Bootstrapping a Conversational Guide for Colonoscopy Prep

Pulkit Arya Madeleine Bloomquist Subhankar Chakraborty Andrew Perrault William Schuler Eric Fosler-Lussier Michael White
The Ohio State University
Columbus, OH, USA
{arya.35,bloomquist.7,perrault.17,schuler.77,fosler-lussier.1,white.1240}@osu.edu
subhankar.chakraborty@osumc.edu

Abstract

Creating conversational systems for niche domains is a challenging task, further exacerbated by a lack of quality datasets. We explore the construction of safer conversational systems for guiding patients in preparing for colonoscopies. This has required a data generation pipeline to generate a minimum viable dataset to bootstrap a semantic parser, augmented by automatic paraphrasing. Our study suggests large language models (e.g., GPT-3.5 & GPT-4) are a viable alternative to crowd sourced paraphrasing, but conversational systems that rely upon language models' ability to do temporal reasoning struggle to provide accurate responses. A neural-symbolic system that performs temporal reasoning on an intermediate representation of user queries shows promising results compared to an end-to-end dialogue system, improving the number of correct responses while vastly reducing the number of incorrect or misleading ones.

1 Introduction

Colorectal cancer is the second leading cause of cancer-related deaths worldwide. Colonoscopy is a safe and effective strategy to screen asymptomatic individuals for precursors of colorectal cancer, but it requires a precisely timed multi-day, multi-step procedure to clear the colon. In today's standard practice, patients are given information sheets to help them prepare for the procedure, which instruct them to follow a low-fiber diet for several days prior to the procedure (among other restrictions) and to drink a preparatory mix that cleanses the colon. Unfortunately, these information sheets are frequently ineffective, resulting in rescheduled procedures with large economic, health-related and social costs.

In this paper, we report on our initial steps to develop a conversational assistant to improve the ease of following colonoscopy preparation instructions. To avoid information overload, the assistant is designed to coach patients through the process (known as "prep"), reminding patients when it is time to carry out each step in the instructions and allowing them to **ask questions at any time** about the procedure and the diet changes they need to make at different stages of the preparatory period. Additionally, the assistant will escalate questions to health-care providers when necessary to answer complex questions or reschedule.

Existing efforts to make it easier to follow colonoscopy prep instructions give strong evidence that our approach can greatly enhance patient success. Engaging patients with automatic text reminders greatly improved colonoscopy prep adherence (90% vs. 62%) when patients were invited to ask follow up questions with health-care providers (Mahmud et al., 2019), but a larger scale trial where patients were not invited to reply to the text messages (for lack of personnel) found no improvements over the control group (Mahmud et al., 2021). The **capacity to answer questions**—which we seek to automate for the first time—appears to have been the crucial difference (Clancy and Dominitz, 2021).

Embodied conversational agents (ECA) from the Northeastern Relational Agents Lab have been developed for a variety of health-care communication scenarios over many years. In particular, Ehrenfeld et al. (2010) develop an ECA for counseling patients on their options for anesthesia prior to surgery, but the system cannot answer specific questions patients ask in their own words.

With no existing data in this domain, we seek to take advantage of pretrained and large language models (PLMs/LLMs) to develop our system in a data-efficient way while **robustly avoiding unsafe behavior**. Recent years have witnessed enormous progress on a wide range of NLP tasks, including conversational AI ones, thanks to engineering advances in training large scale, transformer-based neural language models (Bowman and Dahl, 2021;

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 413–420

et al., 2022; Wei et al., 2022; OpenAI, 2023; Laskar et al., 2023; Hosseini-Asl et al., 2020). However, their deployment for practical tasks has been hindered by concerns about safety, such as the propensity of these models to regurgitate toxic language or hallucinate fake news (Bender et al., 2021; Weidinger et al., 2021; Dinan et al., 2022). In healthcare settings, these concerns are especially problematic, as with insufficient controls PLMs could give harmful or even deadly advice (Bickmore et al., 2018).

To address these safety concerns, we have designed a neuro-symbolic system that uses PLMs for contextual natural language understanding (NLU) together with a rule-based dialogue manager and knowledge base. To bootstrap the system, we have used state machines to create simulated dialogues (Campagna et al., 2020) together with LLMs for paraphrasing, rather than crowdworkers as in the overnight method (Wang et al., 2015); further enhancement using Wizard-of-Oz (Kelley, 1984) methods is left to future work.

2 Methods

2.1 Conversational State Machine

A state machine can be used to model a multi-turn conversation for simulation purposes (Jurafsky et al., 1997; Campagna et al., 2020). Our implementation of conversational state machine models different conversational states as states in the state machine. The transitions in the state machine represent user and agent utterances that are possible for the given conversation state.

Figure 1 illustrates the overall structure of our data generation pipeline. Each transition can yield multiple synthetic user utterances via random choices in an attribute grammar, along with a canonical, context-independent version of the user utterance. An SCFG translates the canonical string into a JSON string that represents the meaning in intent-and-slot style. The dialogue manager uses this formal representation to determine the system response, expressing it with simple templates.

By polling the state machine and recording the utterance emissions for each transition, we can generate a diverse dataset of conversations that a patient preparing for a colonoscopy might have with the patient prep system (Figure 2). The state machine also encodes the dialogue context, which allows the system to reference previous utterances. This allows for insertion of coreferential anaphors ("it" or "that") as well as generating follow-up questions: "Agent: You can't eat strawberries so close to the procedure. User: How about bananas?" Similarly, we expect "why?" questions to be very elliptical and only interpretable in context (Figure 2).

2.2 System Design

Contextual NLU via semantic parsing converts the user utterance into a valid canonical string, taking the previous context into account (Shin et al., 2021). As detailed in the next section, we train neural models for this task, without using an explicit module for dialogue state tracking. If the semantic parser does not return a valid canonical string (e.g., for an out-of-scope user question), the SCFG translation will fail to return a formal representation, triggering a request for the user to rephrase.

Once the semantic parsing module correctly parses the user utterance to a formal representation, it is processed by the dialogue manager. The dialogue manager has 4 modules to respond to user questions, one for each of the categories Food, Procedure, Task, and General. Each module has predefined rule-based templates that ensure the information provided to the patients is accurate, safe, and not misleading (Table 4, Appendix A).

Questions in the food category are time sensitive and thus the most challenging to handle. The food module answers questions after first consulting a food knowledge graph to calculate when a patient must stop consuming the item relative to the procedure. For example, in permission questions like "Can I eat strawberries?" the answer is "no" if the procedure is less than 5 days away, but "Yes, but you must stop eating strawberries on [stop_date]" if it is 5 or more days away.

Our knowledge graph is used to store the stoppage time for different food (or more generally, ingestible) items. Each item, based on its entity type (solid, liquid, medicine, supplement), has attributes such as "has seeds" or "has leaves" which determine the stoppage time. The existing FoodOn (He et al., 2018) and FoodKG (Fernández et al., 2020) resources do not cover relations such as "has_skin" or "has_seeds", so we augmented our knowledge graph by asking ChatGPT to list the 200 common food items and beverages, along with 25 common over-the-counter medications and supplements including items that are mentioned in the information sheet provided to the patients. We then used ChatGPT to provide values for the essential food at-

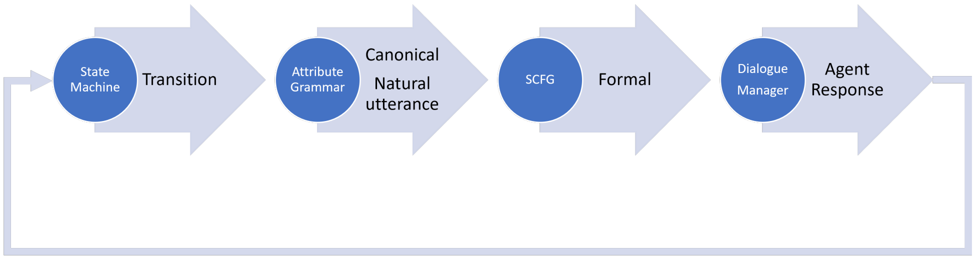

Figure 1: Generating a simulated conversation cycles through four stages. (1) Transitioning in the state machine, which triggers a unique attribute grammar production rule. (2) The production rule translates to a canonical production, which is (3) transformed into a JSON formal representation. (4) The dialogue manager utilizes this representation to create an agent response, and the cycle begins again.

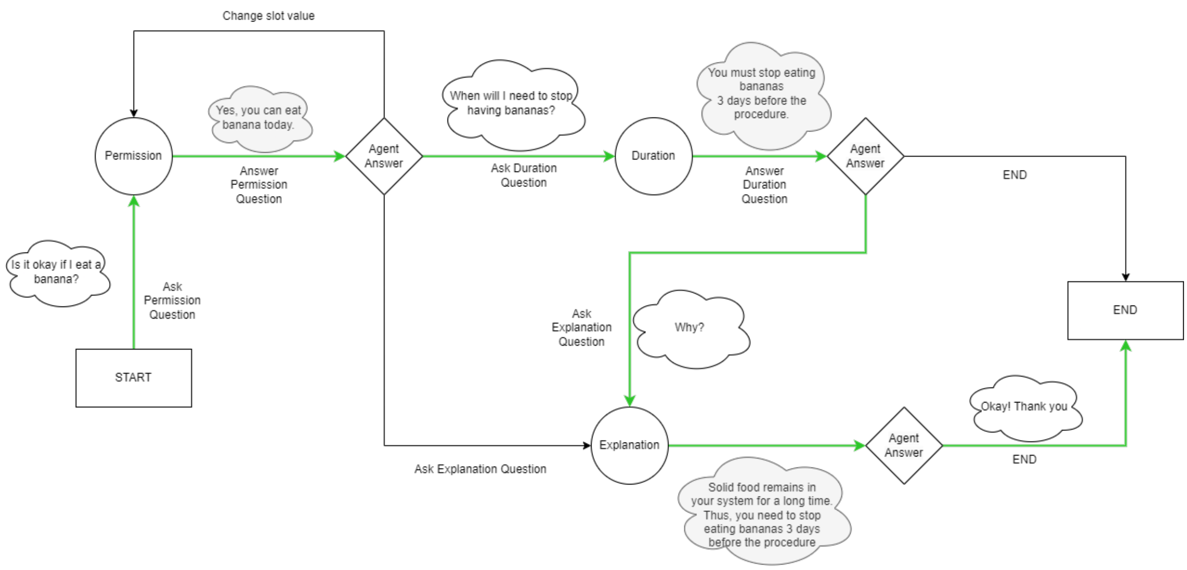

Figure 2: An example conversation generation using the state machine, with utterances emitted on transitions. “Why?” is an incomplete query in isolation, requiring conversation context for full interpretation.

tributes by asking it yes/no questions (e.g., whether apples have seeds), followed by manual inspection to remove erroneous information.

2.3 Simulated and Challenge Datasets

To create a dataset of simulated conversations, we ran the conversational state machine 25,000 times, yielding 11,388 unique conversations that were split 80/5/15 into training, validation, and test sets, respectively; at the turn level, there is 3.84% overlap between our training and test set. For each conversation, the procedure date is set randomly 1 to 10 days in the future.

While the simulated conversations include a variety of synthetic user utterances, they lack linguistic diversity. To enrich these utterances, we used GPT-3.5 and GPT-4 (OpenAI, 2023) to paraphrase 200 conversations from the test set.[1] Since we found the paraphrases from GPT-3.5 to be as good or better than those of GPT-4, we then used GPT-3.5 to paraphrase the entire training set, for a total cost of approximately $10.

To aid in the analysis of our system, we also created a handcrafted dataset of 25 conversations that cover all possible use cases of our system, which we refer to as the challenge set. This set was created by one of the authors without access to the attribute grammar or the automatic paraphrases in the simulated dataset. Of these 25 conversations, 15 are within the scope of the current system, though the conversations often diverge from the simulated ones, especially in their use of follow-up questions.

Sample paraphrased conversations appear in a supplement to the paper along with challenge ones.

3 Experiments

3.1 Models

The goal of the system is to reliably provide accurate and safe answers to user questions. Before training our own models, we first qualitatively tested ChatGPT in a zero-shot setting for our task by providing it relevant information (patient information sheet and procedure date) and asking it questions we envisioned patients asking our system. We found that it did not reliably provide accurate answers to questions requiring temporal reasoning, and that its guardrails against providing medical advice often prevented it from answering questions that the system should be able to answer. We thus moved on to training our own smaller, faster models, which also come with fewer privacy concerns. We used the Hugging Face implementation of pretrained BART (Lewis et al., 2020), fine-tuning the base model (with 140M parameters) for 2 epochs with a learning rate of 1e-5. We compared a semantic parsing model trained on the synthetic user utterances against one trained on the GPT-3.5 paraphrases, and found that the latter achieved 95.0% accuracy on the PARA-GPT-3.5 test set, a 6.5% absolute gain over the former. As a baseline for comparison, we also trained an end-to-end question answering model on the user inputs and system outputs; this model performs NLU implicitly, bypassing the dialogue manager and KB.

[1]We used gpt-4-0314 and gpt-3.5-turbo-0301 model checkpoints via OpenAI's API.

NLU	Soft Match Acc.	BLEU
Explicit	88.4	0.918
Implicit	56.3	0.206

Table 1: Our system with explicit NLU dramatically outperforms the end-to-end, implicit NLU baseline on the PARA-GPT-3.5 test set according to the automatic measures of soft match accuracy (see text) and BLEU.

	Explicit	Implicit
Correct	57	22
Nonresponsive	0	2
Misleading	0	0
Incorrect	3	36

Table 2: The explicit NLU system has only a handful of incorrect responses according to a manual analysis of a test set sample, whereas the end-to-end implicit NLU system responds incorrectly more than half the time, reflecting the inability of pretrained language models to reliably perform temporal reasoning.

3.2 Explicit vs. Implicit NLU

To evaluate the accuracy of our paraphrase-trained model against the implicit NLU baseline on the PARA-GPT-3.5 test set, we employed a soft match for answer polarity, checking if "yes" or "no" is mentioned in the gold answer and also in the predicted answer. We qualitatively checked this soft match metric on a handful of conversations and found it to be generally effective at identifying correct/incorrect matches when the gold answer contains a polarity particle. (Note that when the gold answer does not contain a polarity particle, the soft match metric simply returns false, thereby underestimating true accuracy for both systems.) Table 1 shows that the soft match accuracy for the

	Explicit	Implicit
Correct	17	10
Nonresponsive	34	14
Misleading	4	8
Incorrect	4	27

Table 3: The explicit NLU system has many fewer incorrect responses on the in-scope challenge set in comparison to the end-to-end Implicit NLU system according to a manual analysis.

explicit NLU model is dramatically higher (30% absolute) than the implicit NLU baseline, and has much higher BLEU scores as well.

To verify the results of the automatic evaluation, we conducted a manual analysis of a random sample of 60 items from the test set. Two authors judged the responses as correct, nonresponsive, misleading or incorrect; Table 2 shows the counts of the stricter judge. Without defining these terms, chance-corrected agreement as measured by Krippendorff's α was an acceptable 0.72. On the stricter judge's annotations, a highly significant difference was found between the two systems (ignoring the "misleading" category, which had zero counts for both systems), χ^2 (2, N=60) = 45.4, $p < .001$.

Looking at the answers provided by the implicit NLU baseline, we find that it can reliably answer questions that can be memorized as static FAQs, but it does not reliably answer questions requiring temporal reasoning. For example, whether orange juice is allowed depends on how close one is to the procedure date, and thus the baseline model will sometimes respond to a question like "Can I have orange juice tomorrow?" with "Yes, you may drink orange juice tomorrow" when the correct answer is "No, you cannot have orange juice tomorrow." Such incorrect answers could easily lead to a user being inadequately prepared for a colonoscopy. By contrast, with the explicit NLU model, when the system does not respond correctly, the response is usually recognizable as a non-sequitur, with only a small number of requests to rephrase.

3.3 Challenge Set

We also conducted an exploratory analysis of our system compared to the implicit NLU baseline on the in-scope subset of the challenge set.[2] Three authors judged the responses as correct, nonresponsive, misleading or incorrect; Table 3 shows the majority counts. Without defining these terms, chance-corrected agreement as measured by Krippendorff's α was only 0.48. Nevertheless, a highly significant difference was found between the two systems, χ^2 (3, N=59) = 28.5, $p < .001$. While both systems fared rather poorly overall, as the challenge set included a variety of unanticipated questions and made richer use of the context, the implicit NLU system clearly had many more incorrect responses. Although looking at the handful of incorrect responses our system made turned up some fixable bugs, we expect the misleading responses to be the more serious research challenge, as they depend on how patients might interpret responses in context (Tables 5–6, Appendix A).

[2] On the out-of-scope conversations, the system mostly yielded safe requests to rephrase.

4 Conclusions and Future Work

Our initial development of a neuro-symbolic conversational guide for colonoscopy prep demonstrated that automatic paraphrasing of simulated conversations using GPT models can be successfully used to generate a diverse dataset for drawing meaningful insights into model behavior. We found that GPT and BART language models struggle with temporal reasoning; thus systems that rely upon explicit NLU and temporal reasoning are better suited for answering critical, time-sensitive questions. Further, we found few incorrect responses generated from our system under novel and out-of-scope situations, but misleading ones remain a challenging concern.

In future work, we plan to enhance and extend the system after collecting Wizard-of-Oz data with patients. We expect these system improvements to greatly reduce the prevalence of nonresponsive answers when patients use more contextually varied language, as in our current challenge set experiments. We will also re-evaluate the prevalence of misleading responses and consider implementing steps to filter out such responses. We also plan to experiment with making the system more proactive by quizzing patients on their understanding of the instructions, in order to investigate whether this yields improved understanding leading to improved adherence to the prep protocol.

Acknowledgements

This research was partially supported by an Accelerator award from the Ohio State President's Research Excellence Program.

References

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Timothy W Bickmore, Ha Trinh, Stefan Olafsson, Teresa K O'Leary, Reza Asadi, Nathaniel M Rickles, and Ricardo Cruz. 2018. Patient and consumer safety risks when using conversational assistants for medical information: An observational study of Siri, Alexa, and Google Assistant. *J Med Internet Res*, 20(9):e11510.

Samuel R. Bowman and George Dahl. 2021. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855, Online. Association for Computational Linguistics.

Giovanni Campagna, Agata Foryciarz, Mehrad Moradshahi, and Monica Lam. 2020. Zero-shot transfer learning with synthesized data for multi-domain dialogue state tracking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 122–132, Online. Association for Computational Linguistics.

Carolyn M. Clancy and Jason A. Dominitz. 2021. Texting to Improve Colonoscopy Preparation and Adherence Needs More Study. *JAMA Network Open*, 4(1):e2035720–e2035720.

Emily Dinan, Gavin Abercrombie, A. Bergman, Shannon Spruit, Dirk Hovy, Y-Lan Boureau, and Verena Rieser. 2022. SafetyKit: First aid for measuring safety in open-domain conversational systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4113–4133, Dublin, Ireland. Association for Computational Linguistics.

Jesse M. Ehrenfeld, Warren S. Sandberg, Lisa Warren, Jean Kwo, and Timothy Bickmore. 2010. Use of a computer agent to explain anesthesia concepts to patients. In *Annual meeting of the American Society of Anesthesiologists*, volume 3.

Aarohi Srivastava et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models.

Javier Fernández, Marta Villegas, Maria-Esther Vidal, and Albert Meroño-Peñuela. 2020. FoodKG: A new linked open data resource for food data science. In *Proceedings of the 19th International Semantic Web Conference (ISWC 2020)*, pages 428–444.

Yongqun He, Chris Mungall, Elizabeth A. F. Mejino Jr., James A. Blake, Michael Bada, and Barry Smith. 2018. FoodOn: a harmonized food ontology to increase global food traceability, quality control and data integration. *NPJ Science of Food*, 2(1):23.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue.

Dan Jurafsky, Liz Shriberg, and Debra Biasca. 1997. *Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coders Manual, Draft 13*. University of Colorado at Boulder & SRI International. Available online: `https://web.stanford.edu/~jurafsky/ws97/manual.august1.html`.

J. F. Kelley. 1984. An iterative design methodology for user-friendly natural language office information applications. *ACM Trans. Inf. Syst.*, 2:26–41.

Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Huang. 2023. A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 431–469, Toronto, Canada. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, and Veselin Stoyanov. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 610–621, Online. Association for Computational Linguistics.

Nadim Mahmud, David A. Asch, Jessica Sung, Catherine Reitz, Mary S. Coniglio, Caitlin McDonald, Donna Bernard, and Shivan J. Mehta. 2021. Effect of Text Messaging on Bowel Preparation and Appointment Attendance for Outpatient Colonoscopy: A Randomized Clinical Trial. *JAMA Network Open*, 4(1):e2034553–e2034553.

Nadim Mahmud, Sahil D. Doshi, Mary S. Coniglio, Michelle Clermont, Donna Bernard, Catherine Reitz, Vandana Khungar, David A. Asch, and Shivan J. Mehta. 2019. An automated text message navigation program improves the show rate for outpatient colonoscopy. *Health Education & Behavior*, 46(6):942–946. PMID: 31431077.

OpenAI. 2023. GPT-4 technical report.

Richard Shin, Christopher Lin, Sam Thomson, Charles Chen, Subhro Roy, Emmanouil Antonios Platanios, Adam Pauls, Dan Klein, Jason Eisner, and Benjamin Van Durme. 2021. Constrained language models yield few-shot semantic parsers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7699–7715, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yushi Wang, Jonathan Berant, and Percy Liang. 2015. Building a semantic parser overnight. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1332–1342, Beijing, China. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and social risks of harm from language models.

A Appendix

Food Category

Permission U: Can I eat a strawberry tomorrow? S: No, you may not eat strawberries tomorrow.

Duration U: When can I eat a strawberry? S: You can have strawberries up to 5 days before the procedure. Your procedure is on 6th April 2023 at 5:00pm, you can have strawberries up to 5:00pm on 1st April 2023.

Explanation U: Why can I not eat a strawberry tomorrow? S: Your procedure is 3 days away and you need to stop having strawberries, as it can cause errors in the procedure and hence should not be consumed.

Task Category

Description U: What is in the instructions? S: The instructions guide you through the preparation for the procedure.

Duration U: By when should I read the instructions? S: You should read the instructions at least a week before your procedure.

Explanation U: Why should I read the instructions? S: The instructions help you prepare for the procedure. It is important that you understand the instructions and follow them to ensure that you are prepared for the procedure.

Procedure Category

Description U: What is a colonoscopy? S: A colonoscopy is an inspection of your colon and your large intestine.

Duration U: When is my procedure? S: Your procedure is on 6th April 2023 at 5:00pm.

Explanation U: How long is my procedure? S: The procedure should be 2 to 4 hours long.

Table 4: Sample responses for different categories and subtypes of user questions. Notice how food related questions are time sensitive in nature, while task questions can be answered without logical reasoning.

User:	Can I have ice cream today?
System:	Your procedure is four days away. Yes, you can still have ice cream.
User:	What if it has nuts in it? You told me to avoid nuts.
System:	You can have ice cream up to 2 days before the procedure. Your procedure is on Friday, 19 May at 09:30pm, you can have ice cream up to 09:30pm on Wednesday, 17 May.

Table 5: Example misleading system response from challenge set. After the system correctly says that the patient may have ice cream 4 days in advance of the procedure, the patient asks whether ice cream is still allowed if it contains nuts. Here, the system should say no, but instead says that the patient can have ice cream up to 2 days before the procedure. While this response is literally correct, it does not directly answer the patient's question; as such, it could easily mislead the patient into thinking an indirect positive answer had been given (via a Gricean quantity implicature).

User:	Do I have to stop taking my insulin before my colonoscopy?
System:	I am sorry, but I couldn't understand. Could you please rephrase?
User:	Can you contact my doctor?
System:	I am sorry, but I cannot provide information about insulin.
User:	Thank you. What should I do until I hear from them?
System:	The prescription contains Mirolax, Dulcolax and Gatorade.

Table 6: Example from out-of-scope portion of challenge set with incorrect responses that are not misleading. In this scenario, the user asks the system about adjusting their insulin dosage. Ideally our system should direct the user to seek approval from the doctor. However, for the first question, our system creates an illegal canonical string and replies that it does not understand the user query. For the next question, it creates a legal canonical string, and responds that it cannot provide information about insulin as it was not present in our knowledge graph. In the final turn, the system misunderstands the question as pertaining to the prep mix and responds with a non-sequitur.