EVALUATING LARGE LANGUAGE MODELS FOR COLONOSCOPY PREPARATION ASSISTANCE: CORRECTNESS AND DIVERSITY IN SYNTHETIC DIALOGUES

A PREPRINT

Tomiris Kaumenova, Subhankar Chakraborty, Eric Fosler-Lussier, Kebire Gofar, Isaiah Metcalf, Andrew Perrault, Michael White
The Ohio State University
Columbus, OH, USA
{kaumenova.1, fosler-lussier.1, perrault.17, white.1240}@osu.edu
{subhankar.chakraborty, kebire.gofar, isaiah.metcalf}@osumc.edu

November 19, 2025

ABSTRACT

Background: Colorectal cancer is the third leading cause of cancer-related deaths in the United States, and colonoscopy remains the gold standard for early detection and prevention. However, many procedures are postponed due to inadequate bowel preparation, a preventable failure often caused by patients' difficulty in understanding or following written prep instructions. Prior interventions such as reminder apps and instructional videos have improved adherence only modestly, largely because they cannot answer patients' specific questions. Recent advances in large language models (LLMs) raise the possibility of developing conversational assistants that can provide an interactive support to patients in procedure preparation.

Objective: This study evaluated correctness and diversity of synthetic dialogues generated by leading LLMs acting as both simulated AI Coaches and patients for colonoscopy preparation.

Methods: Four leading LLMs, OpenAI's o3 and GPT-4.1, Meta's Llama 3.3 70B, and Mistral's Large-2411 were used to generate 250 patient-AI Coach dialogues per model. Prompts were designed to elicit diverse patient questions about diet, medications, and other prep-related topics. Human raters, including medical experts, evaluated responses for correctness, error type, and potential harmfulness. Automatic evaluation using an LLM-as-a-judge approach complemented human evaluation.

Results: Leading models approached but did not achieve adequate performance. Closed-weight models (GPT-4.1, o3) outperformed open-weight models (Llama, Mistral) on correctness, while multi-prompt generation substantially improved question diversity. All models produced harmful errors, primarily due to omissions or misinterpretations of prep instructions.

Conclusions: While LLMs demonstrate strong potential for colonoscopy preparation support, none are yet reliable enough for unsupervised deployment in patient-facing contexts without effective safety layers.

Keywords Colorectal cancer · Colonoscopy · Health communication · Artificial intelligence · Virtual assistants · Large language models

1 Introduction

Colorectal cancer is the third leading cause of cancer-related deaths in the United States, and colonoscopy remains the gold standard for early detection and prevention. Tens of thousands of colonoscopy and endoscopy procedures are performed each year just at Ohio State's Wexner Medical Center. Yet despite their efficacy, around 20% of these tests are postponed because patients have not read, understood, and correctly carried out prep instructions. Inadequate colonoscopy prep has

economic, health-related and social costs. Economic costs include missed days from taking the prep, taking time off for the procedure, travel costs to get to the appointment and back and the cost of having an accompanying person to drive (which is a requirement for the procedure). For the healthcare system, they lead to a missed appointment incurring resultant wasted resources. Health-related costs include potentially reduced compliance with screening guidelines (only about 70% comply [1]) and thus risk of missing polyps. Social costs include frustration on the part of patients, in addition to patients' sharing their experience with others, which can deter others from getting a colonoscopy.

A key driver of this preventable problem is information overload: patients receive lengthy and complex written instructions days or weeks before their procedure, making it difficult to recall and correctly execute each step at the right time. For example, the information sheets patients are given typically instruct them to drink half of a prep solution for clearing out the colon at 6 p.m. the evening before their scheduled procedure and the other half six hours before the procedure time. However, some patients nevertheless show up for their procedure with the second half of their prep solution in hand—falsely assuming that they are supposed to take it after arriving at the procedure facility—and end up having to reschedule the procedure.

Past interventions, such as reminder apps [2], instructional videos, and automated text messages [3], have improved prep adherence only modestly [4], largely because they lack the capacity for interactive question answering. The ability to answer questions appears critical: when automated systems could not respond to patients [5], improvements in adherence disappeared [6]. Thus, a conversational assistant that can safely respond to patient queries is a promising next step toward supporting patients in colonoscopy preparation.

Recent progress in large language models (LLMs) makes this prospect newly feasible. Frontier models such as o3 have demonstrated strong reasoning on clinical benchmarks like OpenAI's HealthBench [7] and MedQA [8]. If these models can perform complex diagnostic reasoning, a natural question arises: how well can they handle the simpler task of colonoscopy preparation coaching? This task primarily tests prompt adherence, prep instruction understanding, temporal and common sense reasoning, and dialogue communication, rather than clinical inference. Our study attempts to address that question. Building on prior work by Arya et al. [9], which introduced a neuro-symbolic conversational guide for colonoscopy prep, we explore how LLMs perform on the same challenge. Our preliminary experiments show that recent LLMs substantially outperform earlier models, particularly on temporal reasoning and conversational abilities, which are core difficulties in the task. This motivates a systematic evaluation of LLMs.

In our study, we selected frontier LLMs to generate synthetic dialogues, where models simulated both patients and "AI Coaches" aka colonoscopy prep assistants. These dialogues were evaluated by both human raters and LLM-based raters, enabling us to evaluate not only the dialogue quality but also the ability of LLMs to stand in for human raters. The main benefit of synthetic data is the preservation of patient privacy and safety. Although LLM-generated questions may not fully capture the complexity and variability of questions that real patients may ask, they cover a broad range of topics, offering a challenge and a testing ground for models in our task. The study evaluated dialogues along two dimensions: (1) correctness of AI Coach responses relative to prep instructions, and (2) diversity and difficulty of patient questions. We aim to better understand how close current LLMs are to serving as safe and reliable conversational assistants for real-world healthcare communication tasks, in the context of a colonoscopy procedure.

2 Methods

2.1 Dialogue Generation

Four leading LLMs[1] with strong reasoning and instruction following capabilities were selected for dialogue generation: OpenAI's o3 and GPT-4.1 Instruct, Meta's Llama 3.3 70B Instruct, and Mistral's Large-2411. Among the models, o3 uses explicit reasoning, while the others do not; GPT-4.1 and o3 are closed-weight models, whereas Mistral Large and Llama 3.3 are open-weight. Dialogue generation was conducted in a turn-by-turn manner, with separate prompts for patient questions and AI Coach responses (see examples in A1). For AI Coach responses, we incorporated few-shot dialogues and chain-of-thought reasoning, as preliminary experiments indicated that these strategies reduced errors. Few-shot examples modeled fully correct responses and illustrated how to avoid common error types. The AI Coach prompt contained full prep instructions (see A2). Patient prompts were designed to encourage diversity and realism, so at each patient's turn, one of several prompt types was selected at random:

- **Food-related** prompts generated questions about eating various foods at a specific time before the procedure, or about mistakenly eating restricted foods.

[1]GPT-5 was released after our data generation and human evaluation were completed. Therefore, we report results for this newer model from automatic evaluation only.

- **Theme-based** prompt generated questions on a specific topic unrelated to foods, such as sedation, allergies, medication, etc.
- **Follow-up** prompt generated questions based on the previous turn.

Each prompt generated 10 questions per turn, and one of them was selected at random to continue the dialogue; this is a simple approach to addressing a well-known mode collapse problem [10], where a model repeatedly produces a limited set of highly similar outputs. Each dialogue consisted of 3–7 question–answer pairs. In the prep instructions, dietary restrictions begin five days before the procedure, marking the start of the preparation period most relevant to patients. Accordingly, each dialogue was assigned a time point between 21 and 1 day before the procedure, with dialogues set within the 5-day window occurring twice as frequently as those set earlier, as these tend to involve more challenging or clinically relevant questions. A total of 250 dialogues were generated by each model, with the same model acting as both patient and AI Coach in a dialogue.

2.2 Human Evaluation

Four human raters participated in dialogue evaluation: one attending physician (expert), one medical resident, one medical student (experienced raters), and one layperson (the first author). For each model, 50 dialogues were sampled at random for evaluation. A subset of 50 dialogues was annotated by all four raters, with dialogues evenly distributed across models. The rest of the dialogues were split evenly between experienced raters and layperson. To support consistency, raters were provided with detailed evaluation guidelines, which were refined following a pilot trial. All raters except the lead author were blinded to the model source of the dialogues.

Raters were instructed to determine whether each AI Coach response was correct or incorrect, in reference to prep instructions. Incorrect responses were classified into one of the following categories (see examples in A3):

- **Temporal** errors: inconsistencies with the procedure timeline.
- **Extraneous** information: inclusion of content outside the scope of prep instructions (subcategorized as factually correct vs. factually incorrect).
- **Reasoning** errors: contradictions with prep instructions or faulty reasoning (e.g., about restricted foods).
- **Omissions**: exclusion of relevant or helpful information.
- **Other** errors: irrelevant, disfluent, or ambiguous responses.

Each error was also judged for **harmfulness**. **Harmful** errors were those potentially leading to a rescheduled procedure or more severe consequences, and **harmless** errors were those that did not significantly impact the patient or the procedure (see examples in A3). In addition, the lead author flagged **false refusals**, defined as instances where a model refused to answer despite the question being answerable. Disagreements among raters were resolved by majority vote; in cases without a consensus, the rating of the most experienced rater was used.

Unlike AI Coach responses, patient questions were not judged for correctness, since in practice patients may ask irrelevant or ambiguous questions, and providers are still expected to respond appropriately. Instead 30 questions from each model were sampled at random and categorized into one of three levels:

- **Easily answerable** from prep instructions;
- **Not easily answerable**, such as those that require reasoning about the timeline or fiber content of various foods;
- **Not answerable** at all (e.g sedation details, anxiety management, etc.), which require deferral to a provider.

Such classification aimed to determine whether models were presented with questions of varying degrees of difficulty, as might be expected in a real life scenario.

2.3 Automatic Evaluation

We employed an LLM-as-a-Judge approach to evaluate a large set of dialogues, in which each AI Coach response was independently evaluated by a language model. The evaluation prompt (A1) included full prep instructions, few-shot examples of common errors, and correct responses, and it required the model to provide an explanation whenever it judged a response as incorrect. These explanations were retained for qualitative error analysis.

Additionally, we assessed the diversity of patient questions between models using the type–token ratio (percentage of unique unigrams and bigrams, referred to as Distinct-1 and Distinct-2, respectively) and the entropy measure. We compared question diversity generated from our multi-prompt, 10-question strategy to a single-prompt, single-question approach,

which served as our baseline. Baseline dialogues were generated using Llama. Most frequent tokens are listed in the appendix (A4).

2.4 Statistical Analysis

We estimated the required sample size for automatic evaluation using a chi-square power analysis with a significance level of .05 and statistical power of 0.80. Pilot estimates of observed proportions yielded a Cohen's effect size of 0.094, resulting in a minimum sample size of 1,228 responses per model. Assuming an average of 5 responses per dialogue, this corresponds to 245.6 dialogues per model, which was rounded to 250. A Bonferroni-adjusted significance level was applied to maintain the desired power when conducting multiple comparisons. Differences in proportions of correct responses across models were tested using overall chi-square test of independence and pairwise tests of proportion with Bonferroni-adjusted significance level of .05. Pairwise tests were conducted against the best-performing model. When comparing more than two models as evaluators, we again used overall chi-square test of independence as well as Cochran's Q test, with Bonferroni correction. Pairwise tests were not needed.

For human evaluation, we estimated the sample size required to detect significant inter-rater agreement using Bloch and Kraemer's formula for Krippendorff's alpha ([11]), assuming a significance level of .05. However, we did not anticipate such high correctness rates observed in our data and instead calculated Gwet's AC1 as the measure of inter-rater agreement. This statistic is more robust to skewed category distributions [12] and is reported in the Discussion (4).

Dialogue example: 1 | **Date**: 12 days before procedure | **Model**: GPT-4.1 Instruct

Turns	**Note**
AI Coach: Hello! I'm here to help you prepare for your colonoscopy. What questions or concerns can I assist you with today?	Greeting.
Patient: Can I still eat spinach salad with vinaigrette?	Not easily answerable from instructions.
AI Coach: Yes, you can still eat spinach salad with vinaigrette. Dietary changes, such as avoiding leafy greens like spinach, start 5 days before your procedure.	Correct: reminds patient that spinach is restricted within a 5-day window.
Patient: Got it, thanks for clarifying. What exactly is in the bowel prep drink?	Easily answerable from instructions.
AI Coach: Your bowel prep will include MiraLAX (polyethylene glycol 3350, 8.3 oz or 238 g), 4 Dulcolax tablets (5 mg each), and 64 oz of sports drink (not red or purple). These ingredients are mixed to clean your colon before the procedure.	Correct.
Patient: Will the MiraLAX irritate my ulcerative colitis?	Unanswerable from instructions.
AI Coach: Please contact your health care provider to discuss your specific medical condition and get guidance on your bowel prep.	Correct: directs patient to a provider.

3 Results

3.1 Question Difficulty & Diversity

Human evaluation of question difficulty (Table 1) shows that questions are balanced across difficulty levels for all models, with o3 and Mistral Large generating slightly more challenging or unanswerable ones. Automatic diversity metrics, such as entropy and Distinct-1/2 (Table 2), indicate that using Llama with elaborate prompts resulted in nearly twice as diverse questions as with baseline prompts. While o3 cannot be directly compared to the baseline condition (since the baseline was generated using Llama), both o3 and GPT-5 produced the most lexically diverse patient questions, achieving the highest Distinct-1/2 and entropy scores among all models. Appendix (A4) lists the most frequent tokens across models, revealing a clear qualitative pattern: stronger models such as o3, GPT-5, and GPT-4.1 generated more specific terms (e.g., chicken, coffee, white), whereas Llama and the baseline relied more on generic ones (e.g., colonoscopy, procedure). Although Mistral produced a range of specific items (e.g., popcorn, wine, salad), its token frequencies were more uneven, whereas stronger models exhibited more balanced lexical distributions.

Table 1: Human evaluation of patient question difficulty relative to prep instructions (N=30 per model).

Model	Easily answerable, n	Not easily answerable, n	Unanswerable, n
Mistral Large	6	12	12
Llama 3.3 70B	10	8	12
GPT-4.1	11	10	9
o3	5	12	13

Table 2: Automatic evaluation of question diversity across models.

Model	Distinct-1	Distinct-2	Entropy	N
Baseline	0.03	0.10	7.22	1227
Llama 3.3 70B	0.07	0.25	8.00	1220
Mistral Large	0.11	0.40	8.70	1250
GPT-4.1	0.11	0.39	8.59	1292
o3	**0.16**	**0.51**	**9.14**	1262
GPT-5	**0.16**	0.47	8.93	1219

3.2 Overall Response Accuracy

Table 3 presents human evaluation results for dialogue- and turn-level accuracy relative to colonoscopy preparation instructions. At the dialogue level, o3 achieved the highest accuracy (76%), followed by GPT-4.1 (56%), Llama 3.3 70B (34%), and Mistral Large (14%). A similar pattern was observed at the turn level, with o3 (92.6%) outperforming GPT-4.1 (85.2%), Llama 3.3 70B (76.8%), and Mistral Large (67.3%). When counting responses that were extraneous but still correct, absolute correctness rose just slightly for all models.

To extend these findings to a larger dataset, we conducted automatic evaluation using the LLM-as-a-Judge approach. We employed four LLMs as evaluators: DeepSeek-R1, OpenAI's GPT-5 and o3, and Meta's Llama 3.3 70B (Table 4). While none of the models differed significantly (Cochran's Q test: Q=9.4, P=.02; chi-square test: χ^2_3=5.2, P=.16), we ultimately selected DeepSeek-R1 as the evaluator, as it achieved the highest accuracy and F_1 score against human judgments. Automatic evaluation (Table 5) followed similar trends to those observed in human evaluation. At both dialogue- (72.8%) and turn-level (93.1%), GPT-5 was the best-performing model, followed by o3 (63.2%, 90.7%), with no significant difference between the two (χ^2_1=4.4, P=.36), while GPT-4.1 (58.8%, 88.9%, P=.003) and Llama 3.3 70B (41.6%, 81.1%, P<.001) performed significantly worse. Mistral Large showed the lowest performance (60%, 61.8%, P<.001).

Table 3: Human evaluation of model performance on dialogue- and turn-level accuracy relative to prep instructions, and absolute correctness including extraneous but correct information. Pairwise comparisons against the best model (o3) were conducted at turn-level using two-sample proportion test with Bonferroni correction.

Model	Dialogue-level Accuracy, n/N (%)	Turn-level Accuracy, n/N (%)	Absolute Correctness, n/N (%)	χ^2_1	adj. P value (vs. best)
Mistral Large	7/50 (14)	169/251 (67.3)	177/251 (70.5)	45.7	<.001
Llama 3.3 70B	17/50 (34)	182/237 (76.8)	187/237 (78.9)	21.4	<.001
GPT-4.1	28/50 (56)	213/250 (85.2)	217/250 (86.8)	5.9	.048
o3	**38/50** (76)	**214/231 (92.6)**	**217/231 (93.9)**		

Table 4: LLM performance as error predictor on 50 cross-annotated dialogues (N=252 AI Coach responses).

Model	Accuracy, n (%)	Precision	Recall	F_1 score
Llama 3.3 70B	199 (79)	**0.63**	0.40	0.49
GPT-5	207 (82.1)	0.45	0.55	0.49
o3	213 (84.5)	0.51	**0.58**	0.54
DeepSeek-R1	**217 (86.1)**	0.56	**0.58**	**0.57**

Table 5: Automatic evaluation of model performance at dialogue- and turn-level accuracy. Pairwise comparisons against the best model (GPT-5) were conducted at turn-level using two-sample proportion test with Bonferroni correction.

Model	Dialogue-level Accuracy, n/N (%)	Turn-level Accuracy, n/N (%)	χ^2_1	adj. P value (vs. best)
Mistral Large	150/250 (60)	773/1250 (61.8)	341.9	<.001
Llama 3.3 70B	104/250 (41.6)	990/1220 (81.1)	76.7	<.001
GPT-4.1	147/250 (58.8)	1148/1292 (88.9)	13.2	.001
o3	158/250 (63.2)	1145/1262 (90.7)	4.4	.14
GPT-5	**182/250 (72.8)**	**1135/1219 (93.1)**		

3.3 Error Analysis

Error breakdowns are reported in Table 6. The most frequent error type across models was omission, ranging from 2.6% (o3) to 13.9% (Mistral Large). o3 and GPT-4.1 consistently had the lowest error rates across categories. Table 7 shows that both number and proportion of harmful errors varied notably by model. Mistral Large produced the highest number of total errors but surprisingly a lower proportion of harmful ones (37.5%). In contrast, o3 had the fewest total errors but the highest share that were harmful (63.2%).

Table 6: Error types as rates across models, where n means the number of turns with at least one error of such type.

Model	Temporal, n (%)	Extraneous-correct, n (%)	Extraneous-incorrect, n (%)	Reasoning, n (%)	Omission, n (%)	Other, n (%)	N
Mistral Large	21 (8.4)	8 (3.2)	7 (2.8)	26 (10.4)	35 (13.9)	3 (1.2)	251
Llama 3.3 70B	10 (4.2)	5 (2.1)	8 (3.4)	8 (3.4)	24 (10.1)	3 (1.3)	237
GPT-4.1	8 (3.2)	4 (1.6)	**2 (0.8)**	**1 (0.4)**	23 (9.2)	**1 (0.4)**	250
o3	**1 (0.4)**	**3 (1.3)**	4 (1.7)	3 (1.3)	**6 (2.6)**	2 (0.9)	231

Table 7: Proportion of harmless vs. harmful errors per model.

Model	N	Harmless, n (%)	Harmful, n (%)
Mistral Large	136	85 (62.5)	**51 (37.5)**
Llama 3.3 70B	60	27 (45.0)	33 (55.0)
GPT-4.1	39	20 (51.3)	19 (48.7)
o3	**19**	7 (36.8)	12 (63.2)

4 Discussion

4.1 Diversity of Patient Questions

Prompt design had a clear effect on the diversity of patient queries. A simple baseline prompt produced repetitive questions, whereas our multi-prompt strategy, where each turn generated ten candidate questions from food-related, thematic, or follow-up prompts, led to greater lexical and topical variety, helping mitigate the mode collapse problem. Automatic metrics (distinct-n and entropy) confirmed this increase in diversity. Human evaluation further showed that our approach yielded a balanced distribution of questions: some were easily answerable from the preparation instructions, others required indirect reasoning (e.g., about timing or food composition), and some were not answerable at all (e.g., anxiety or sedation concerns). This suggests that our generation setup exposed models to a realistic range of question types and difficulty, mirroring those encountered in actual patient–provider interactions. Consequently, the resulting dialogues provide a more robust basis for evaluating model's performance in this task.

4.2 Correctness of AI Coach Responses

This task is relatively straightforward: responses are evaluated against a fixed set of colonoscopy preparation instructions. However, the best-performing models (o3, GPT-4.1, GPT-5) approach but do not reach 100% accuracy, with differences

between them often not statistically significant. Smaller open-source models (Llama and Mistral) are not a viable alternative due to a substantially higher number of errors.

Error patterns reveal important distinctions across models. Temporal errors are much less frequent in stronger models; however, instruction-following errors (omission, extraneous information, faulty reasoning) remain prevalent even when temporal accuracy is high, suggesting that models could benefit from further fine-tuning for instruction adherence. Additionally, o3 produces a large proportion of harmful errors despite high overall accuracy, showing that accuracy and safety are not equivalent. In fact, the existence of harmful errors demonstrates the potential risks of deploying even highly accurate models in patient-facing contexts without additional safeguards. This concern is especially relevant in scenarios where patients already have a solid understanding of preparation instructions, since models could inadvertently misdirect otherwise well-prepared patients. For less-prepared patients, however, model assistance could still be mildly beneficial, although not enough on its own to ensure a successful prep. Patients may need reminders [3] to complete preparation steps which would in turn require a more complex interactive system capable of supporting these mechanisms.

Common harmful errors involved omissions of critical safety or procedural information. Examples include failing to mention the need for a designated driver at check-in, omitting the two-hour restriction on liquids (with the exception of small sips of water for medication), or failing to warn about future dietary restrictions (e.g., avoiding red or purple liquids and dairy products on the day before the procedure). Other errors concerned medication guidance, such as not advising patients to consult their provider about medication changes or neglecting to state specific rules for oral diabetes medications. In several cases models failed to instruct patients to contact a provider after mistakenly consuming a prohibited food item. There were also some patient questions phrased in terms of weekdays (e.g., "this Sunday") rather than in terms of the number of days left before the procedure (e.g., "3 days before"); such questions create ambiguity that a robust system should detect, but even stronger models were not able to do so.

Model behavior differed in style, too: Mistral's responses were often extremely brief, offering no explanation derived from the instructions and sometimes consisting of single-word replies. Mistral refused many dietary questions when dialogues were set many days before the procedure and conflated patient and provider roles more often than others. Llama tended to be verbose and occasionally unnatural in phrasing, whereas GPT and o3 produced more natural responses, which were lengthy only when warranted.

4.3 Evaluation

Both automatic and human evaluation were essential and complementary. Human evaluation provided insights into error types and their potential harmfulness, while automatic evaluation provided a scalable way to evaluate large numbers of dialogues beyond what's feasible with expert raters. Interrater agreement among human raters was substantial (expert and experienced raters: AC1=0.74, 95% CI (0.68, 0.80), percent agreement=0.82; all raters: AC1=0.68, 95% CI (0.62, 0.75), percent agreement=0.79), consistent with our expectation that evaluating factual correctness in this task is easier than clinical diagnosis.

Interestingly, we found systematic differences between DeepSeek-R1 (R1) and expert raters. R1's strict adherence to prep instructions led it to identify certain erroneous responses that experts judged acceptable, suggesting that automatic evaluators may in fact be more reliable at guideline and prep fidelity. Some of these cases involved minor issues: responses that were incomplete or lacked explanation. In one case, the response model misclassified Tylenol as a non-steroidal anti-inflammatory drug (NSAID), which constitutes a factual error but not a harmful one, as experts agreed the real safety concern lies with anticoagulants such as Warfarin. In another, the response model failed to specify that instant ramen should be made from low-fiber noodles; while technically correct, this omission was considered inconsequential, since most instant ramen products already meet that criterion. R1 also penalized a few responses that were truthfully correct, reflecting an overly rigid standard of adherence.

This finding does not undermine our earlier results from human evaluation. Importantly, evaluating a large set of responses offline in a spreadsheet differs from interacting with real patients, where pragmatic communication takes precedence. The lay rater showed similar tendencies to R1, prioritizing strict adherence and making only a single factual mistake (incorrectly believing that gummy bears were allowed the day before the procedure). This further suggests that factual evaluation of dialogues in reference to prep instructions is a relatively easy task for humans to learn, yet even the strongest models are not perfect at it.

4.4 Safety Filter & False Refusals

Applying an LLM-as-judge filter (Table 8) reduced both overall error and harmfulness rates for all models. The filter consisted of replacing all AI Coach responses that the LLM judge classified as incorrect with a deferral to a provider, then recomputing turn-level error rate, harmfulness and false refusal rate on the responses. The relative improvement was

greatest for Mistral Large (error rate reduced from 32.7% to 11.2%) and Llama 3.3 70B (23.2% to 13.1%). o3 remained the most accurate before and after filtering (7.4% to 5.6%). Similarly for harmfulness, the filter helped weaker models like Llama and Mistral, but offered little or no benefit for stronger models such as GPT-4.1 or o3, where it removed only one harmful error per model. False refusals were rare across all dialogues and models; however, for weaker models, the false refusal rate increased substantially after filtering. If such a filter were deployed in patient interactions, it would likely be counterproductive: if patients are redirected to contact a provider for questions that could have been answered from the prep instructions, the usefulness of the AI Coach as an assistant for colonoscopy prep is undermined. Most importantly, the filter failed to eliminate a sufficient number of harmful errors, making it impractical as a safety mechanism.

False refusal example: 1 | **Date**: 9 days before procedure | **Model**: Llama 3.3 70B Instruct

Turns	**Note**
Patient: What about a bowl of chili 3 days before?	Not easily answerable.
AI Coach: Sorry, I don't have such information. Please contact your doctor for such questions.	**False refusal**: Chili is permissible 3 days before procedure as long as it does not contain high-fiber ingredients.

Table 8: Error rate, harmfulness and false refusal rate before and after filtering using an LLM-as-judge approach, where n denotes the number of turns with at least one error, the number of turns with at least one harmful error, and the number of turns with a false refusal, respectively.

Filtering		Before			After		
Model	N	Error rate, n (%)	Harmfulness, n (%)	False refusal rate, n (%)	Error rate, n (%)	Harmfulness, n (%)	False refusal rate, n (%)
Mistral Large	251	82 (32.7)	36 (14.3)	13 (5.2)	28 (11.2)	13 (5.2)	67 (26.7)
Llama 3.3 70B	237	55 (23.2)	30 (12.7)	8 (3.4)	31 (13.1)	18 (7.6)	32 (13.5)
GPT-4.1	250	37 (14.8)	18 (7.2)	2 (0.8)	28 (11.2)	17 (6.8)	11 (4.4)
o3	231	**17 (7.4)**	**11 (4.8)**	**0 (0)**	**13 (5.6)**	**10 (4.3)**	**4 (1.7)**

4.5 Limitations

This study has several limitations. First, the distribution of synthetic patient questions may not fully capture the range or underlying motivations behind real-world patient inquiries. The factors driving patient unpreparedness are not fully understood and can stem from causes beyond informational gaps, such as anxiety, or noncompliance, that are difficult to model through prompt-based generation. Additionally, real patient questions can prove to be more challenging, as they can be unclear or require follow-up clarification, although expert raters qualitatively found the synthetic questions to be otherwise plausible. Second, while model responses were evaluated for factual accuracy relative to preparation instructions, other aspects of communication quality, for instance empathy, were not assessed. Third, our analysis did not explore the relationship between question difficulty or topic and response correctness or harmfulness. Finally, diversity measures like type–token ratio and entropy quantify surface-level lexical variety but do not necessarily reflect semantic or pragmatic diversity. We did not have access to a corpus of messages or phone call transcripts between patients and providers as these are privacy-protected; consequently, we were not able to quantitatively compare the synthetic conversations with real ones.

4.6 Conclusions

Taken together, our results demonstrate that LLMs approach but do not yet achieve adequate performance in this task, given the number of harmful errors. Human and automatic evaluations together provide a more nuanced understanding of model behavior, balancing interpretability and scalability. Prompt-based automatic filtering improves performance only for open models and does not fully prevent harmful errors, suggesting that the practical benefits of such straightforward methods remain limited. Future work should explore improving evaluator models through self-training [13]. Another promising direction is calibrating model's confidence to improve automatic dialogue evaluation. Overly cautious models risk unnecessary refusals, while overconfident ones can introduce harmful misinformation. Future research should aim to develop adaptive systems capable of calibrating their confidence based on the context and risk level of a patient's question. Finally, testing on real patient queries will be necessary to validate model evaluation and help align models more closely with real patient expectations.

4.7 Acknowledgements

We thank Arkobrato Gupta for his guidance and feedback on the statistical analysis.

4.8 Data Availability

The synthetic patient–AI Coach dialogues generated and evaluated in this study will be made publicly available at: https://github.com/sirimott/prep-coach-llms.

4.9 Authors' Contributions

TK and MW conceived and designed the study. TK conducted data generation, evaluation, and analysis, and drafted the manuscript, under the supervision and guidance of MW. SC, KG, and IM contributed to dialogue annotation and provided feedback on annotation guidelines. AP and EF contributed through prior discussions and general support. All authors reviewed and approved the final version of the manuscript.

4.10 Funding

This research received no external funding.

4.11 Conflicts of Interest

None declared.

References

[1] Richardson LC, King JB, Thomas CC, Richards TB, Dowling NF, King SC. Adults who have never been screened for colorectal cancer, behavioral risk factor surveillance system, 2012 and 2020. *Prev Chronic Dis.* 2022;19. doi: 10.5888/pcd19.220001.

[2] Zander QEW, Reumkens A, van de Valk B, Winkens B, Masclee AAM, de Ridder RJJ. Effects of a personalized smartphone app on bowel preparation quality: Randomized controlled trial. *JMIR Mhealth Uhealth.* 2021;9:e26703. doi: 10.2196/26703.

[3] Mahmud N, Doshi SD, Coniglio MS, et al. An Automated Text Message Navigation Program Improves the Show Rate for Outpatient Colonoscopy. *Health Educ Behav.* 2019;46:942-946. PMID: 31431077.

[4] Sharara AI, Chalhoub JM, Beydoun M, et al. A customized mobile application in colonoscopy preparation: A randomized controlled trial. *Clin Transl Gastroenterol.* 2017;8:e211. doi: 10.1038/ctg.2016.65.

[5] Mahmud N, Asch DA, Sung J, et al. Effect of Text Messaging on Bowel Preparation and Appointment Attendance for Outpatient Colonoscopy: A Randomized Clinical Trial. *JAMA Netw Open.* 2021;4. doi: 10.1001/jamanetworkopen.2020.34553.

[6] Clancy CM, Dominitz JA. Texting to Improve Colonoscopy Preparation and Adherence Needs More Study. *JAMA Netw Open.* 2021;4. doi: 10.1001/jamanetworkopen.2020.35720.

[7] Arora RK, Wei J, Hicks RS, et al. HealthBench: Evaluating Large Language Models Towards Improved Human Health. *aRxiv.* Preprint posted online May 13, 2025. doi: 2505.08775.

[8] Jin D, Pan E, Oufattole N, Weng WH, Fang H, Szolovits P. What Disease does this Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams. *aRxiv.* Preprint posted online Sep 28, 2020. doi: 2009.13081.

[9] Arya P, Bloomquist M, Chakraborty S, et al. Bootstrapping a Conversational Guide for Colonoscopy Prep. *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue.* Sep 2023:413–420. Prague, Czechia. doi: 10.18653/v1/2023.sigdial-1.38.

[10] Zhang J, Yu S, Chong D, et al. Verbalized Sampling: How to Mitigate Mode Collapse and Unlock LLM Diversity. *arXiv.* Preprint posted online Oct 10, 2025. doi: 2510.01171.

[11] ATLAS.ti Team. *Krippendorff's Alpha: Sample Size and Decision Rules*. Lumivero. Berlin, Germany. 2025.

[12] Gwet KL. *Handbook of Inter-Rater Reliability*. Gaithersburg, MD: Advanced Analytics, LLC;9 ed. 2014. ISBN:978-0970806284.

[13] Wang T, Kulikov I, Golovneva O, et al. Self-Taught Evaluators. *arXiv.* Preprint posted online Aug 8, 2024. doi: 2408.02666.

[14] OSU WMC. 1 Day Bowel Prep With MiraLAX and Dulcolax. `https://www.healthwise.net/osumychart/Content/StdDocument.aspx?DOCHWID=custom.hs0146` 2025. Accessed: 2025-11-06.

Abbreviations

AI Artificial Intelligence

LLM Large Language Model

NSAID Non-Steroidal Anti-Inflammatory Drug

A1 Prompt Examples

Three of the six prompts used for generating patient questions are presented below, including the baseline prompt; the remaining two prompts are for AI Coach response generation and evaluation, each incorporating few-shot dialogue examples. In the generation prompt, chain-of-thought reasoning is included within square brackets. However, in preliminary experiments with Llama, including chain-of-thought reasoning during evaluation was found to confuse the model. All prompts were iteratively refined based on these experiments.

Prompt example: 1 | Prompt for thematic patient questions

Let's role play. You are a patient who will undergo a colonoscopy, and you are talking to an AI Coach. Write 10 questions you could ask next based on the theme. Output only the questions, do not output "Here are 10...", "Here are the questions" or theme headings. Be creative, but specific! Be concise - the questions must be short and natural. The theme is:

<choose one at random from the list below>

biopsy, risks and complications, finance, effectiveness, family history, privacy, follow-up appointments, results, alternatives, pre-test activities (other than diet, medications and travel to the hospital), activities after the test, sedation, side effects, existing medical conditions, medication, allergies, emergency, electrolyte imbalance, stoma, travel, bowel prep ingredients, meals, hospital travel, lifestyle, mental health, phobias, anxiety, doctors, nurses, ulcer, abnormality, insurance, facility, scheduling an appointment, discharge, recovery, risk factors, instruments, terminology.

Prompt example: 2 | Prompt for food-related patient questions

Let's role play. You are a patient who will undergo a colonoscopy. Write 10 questions you could ask next about specific foods, dishes or drinks which you have already eaten. Make sure that the time refers to the past. Be creative with your phrasing. Output only the questions, do not include any other text like "Here are 10 questions...", "Here are the questions" or "Here they are". Be clear and concise - the questions must be short and natural!

Prompt example: 3 | Baseline prompt for patient questions

Let's role play. You are a patient who will undergo a colonoscopy. Write a question you would ask. Be creative with your phrasing. Output only the question. Be clear and concise - the question must be short and natural!

Prompt example: 4 | Prompt for AI Coach responses

Let's role play. You are a friendly AI Coach who helps patients before colonoscopies. Answer the following patient's question only using the information from instructions below. Do not make up information or make your own recommendations. If the question cannot be answered by instructions, direct the patient to their doctor. Be relevant, clear and concise. Make sure your answer is consistent with the number of days before the procedure. The procedure is in <choose a random integer between 1 to 21> days.

INSTRUCTIONS

<prep instructions>

DIALOGUE 1 (4 days before the procedure)
AI Coach: Hello, how can I assist you?
Patient: Can I have cottage cheese today?
AI Coach: [Starting 5 days before the procedure, patients should not eat high fiber foods. The procedure is in 4 days, and cottage cheese is not high fiber. Milk products are disallowed only the day before the procedure. The patient may have it now.] Cottage cheese is low fiber, so you can have it now. Remember to avoid dairy products the day before the procedure.
Patient: I am worried that the restrictive diet and the prep will cause me to lose electrolytes and other nutrients. What should I do?
AI Coach: [Instructions do not advise on electrolyte or nutrient replenishment. The patient should refer to their provider.] Please consult your provider about any concerns on electrolyte or nutrient replenishment.
Patient: Alright, thanks. By the way, can I take an Uber after the test?
AI Coach: [Patients are given medicine to help them relax during the procedure. For safety, if the patient takes a cab, bus, or medical transportation, an adult other than the driver needs to ride with the patient. The accompanying person needs to check in for the procedure with the patient. Otherwise the procedure may be delayed or canceled. So the patient may take an Uber as long as an adult is with them.] You may take an Uber as long as there is an adult accompanying you. You will be sedated during the procedure, so it is crucial for someone to take care of you on your way home. If there is no one present with you at the time of check-in for the procedure, you will need to reschedule it.

DIALOGUE 2 (1 day before the procedure)
AI Coach: Hi, I am happy to answer any questions about your upcoming colonoscopy test!
Patient: Hi, I was wondering if I need to bring my insurance card with me?
AI Coach: [Patients should bring these to the procedure: list of all medicines, including prescription and over the counter medicines, and any vitamin or herbal products they are taking; list of allergies; list of medical conditions and previous surgeries; copy of their advanced directive, such as a living will or power of attorney; photo identification, insurance card, and co-payment, if needed. So the patient needs to bring their insurance card.] Yes, that will certainly be helpful! Please also remember to take a list of all your medicines, a list of all your medical conditions, an ID and a valid payment method for copayment, if needed.
Patient: OK. I can bring my water bottle, right?
AI Coach: [Patients can drink clear liquids up to 4 hours before their procedure. Patients can take a small sip of water to take any allowed medicines within 2 hours of their procedure. So the patient can bring their water bottle.] Yes, but you may only drink clear liquids up to 4 hours before the procedure. Your procedure will be delayed or canceled if you drink anything other than a sip of water with medicines within that time frame.
Patient: Do I have to finish all of the prep mixture? It tastes so bad.
AI Coach: [The prep medicine may not taste good but patients need to take all of it. If a patient vomits up the prep medicine, they should contact the health care provider who is doing their procedure. The patient needs to finish the prep mixture but be mindful of their symptoms.] Yes, finishing the mixture ensures that your colon is clean and ready for the procedure. However, if you experience severe symptoms because of the mixture, please contact your health care provider immediately.
Patient: I had some food in the morning. Is it still okay for me to come for the test tomorrow?
AI Coach: [The day before their procedure patients start a clear liquid diet. Patients may not eat any solid food or eat or drink any milk products until after their procedure is done. Patient's procedure is tomorrow, which means they should be on the liquid diet today. Instructions do not specify what to do if a patient violated them. The patient should consult their doctor.] You should consume only clear liquids the day before your test. Please consult your doctor whether you need to reschedule your test.

DIALOGUE 3 (3 days before the procedure)
AI Coach: Hello, how can I help you?
Patient: Do I have to start the liquid diet tomorrow?
AI Coach: [The day before their procedure patients start a clear liquid diet. The procedure is in 3 days. Tomorrow is 2 days away from their test. The patient should not start the liquid diet tomorrow.] No, tomorrow is 2 days before your test. You should start the liquid diet the day before your test.
Patient: Okay, when do I start the liquid diet then?
AI Coach: [The day before their procedure patients start a clear liquid diet. The procedure is in 3 days. The patient should start the liquid diet in 2 days.] You should start the liquid diet the day before your procedure. In your case, that will be 2 days from now.
Patient: What if I get hungry? Is there anything I can eat then?
AI Coach: [During the liquid diet phase, the patient may not eat any solid food or eat or drink any milk products until after their procedure is done. The patient may drink only clear or light colored liquids. They need to clean out their colon before the procedure so the doctor can see it.] No, you should stick to liquids only. Liquid diet ensures your intestines can be seen by a doctor.
Patient: So I can't drink Fanta that day, I am assuming?
AI Coach: [During the liquid diet phase, the patient may not eat any solid food or eat or drink any milk products until after their procedure is done. The patient may drink only clear or light colored liquids. The patient should not have anything with red or purple dye. Fanta has a deep color with red hues. It is not allowed.] Correct, you may not drink Fanta. Only clear or light-colored liquids are allowed.
Patient: Alright, can I at least eat after my procedure?
AI Coach: [Instructions do not give dietary recommendations after the procedure. The patient should consult their healthcare provider.] Please consult your doctor who will tell you when and what you can eat after your test.
Patient: Is it okay if I had lamb shawarma on Sunday?
AI Coach: [Starting 5 days before the procedure, patients should not eat high fiber foods. Shawarma may contain high fiber ingredients. It is unclear whether Sunday was within 5 days of the procedure. So the patient needs to make sure it wasn't. Otherwise they should consult their doctor.] If you had it within 5 days of the procedure, it may be a problem, if the shawarma has high fiber ingredients. Please check with your doctor.

DIALOGUE 4 (2 days before the procedure)
AI Coach: Greetings! How may I be of help?
Patient: Is it okay to cook pasta today? My procedure is 2 days away.
AI Coach: [Starting 5 days before the procedure, patients should not eat whole grain or high fiber foods. Pasta may or may not be whole grain. Pasta often comes with vegetables which are high fiber. Therefore, the patient may eat pasta as long as it is not whole grain or high fiber and doesn't have vegetables.] That depends! High fiber foods are not allowed 5 days before your procedure. So you can only have regular (white) pasta. But no veggies with seeds in it!
Patient: Alright, can I have a glass of wine with it?
AI Coach: [Patients may not drink alcohol the day before their procedure. The procedure is in 2 days. The patient may have alcohol now.] Yes, but you should stop drinking wine on the day before your procedure.
Patient: Even white wine?
AI Coach: [Patients may not drink alcohol the day before their procedure.] Yes, no alcohol is allowed.
Patient: Can you tell me if I can eat steak?
AI Coach: [Starting 5 days before the procedure, patients should not eat high fiber foods. Steak may or may not be high fiber. The patient should consult their provider.] Meat is typically low fiber. However, some meats may be hard to digest. Please check with your doctor about it.
Patient: Thanks.
AI Coach: You are welcome! Let me know if you have more questions.

DIALOGUE 5 (6 days before the procedure)
AI Coach: Hi, what is on your mind? I am here to help you with test preparation!
Patient: I accidentally had some tiramisu - what do I do now?
AI Coach: [The procedure is 6 days away, and dietary restrictions start 5 days before the test. The patient did not violate the instructions.] Your dietary restrictions start tomorrow, so you are fine! Besides, tiramisu is low in fiber. So you may continue to enjoy it!
Patient: Phew. Great. Can I eat desserts 3 days before my test? I have a celebration.
AI Coach: [Starting 5 days before the procedure, patients should not eat high fiber foods, which means patients may eat low fiber foods. Some desserts are low fiber.] You may eat low fiber desserts, such as plain cake, plain pudding, popsicles, and Jell-O.
Patient: I was thinking ice cream or cookies - are they allowed?
AI Coach: [Starting 5 days before the procedure, patients should not eat fruit with skin or seeds, nuts, and high fiber foods. Plain ice cream is low fiber. Cookies may be low fiber if made from white flour. Cookies often contain seeds, nuts or fruit, so the patient should check before eating.] You may eat plain ice cream. Cookies are allowed if they are made from white flour and don't contain nuts, seeds or fruit.
Patient: Perfect! Thank you!
AI Coach: No problem! Have a great day.

Prompt example: 5 | Evaluation prompt for AI Coach responses

Consider DIALOGUE 8. Verify that the AI Coach's response to the patient's question is correct based on these guidelines. If the response is fully correct, output "Correct". Otherwise, output "Incorrect" and show step-by-step reasoning.

GUIDELINES

- A correct response should be based only on the instructions.
- If the instructions cannot answer the question, then a correct response should direct the patient to their provider.
- If the patient has already violated the instructions, then a correct response should direct the patient to their provider.
- A correct response should be consistent with the number of days before the procedure, or with the time mentioned in the question.
- A correct response should be relevant, unambiguous, natural and fluent.

INSTRUCTIONS

<prep instructions>

DIALOGUE 1 (3 days before the procedure)
Patient: Can I have a cup of plant-based milk after my lunch today?
AI Coach: Yes, you are not on a clear liquid diet yet. However, please check again tomorrow.
VERIFICATION: Incorrect. A correct response should be consistent with the number of days before the procedure. The procedure is in 3 days, and the patient should avoid dairy only the day before the procedure. So the patient may still have a cup of plant-based milk tomorrow.

DIALOGUE 2 (7 days before the procedure)
Patient: I realized my ADHD pills are dark blue when I took them this morning. Should I cancel my test?
AI Coach: Don't worry! You are still allowed to have dark-colored products. This restriction applies on the day before your procedure.
VERIFICATION: Correct. The response is based on the instructions.

DIALOGUE 3 (5 days before the procedure)
Patient: I am worried that the restrictive diet and the prep will cause me to lose electrolytes and other nutrients. What should I do?
AI Coach: You may consume sports drinks, like Gatorade, which are known for replenishing electrolytes.
VERIFICATION: Incorrect. The instructions do not advise patients on replenishing nutrients or electrolytes. Therefore, a patient must be directed to their provider.

DIALOGUE 4 (1 day before the procedure)
Patient: Hey, I am still taking a cough syrup for my throat infection. I hope it's alright.
AI Coach: Depending on its color, it may be alright – make sure it is light and does not contain red or purple hues. Also it should be a clear liquid, as you are on a clear liquid diet today.
VERIFICATION: Incorrect. The instructions say that the patient must inform their provider of all medications before taking them.

DIALOGUE 5 (4 days before the procedure)
Patient: I can have cottage cheese both now and after the procedure, right?
AI Coach: Cottage cheese is low in fiber, so you can have it now. Remember to avoid dairy products the day before the procedure. As for the diet after the procedure, please ask your doctor.
VERIFICATION: Correct. The response is based on the guidelines.

DIALOGUE 6 (2 days before the procedure)
Patient: Are laxatives the primary ingredient in the bowel prep?
AI Coach: Dulcolax is indeed a laxative and part of your bowel prep.
VERIFICATION: Incorrect. The response omits the fact that Miralax is also a laxative and part of the bowel prep.

DIALOGUE 7 (10 days before the procedure)
Patient: Why do I need to undergo colonoscopies?
AI Coach: The prep will help you cleanse your bowels, so the doctor can see it.
VERIFICATION: Incorrect. The response is irrelevant to the question.

A2 Prep Instructions

The Wexner Medical Center at The Ohio State University has published several colonoscopy preparation instruction sets. The version used in this study corresponds to the one-day bowel preparation with MiraLAX and Dulcolax ([14]). It was converted into plain text for use in the prompts in June 2025, at the time of the study.

[1] You will need to do a bowel prep with MiraLAX and Dulcolax before your colonoscopy. As part of the prep, you may need to make changes to your medicines many days before your procedure. These include blood thinners, medicines for diabetes, weight loss, heart conditions, and others. If you do not make the needed medicine changes, your procedure may need to be rescheduled. You will need to change your diet starting 5 days before the procedure and start the bowel prep at 3 p.m. the day before the procedure. You need to have an adult come with you to your procedure. If there are major changes to your health, please contact the health care provider who is doing your procedure. Your procedure may need to be changed for your safety. If you have other questions or concerns about your prep for your procedure, please also contact the health care provider who is doing your procedure.

[2] You need to clean out your colon before the procedure so the doctor can see your colon. This means you will need to purchase in advance a large bottle of laxative polyethylene glycol 3350 such as Miralax (8.3 ounces or 238 grams), 4 tablets of Dulcolax laxative (5 mg each) and a total of 64 ounces of sports drinks, like Gatorade. The drinks cannot be red or purple in hue. Select a sugar-free option if you have diabetes. You will also need to adjust your diet before the procedure. Starting 5 days before your procedure, you should not eat: corn, nuts, leafy greens, popcorn, fruit with skin or seeds, whole grain or high fiber foods.

[3] You may need to make changes to your medicines many days before your procedure. Tell your doctor all the medicines you take, including vitamins, supplements, and herbal remedies. Your doctor will tell you if you should stop taking any of them before this procedure and how soon to do it. If you take aspirin or NSAIDs, such as ibuprofen (Advil, Motrin, and Nuprin), naproxen (Aleve), or celecoxib (Celebrex) for pain, you do not need to stop taking these medicines before this procedure. You may not take vitamins with iron, iron tablets, or fiber supplements. If you take oral medicines for diabetes,

you should not take oral diabetes medicines on the day of your procedure. If you take injection medicines for diabetes, such as insulin, ask the doctor who ordered the medicines how to adjust your medicines for this procedure.

[4] If you take a blood thinner (anticoagulant or antiplatelet) medicine, ask the doctor who ordered this medicine if it is safe for you to stop taking it before this procedure. The doctor may or may not tell you to stop taking it a certain number of days before your procedure. Make sure that you understand exactly what your doctor wants you to do. If you have had a stent, especially a stent in your heart or brain, you should not stop taking your blood thinner/stent medicine until you are instructed by the doctor who placed the stent. If your doctor has told you not to stop taking these medicines before this procedure, please contact the health care provider who is doing your procedure.

[5] You will drink a medicine mixture to clear your bowels of all solid matter. You will need to go to the bathroom often, and your stool will get very watery. The prep may cause you to have cramps or feel bloated. The prep medicine may not taste good. You need to take all of it so your bowels are clear. Your bowels are clear when you are passing pale yellow liquid without any stool. If you vomit up your prep medicine, have not had any bowel movements, or your bowels are not clear after this prep, please contact the health care provider who is doing your procedure. The procedure may need to be rescheduled. If you have a history of heart failure, kidney disease, cirrhosis of the liver, electrolyte problems, or chronic constipation, you may need a different bowel prep medicine. Please contact the nurse for instructions.

[6] The day before your procedure you will start a clear liquid diet. You may not eat any solid food or eat or drink any milk products until after your procedure is done. You may drink only clear or light colored liquids. You should not have anything with red or purple dye. It is okay to have: water, flavored water, or ice chips; coffee or tea (no milk or cream); ginger ale or lemon-lime soda; fruit juices you can see through, such as apple or white grape; Gatorade, other sports drinks, or drink mixes like Kool-Aid; popsicles that aren't red or purple (no fruit or cream); Jell-O or other gelatin, without fruit (no red or purple); clear broth or bouillon. You may not have: alcoholic drinks; milk; smoothies; milkshakes; cream; yogurt; orange juice; grapefruit juice; tomato juice; soup (other than clear broth); red or purple drinks. You should also not smoke or vape the day before the procedure and on the day of the procedure.

[7] At 3 p.m. on the day before your procedure, take 4 Dulcolax tablets. Pour all of the MiraLAX and the two bottles of sports drink into a larger container and shake well. Put it into the refrigerator to keep cool. At 6 p.m. on the day before your procedure, drink one (8-ounce) glass of the mixture every 10 to 15 minutes until you finish half of it. It may help to use a straw. Put the rest back in the refrigerator. 6 hours before your procedure, drink the rest of the mixture. You may need to set your alarm to get up to finish your prep medicine. Drink two (8-ounce) glasses of clear liquids like water after you finish the mixture. You can drink clear liquids up to 4 hours before your procedure.

[8] You can take a small sip of water to take any allowed medicines within 2 hours of your procedure. Your procedure will be delayed or canceled if you drink anything other than a sip of water with medicines. This is important to make sure it is safe to give you sedation or anesthesia. If you have alcohol, marijuana or illegal drugs in your system, your procedure may be rescheduled for your safety.

[9] Arrive by the arrival time on your letter, so the staff can get you ready. Expect to be here 2 to 4 hours for your procedure and recovery time. Please leave all valuables and jewelry, including piercings, at home. Patients who can become pregnant should come prepared to give a urine sample in case a pregnancy procedure is needed. Bring these with you to your procedure: list of all medicines, including prescription and over the counter medicines, and any vitamin or herbal products you are taking; list of your allergies; list of medical conditions and previous surgeries; copy of your advanced directive, such as a living will or power of attorney; photo identification, insurance card, and co-payment, if needed.

[10] You will be given medicine to help you relax during the procedure. For your safety, you will need a responsible adult to drive you home after the procedure. Your driver needs to check in for the procedure with you. If your driver doesn't check in with you, we will need to speak to your driver before we can start your procedure. Your procedure will be delayed or canceled if we cannot speak to your driver. If your driver leaves the facility during the procedure, they need to give the staff a phone number where they can be reached. Your driver needs to be within 30 minutes of the procedure location. If you take a cab, bus, or medical transportation, an adult other than the driver needs to ride with you for your safety. You should have an adult with you to help you at home after the procedure for at least 6 hours. You should not drive, operate machinery, drink alcohol, or make any legal decisions until the day after your procedure.

A3 Error Examples

Error example: 1 | **Date**: 14 days before procedure | **Model**: Mistral Large-2411

Turns	Note
Patient: Oh, alright then... Can I have a salad with nuts in 6 days?	Not easily answerable.
AI Coach: No, you should not eat corn, nuts, leafy greens, popcorn, fruit with skin or seeds, whole grain or high fiber foods.	**Temporal error**: 8 days before procedure the patient doesn't have any dietary restrictions. The error is clinically **harmful**, as it could restrict the patient's diet too much too early.

Error example: 2 | **Date**: 5 days before procedure | **Model**: Mistral Large-2411

Turns	Note
Patient: Oh, that's good to know. How do they test the tissue after removing a polyp?	Unanswerable.
AI Coach: A doctor called a pathologist will look at it under a microscope.	**Extraneous & correct information**: The response provided out of scope information that is factually correct. The error is **harmless**.

Error example: 3 | **Date**: 13 days before procedure | **Model**: Llama 3.3 70B Instruct

Turns	Note
Patient: Can I go back to work the next day?	Unanswerable.
AI Coach: It's not recommended. You should have someone to take care of you for at least 6 hours after the procedure, and you should not drive or operate machinery until the day after the procedure.	**Extraneous & incorrect information**: The response provided out of scope erroneous information, as prep instructions do not give recommendations for the day after procedure. The error is clinically **harmless**.

Error example: 4 | **Date**: 10 days before procedure | **Model**: GPT-4.1 Instruct

Turns	Note
Patient: Just want to make sure I'm prepared for this. Could I have black coffee in the morning seven days before my colonoscopy?	Easily answerable.
AI Coach: Yes, you can have black coffee seven days before your colonoscopy. Dietary restrictions on liquids such as black coffee don't begin until the day before your procedure. Starting five days before your procedure, you should avoid certain high-fiber foods, but black coffee is permitted until then.	**Reasoning error**: Contradicts prep instructions which say black coffee is allowed. The error is **harmless**, as restricting coffee will not impact patient's health or procedure.

Error example: 5 | **Date**: 2 days before procedure | **Model**: Llama 3.3 70B Instruct

Turns	Note
Patient: Can I drive myself home after the procedure?	Easily answerable.
AI Coach: No, you will need a responsible adult to drive you home after the procedure. The medicine you'll be given may impair your ability to drive safely.	**Omission**: The requirement of patient's driver to check-in before the procedure is missing. The procedure will be rescheduled if a driver does not show up, so the response is **harmful**.

Error example: 6 | **Date**: 17 days before procedure | **Model**: o3

Turns	Note
Patient: Do you have any tricks to make the bowel prep more bearable? **AI Coach:** Two simple ideas built into the instructions can help: • Mix the MiraLAX with a clear, non-red/purple sports drink flavor you like (choose a sugar-free one if you have diabetes). • After you mix it, keep the container in the refrigerator; the prep tastes better when it is cold, and many people find it easier to drink through a straw. Be sure to follow the exact schedule: 4 Dulcolax tablets at 3 p.m. the day before, then one 8-ounce glass of the chilled mixture every 10–15 minutes starting at 6 p.m. until you finish half, and the other half 6 hours before the procedure.	Easily answerable. **Other error**: The description of color in drinks is ambiguous, as it could mean that purple is an allowed hue. The error is **harmful**, as such implication could lead to a rescheduled procedure.

A4 Most Frequent Tokens

In processing the patient questions, we filtered out articles (e.g., the, a, an), prepositions (e.g., about, on, in, except those semantically relevant to the task, such as before), pronouns (e.g., she, I, you), auxiliary and modal verbs (e.g., would, can, should), interrogatives (e.g., who, what), common interjections (e.g., thanks, okay), and other frequent non-informative words.

Table 9: Top 25 most frequent tokens for each model with counts.

Baseline	Llama 3.3 70B	Mistral Large	GPT-4.1	o3	GPT-5
procedure: 859	procedure: 499	days: 304	colonoscopy: 165	feeling: 128	nervous: 129
after: 690	days: 287	before: 185	procedure: 141	prep: 100	prep: 100
take: 319	before: 248	procedure: 183	prep: 126	nervous: 87	clear: 100
colonoscopy: 276	after: 179	see: 153	eat: 118	clear: 61	drink: 79
home: 234	ago: 178	eat: 141	nervous: 117	plain: 59	coffee: 67
long: 225	eat: 178	glass: 118	feeling: 116	eat: 55	chicken: 60
drive: 221	eating: 168	bowl: 111	before: 114	procedure: 54	plain: 59
feeling: 213	whole: 161	salad: 107	feel: 88	p.m: 49	before: 58
completely: 203	process: 152	dinner: 102	salad: 83	after: 49	white: 57
sedation: 196	colonoscopy: 152	prep: 97	breakfast: 74	before: 49	seeds: 57
specific: 193	anxious: 150	chicken: 94	relief: 73	morning: 48	ate: 56
follow: 187	prep: 144	coffee: 94	avoid: 73	allowed: 47	day: 56
discomfort: 181	feeling: 143	enjoy: 88	tomorrow: 72	helps: 46	red: 55
whole: 181	take: 134	ago: 85	make: 69	sip: 45	black: 54
aware: 170	able: 125	good: 83	clear: 69	call: 45	allowed: 50
anxious: 161	getting: 124	patient: 83	after: 66	tomorrow: 44	procedure: 49
process: 148	drink: 119	after: 82	problem: 63	water: 44	sports: 48
complications: 130	foods: 107	popcorn: 81	days: 62	drink: 44	anxious: 47
normal: 126	relief: 104	drink: 78	know: 60	days: 43	broth: 47
pain: 123	like: 100	wine: 77	chicken: 58	soup: 42	pm: 46
experience: 122	issues: 100	breakfast: 77	start: 57	great: 42	helps: 44
asleep: 120	avoid: 98	day: 75	drink: 56	coffee: 42	days: 43
ensure: 115	specific: 97	morning: 73	get: 56	acceptable: 41	cream: 43

kind: 114	affect: 95	soup: 72	right: 56	anxious: 40	tomorrow: 42
restrictions: 113	worried: 95	avoid: 70	oatmeal: 55	day: 40	juice: 41