule-based vs. data-driven dialog management

In the evaluation study, we compared multiple approaches to dialog management for micro-coaching dialogs, including scripted, rule-based, and data-driven approaches. The scripted chatbot always asked the same goal-relevant questions, regardless of the meal and responses. The rule-based chatbot took advantage of the goal-relevant food features identified with the expert system to determine the next question with a small set of rules. The data-driven system used the same features as the rule-based system, but instead selected the next question based on a reinforcement learning (RL) algorithm. The RL algorithm, q-learning, was trained on a sample corpus of 300 dialogs created through crowdsourcing and learned which questions to ask to most quickly learn the goal-relevant aspects of the meal.

We evaluated these four different chatbots comparing the length of conversations; we also assessed individuals' perceptions of the *strategy*, *coherence*, *naturalness*, and *usability* of the different chatbots, by asking crowd workers to rate the conversations and complete a usability assessment in a survey study.

The results of the evaluation study suggest that each chatbot approach had distinct strengths. Principally, the RL chatbot succeeded in its intended purpose of completing conversation with the fewest number of questions asked. However, performance of the quality assessment was mixed across the four chatbots, which we discuss below.



We expected the RL chatbot to be perceived as having a superior question-asking *strategy* because it accomplished the aim of shorter conversations on average. However, the scripted and rule-based chatbots were rated as having a better strategy, depending on the goal; in contrast, the RL chatbot was never rated higher on its strategy. In addition, considering all of the non-scripted dialogs (which varied in length) raters consistently assessed longer dialogs as more strategic compared to shorter dialogs. There are a number of potential explanations for this unexpected result. First, it is possible that longer dialogs may have allowed for a more gradual exploration of different properties of each meal, thus creating an appearance of a better strategy. Furthermore, individuals may have perceived the order of questions with the RL chatbot to be less intuitive compared to rule-based or scripted strategies. While RL-based dialog management did result in more efficient conversations, it may not have aligned with participants' intuition on successful question-asking strategies. This aligns with research in conversational symptom checkers, which found that individuals dislike when questions are asked in a seemingly random or nonsensical order [66]. This is also consistent with arguments in clinical decision support that models and explanations ought to align with the way humans think about a problem to be adopted and trusted [8].

Considering how *coherent* the dialogs were, we expected the RL or rule-based chatbots to perform well, because the follow-up questions asked would be more specific to food items mentioned by the user. In contrast, the scripted chatbot's responses not connected to the user's previous replies, for example asking about fruit portions if the user said they had not eaten any fruit. However, we found mixed results, with either the RL or scripted chatbots rated as more coherent. It is possible that the consistent pattern followed by the scripted dialogs may have better aligned with rater's expectations, making these dialogs read at times as more coherent than RL-generated ones.

We expected all of the dialogs to perform similarly on *naturalness*, with perhaps scripted dialogs being rated as more natural because there were no chances for small grammatical errors or unnatural phrasing that could occur in the conditions with dynamically generated responses from FoodNLU. We found that performance was mixed across all of the different goals, which was generally in line with our expectations.

Overall, the scripted chatbot performed surprisingly well across all the quality assessments. Importantly, however, while the rule-based and RL chatbots both collected necessary information and reached the stop criteria 100% of the time, the fully-scripted chatbot had several important limitations. First, it only succeeded in collecting information needed to assess goal attainment 65% of the time. In addition, in this evaluation study, participants were exposed to each type of dialog only once; while the scripted dialogs may have felt more natural in their first occurrence, they could be perceived as increasingly repetitive overtime, as compared to the more dynamic chatbots.

These findings have several implications for future design and research in conversational agents in health. First, they suggest that there exist trade-offs between different dialog models and that different models may be more or less appropriate in different circumstances.



First, while the RL model was designed to minimize the length of dialogs, shorter dialogs were not consistently rated as higher in the assessed quality dimensions. This presents a potential tension between the empirical efficiency of a dialog management approach, and its perceived quality. Longer, more comprehensive dialogs, or dialogs where the questions are asked in a more intuitive order may have fewer benefits in the context of efficient dialogs for nutritional micro-coaching but may be advantageous for user experience in other contexts, for example in more general health coaching. Future research can more directly examine the relationship between dialog efficiency and perceived quality in different health contexts. This tension also suggests a particular direction for future research in RL for health chatbots. In this work, RL reward function considered only the length of conversations, but other approaches could incorporate additional components to the reward, for example considering the perceived user quality of resulting conversations in addition to the dialog length [40, 42]. To inform such a reward function, future work could more directly examine the relationship between conversation length and user perceptions of the chatbot, as well as considering the quality ratings from those with more coaching expertise. Alternatively, if the dialog management stays focused on dialog length, another approach could be to incorporate elements of explainable-AI to offer explanations for why the chatbot is asking a particular question [66].

Second, the choice of an approach to dialog modeling may be impacted by practical considerations as well. The RL-based chatbot in this study resulted in shorter conversations; however, it did require the use of crowdsourcing to create a dialog corpus to learn from. While the resources for such a corpus were relatively modest (about \$200 per 100 dialogs), the data set was not necessary at all for the rule-based approach. Still, both approaches were relatively simple, considering only a small number of features about the meal in question. To scale up either approach, either a more complex rule-based system to handle more cases, or a more sophisticated RL algorithm, would require additional resources. For the rule-based system, expert input would be needed to craft the additional rules and features in a more complex system. More complex rule-based systems, for example for motivational interviewing, can require hundreds or thousands of rules [57], and expert input to create a large number of rules could be more resource intensive than crowdsourcing. In contrast, scaling up the RL algorithm with more features in the state space, or a more sophisticated algorithm may require an incrementally larger corpus to learn from [76], but there is no need for additional expert input. Because these results demonstrate the feasibility of using RL to manage follow-up question asking in dialogs, pursuing more complex RL approaches is a promising vein for future work. RL-based approaches also have the advantage of being able to continue to learn and adapt their approach once deployed [63], whereas a rule-based system would need to be explicitly redesigned and revised [46]. These results are consequential, in part, because little research has compared user perceptions of rule-based vs. data-driven dialog management systems side-by-side.

## 8.2 Alternatives to knowledge-based natural language understanding

In order to design a chatbot that can converse intelligently with users about their meals, we needed to integrate food-related knowledge. To this end, we designed and evaluated FoodNLU, a knowledge-based system for natural language understanding (NLU) of meal-



related conversations. The system incorporated existing tools for named entity recognition (NER) of food items, as well as a food ontology (FoodOn), to tag foods with relevant attributes like their primary macronutrient, and whether they likely contained sub-foods within them. This representation was used to inform both a set of possible follow-up questions about the meal, as well as for a rule-based criteria to assess whether the meal was likely consistent with a nutrition goal.

This system was able to assess when there was sufficient information to determine if a meal was consistent with a goal with more than 80% accuracy, and was also 80% accurate at making predictions about whether meals were consistent with a health goal. These results suggest the feasibility of such an approach, which is generally in line with previous NLU systems for health-related chatbots [10, 24, 38, 49, 68]. However, there are many potential directions to explore to improve the performance of FoodNLU.

Because FoodNLU was designed as a combination of existing resources, the performance of the overall system was limited by its component parts. For instance, the underlying knowledge base, FoodOn [19], could be expanded and refined to handle more types of food. An alternative approach could build on recent advances in data-driven NLP, for example leveraging the success of pre-trained language models like BERT [17] and GPT [6] as a starting point for a model to interpret user utterances and perform named entity recognition. Achieving high performance on food-related dialogs would likely require re-training such models on nutrition, food, and health-related corpora, due to the amount of domain-specific knowledge, similar to BioBERT for biomedical text [39]. Additional work would be necessary to curate or create the necessary corpora for pre-training with coverage of domain-specific terminology.

There are also alternative approaches to meal logging that are not text-based, for example food photo diaries. Researchers have examined photo-based food logs as a lightweight approach to logging, but photos by themselves do not contain the features necessary to assess goal achievement [15, 20]. Considering the difference in performance across the three different goals, the results were not uniform — in particular the goal to “make 1/2 of my meal fruits and/or non-starchy vegetables” was less accurate. Since it is based on the visual plate proportions in the USDA MyPlate guidelines [67], a visual approach may be more successful for this goal. ML can be applied to food photos to detect component food items, or estimate nutrient values through comparison with other meal photos [32, 33, 45, 74]. However, these systems are often inaccurate, or require additional database lookup and confirmation from users, which can increase the burden of logging. In addition, requiring a photo negates the ability to engage in meal *planning*, which the text-based micro-coaching approach facilitates. Once a meal is ready to eat, there’s less that can be done to help support changes in-the-moment. In addition, text-based approaches can connect explanations and feedback back to the exact words people used to describe their own meals, which could facilitate more understanding and learning than the food items detected from a meal image. Future work could directly compare text- and photo-based approaches for lightweight logging as input to micro-coaching support.



### 8.3 Implications and future directions for micro-coaching

This research constitutes an initial step towards enabling a larger proposed vision for micro-coaching dialogs. The results suggest feasibility of AI-based approaches for the first component, assessing the consistency of a planned meal with a nutrition goal. Additional proposed components of micro-coaching include offering feedback based on the goal assessment, as well as support in the form of personalized suggestions to modify the plan.

Feedback was something that participants in all of the prior studies of this thesis expressed a keen interest for. This applied to feedback on achieving particular goals, as well as overall improvements to self-management and health outcomes. Feedback and explanations are also important part of learning [7]. Considering the theoretical foundations of health coaching, feedback helps to establish accountability, as well as an opportunity for education and increasing an individual's nutrition knowledge [52, 73]. The rule-based approach to assessing meal dialogs against goals enables feedback with explanations as well, because the connection between each food item mentioned and the systems assessment is clear. Considering, for example, the goal to choose lean proteins, this would enable the system to explain to an individual that they did achieve their goal by eating "chicken breast without skin," or that they did not because they ate "bacon." Future work could explore additional ways of delivering feedback during micro-coaching conversations, and their impact on motivation and engagement.

Such an approach would also require nutrition knowledge, but of a different form. Specifically, knowledge of what foods go well with each other, how meals could be adjusted to be more consistent with a goal, as well as similar, alternative meals would all be useful. In addition, personalizing suggestions would necessitate a representation of the user's preferences and context. Given these constraints, *conversational recommender systems* may offer a promising direction for future research. Conversational recommender systems are dialog systems that search among alternatives in a database (for example of restaurants or products) taking into account a user's preferences across multiple sessions. Such an approach could be applied to a database of recommendations, and research in meal similarity and ingredient substitution could also be applied in crafting suggestions [27, 45, 75].

While the focus of this research was Specifically on nutrition-related micro-coaching, the results have implications for chatbot design in health, healthcare, and coaching more broadly. A similar approach could be applied for goal assessment in other domains, like physical activity or medication adherence [13, 21]. To apply the approach to another coaching domain would require a knowledge base for the NLU component of the system. In addition, the findings related to user perceptions of dialogs generated by chatbots with different dialog management approaches has implications for researchers and designers interested in applying data-driven approaches like reinforcement learning for dialog management. The tensions between the length, efficiency, and perceived quality has relevance not just for coaching, but also for other areas like conversational symptom checking [66].



## 8.4 Limitations

This work has notable limitations. The examination of micro-coaching considered a sample of only three nutrition goals. While these goals were chosen to be representative of a diverse set of nutrition goals, it's possible that the findings and approach may not generalize to other nutrition goals. In addition, the assessments of quality and user experience come from lay-individuals reviewing complete dialogs from one of four chatbots. However, the perceived user experience from reading a completed dialog may not capture the perceived user experience of directly interacting with a chatbot, and may have limited the ability of the evaluation to detect meaningful differences in user experience. Finally, this study only focused on perceptions of the coaching chatbot, and not on their impact on individuals' behaviors and health.

## 8.5 Conclusion

In this paper, we explored AI-methods for chatbot design, including the application of data-driven, RL-based dialog management, which has rarely been applied in a health context, but may be necessary to enable more intelligent automated coaching interventions. Despite the success of the RL chatbot in enabling shorter conversations, its assessments on dialog quality were mixed. This suggests the need for additional research into ways to combine the efficiency of data-driven approaches with intuitiveness and transparency of scripted approaches to chatbots in health.

## ACKNOWLEDGMENTS

Thank you to David J Albers and Iñigo Urteaga for their invaluable input to the ideation of this work, and without whom this manuscript would not have been possible. This research was funded by the National Institute of Diabetes and Digestive and Kidney Diseases award number R01DK113189 and the National Library of Medicine award number T15LM007079.

## APPENDIX

### A. : SUPPLEMENTARY TABLES AND FIGURES

**Table 10:**

Script for the fully-scripted chatbot

Goal	Question type	Question text
Choose lean proteins	what_else	What else will you have with your meal?
	any(lean protein)	Will you have any lean proteins with your meal, like chicken breast or egg whites?
Eat no more than 2 portions of carbs (30g)	what_else	What else will you have with your meal?
	how_much(carbs)	What portion of carbohydrates like rice, pasta, or bread will you eat? For example, one fist is about one cup
	how_much(fruit)	What amount of fruit will you eat? For example, one fist is about one cup
Make ½ my meal fruits and/or non-starchy vegetables	what_else	What else will you have with your meal?
	how_much(fruit)	What amount of fruit will you eat? For example, one fist is about one cup



Goal	Question type	Question text
	how_much(non-starchy veg)	What amount of non-starchy vegetables will you eat? For example, one fist is about one cup
	how_much(protein)	All fruit and vegetables have amounts?
	how_much(carbs)	What portion of carbohydrates like rice, pasta, or bread will you eat? For example, one fist is about one cup

**Table 11:**

Summary of stop criteria logic for each of the three goals.

Goal	Stop criteria
Choose lean proteins	[any(proteins) and none(ambiguously_fatty_protein)] or [none(proteins) and (n_food_items > 2) and asked_what_else]
Eat no more than 2 portions of carbs (30g)	[any(carbs) and all(has_amount(carbs))] or [none(carbs) and (n_food_items > 2) and asked_what_else]
Make ½ my meal fruits and/or non-starchy vegetables	[all(has_amount(fruit_veg)) and all(has_amount(non_fruit_veg))] or [none(fruit_veg) and (n_food_items > 2) and asked_what_else] or [none(non_fruit_veg) and (n_food_items > 2) and asked_what_else]

**Table 12:**

Error types, examples, and counts from the error analysis of the natural language understanding (NLU) system

Label type	Error type	Examples	Count
Enough information to assess meal/goal achievement	Unasked drill-down question	•Amount of fruit in a smoothie (Carb and Fruit/Veg goals)	4
	Disagreement about food item attribute	•Soy milk (Lean proteins) •Milk (Carbohydrate)	3
	Nutritionix missing sub-recipe	•System does not know that “Chicken noodle soup” contains “chicken” “noodles” or “vegetables”	1
	Assumed amount of food items	•Assumed quantity of carrots and onions would be less than the amount of shrimp, lima beans, and corn already stated (3 cups)	1
Meal/goal achievement	Differing amount estimates	•Is “1 cup of noodles” more or less than 30 grams?	3



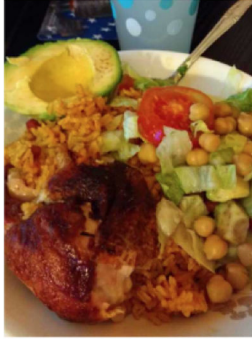
### Conversation History

What are you thinking of having for dinner?

I'm cooking some a roasted chicken thigh, rice, beans, along with a salad and half an avacado.

How much beans will you eat? For example, one fist equals about one cup.

### The meal you're planning to eat



**Ingredients:** Roasted chicken thigh; rice; beans; salad (romaine lettuce, chickpeas, tomato, avocado); half avocado

**How would you answer this question from your health coach as a short text message?**

*Please keep your reply as brief as possible.*

Type how you would answer this question as a short text m

**Submit**

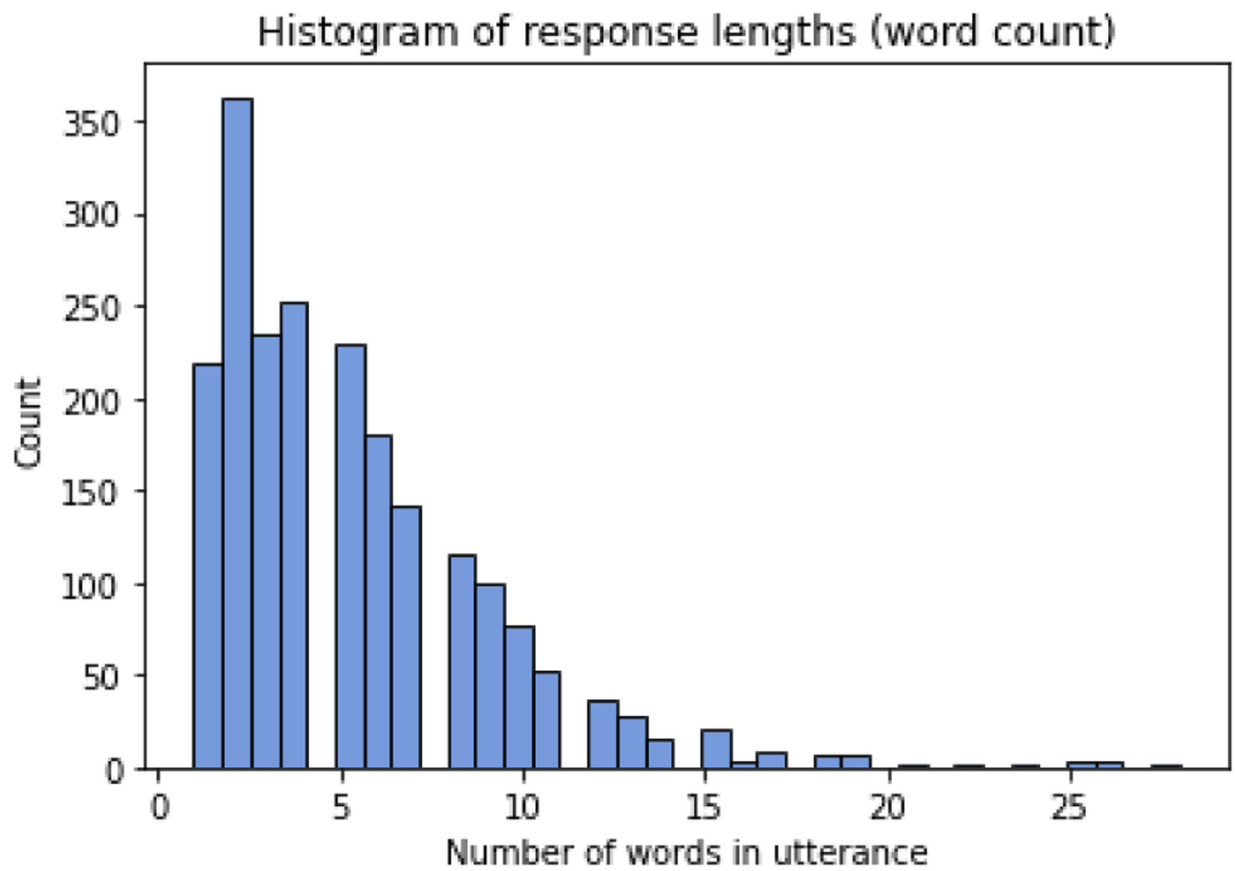
**Figure 12:**  
Example crowdsourcing task to create crowdsourced dialogs.

## B. : CROWDSOURCED CORPUS DESCRIPTIVE STATISTICS

Figure 13 and Figure 14 summarize the length of messages from crowd workers in the corpus, with word count and character counts. There is a diversity of response lengths, and importantly, all of the responses are fewer than 160 characters, suggesting they are a reasonable length for SMS messaging.

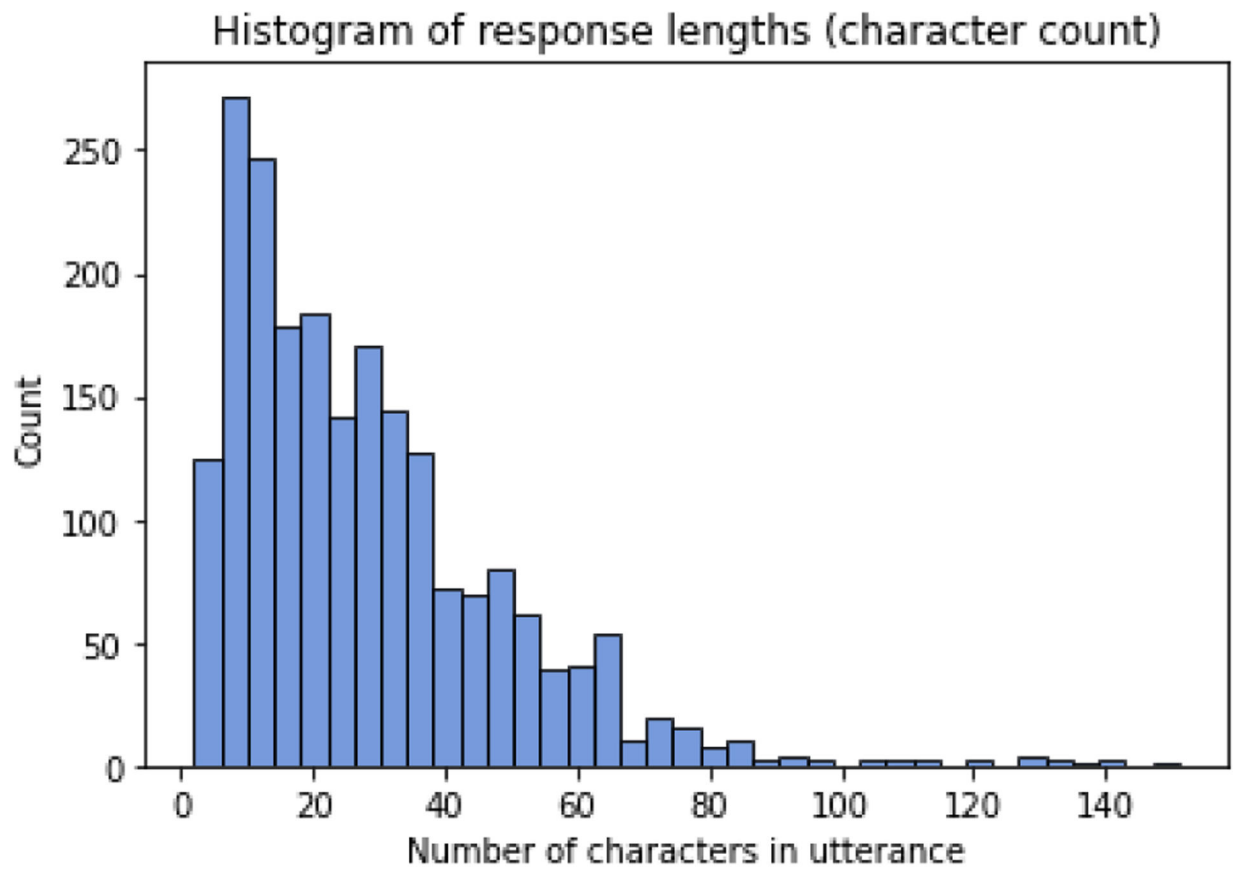
Figure 15 shows how the count of food items parsed by *Nutritionix* increases as conversations increase in length. The number of food items identified increases most after the first turn, and then gradually increases in subsequent turns.





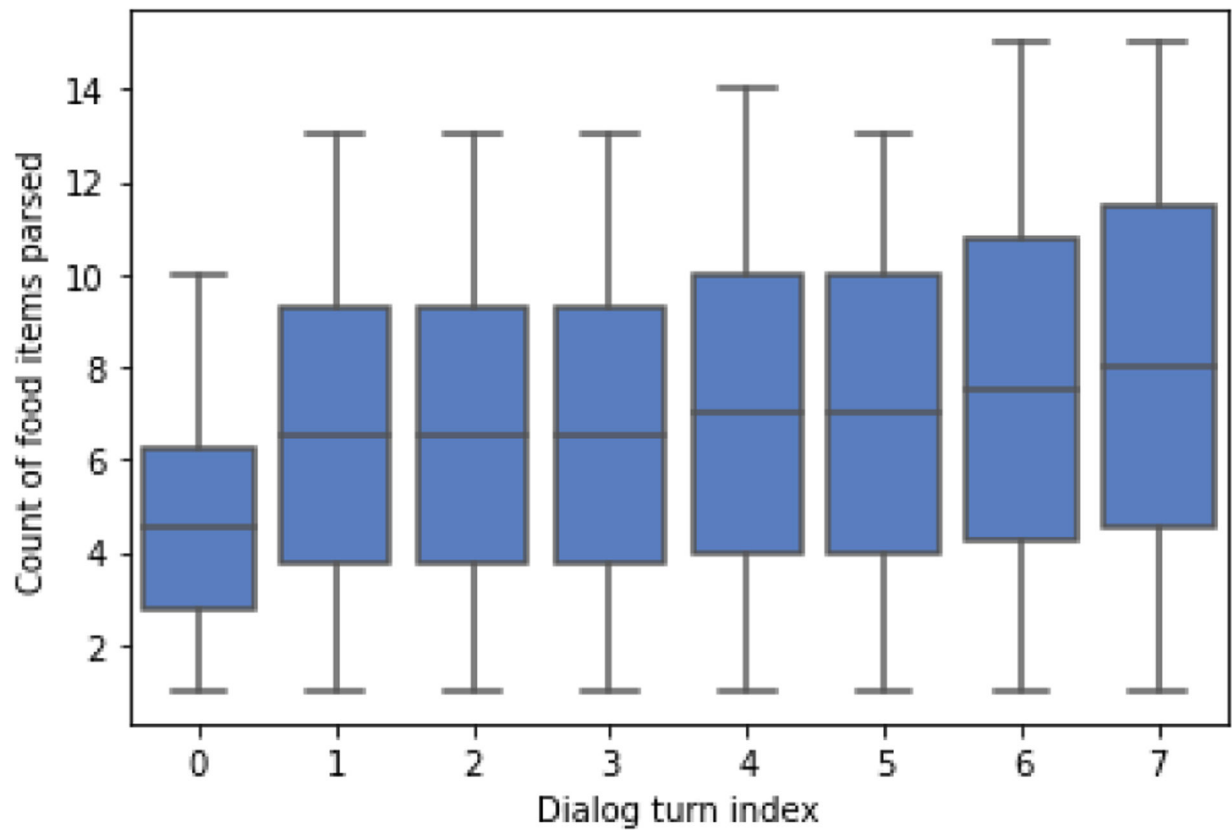
**Figure 13:**  
Histogram of crowd worker response lengths (word count)





**Figure 14:**  
Histogram of crowd worker response lengths (character count)





**Figure 15:**

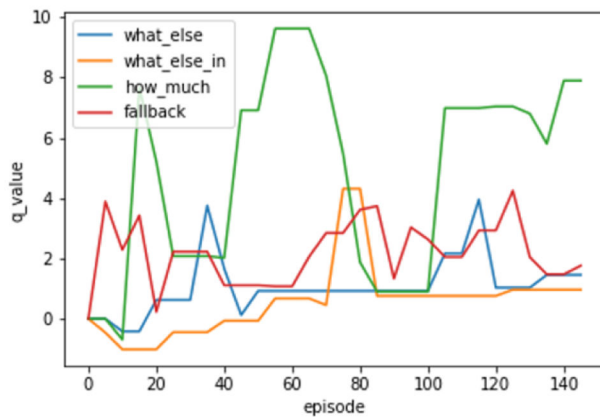
Box-and-whisker plot of the cumulative count of food items parsed by the depth of the conversation.

### C. : SUPPLEMENTARY Q-LEARNING RESULTS

For the second goal (Figure 16), “Eat no more than 2 portions of carbs in each meal (30g)”, we similarly found that the RL agent would correctly favor asking “how much” questions to quantify the carbohydrate content of the meal when at least one carbohydrate was present, but would instead search by asking “what else in” questions when no carbohydrates had been mentioned yet.



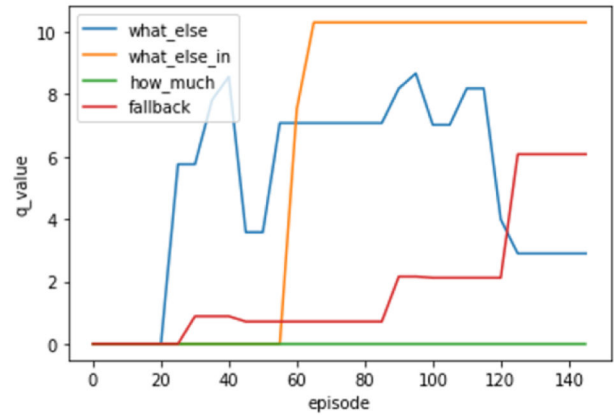
**History of q-values when “how much” is a logical action; at least one carbohydrate has been mentioned with no quantity**



State:

```
n_food_items = 4
any_carbs = 1
amount_carbs_all = 0
```

**History of q-values when “what else in” is a logical action; there is only one food item present and it is not a carbohydrate**



State:

```
n_food_items = 1
any_carbs = 0
amount_carbs_all = 0
```

**Figure 16:**

Change in q-values over 150 training episodes for two different states, for the goal “Eat no more than 2 portions of carbs in each meal (30g).” Higher q-values suggest an action will be more valuable in a given state.

For the third goal (Figure 17), “Make 1/2 of my meal fruit and/or non-starchy vegetables”, we found a similar pattern: the RL agent learned to prioritize asking for amounts of fruits and non-starchy vegetables when at least one had been mentioned without an amount. If amounts were present for all fruits and vegetables, it would instead prioritize asking about non-fruits and non-vegetables, like carbohydrates and proteins.

Considering the state-space coverage for the first goal, “Choose lean proteins” (Figure 18), all states are well represented except for one: when only one food item has been mentioned, and it is a protein, but it is ambiguous. For example, the user stating “I’m eating chicken” would result in this state.

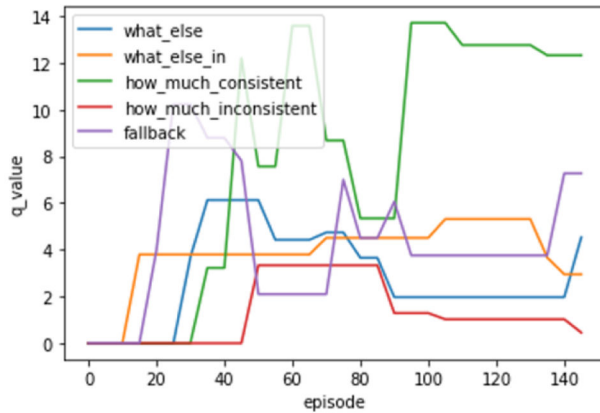
For the second goal, “Eat no more than 2 portions of carbs” (Figure 19), there is relatively low coverage for states with a large number of food items (3 or more), but none of them are carbohydrates.

The third goal “Make 1/2 my meal fruits and/or non-starchy vegetables” (Figure 20), has a considerably larger state space than the other two goals. Coverage was spotty when there were two food items identified, and exactly one was a fruit/vegetable and the other was non-fruit/vegetable. For example, “an apple and peanut butter,” or “chicken and broccoli” would be examples of meal descriptions with low coverage in the corpus.



Overall, these results suggest reasonable coverage, with the caveat that if some states appear in the test set, q-learning may not have had the opportunity to learn reasonable q-values for that state.

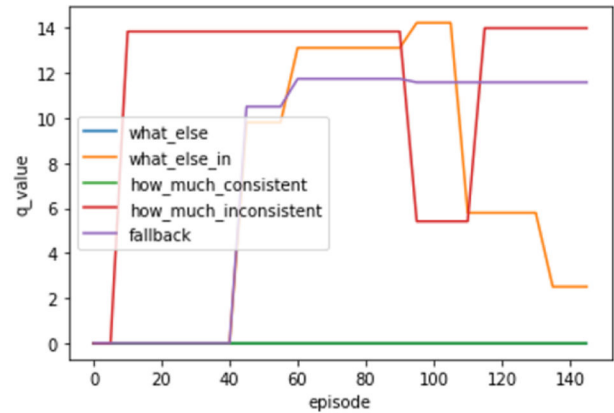
**History of q-values when “how much consistent” is a logical action; at least one fruit or non-starchy vegetables is missing amounts and 4 food items have been identified**



State:

```
n_food_items = 4
any_fruit_veg = 1
any_non_fruit_veg = 1
amt_fruit_veg_all = 0
amt_non_fruit_veg_all = 0
```

**History of q-values when “how much inconsistent” is a logical action; amounts are present for all fruits/vegetables, and 4 food items have been identified**



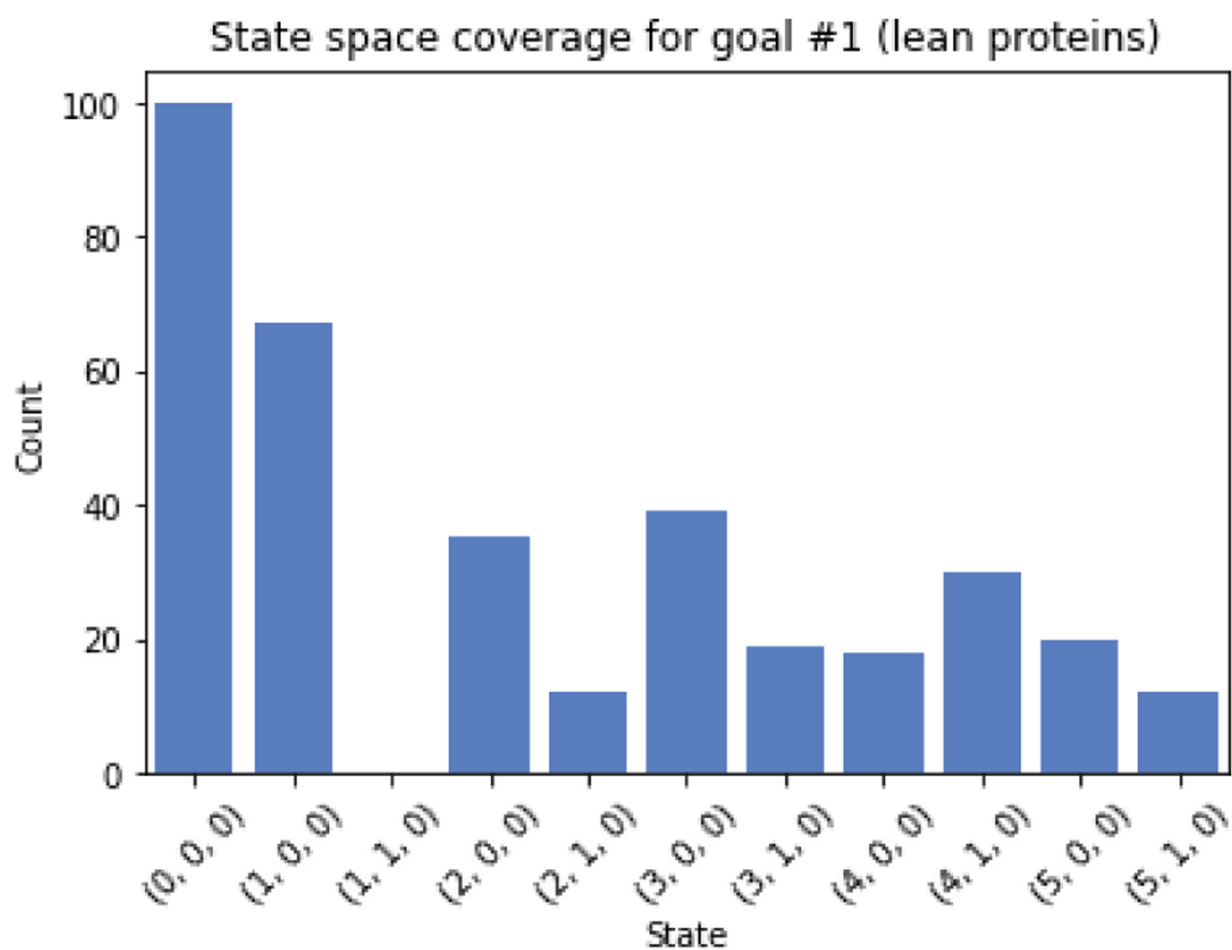
State:

```
n_food_items = 4
any_fruit_veg = 1
any_non_fruit_veg = 1
amt_fruit_veg_all = 1
amt_non_fruit_veg_all = 0
```

**Figure 17:**

Change in q-values over 150 training episodes for two different states, for the goal “Make  $\frac{1}{2}$  of my meal fruit and/or non-starchy vegetables.”

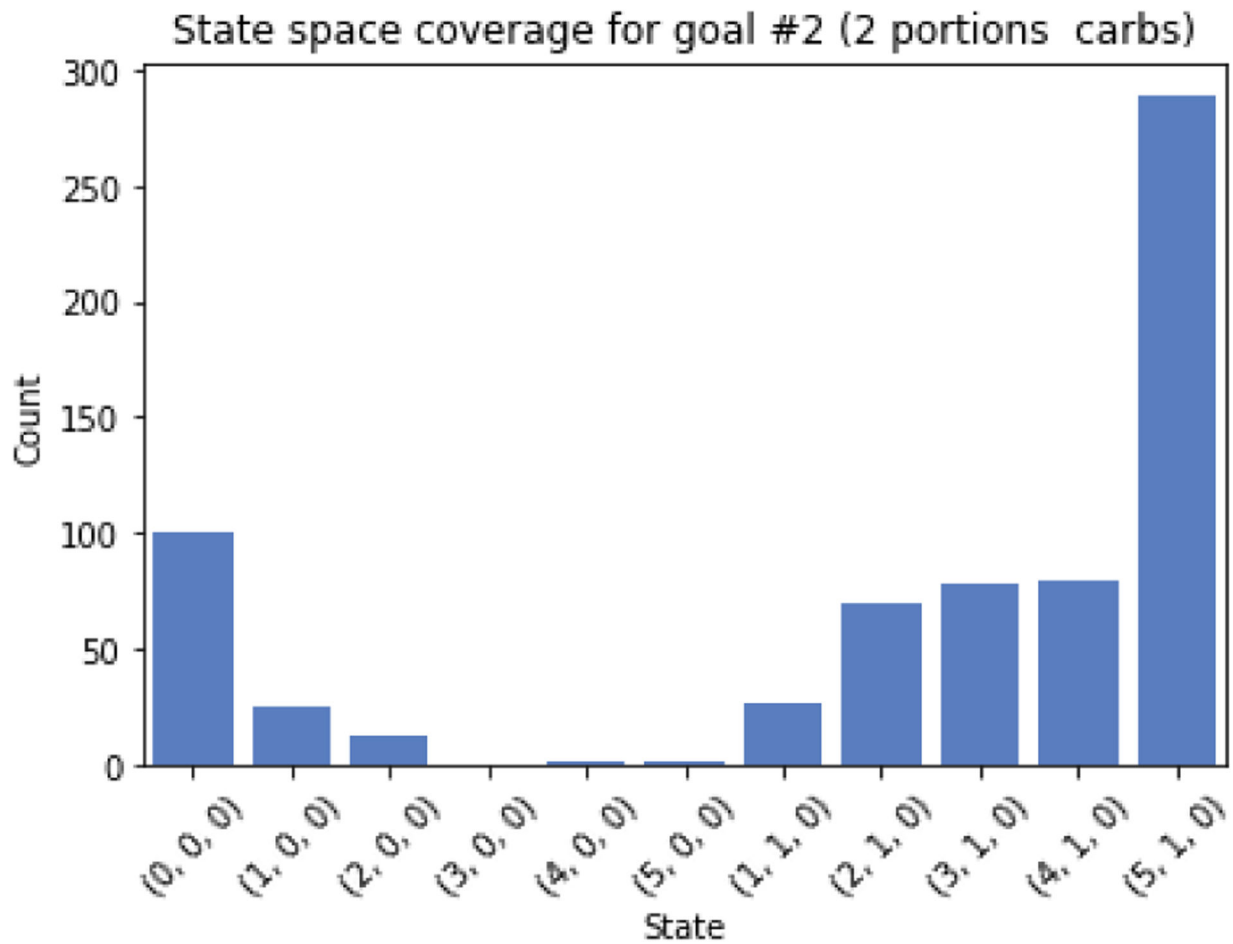




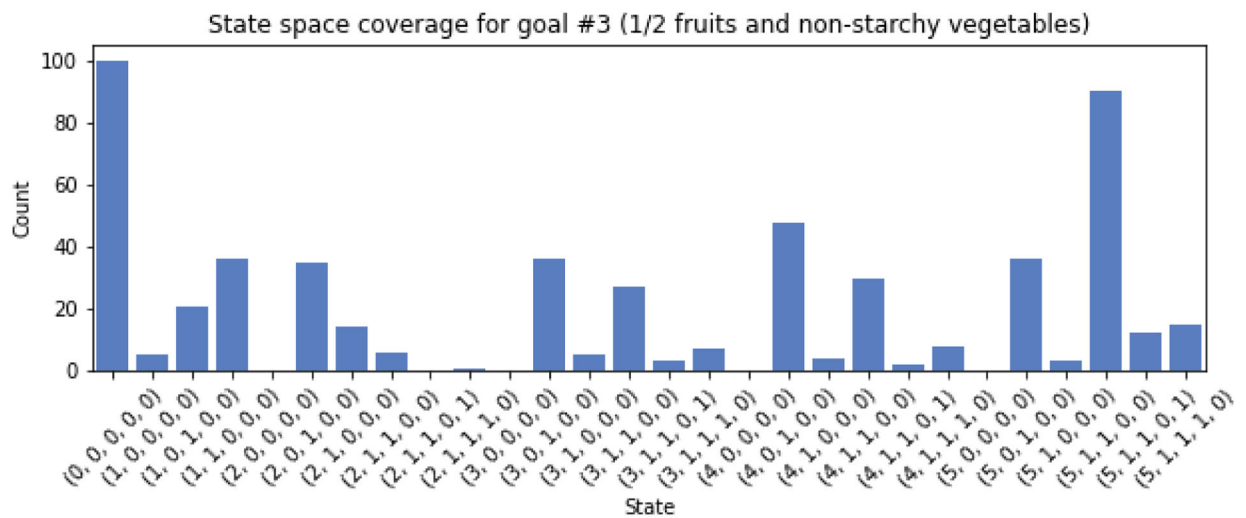
**Figure 18:**

State space coverage for “Choose lean proteins” The x-axis is labeled with the value tuples for the 3 state features: (n\_food\_items, any\_protein, all\_protein\_non\_ambiguous)



**Figure 19:**

State space coverage for “Eat no more than 2 portions of carbs” The x-axis is labeled with the value tuples for the 3 state features: (n\_food\_items, any\_carbs, amt\_carbs\_all)

**Figure 20:**



State space coverage for “Make ½ my meal fruits or non-starchy vegetables” The x-axis is labeled with the value tuples for the 3 state features: (n\_food\_items, any\_fruit\_veg, any\_non\_fruit\_veg, amt\_fruit\_veg\_all, amt\_non\_fruit\_veg\_all)

## REFERENCES

- [1]. Abacha Asma Ben and Zweigenbaum Pierre. 2015. MEANS: A medical question-answering system combining NLP techniques and semantic Web technologies. *Information Processing & Management* 51, 5: 570–594. 10.1016/j.ipm.2015.04.006
- [2]. Adiwardana Daniel, Luong Minh-Thang, So David R, Hall Jamie, Fiedel Noah, Thoppilan Romal, Yang Zi, Kulshreshtha Apoorv, Nemade Gaurav, Lu Yifeng, and Le Quoc V.. 2020. Towards a Human-like Open-Domain Chatbot. Retrieved February 6, 2020 from <http://arxiv.org/abs/2001.09977>
- [3]. Bickmore Timothy, Gruber Amanda, and Picard Rosalind. 2005. Establishing the computer–patient working alliance in automated health behavior change interventions. *Patient Education and Counseling* 59, 1: 21–30. 10.1016/J.PEC.2004.09.008 [PubMed: 16198215]
- [4]. Bickmore Timothy W., Pfeifer Laura M., Byron Donna, Forsythe Shaula, Henault Lori E., Jack Brian W., Silliman Rebecca, and Paasche-Orlow Michael K.. 2010. Usability of Conversational Agents by Patients with Inadequate Health Literacy: Evidence from Two Clinical Trials. *Journal of Health Communication* 15, sup2: 197–210. 10.1080/10810730.2010.499991 [PubMed: 20845204]
- [5]. Bodenheimer Thomas, Lorig Kate, Holman Halsted, and Grumbach Kevin. 2002. Patient Self-management of Chronic Disease in Primary Care. *JAMA* 288, 19: 2469. 10.1001/jama.288.19.2469 [PubMed: 12435261]
- [6]. Paweł Budzianowski and Vuli Ivan. 2019. Hello, It’s GPT-2 – How Can I Help You? Towards the Use of Pretrained Language Models for Task-Oriented Dialogue Systems. *EMNLP-IJCNLP 2019 - Proceedings of the 3rd Workshop on Neural Generation and Translation*: 15–22. 10.18653/v1/d19-5602
- [7]. Burgermaster Marissa, Gajos KZ, and Mamykina L. 2016. Explanations Improve Nutrition Learning Among Lab in the Wild Quiz-Takers. *Journal of Nutrition Education and Behavior* 48, 7: S52–S53. 10.1016/j.jneb.2016.04.142
- [8]. Cabitza Federico, Ciucci Davide, and Rasoini Rafaele. 2019. A Giant with Feet of Clay: On the Validity of the Data that Feed Machine Learning in Medicine.. *Springer, Cham*, 121–136. 10.1007/978-3-319-90503-7\_10
- [9]. Carbone Elena T. and Zoellner Jamie M.. 2012. Nutrition and Health Literacy: A Systematic Review to Inform Nutrition Research and Practice. *Journal of the Academy of Nutrition and Dietetics* 112, 2: 254–265. 10.1016/J.JADA.2011.08.042 [PubMed: 22732460]
- [10]. Cheng Amy, Raghavaraju Vaishnavi, Kanugo Jayanth, Handrianto Yohanes P, and Shang Yi. 2018. Development and evaluation of a healthy coping voice interface application using the Google home for elderly patients with type 2 diabetes. In *2018 15th IEEE Annual Consumer Communications & Networking Conference (CCNC)*, 1–5. 10.1109/CCNC.2018.8319283
- [11]. Chyung Seung Youn (Yonnie), Hutchinson Douglas, and Shamsy Jennifer A.. 2020. Evidence-Based Survey Design: Ceiling Effects Associated with Response Scales. *Performance Improvement* 59, 6: 6–13. 10.1002/PFI.21920
- [12]. CIRP. 2019. Report: Smart speaker adoption in US reaches 66M units, with Amazon leading. Retrieved February 12, 2019 from <https://techcrunch.com/2019/02/05/report-smart-speaker-adoption-in-u-s-reaches-66m-units-with-amazon-leading/>
- [13]. Clavel Céline, Whittaker Steve, Blacodon Anaïs, and Martin Jean-Claude. 2018. WEnner: A Theoretically Motivated Approach for Tailored Coaching About Physical Activity. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers - UbiComp ‘18 (UbiComp ‘18)*, 1669–1675. 10.1145/3267305.3274190
- [14]. Cole-Lewis Heather J., Smaldone Arlene M., Davidson Patricia R., Kukafka Rita, Tobin Jonathan N., Cassells Andrea, Mynatt Elizabeth D., Hripcsak George, and Mamykina Lena.



2016. Participatory approach to the development of a knowledge base for problem-solving in diabetes self-management. *International Journal of Medical Informatics* 85, 1: 96–103. 10.1016/J.IJMEDINF.2015.08.003 [PubMed: 26547253]
- [15]. Cordeiro Felicia, Bales Elizabeth, Cherry Erin, and Fogarty James. 2015. Rethinking the Mobile Food Journal: Exploring Opportunities for Lightweight Photo-Based Capture. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*, 3207–3216. 10.1145/2702123.2702154
- [16]. Denecke Kerstin, Tschanz Mauro, Dörner Tim Lucas, and May Richard. 2019. Intelligent Conversational Agents in Healthcare: Hype or Hope? *Studies in health technology and informatics* 259: 77–84. Retrieved August 9, 2019 from <http://www.ncbi.nlm.nih.gov/pubmed/30923277> [PubMed: 30923277]
- [17]. Devlin Jacob, Chang Ming Wei, Lee Kenton, and Toutanova Kristina. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL HLT 2019 – 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference 1*: 4171–4186. Retrieved December 28, 2021 from <https://arxiv.org/abs/1810.04805v2>
- [18]. Diabetes Prevention Program Research Group. 2009. 10-year follow-up of diabetes incidence and weight loss in the Diabetes Prevention Program Outcomes Study. *The Lancet* 374, 9702: 1677–1686. 10.1016/S0140-6736(09)61457-4
- [19]. Dooley Damion M., Griffiths Emma J., Gosal Gurinder S., Buttigieg Pier L., Hoehndorf Robert, Lange Matthew C., Schriml Lynn M., Brinkman Fiona S.L., and Hsiao William W.L.. 2018. FoodOn: A harmonized food ontology to increase global food traceability, quality control and data integration. *npj Science of Food* 2, 1: 1–10. 10.1038/s41538-018-0032-6 [PubMed: 31304251]
- [20]. Epstein Daniel A., Cordeiro Felicia, Fogarty James, Hsieh Gary, and Munson Sean A.. 2016. Crumbs: Lightweight Daily Food Challenges to Promote Engagement and Mindfulness. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*: 5632–5644. 10.1145/2858036.2858044
- [21]. Fadhil Ahmed, Schiavo Gianluca, and Wang Yunlong. 2019. CoachAI: A Conversational Agent Assisted Health Coaching Platform. Retrieved June 13, 2019 from <http://arxiv.org/abs/1904.11961>
- [22]. Fadhil Ahmed, Wang Yunlong, and Reiterer Harald. 2019. Assistive Conversational Agent for Health Coaching: A Validation Study. *Methods of Information in Medicine*. 10.1055/s-0039-1688757
- [23]. Fitzpatrick Kathleen Kara, Darcy Alison, and Vierhile Molly. 2017. Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. *JMIR mental health* 4, 2: e19. 10.2196/mental.7785 [PubMed: 28588005]
- [24]. Fulmer Russell, Joerin Angela, Gentile Breanna, Lakerink Lysanne, and Rauws Michiel. 2018. Using Psychological Artificial Intelligence (Tess) to Relieve Symptoms of Depression and Anxiety: Randomized Controlled Trial. *JMIR Mental Health* 5, 4: e64. 10.2196/mental.9782 [PubMed: 30545815]
- [25]. Gao Jianfeng, Galley Michel, and Li Lihong. 2018. Neural Approaches to Conversational AI. 10.1145/3209978.3210183
- [26]. Hansen Paul and Ombler Franz. 2008. A new method for scoring additive multi-attribute value models using pairwise rankings of alternatives. *Journal of Multi-Criteria Decision Analysis* 15, 3–4: 87–107. 10.1002/MCDA.428
- [27]. Haussmann Steven, Seneviratne Oshani, Chen Yu, Ne’eman Yarden, Codella James, Chen Ching-Hua, McGuinness Deborah L., and Zaki Mohammed J.. 2019. FoodKG: A Semantics-Driven Knowledge Graph for Food Recommendation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 11779 LNCS: 146–162. 10.1007/978-3-030-30796-7\_10
- [28]. Hone Kate S. and Graham Robert. 2000. Towards a tool for the Subjective Assessment of Speech System Interfaces (SASSI). *Natural Language Engineering* 6, 3&4: S1351324900002497. 10.1017/S1351324900002497



- [29]. Inkster Becky, Sarda Shubhankar, and Subramanian Vinod. 2018. An Empathy-Driven, Conversational Artificial Intelligence Agent (Wysa) for Digital Mental Well-Being: Real-World Data Evaluation Mixed-Methods Study. *JMIR mHealth and uHealth* 6, 11: e12106. 10.2196/12106 [PubMed: 30470676]
- [30]. Jin Lifeng, White Michael, Jafe Evan, Zimmerman Laura, and Danforth Douglas. 2017. Combining CNNs and Pattern Matching for Question Interpretation in a Virtual Patient Dialogue System. 11–21. 10.18653/V1/W17-5002
- [31]. Jin Ying, Yang Zhuoran, and Wang Zhaoran. 2020. Is Pessimism Provably Efficient for Offline RL? Retrieved July 5, 2021 from <http://arxiv.org/abs/2012.15085>
- [32]. Kitamura Keigo, de Silva Chaminda, Yamasaki Toshihiko, and Aizawa Kiyoharu. 2010. Image processing based approach to food balance analysis for personal food logging. In 2010 IEEE International Conference on Multimedia and Expo, 625–630. 10.1109/ICME.2010.5583021
- [33]. Kitamura Keigo, Yamasaki Toshihiko, and Aizawa Kiyoharu. 2008. Food log by analyzing food images. In Proceeding of the 16th ACM international conference on Multimedia - MM '08, 999. 10.1145/1459359.1459548
- [34]. Klasnja Predrag and Pratt Wanda. 2012. Healthcare in the pocket: Mapping the space of mobile-phone health interventions. *Journal of Biomedical Informatics* 45, 1: 184–198. 10.1016/J.JBI.2011.08.017 [PubMed: 21925288]
- [35]. Kocaballi A. Baki, Quiroz Juan C., Laranjo Liliana, Rezazadegan Dana, Kocielnik Rafal, Clark Leigh, Liao Q. Vera, Park Sun Young, Moore Robert J., and Miner Adam. 2020. Conversational agents for health and wellbeing. In Conference on Human Factors in Computing Systems - Proceedings, 1–8. 10.1145/3334480.3375154
- [36]. Kocaballi Ahmet Baki, Berkovsky Shlomo, Quiroz Juan C, Laranjo Liliana, Tong Huong Ly, Rezazadegan Dana, Briatore Agustina, and Coiera Enrico. 2019. The Personalization of Conversational Agents in Health Care: Systematic Review. *Journal of medical Internet research* 21, 11: e15360. 10.2196/15360 [PubMed: 31697237]
- [37]. Kocielnik Rafal, Xiao Lillian, Avrahami Daniel, and Hsieh Gary. 2018. Refection Companion: A Conversational System for Engaging Users in Refection on Physical Activity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 2: 1–26. 10.1145/3214273
- [38]. Laranjo Liliana, Dunn Adam G, Tong Huong Ly, Kocaballi Ahmet Baki, Chen Jessica, Bashir Rabia, Surian Didi, Gallego Blanca, Magrabi Farah, Lau Annie Y S, and Coiera Enrico. 2018. Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association* 25, 9: 1248–1258. 10.1093/jamia/ocy072 [PubMed: 30010941]
- [39]. Lee Jinhyuk, Yoon Wonjin, Kim Sungdong, Kim Donghyeon, Kim Sunkyu, So Chan Ho, and Kang Jaewoo. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 4: 1234–1240. 10.1093/bioinformatics/btz682
- [40]. Li Jiwei, Monroe Will, Ritter Alan, Galley Michel, Gao Jianfeng, and Jurafsky Dan. 2016. Deep Reinforcement Learning for Dialogue Generation. Retrieved October 21, 2018 from <http://arxiv.org/abs/1606.01541>
- [41]. Li Jiwei, Monroe Will, Ritter Alan, Galley Michel, Gao Jianfeng, and Jurafsky Dan. 2016. Deep Reinforcement Learning for Dialogue Generation. *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*: 1192–1202. Retrieved July 5, 2021 from <http://arxiv.org/abs/1606.01541>
- [42]. Li Xiujun, Chen Yun-Nung, Li Lihong, Gao Jianfeng, and Celikyilmaz Asli. 2017. End-to-End Task-Completion Neural Dialogue Systems. Retrieved July 5, 2021 from <http://arxiv.org/abs/1703.01008>
- [43]. Lowe Ryan, Pow Nissan, Serban Iulian, and Pineau Joelle. 2015. The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. Retrieved November 14, 2018 from <http://arxiv.org/abs/1506.08909>
- [44]. McTear Michael F.. 2002. Spoken dialogue technology: enabling the conversational user interface. *ACM Computing Surveys* 34, 1: 90–169. 10.1145/505282.505285
- [45]. Merler Michele, Wu Hui, Uceda-Sosa Rosario, Nguyen Quoc-Bao, and Smith John R.. 2016. Snap, Eat, RepEat: a Food Recognition Engine for Dietary Logging. In *Proceedings of the 2nd*



- International Workshop on Multimedia Assisted Dietary Management - MADiMa '16, 31–40. 10.1145/2986035.2986036
- [46]. Middleton B, Sittig DF, and Wright A. 2016. Clinical Decision Support: a 25 Year Retrospective and a 25 Year Vision. *Yearbook of Medical Informatics* 25, S 01: S103–S116. 10.15265/IYS-2016-s034
- [47]. Mitchell Elliot G., Heitkemper Elizabeth M., and Burgermaster Marissa. 2021. From refection to action: Combining machine learning with expert knowledge for nutrition goal recommendations. *Conference on Human Factors in Computing Systems - Proceedings*: 17. 10.1145/3411764.3445555
- [48]. Mitchell Elliot G., Maimone Rosa, Cassells Andrea, Tobin Jonathan N., Davidson Patricia, Smaldone Arlene M., and Mamykina Lena. 2021. Automated vs. Human Health Coaching. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1: 1–37. 10.1145/3449173
- [49]. Montenegro Joao Luis Zeni, da Costa Cristiano André, and da Rosa Righi Rodrigo. 2019. Survey of conversational agents in health. *Expert Systems with Applications* 129: 56–67. 10.1016/J.ESWA.2019.03.054
- [50]. Naphtal Rachael. 2015. Natural Language Processing Based Nutritional Application. Massachusetts Institute of Technology.
- [51]. Ni Lin, Lu Chenhao, Liu Niu, and Liu Jiamou. 2017. MANDY: Towards a Smart Primary Care Chatbot Application. *Communications in Computer and Information Science* 780: 38–52. 10.1007/978-981-10-6989-5\_4
- [52]. Olsen Jeanette M.. 2014. Health Coaching: A Concept Analysis. *Nursing Forum* 49, 1: 18–29. 10.1111/nuf.12042 [PubMed: 24456550]
- [53]. Raj Shriti, Toporski Kelsey, Garrity Ashley, Lee Joyce M., and Newman Mark W.. 2019. “My blood sugar is higher on the weekends”: Finding a role for context and context-awareness in the design of health self-management technology. In *Conference on Human Factors in Computing Systems - Proceedings*, 1–13. 10.1145/3290605.3300349
- [54]. Rutjes Heleen, Willemsen Martijn C., and IJsselstein Wijnand A.. 2019. Beyond Behavior: The Coach’s Perspective on Technology in Health Coaching. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI ‘19*, 1–14. 10.1145/3290605.3300900
- [55]. Saaty Thomas L. 2008. Relative measurement and its generalization in decision making why pairwise comparisons are central in mathematics for the measurement of intangible factors the analytic hierarchy/network process. *Revista de la Real Academia de Ciencias Exactas, Fisicas y Naturales - Serie A: Matematicas* 102, 2: 251–318. 10.1007/BF03191825
- [56]. Schroeder Jessica, Karkar Ravi, Murinova Natalia, Fogarty James, and Munson Sean A.. 2019. Examining Opportunities for Goal-Directed Self-Tracking to Support Chronic Condition Management. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 4: 1–26. 10.1145/3369809
- [57]. Schulman Daniel, Bickmore Timothy, and Sidner Candace L. 2011. An Intelligent Conversational Agent for Promoting Long-term Health Behavior Change Using Motivational Interviewing. 2011 AAAI Spring Symposium Series. Retrieved April 26, 2017 from <http://relationalagents.com/publications/AAAI2011-schulman.pdf>
- [58]. Serban Iulian V., Sankar Chinnadhurai, Germain Mathieu, Zhang Saizheng, Lin Zhouhan, Subramanian Sandeep, Kim Taesup, Pieper Michael, Chandar Sarath, Ke Nan Rosemary, Rajeshwar Sai, de Brebisson Alexandre, Sotelo Jose M. R., Suhubdy Dendi, Michalski Vincent, Nguyen Alexandre, Pineau Joelle, and Bengio Yoshua. 2017. A Deep Reinforcement Learning Chatbot. Retrieved December 28, 2021 from <https://arxiv.org/abs/1709.02349v2>
- [59]. Serban Iulian Vlad, Lowe Ryan, Henderson Peter, Charlin Laurent, and Pineau Joelle. 2015. A Survey of Available Corpora for Building Data-Driven Dialogue Systems. *arXiv preprint*. 10.5087/dad
- [60]. Shah Pararth, Hakkani-Tür Dilek, Tür Gokhan, Rastogi Abhinav, Bapna Ankur, Nayak Neha, and Heck Larry. 2018. Building a Conversational Agent Overnight with Dialogue Self-Play. Retrieved July 11, 2019 from <http://arxiv.org/abs/1801.04871>



- [61]. Silver David, Schrittwieser Julian, Simonyan Karen, Antonoglou Ioannis, Huang Aja, Guez Arthur, Hubert Thomas, Baker Lucas, Lai Matthew, Bolton Adrian, Chen Yutian, Lillicrap Timothy, Hui Fan, Sifre Laurent, Van Den Driessche George, Graepel Thore, and Hassabis Demis. 2017. Mastering the game of Go without human knowledge. *Nature* 550, 7676: 354–359. 10.1038/nature24270 [PubMed: 29052630]
- [62]. Su Pei Hao, Paweł Budzianowski, Ultes Stefan, Gaši Milica, and Young Steve. 2017. Sample-efficient actor-critic reinforcement learning with supervised data for dialogue management. In *SIGDIAL 2017 – 18th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Proceedings of the Conference*, 147–157. 10.18653/v1/w17-5518
- [63]. Sutton Richard S and Barto Andrew G. 2018. *Reinforcement learning: An introduction*. MIT press.
- [64]. Tanaka Hiroki, Adachi Hiroyoshi, Ukita Norimichi, Ikeda Manabu, Kazui Hiroaki, Kudo Takashi, and Nakamura Satoshi. 2017. Detecting Dementia Through Interactive Computer Avatars. *IEEE Journal of Translational Engineering in Health and Medicine* 5: 1–11. 10.1109/JTEHM.2017.2752152
- [65]. Tennenholtz Guy. 2021. Offline Reinforcement Learning. Conference on Health, Inference, and Learning (CHIL 2021). Retrieved July 5, 2021 from [https://www.chilconference.org/tutorial\\_T03.html](https://www.chilconference.org/tutorial_T03.html)
- [66]. Tsai Chun-Hua, You Yue, Gui Xinning, Kou Yubo, and Carroll John M.. 2021. Exploring and Promoting Diagnostic Transparency and Explainability in Online Symptom Checkers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–17. 10.1145/3411764.3445101
- [67]. United States Department of Agriculture (USDA). ChooseMyPlate. Retrieved September 16, 2020 from <https://www.choosemyplate.gov/>
- [68]. Vaira Lucia, Bochicchio Mario A, Conte Matteo, Casaluci Francesco Margiotta, Melpignano Antonio, Vaira L, Bochicchio MA, Conte M, and Casaluci F Margiotta. 2018. MamaBot: a System based on ML and NLP for supporting Women and Families during Pregnancy. 10.1145/3216122.3216173
- [69]. Vaswani Ashish, Shazeer Noam, Parmar Niki, Uszkoreit Jakob, Jones Llion, Gomez Aidan N, Kaiser Łukasz, and Polosukhin Illia. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, 5999–6009. Retrieved October 23, 2018 from <https://arxiv.org/pdf/1706.03762.pdf>
- [70]. Watkins Christopher J C H and Dayan Peter. 1992. Q-Learning. 8: 279–292.
- [71]. Watkins CJCH. 1989. Learning from delayed rewards. Retrieved July 18, 2021 from [https://www.academia.edu/download/50360235/Learning\\_from\\_delayed\\_rewards\\_20161116-28282-v2pwwq.pdf](https://www.academia.edu/download/50360235/Learning_from_delayed_rewards_20161116-28282-v2pwwq.pdf)
- [72]. Weizenbaum Joseph. 1966. ELIZA – a computer program for the study of natural language communication between man and machine. *Communications of the ACM* 26, 1: 36–45. 10.1145/357980.357991
- [73]. Wolever Ruth Q. and Eisenberg David M.. 2011. What is health coaching anyway? Standards needed to enable rigorous research. *Archives of Internal Medicine* 171, 2017–2018. 10.1001/archinternmed.2011.508 [PubMed: 21986348]
- [74]. Yang Longqi, Cui Yin, Zhang Fan, Pollak John P., Belongie Serge, and Estrin Deborah. 2015. PlateClick. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management - CIKM '15*, 183–192. 10.1145/2806416.2806544
- [75]. Yang Longqi, Hsieh Cheng-Kang, Yang Hongjian, Dell Nicola, Belongie Serge, Cole Curtis, and Estrin Deborah. 2016. Yum-me: A Personalized Nutrient-based Meal Recommender System. *ACM Transactions on Information Systems* 36, 1: 7. 10.1145/3072614
- [76]. Yasavur Ugan, Lisetti Christine, and Rishe Naphtali. 2014. Let's talk! speaking virtual counselor offers you a brief intervention. *Journal on Multimodal User Interfaces* 8, 4: 381–398. 10.1007/S12193-014-0169-9/TABLES/6
- [77]. Zhang Jingwen, Oh Yoo Jung, Lange Patrick, Yu Zhou, and Fukuoka Yoshimi. 2020. Artificial Intelligence Chatbot Behavior Change Model for Designing Artificial Intelligence Chatbots to



Promote Physical Activity and a Healthy Diet: Viewpoint. J Med Internet Res 2020;22(9):e22845  
<https://www.jmir.org/2020/9/e22845> 22, 9: e22845. 10.2196/22845

[78]. Nutrition API by Nutritionix. Retrieved March 26, 2018 from <https://www.nutritionix.com/business/api>

Author Manuscript

Author Manuscript

Author Manuscript

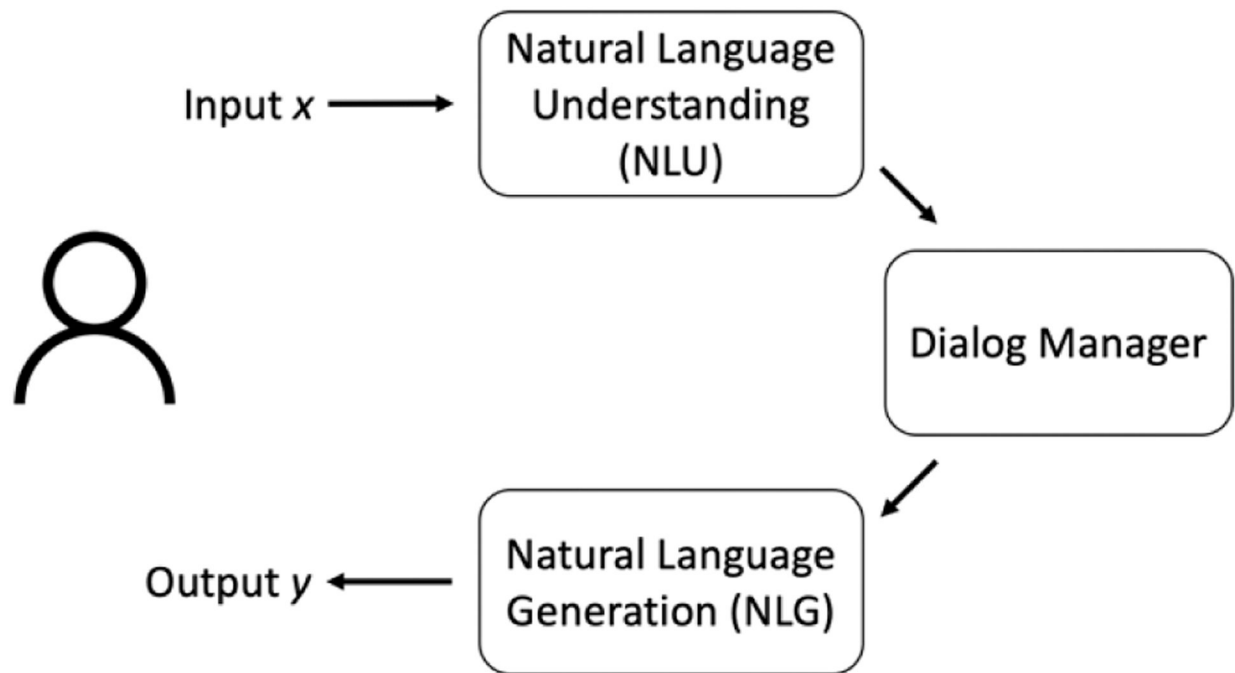
Author Manuscript



**CCS CONCEPTS**

• **Applied computing** → Life and medical sciences; Consumer health; • **Theory of computation** → Theory and algorithms for application domains; Machine learning theory; Reinforcement learning; • **Computing methodologies** → Artificial intelligence; Natural language processing; Artificial intelligence; Distributed Artificial intelligence; Intelligent agents.

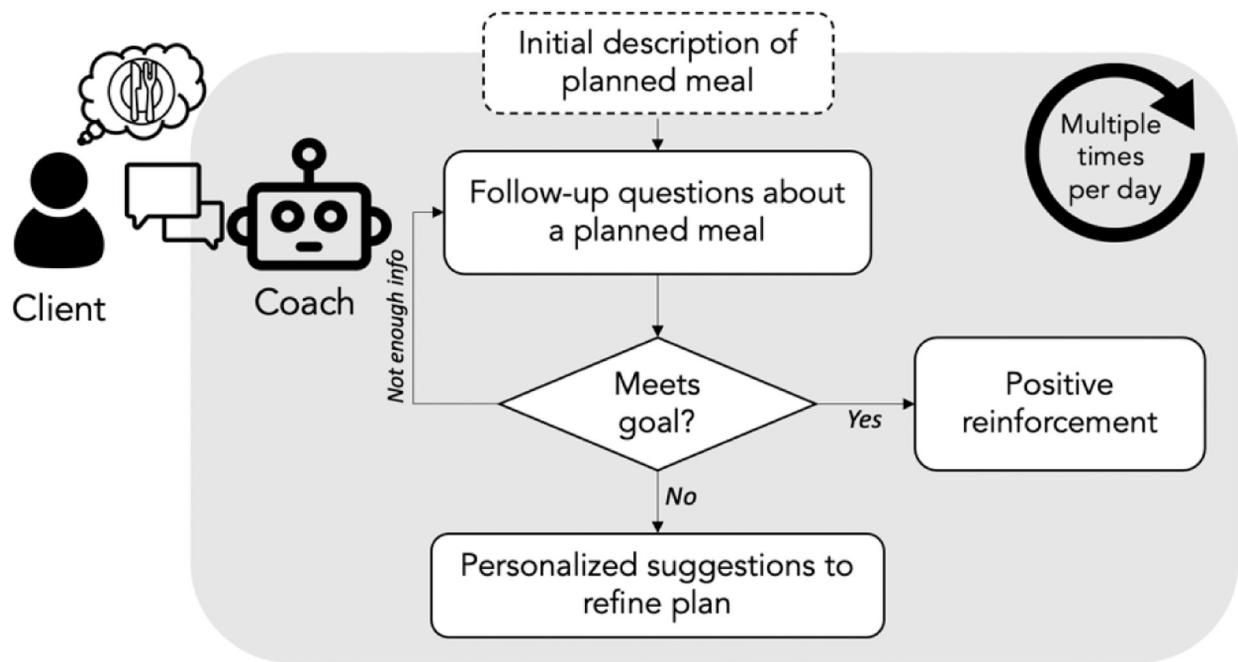




**Figure 1:**

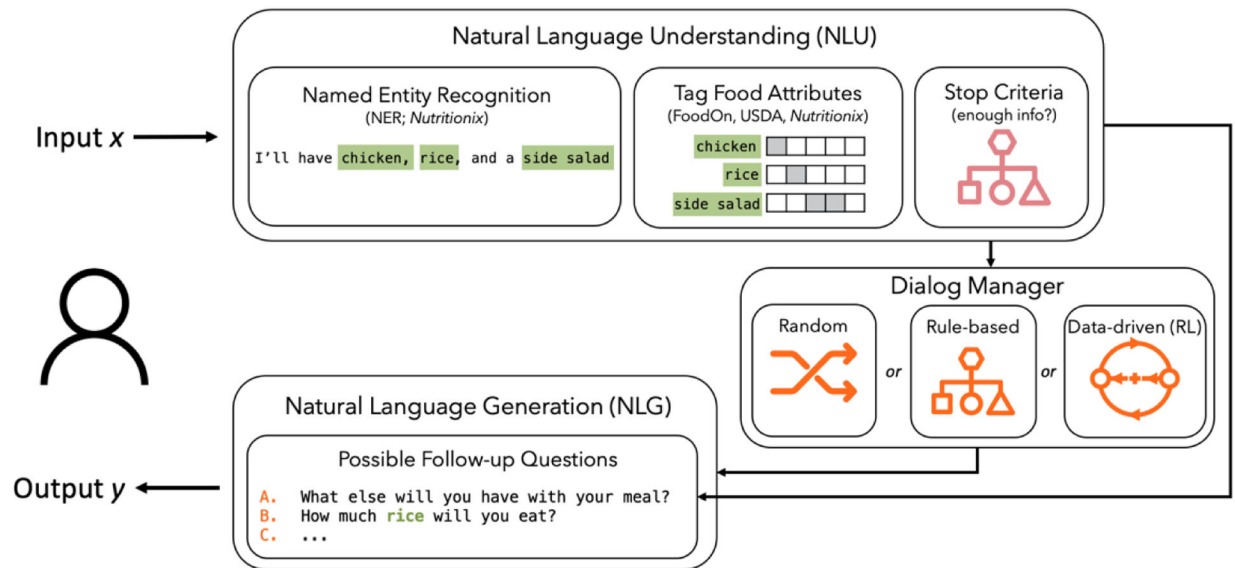
A common chatbot architecture separates natural language understanding and generation from dialog management





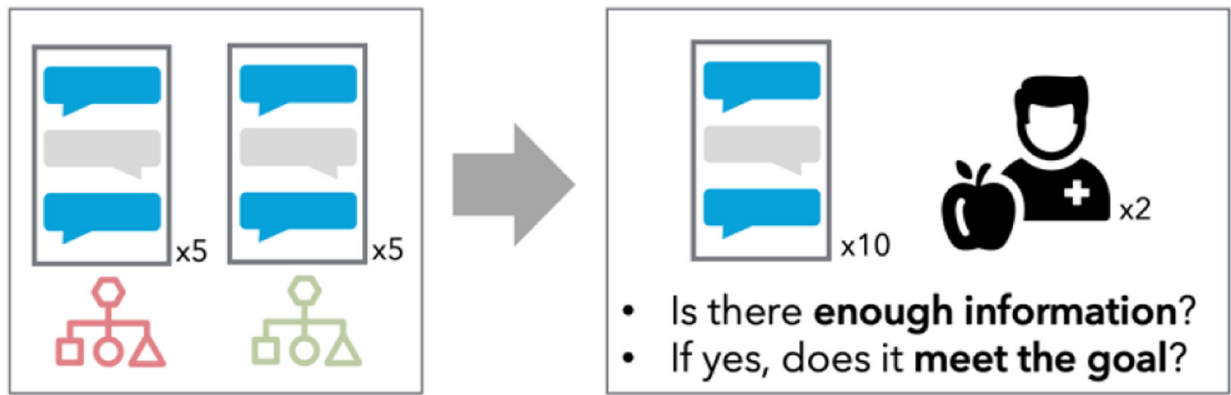
**Figure 2:**  
Proposed structure for micro-coaching dialogs.





**Figure 3:**  
Outline of the process of parsing meal descriptions from input dialog utterances.

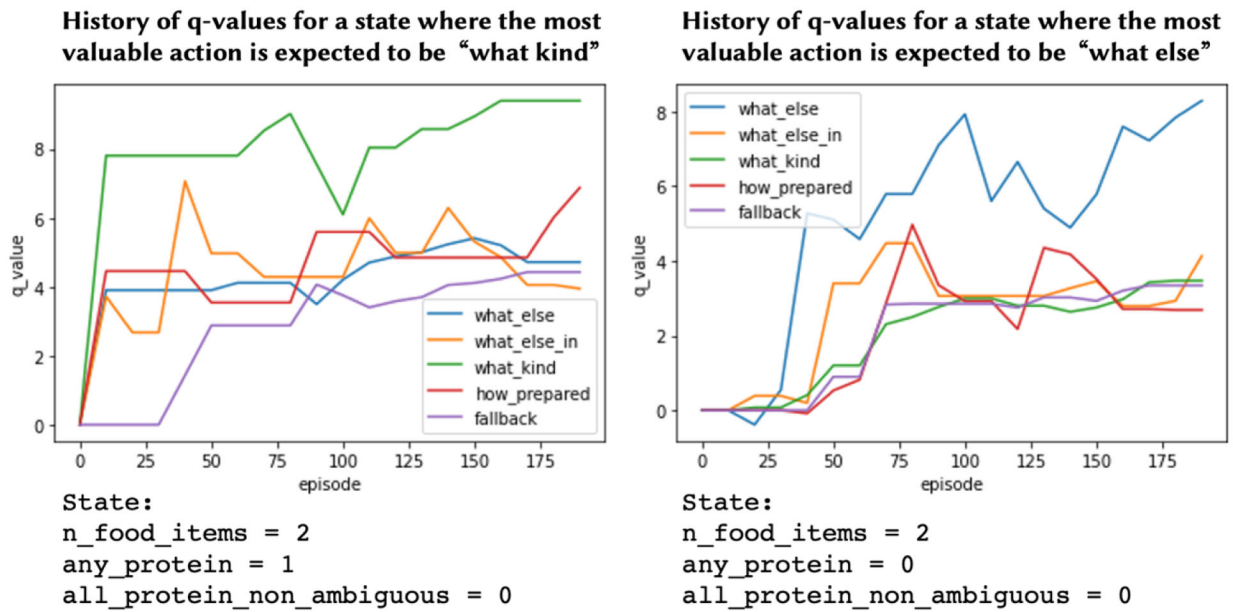




**Figure 4:**

Study design for the evaluation of the natural language understanding (NLU) system. Two dietitians assessed 10 dialogs each, 5 dialogs that reached the stop criteria and 5 that did not.



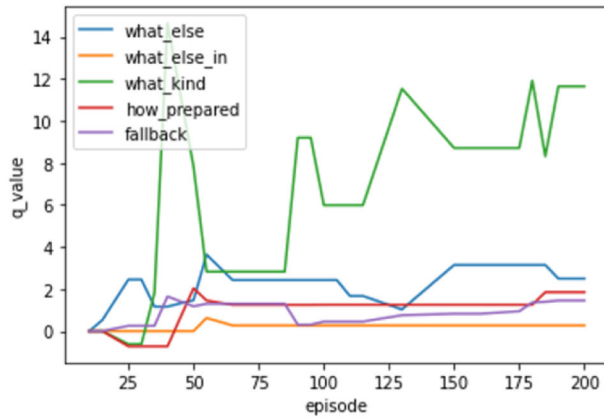


**Figure 5:**

Comparison of change in q-values over training between two different states in offline learning with simulated data. Higher q-values suggest an action will be more valuable in a given state. The only difference between the two states is whether any proteins have been mentioned by the user — `any_protein` is equal to 1 on the left and 0 on the right. If a protein has been mentioned, then most valuable action per the simulation is to ask “what kind” of protein to determine if it’s fatty or lean. The graph on the left shows that the q-value for “what kind” questions (green) quickly becomes the most valuable after a few dozen training episodes. In contrast, when there are no proteins mentioned yet, as on the right, that question is not valuable and instead asking “what else” to find addition food items that might be proteins should be more valuable. The graph on the right shows that the q-value for “what else” questions quickly and appropriately becomes the most valuable action for that state.

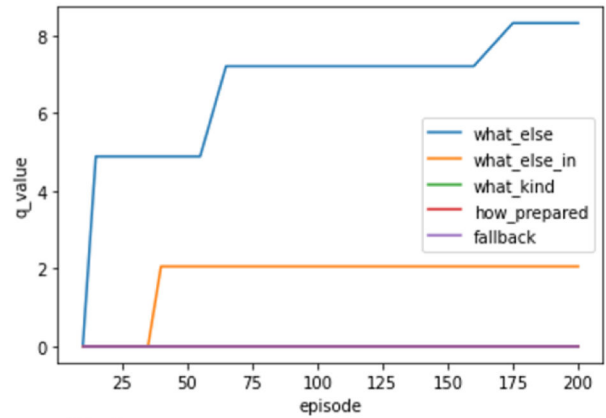


**History of q-values when “what kind” is a logical action; multiple food items have been identified, and at least one is an ambiguously fatty protein**



**State:**  
 n\_food\_items = 4  
 any\_protein = 1  
 all\_protein\_non\_ambiguous = 0

**History of q-values when “what else” is a logical action; multiple food items have been identified, and no proteins have been mentioned**

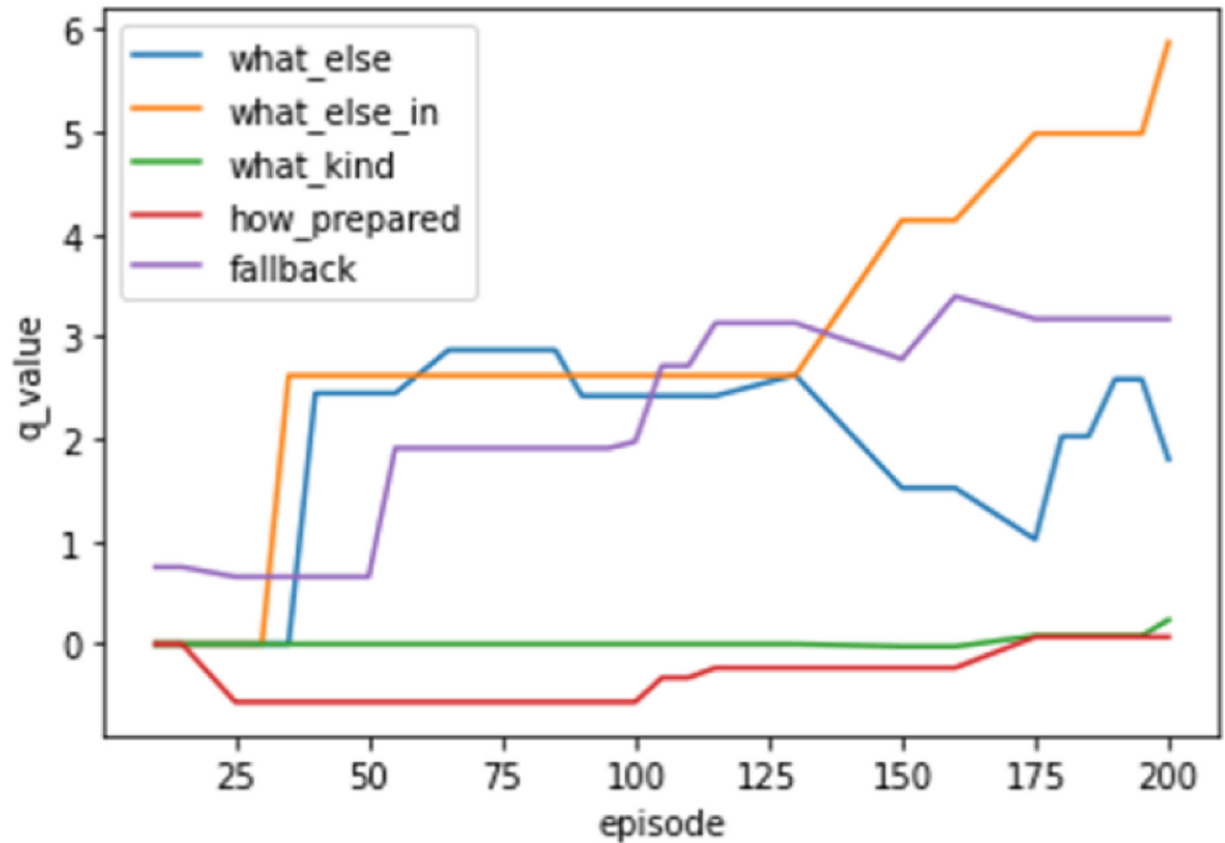


**State:**  
 n\_food\_items = 4  
 any\_protein = 0  
 all\_protein\_non\_ambiguous = 0

**Figure 6:**

Change in q-values over 200 training episodes for two different states, for the goal “Choose lean proteins.” The high q-values for “what kind” and “what else” questions on the right- and left-hand graphs, respectively, mirror the patterns found in the simulated data set, as shown in Figure 5





**State:**

`n_food_items = 2`

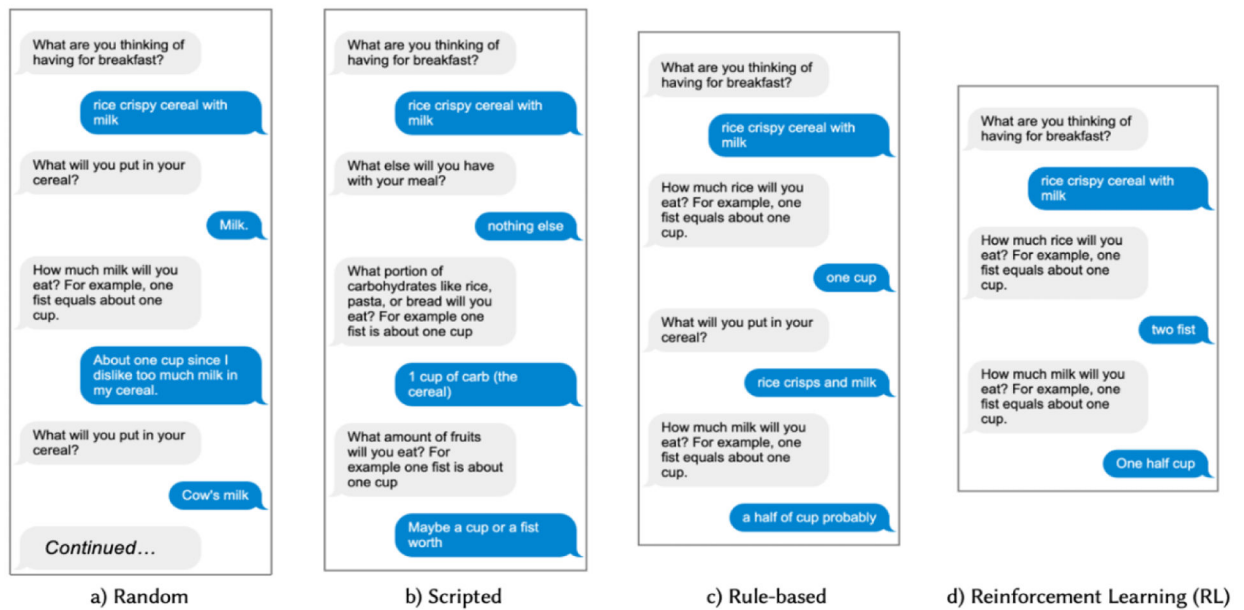
`any_protein = 1`

`all_protein_non_ambiguous = 0`

**Figure 7:**

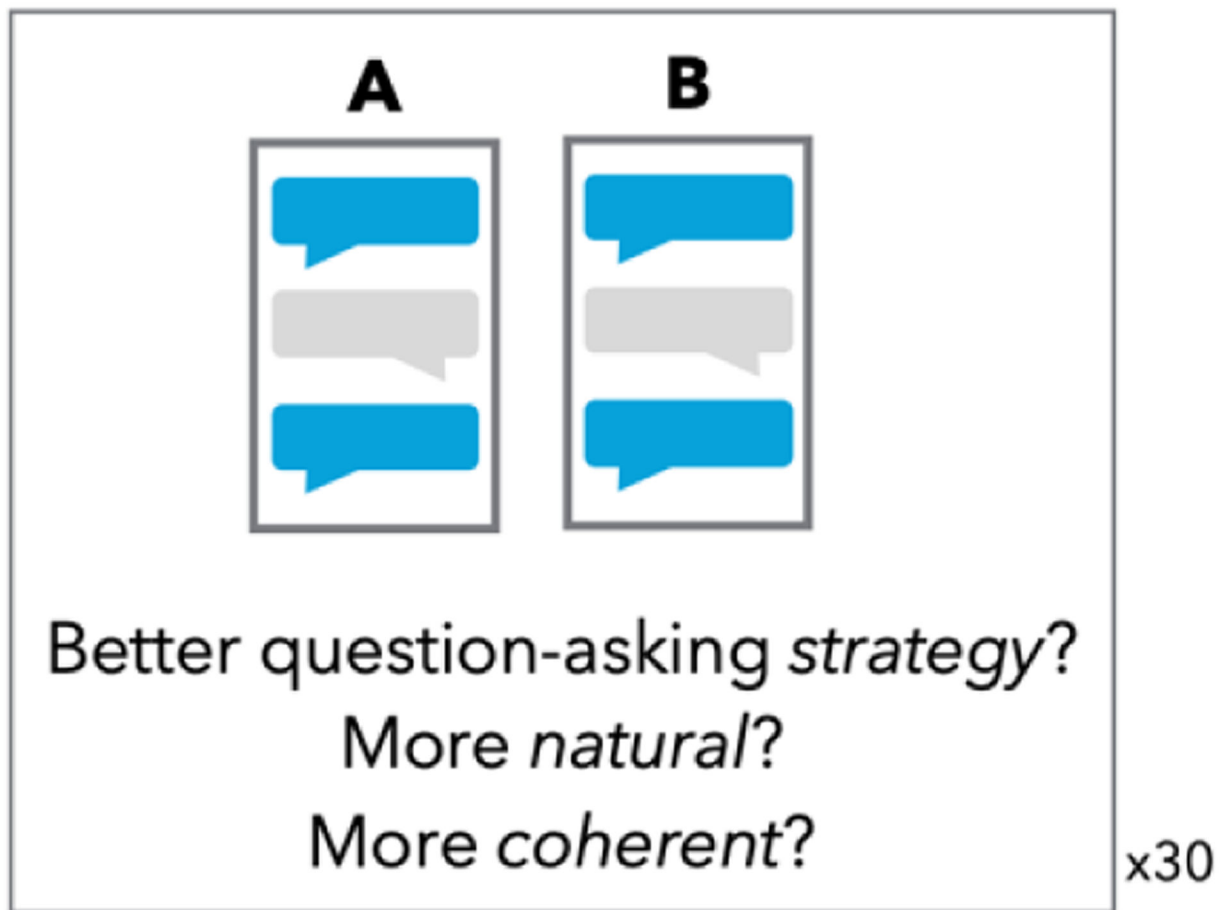
Change in q-values over 200 training episodes for the goal “Choose lean proteins,” when only two foods are mentioned and one is a protein. In this state, even though there is an ambiguous protein to ask a follow-up “what kind” question about, the q-values reward asking “search” questions like “what else is in  $\langle x\_food\_item \rangle$ ”, demonstrating a balance between valuing “search” and “drill-down” question types.



**Figure 8:**

Side-by-side comparison of chatbots with multiple dialog management approaches for the same starting meal description for the goal to “Eat no more than 2 portions of carbs (30g).” The Random chatbot dialog (a) continues for 2 more turns, and is longest of the four dialogs. The Scripted dialog (b) consistently asks the same 3 follow-up questions for every meal, based on the user’s goal. In this example, the Rule-based (c) and RL (d) dialogs include similar follow-up questions, with the exception that the rule-based dialog includes at least one “*Search*” question (“What else will you put in your cereal?”), whereas the RL chatbot does not. Note that while the dialogs start with the same seed meal description, each dialog is continued by a different crowd worker, so the final description of meals diverge by the end of each dialog.

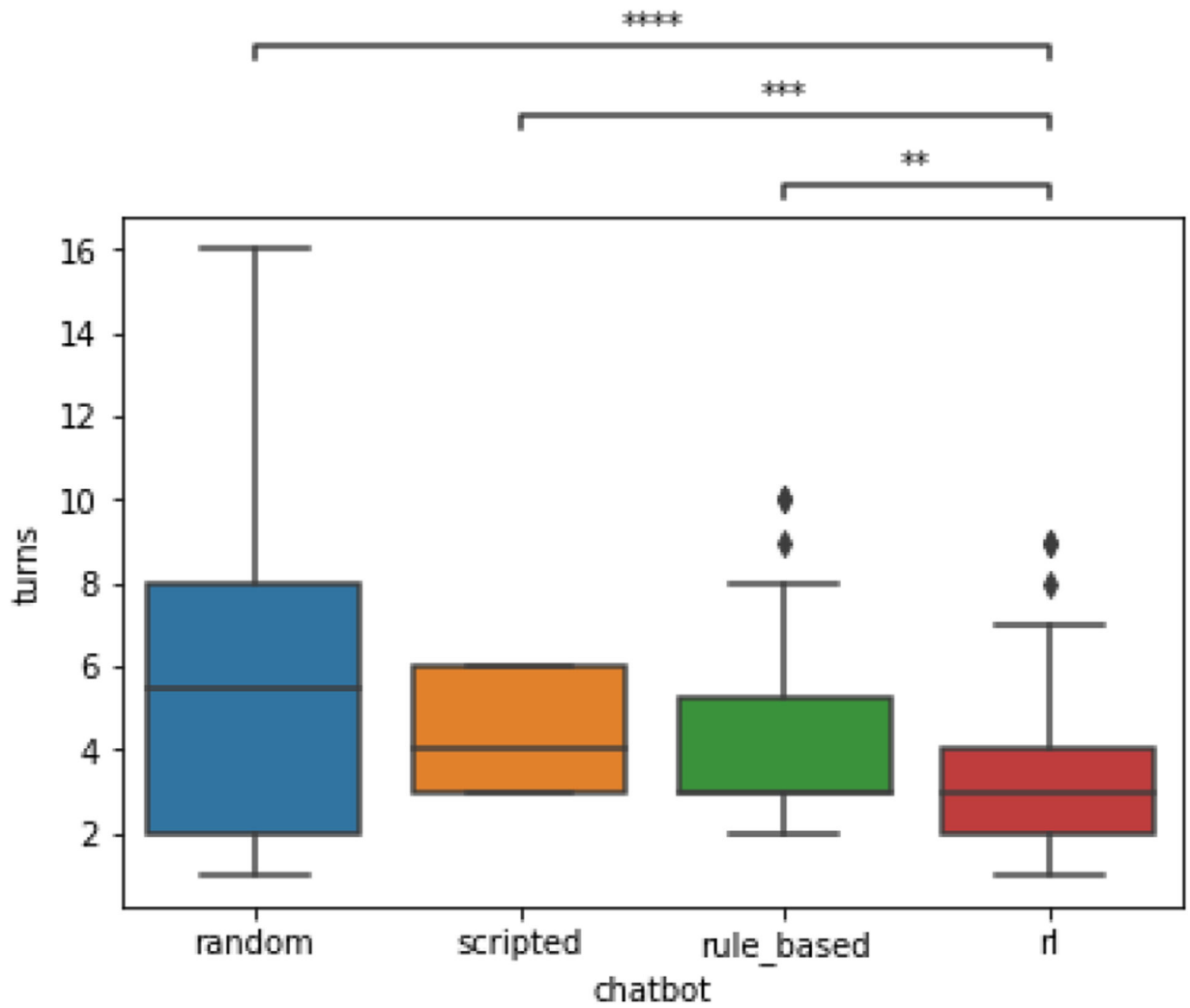




**Figure 9:**

Illustration of the pairwise comparison task to evaluate dialog quality. A and B are dialogs generated from two different chatbots, for the same seed meal description.

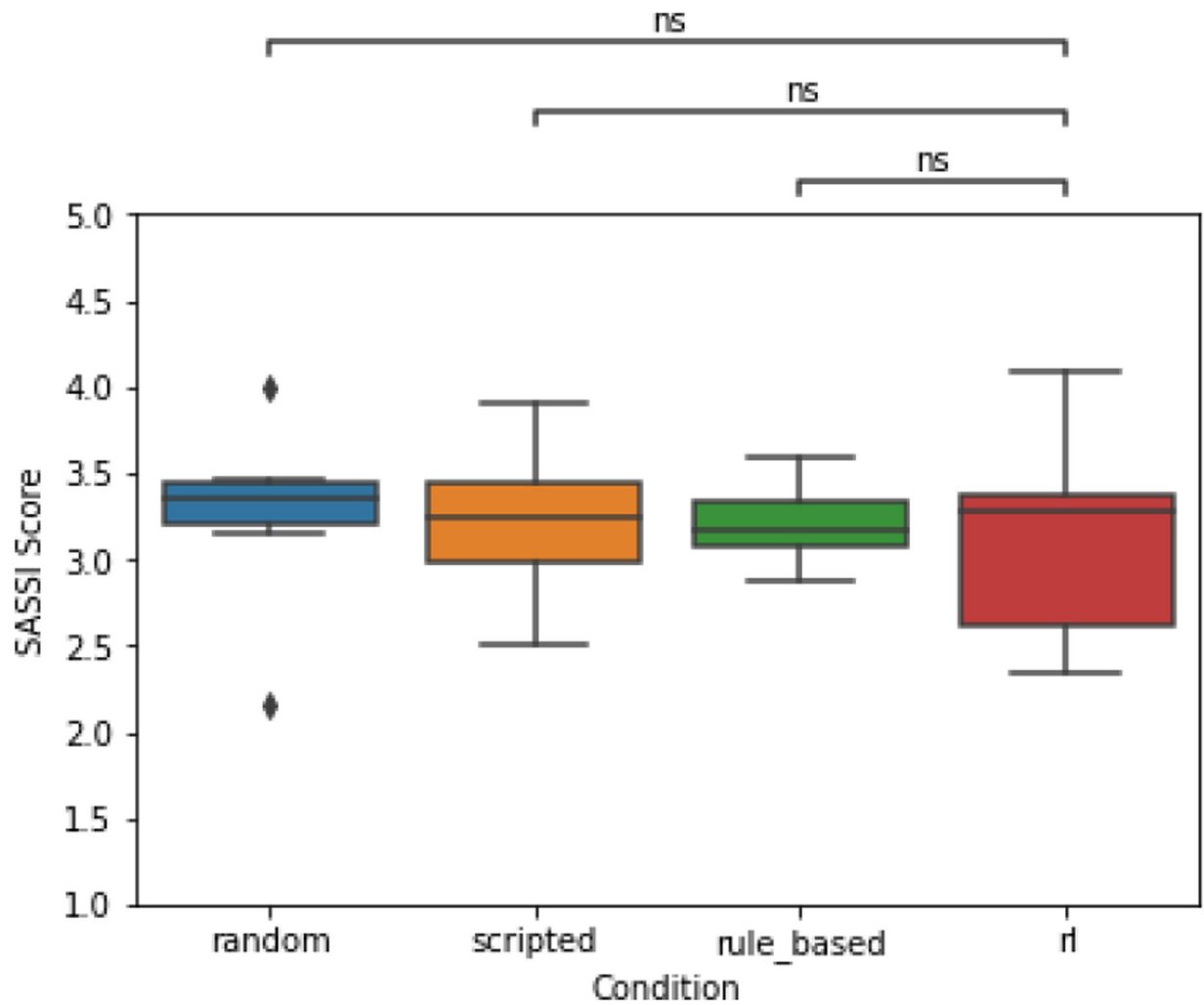




**Figure 10:**

Box-and-whisker plot comparing the number of conversational turns per dialog across the four chatbot conditions. Reinforcement learning (RL) dialogs were the shortest, followed by rule-based, scripted and random. \*\*p<0.01; \*\*\*p<0.001, \*\*\*\*p < 0.0001





**Figure 11:**

Average user experience scores across the four chatbot conditions, measured with the Subjective Assessment of Speech Systems Interfaces (SASSI; [28]) indicate no detectable differences in perceived user experience across the four chatbots.



Table 1:

Types of meal-related questions asked my health coaches

Question Category	Question Type	Example
Search	What else?	“What else will you have with your meal?”
Drill-down	What kind?	“What kind of chicken will you have?”
	How much?	“What portion of rice will you eat?”
	How prepared?	“How was your spinach prepared?”



Table 2:

Nutrition goals selected for crowdsourcing experiments

Nutrition Goal	Qualitative vs. quantitative	Presence vs. absence
Choose lean proteins	Qualitative	Presence/increase
Eat no more than 2 portions of carbs in each meal	Quantitative (amounts)	Absence/decrease
Make ½ my meal fruits and/or non-starchy vegetables	Quantitative (proportions)	Both/replace



Table 3:

Summary of the follow-up questions for each nutrition goal, with examples.

Goal	Question types	Example
All goals	What else?	“What else will you have with your meal?”
	What else in <container-food>?	“What will you put in your burrito?”
	Fallback	“Could you please describe your meal using different words?”
Choose lean proteins	What kind <ambiguous_protein>?	“What kind of chicken? (for example breast or thigh, with or without skin)”
	How prepared <preparable_food>?	“How will your chicken be prepared?”
Eat no more than 2 portions of carbs (30g)	How much <goal_related>?	“How much rice will you eat? (one fist is about the size of one cup)”
Make ½ my meal fruits and/or	How much <goal_consistent>?	“How much broccoli will you eat? (one fist is about the size of one cup)”
non-starchy vegetables	How much <goal_inconsistent>?	“How much rice will you eat? (one fist is about the size of one cup)”



**Table 4:**

Average accuracy of stop criteria from the rule-based system with expert registered dietitian (RD) annotations

Goal	Accuracy
All goals	83%
Choose lean proteins	95%
Eat no more than 2 portions of carbs (30g)	80%
Make ½ my meal fruits and/or non-starchy vegetables	75%

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript



Table 5:

State features and state space size for the three nutrition goals

Goal	State Space Feature	Values	N States
Choose lean proteins	Number of food items	(0, 5)	24
	Any proteins?	(0, 1)	
	All proteins non-ambiguous?	(0, 1)	
Eat no more than 2 portions of carbs (30g)	Number of food items	(0, 5)	24
	Any carbohydrates?	(0, 1)	
	All carbohydrates with amounts?	(0, 1)	
Make ½ my meal fruits and/or non-starchy vegetables	Number of food items	(0, 5)	96
	Any fruits and/or non-starchy vegetables?	(0, 1)	
	Any carbohydrates or proteins?	(0, 1)	
	All fruit and vegetables have amounts?	(0, 1)	
	All carbs and proteins have amounts?	(0, 1)	



Average conversation lengths (number of turns) and reward earned per episode in the experiment with simulated data

**Table 6:**

	Greedy-q policy	Random policy
Conversation length (turns)*	<b>2.36</b> (SD = 1.88)	3.34 (SD = 2.67)
Reward per episode*	<b>9.99</b> (SD = 1.62)	9.41 (SD = 1.91)

\*  $p < 0.001$



**Table 7:**

Average turn length across the four conditions, by nutrition goal

	Random	Scripted	Rule-based	RL
Overall	5.75 ( $\pm$ 3.65)	4.33 ( $\pm$ 1.24)	4.18 ( $\pm$ 2.22)	3.56 ( $\pm$ 2.30)
Goal 1 “Choose lean proteins”	3.75 ( $\pm$ 2.59)	3.00 ( $\pm$ 0)	3.60 ( $\pm$ 2.20)	3.10 ( $\pm$ 2.41)
Goal 2 “Eat no more than 2 portions of carbs (30g)”	6.60 ( $\pm$ 4.07)	4.00 ( $\pm$ 0)	3.45 ( $\pm$ 1.35)	2.55 ( $\pm$ 1.02)
Goal 3 “Make ½ my meal fruits and/or non-starchy vegetables”	6.90 ( $\pm$ 3.28)	6.00 ( $\pm$ 0)	5.50 ( $\pm$ 2.44)	5.05 ( $\pm$ 2.36)



Table 8:

Quality construct “win percentage” for the four chatbots, by goal.

Goal	Condition	Win Percentage			
		Strategy	Naturalness	Coherence	Composite
Choose lean proteins	RL	34%	45%	34%	38%
	scripted	<b>66%</b>	49%	<b>66%</b>	<b>61%</b>
	rule-based	64%	<b>57%</b>	55%	59%
No more than 2 portions carbs	random	36%	49%	44%	43%
	RL	48%	<b>62%</b>	<b>56%</b>	<b>55%</b>
	scripted	47%	53%	51%	50%
1/2 fruits and non-starchy vegetables	rule-based	<b>58%</b>	47%	53%	53%
	random	47%	36%	40%	41%
	RL	35%	38%	37%	37%
	scripted	<b>71%</b>	<b>66%</b>	<b>66%</b>	<b>68%</b>
	rule-based	42%	46%	46%	45%
	random	50%	49%	50%	50%



Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

**Table 9:**  
Quality construct “win percentage,” by dialog length (excluding the scripted chatbot)

Win Percentage				
	Strategy	Naturalness	Coherence	Composite
Shorter Dialog Wins	32%	<b>46%</b>	40%	39%
Longer Dialog Wins	<b>48%</b>	33%	39%	40%
Tie	21%	21%	21%	21%