From Clinical Need to NLP System: Lessons from Building a Patient Advocate with Large Language Models

Anonymous ACL submission

Abstract

While medical NLP has progressed on many well-defined tasks, many healthcare scenarios still lack a clear path from clinical need to effective NLP systems. Echoing this year's Theme Track "New Missions for NLP Research", we explore one such scenario that is open-ended, patient-facing, and lacks ground truth: building an LLM-driven patient advocate for medical visit preparation. Through iterative collaboration with clinicians and patients, we translate the clinical need into a compound NLP task, structure LLM interaction with a clinically validated framework, and ground generation in vetted sources. We evaluate against a plain LLM with 20 human patients across two disease settings (prostate cancer, type 2 diabetes) using fine-grained cognitive load assessment, and with 90 simulated patients using reference-based metrics and LLM-as-judge evaluation. Our system reduces overall mental effort and improves organized exploration, and generates more relevant questions verified with simulated patients. We share the iterative design process, discuss open problems we encounter including balancing clinical controllability with flexibility, and report a clinician safety evaluation. We believe these lessons transfer to other underexplored healthcare NLP scenarios.

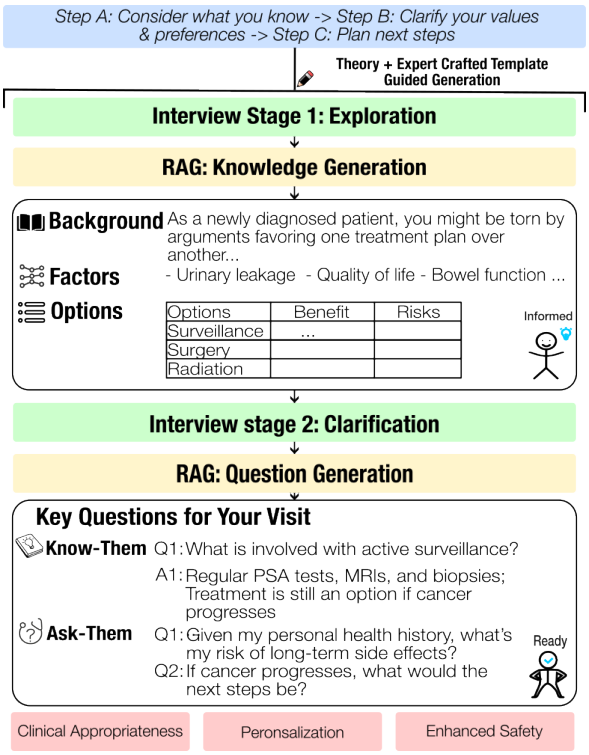

Figure 1: Stages of an agent for medical visit preparation: integrating Ottawa Decision Guide (a shared decision-making framework) with retrieval-augmented generation into conversations.

1 Introduction

Many medical conditions are complex and require shared decision-making between patients and clinicians, yet limited appointment time and passive patient engagement can hinder effective communication (Martinez-Gonzalez et al., 2019). Examples include chronic conditions like eczema with various management strategies, low-risk cancers with multiple treatment options, and other scenarios involving trade-offs between efficacy, side effects, and quality of life. Helping patients articulate questions, clarify preferences, and actively participate is crucial for strong bidirectional patient-provider communication (Martin and Tipton, 2007). Traditionally, Patient Advocates have fulfilled this role, helping patients to have their interests and needs heard (Abbasinia et al., 2020; Martin and Tipton, 2007). However, current healthcare financing structures do not support their widespread adoption, leaving patients to navigate these communication challenges largely alone (Negarandeh et al., 2006; Nsiah et al., 2020).

Motivated to build an intelligent assistant that, like a traditional patient advocate, centers on patients' concerns and needs, we propose developing a Large Language Model (LLM)-driven patient advocate, focusing on one specific advocacy function: helping patients prepare communication for upcoming clinical visits. This task presents unique challenges despite existing medical NLP applications:

it requires addressing patient knowledge gaps, exploring personal values and preferences, and carefully suggesting a visit plan while balancing clinical appropriateness with responsiveness, all within an open-ended, patient-centered context without clear ground truth. This year's Theme Track, "New Missions for NLP Research", asks what new missions exist for NLP research. Among the many possible answers, we share one drawn from our practice: for many healthcare scenarios, there is still no clear path from clinical need to NLP research, and the lessons from exploring such a scenario are themselves a contribution. Rather than presenting our system primarily as a new chatbot or model, we frame this paper as a design-and-evaluation case study of how an underdefined clinical need can be translated into an NLP task, system design, and evaluation strategy.

This paper makes three contributions: **First**, we present a design-and-evaluation case study that translates medical visit preparation from an underdefined clinical need into a compound NLP task and patient-facing system. Our exploration includes using a clinically validated framework, the Ottawa Personal Decision Guide, to structure an LLM-assisted interview that helps patients progress from being uninformed to visit-ready, and integrating retrieval-augmented generation (RAG) to enhance information safety by grounding responses on vetted sources. **Second**, we evaluate across two disease settings (localized prostate cancer, type 2 diabetes) with 20 human patients using cognitive load assessment, 90 simulated patients using reference-based metrics and LLM-as-judge evaluation, and a clinician safety review. Both human and simulated evaluations complement each other, providing insights into the benefits of our design from multiple perspectives. **Lastly**, we surface research problems that deserve attention, including how to balance clinical controllability with conversational flexibility, and how to design evaluation metrics for open-ended tasks without ground truth. We provide design lessons for future patient-facing NLP systems.

2 Related Work

2.1 Well-Translated Medical NLP Tasks

Many well-studied medical NLP tasks have natural translations: NER and ICD coding (Gu et al., 2021) is sequence labeling, diagnosis (Singhal et al., 2023; Nori et al., 2023) can be classification, medical QA (Li et al., 2024) becomes extraction or generation with reference answers, conversational diagnosis (Tu et al., 2024) produces deterministic outputs despite interactive form, and clinical summarization (Lajmi et al., 2026) maps to conditional generation. Patient-facing dialogue like emotional support (Liu et al., 2021) can also rely on widespread generation metrics like BLEU/ROUGE and point-wise human judgment. In each case, what counts as good was already established.

2.2 Underexplored Clinical Scenarios

Recent work has explored LLM applications in realistic clinical contexts, such as shared decision-making (Elwyn et al., 2024), conversational education for public health (Jo et al., 2023), and advance care planning (Hsu et al., 2025). Meanwhile, studies have shown that LLMs can produce responses rated higher quality than those from clinicians (Ayers et al., 2023) and pass medical licensing exams (Singhal et al., 2023; Nori et al., 2023), suggesting that clinical accuracy alone may no longer be the primary challenge. Instead, open problems emerge: whether LLM-based systems actually reduce patients' cognitive burden when delivering help, how to measure system quality in open-ended settings where no single correct answer exists, how to meet clinical regulations with controllability while accommodating diverse patient needs with flexibility, and how to develop scalable evaluation for these dimensions.

Our work focuses on one such scenario: medical visit preparation. Directly related work (Hsu et al., 2025; Amogh Ananda Rao et al., 2025), primarily from the health informatics community, focuses on demonstrating systems holistically and relies on clinician judgment, limiting experimental scale and deeper investigation on specific research problems. We explore these open problems and share what we learned.

3 System Design Process

This section describes how we translated a clinical need into an NLP system through iterative collaboration with clinicians and patients. We first present the clinical starting point and task formulation, then the system architecture, and finally justify key design choices with pilot data, showing how each iteration was informed by stakeholder feedback and empirical evidence.

3.1 Clinical Need and Stakeholder Input

Our starting point is a clinician collaborator's description of the problem: patients often do not know what to ask their doctor because they lack

medical background, visit time is limited, and patients feel unprepared to articulate their concerns. The clinician identified the Ottawa Personal Decision Guide (OPDG) (O'Connor, 2003), a validated shared decision-making framework, as the scaffolding that an AI system should also follow. The OPDG outlines three steps:

STEP A Consider what the patient already knows about the decision;

STEP B Clarify personal values and preferences by reviewing options along with their risks, benefits, and levels of certainty;

STEP C Identify support needs and plan the next steps.

Originally designed as a blank guided worksheet,[1] the OPDG assumes patients need to obtain the medical knowledge to complete it, independently or by doctor. This gap is where NLP tools can help.

3.2 Task Formulation

Working with the clinician, NLP researchers translated the clinical need into a compound NLP task. Given a disease type D and a patient concern type $T \in \mathcal{T}$, the system conducts a multi-turn interview conversation $C = \{(a_i, p_i)\}_{i=1}^{n}$ within a token budget, then generates a question set $Q = Q_{\text{know}} \cup Q_{\text{ask}}$, where:

- **Know-them questions** Q_{know}: five questions answerable using validated sources, optionally paired with answers.
- **Ask-them questions** Q_{ask}: five personalized questions for the patient to bring to their doctor, typically not directly answerable from validated sources.

This is not a single NLP task but a compound one: an interview process followed by structured question generation. Both our system and the plain LLM baseline are constrained to produce outputs in this format.

3.3 System Overview

Since the clinician suggests grounding the interaction on the OPDG, we structure the interview to follow its steps. Between interview stages, patients need medical knowledge to progress (e.g., understanding options before clarifying values), making retrieval-augmented generation a natural choice. The clinician assigns validated patient-facing guidebooks as retrieval sources (see Appx C for sources used in this study). The resulting system follows a four-stage workflow (Fig 1):

I. INTERVIEW STAGE 1 explores what the patient already knows, guided by expert-crafted template questions adapted from OPDG Step A;

II. RAG: KNOWLEDGE GENERATION retrieves from validated patient-facing sources to produce a knowledge panel;

III. INTERVIEW STAGE 2 clarifies personal values and preferences (OPDG Step B);

IV. RAG: QUESTION GENERATION produces know-them and ask-them questions.

We work with a clinician to transform the OPDG into an interactive conversation. At the first stage, we categorize concerns into 8 categories (e.g., Treatment Plan, Long-Term Management). For each category, the expert crafts 3 interview questions. At the second stage, seven questions probe feelings on options, trade-offs, and support. Each interview involves around 9 template questions based on the concern category.

At RAG steps, to ensure safety and reduce hallucination, we limit the LLM to only paraphrasing and organizing retrieved content (full prompt in Appx D.2 and D.3). The system tracks and cites page numbers from source material for transparency.

3.4 Design Iteration: Balancing Controllability and Flexibility

First attempt: template-based interview. We initially assumed that directly using the clinician-crafted template questions as the full interview would be effective, since they closely mirror real clinical practice. We compared this template-based interview against a plain LLM that interviews in free form, using the same evaluation protocol as our main study (Sec 4) with 17 separate pilot participants in the prostate cancer setting (not overlapping with the 20 in the main study). Surprisingly, the crafted interview did not outperform the plain LLM on any cognitive load dimension (Tab 1). This was unexpected: our team felt the crafted version was organized and clean, while the plain LLM asked redundant questions. Yet patients disagreed.

Post-study feedback. A quick feedback collection revealed why. Participants appreciated the framework's organization but felt constrained: template questions were always presented regardless of what the patient said.

One participant notes: "*I preferred System A (Plain LLM) mainly because it felt more supportive, responses were natural and not robotic. System*

[1] OPDG worksheet: https://decisionaid.ohri.ca/docs/das/OPDG.pdf

Metric	Template vs. Plain	Balanced vs. Plain
Overall Mental Effort ↓	−0.18	**−0.90***
Intrinsic Load (IL1)	−0.12	+0.40
Extraneous Load (EL1) ↓	+0.35	**−0.53**
Extraneous Load (EL2) ↓	−0.12	**−0.20**
Germane Load (GL1) ↑	−0.24	**+0.60***
Germane Load (GL2) ↑	+0.12	**+0.50**

Table 1: Difference from plain LLM baseline. Template: rigid expert-crafted interview (n=17). Balanced: template-guided with LLM flexibility (n=20). $^*p < 0.05$.

Output	Dimension	Win	Lose	Tie
Knowledge Panel	Relevance	7/5	2/4	1/1
	Informativeness*	8/8	2/2	0/0
	Helpfulness*	8/8	2/2	0/0
Visit Questions	Relevance	6/6	0/4	4/0
	Informativeness*	6/10	0/0	4/0
	Helpfulness	3/6	2/3	5/1

Table 2: Pairwise preferences comparing RAG vs. no-retrieval baseline. Counts as *Gemini-Pro-3 / GPT-5.2*. *: $p < 0.05$ aggregating across both judges.

B (Template Interview) sometimes felt clunky although informative."

Balanced design: template-guided generation Based on this feedback, we implement a template-guided response generation mechanism that retains clinical structure while allowing LLM flexibility (prompt available in Appx D.1). For each turn, the LLM: (1) must include the template question exactly as provided, (2) may add one personalized segment responding to what the patient said, and (3) may optionally add one additional personalized question. Before each stage ends, two free questions without template constraints allow exploration of patient-specific concerns. Each interview typically involves around 13 turns (see example dialogue in Appx 8). This balanced design significantly outperforms the plain LLM (Tab 1), reducing overall mental effort by 0.90 ($p < 0.05$) and increasing germane load GL1 by 0.60 ($p < 0.05$). Per-system scores are in Tab 3.

A surfaced research problem. This process surfaces a research problem beyond our setting: many healthcare applications must follow clinical frameworks or regulations like the OPDG suggested by clinician collaborators, yet rigid adherence hurts the model's natural advantages. Our current design may not be the optimal balance, but we demonstrate the tradeoff is real and consequential. Finding principled methods to navigate this tradeoff is a future NLP research direction.

3.5 Design Iteration: Retrieval-Augmented Generation

Why retrieve from designated sources. In clinical practice, clinicians refer patients to a fixed set of vetted references (guidebooks, institutional websites) for consistency and regulatory compliance. We adopt the same principle: rather than relying on the LLM's pretrained knowledge, we retrieve from clinician-designated sources. This gives us control over information validity and traceability, as the system can cite specific pages from known materials.

The tradeoff. On the other hand, LLMs already possess a broad knowledge base from pretraining, covering topics that may extend beyond any single guidebook (e.g., local support groups, lifestyle advice, emerging treatments). Pretrained knowledge also offers flexibility in how information is presented. The question is whether constraining generation to designated sources actually improves output quality, or whether it unnecessarily limits the system.

Verification study. To justify this design choice, we took 10 interaction sessions from our final system's human study logs (all from the prostate cancer setting). For each session, we removed the RAG-generated knowledge panel and questions, then prompted a non-retrieval LLM to regenerate them using only the conversation context and its pretrained knowledge. We then asked two LLM judges (Gemini-Pro-3, GPT-5.2) to rate both versions in a pairwise comparison.

Retrieval yields the largest gains on informativeness: RAG wins 16/20 comparisons for the Knowledge Panel ($p = 0.012$) and 16/20 for know-them questions ($p < 0.001$), indicating that retrieval substantially increases decision-relevant content. Effects on relevance and ask-them helpfulness are smaller, likely because the Stage-2 interview is already conditioned on the narrative from the RAG-guided Stage-1 interview.

3.6 Summary of Design Process

Through starting from a clinician's description of the pain point (Sec 3.3), piloting with patients (Sec 3.4), and verifying design choices with evidence (Sec 3.5), we gain the insight that iterative collaboration with stakeholders is an effective way to build task definitions that bridge clinical need and NLP convention, and to reveal which NLP

problems actually matter. Concretely, we gain: (1) a task definition translating a clinical need into an NLP problem, (2) evidence that finding the right balance between controllability and flexibility is important, and (3) verification that retrieval from designated sources is necessary. The remaining open problems from Sec 4 and Sec 5, whether such systems reduce cognitive burden and how to develop scalable evaluation, are addressed in next sections about evaluation.

4 Evaluation with Human Patients

After several design iterations, we arrive at a system that performs well in internal testing. However, a strong flagship model such as GPT-4o used as a plain LLM may already suffice. Both systems produce clinically reasonable outputs, and patients may rate both favorably compared to no AI assistance. Yet a non-specialized system may still introduce unnecessary effort: asking too many questions at once, or requiring patients to seek information elsewhere, while not necessarily helping them organize their concerns or build understanding. These differences are not captured by general satisfaction ratings. We therefore evaluate using fine-grained cognitive load measures (Klepsch et al., 2017) that separate unproductive effort caused by poor system design from productive effort that supports understanding and reflection. We compare our system against the plain LLM under the same token budget.

4.1 Comparison Setup

Comparison Approach. We compare our system against a free-form baseline: a plain LLM with a simple system prompt instructing it to interview a patient for visit preparation. Our system follows a fixed number of turns due to the semi-structured interview design, while the plain LLM may generate longer responses per turn (e.g., asking 10 questions per message).

Fair Comparison Constraints. To ensure fair comparison, we constrain both systems to the same token budget, the same base model (GPT-4o), and require outputs in the same form. We estimate our system's average output length over 5 rounds of testing and set this average (1050 tokens) as the shared token budget. Our system always completes its predefined turns, while the plain LLM is instructed to conclude the interview briefly when approaching the budget. After the interview stage, both systems generate five "know-them" and five "ask-them" questions. However, the plain LLM leaves answers to "know-them" questions blank due to lack of knowledge sources. To prevent the plain LLM from offering visit prep suggestions during the interview, we constrain it to only ask exploration questions and provide information, ensuring both systems only recommend visit prep questions at the end.

4.2 Procedure and Participants

Study Procedure. We evaluate in two disease settings to test generalizability. Localized prostate cancer is a low-risk cancer with complex medical tradeoffs, selected as starting point because a related clinician on our team customized the templates and guidebooks for this condition. Then we select Type 2 diabetes to broaden the scope. It is a highly prevalent chronic condition that allows fast development and recruitment. Using a within-subject design, each patient interacts with both systems in random order. We use a within-subject design because our measurements focus on cognitive experience with ordinal ratings that are difficult to calibrate across subjects. We allow patients to use different concerns for each system to avoid the first experience influencing the second (e.g., patients becoming better at articulating the same concern after the first system). Both systems use identical interface layouts. The study has IRB approval.

Metrics. We assess cognitive load during visit preparation using a framework adapted from prior work (Klepsch et al., 2017; Chen et al., 2025). Unlike general instruments like NASA-TLX (Hart, 2006) that include irrelevant dimensions (e.g., physical load), this framework focuses on three types of cognitive load: *intrinsic load* (inherent complexity of visit preparation with this concern), *extraneous load* (unnecessary cognitive effort caused by poor system design), and *germane load* (beneficial cognitive effort that supports understanding and reflection). We measure the following metrics: (1) **Overall Mental Effort** on a 1-9 scale, (2) **IL1** (intrinsic load): "The medical issues I needed to think through for this visit were inherently complex (regardless of the system)," (3) **EL1** (extraneous load): "I needed extra time to look for information elsewhere because the system didn't provide enough," (4) **EL2** (extraneous load): "I may need external research to validate the information as I don't know the source of the information," (5) **GL1** (germane load): "The system explores my concerns and provides information in an organized way," and (6) **GL2** (germane load): "The system helped me learn

new information or better understand my medical situation and options." All cognitive load questions use 1-7 scales.

Participants. For localized prostate cancer, we recruited 10 male participants from our university's health research center (7) and Prolific (3). Inclusion criteria required participants to be biologically male at birth, over 30 years of age, currently have prostate cancer, and have been scheduled for a relevant medical visit. For type 2 diabetes, we recruited 10 participants with a type 2 diabetes diagnosis and an upcoming medical visit from Prolific[2]. In total, 20 participants complete the study. Participants are compensated at 14-20 US dollars per hour.

4.3 Results

Tab 3 presents cognitive load results comparing our system against the plain LLM across 20 human participants (combined across both disease settings). Our system significantly reduces overall mental effort (4.80 vs. 3.90, $p = 0.014$, Cohen's $d = 0.61$, medium effect) and significantly increases germane load GL1 (5.65 vs. 6.25, $p = 0.010$, Cohen's $d = 0.64$, medium effect), indicating that the balanced design helps patients engage more productively with their preparation. Other dimensions show consistent trends favoring our system (e.g., EL1: $d = 0.32$, GL2: $d = 0.29$) but do not reach statistical significance, likely due to the limited sample size. Intrinsic load varies across concerns by design (within-subject, different concerns per system) and serves as a control. Per-disease breakdowns show consistent patterns across both prostate cancer and diabetes settings (see Appx A).

We include case studies in the Appendix to illustrate the difference. Our system asks one structured question per turn while weaving in personalized responses that build on what the patient said (Appx 8). The plain LLM, given the same budget, asks many questions in a single message covering diagnosis understanding, screening tests, next steps, and symptoms simultaneously, without building on previous answers (Appx 10). Patients report the latter feels overwhelming despite covering more topics.

4.4 Safety Evaluation

A clinician with expertise in cancer reviewed the Knowledge Panel and "know-them" question answers generated by our system for clinical appropriateness with respect to the retrieved evidence in prostate cancer cases. Overall rating is 6.82/7, indicating that the RAG grounding and template structure effectively constrain generation to proper content. All outputs receive full marks except one case where a stage-4 cancer patient, outside our target population of localized cancer, is mistakenly enrolled and receives mismatched content. This highlights the need for stronger eligibility screening in future deployments, but also demonstrates that the observed safety issue in our study was primarily logistical.

[2] https://www.prolific.com

Metric	Plain LLM	Ours	d
Overall Mental Effort ↓	4.80	**3.90***	0.61†
Intrinsic Load (IL1)	3.75	4.15	0.31
Extraneous Load (EL1) ↓	2.75	**2.20**	0.32
Extraneous Load (EL2) ↓	2.20	**2.00**	0.16
Germane Load (GL1) ↑	5.65	**6.25***	0.64†
Germane Load (GL2) ↑	5.25	**5.75**	0.29

Table 3: Cognitive load results (n=20, combined across both diseases). ↓: lower is better; ↑: higher is better. $^*p < 0.05$ (paired t-test/Wilcoxon). d: Cohen's d_z (paired); †medium effect (≥0.5).

5 Evaluation with Simulated Patients

While the human study provides direct evidence of patient experience, recruiting patients for interactive studies remains challenging, hindering follow-up work and scalable verification. Developing scalable evaluation for open-ended medical tasks without ground truth is itself an open problem. We explore one approach: finding similar data from social media and formulating it into a standard test set, using post content as patient simulation materials and questions that these patients ask as references. We then explore whether these references can support both traditional metrics (BLEU/ROUGE) and LLM-as-judge evaluation.

5.1 Comparison Setup

We design to conduct evaluations where each simulated patient interacts with both a plain LLM and our designed system. Each simulated patient completes two interview sessions, one with each system, and both systems generate "ask-them" questions (questions recommended for the patient to bring to their doctor). We then evaluate how well these generated questions align with reference questions from the patient's original profile.

We construct simulated patients from real patient posts on online forums, treating the questions that patients ultimately asked in their posts as golden reference questions. We hypothesize that patients who are able to make detailed posts are well-informed, and the questions they formu-

late represent well-prepared, high-quality questions suited for references. We clean and process these questions to serve as reference questions for evaluation.

5.2 Metrics

We evaluate question quality using ROUGE-1 (Lin, 2004), ROUGE-L, BLEU-1 (Papineni et al., 2002), and an LLM-as-judge metric. Note we also considered using the same cognitive load metrics from the human study with simulated patients by asking LLM Judge to estimate the cognitive loads, but their ratings are highly homogeneous with no difference between variants; see Sec 6 for our lessons. We therefore pursue reference-based approaches, believing that grounding evaluation in real-world patient questions can address LLM-as-judge's weakness in cognitive load estimation.

Since generated questions and references are both multiple, we use macro-oracle evaluation (Sai et al., 2020; Liu and Liu, 2021; Ravaut et al., 2022): for each input with K generated candidates and M references, we compute ROUGE for all (h_i, r_j) pairs, take the maximum over candidates for each reference, then average across references: $\frac{1}{M}\sum_j \max_i \text{ROUGE}(h_i, r_j)$.

We evaluate against two types of reference questions: (1) *original questions*, which are questions as originally posted by patients, often colloquial and incomplete (e.g., "Why waiting for incontinence to improve?"), and (2) *refined questions*, which are professionally refined versions that are clearer and more complete (e.g., "Why does my doctor recommend waiting until my urinary incontinence improves to one pad per day before starting radiation?"), generated using GPT-5 to improve clarity while preserving the original intent.

LLM-as-Judge. In early trials, we found that point-wise rubric-based LLM scoring was not sensitive enough to capture small differences between systems. We therefore use a reference-anchored pairwise LLM-as-judge setup: given a reference question and two candidate questions (Plain LLM vs. Ours), the judge (GPT-5) scores each candidate on a 1–10 scale based on semantic similarity and relevance to the reference. We randomize the presentation order of candidates to mitigate position bias (Wang et al., 2024) and normalize scores into percentages for reporting.

5.3 Simulated Patient Construction

Collection for Profiles and Reference Questions. We construct simulated patients from real patient posts on Reddit forums. For prostate cancer, we crawl posts from r/ProstateCancer where patients discuss their experiences, concerns, and questions. For type 2 diabetes, we follow the same pipeline on r/diabetes_t2. For each post, we extract questions using a structured prompt that identifies patient questions, refines them for clarity and completeness, and categorizes them by type (e.g., Treatment Plan, Long-Term Management). We preserve the surrounding context from the original post as patient profile information. To ensure multiple references for robust evaluation, we filter posts that contain only one question or where questions are unrelated to the post content. This yields 43 prostate cancer profiles and 47 diabetes profiles (90 total). As shown in Tab 4, most profiles correspond with 2–3 questions.

# Profiles	Avg # Q/Profile	Distribution (# Profiles)			
		2 Q	3 Q	4 Q	6 Q
90	2.76	55	21	11	3

Table 4: Simulated Patient Profiles and Question Distribution

Patient System. Following prior work (Li et al., 2024) that has successfully built patient systems for medical benchmarks, we convert the context from each post into a set of atomic facts, discrete, verifiable statements about the patient's situation, which we organize into structured patient profiles. See Appx B for an example data point. We will also release the patient simulation codes and data upon publication. These atomic facts serve as constraints during patient role-play (patient response generation prompt in Appx D.4), ensuring grounding in the original post content and reducing hallucination. Each simulated patient is instructed to respond using only information from their atomic facts, keeping responses faithful to the source while allowing natural conversation exploration.

5.4 Results

Tab 5 presents scores comparing our system against the plain LLM across both disease settings, evaluated against refined reference questions.

For prostate cancer, our system shows consistent improvements: ROUGE-1 increases from 23.1 to 25.0 ($p < 0.05$), ROUGE-L from 18.1 to 20.3 ($p < 0.01$), and BLEU-1 from 14.0 to 15.6 ($p < 0.05$). The LLM-as-judge metric shows even clearer separation (36.4 vs. 44.7, $p < 0.01$). For diabetes, BLEU/ROUGE improvements are mixed, but LLM-as-judge again shows significant gains

Metric	Prostate (n=43)		Diabetes (n=47)	
	Plain	Ours	Plain	Ours
ROUGE-1	23.1	**25.0***	26.7	**27.1**
ROUGE-L	18.1	**20.3****	8.6	**8.9**
BLEU-1	14.0	**15.6***	**18.4**	17.2
LLM-judge	36.4	**44.7****	46.2	**55.1****

Table 5: Simulated patient evaluation (refined questions). $^{*}p < 0.05$, $^{**}p < 0.01$ (paired t-test/Wilcoxon).

(46.2 vs. 55.1, $p < 0.01$). This suggests that surface metrics are less sensitive to the semantic alignment our system provides, while LLM-as-judge captures meaningful differences in question relevance across both diseases.

Together with the human study, these results indicate that our system generates questions better aligned with patient information-seeking needs. The framework-guided interview structure and retrieval-augmented generation help produce more relevant questions for patients to bring to their medical visits.

6 Lessons

We study medical visit preparation across two disease settings as a case of translating clinical need into NLP research. From this process, we identify transferable lessons for patient-facing scenarios where task formulation, system design, and evaluation remain underdefined.

Iterative Stakeholder Collaboration Bridges Clinical Need and NLP. For open-ended clinical scenarios without established NLP formulations, we recommend building prototypes and piloting with real stakeholders early. In our case, the task definition, the importance of balancing controllability with flexibility, and the necessity of retrieval all emerged only through this process. Researchers entering similar underexplored healthcare scenarios should expect that the research problem cannot be fully specified upfront.

Balancing Clinical Controllability with Flexibility is a Consequential Research Problem. Health systems need to meet regulatory and framework constraints, yet rigid adherence hurts the model's conversational advantages. We recommend the community investigate this as a standalone research problem: how to leverage LLM conversational power at maximum while maintaining clinical compliance. Our template-guided generation is one balance point, but principled methods for navigating this spectrum across domains remain needed.

Fine-Grained Cognitive Measurement Captures Differences Beyond Correctness and General Satisfaction. When the baseline is already a strong flagship model, both systems achieve high clinical correctness and general satisfaction. This ceiling effect renders coarse evaluation metrics ineffective for distinguishing system variants. For tasks where the goal is not correctness but reducing cognitive effort during preparation, fine-grained cognitive load measurement, which separates unproductive effort (e.g., needing to verify or search elsewhere) from productive understanding (e.g., organized exploration and learning), captures meaningful differences that general ratings cannot.

Real-World Data as References Enables Scalable Evaluation for Open-Ended Tasks. For automatic evaluation, we first try having simulated patients answer the same cognitive load measurement questions with rubrics. The results are highly homogeneous with no distinguished performance between systems, unlike human results. We conjecture that LLMs are not good at estimating cognitive load in a human-like manner. This leads us to anchor real-world data as references for meaningful evaluation. Reference-based LLM-as-judge, grounded in questions real patients actually ask, captures semantic alignment that rubric-based approaches alone cannot.

More Steps Are Needed Before Real-World Deployment. Participants also raise ethical concerns in feedback forms. We identify several limitations of the current system in terms of safety and discuss them in the Limitations section (Sec 7).

7 Conclusion

We explore building an LLM-driven patient advocate for medical visit preparation, a healthcare scenario with no established path from clinical need to NLP system. Through iterative collaboration with clinicians and patients across two disease settings, we demonstrate that structured design (framework-guided interview with LLM flexibility and retrieval-augmented generation) reduces overall mental effort and improves organized exploration in a small human study, and generates more relevant questions verified with simulated patients. We surface research problems deserving attention, including balancing clinical controllability with flexibility and designing scalable evaluation for open-ended tasks. We hope these lessons transfer to other patient-facing NLP scenarios where the research problem itself must be discovered through practice.

Ethical Considerations

This study received IRB approval from our institution. All 20 participants provided informed consent and were informed of potential risks including technical errors, hallucinations, and incomplete information. We ensured privacy by collecting no personally identifiable information beyond age and race/ethnicity. User interactions were logged securely. Participants were compensated at $20 per hour per Prolific platform standards.

This system represents early-stage research not intended for clinical use without substantial further development and validation. Future work must include improved safety mechanisms, transparent risk communication, clinical integration protocols, and formal randomized controlled trials before deployment.

Limitations

This study has several limitations requiring resolution before real-world deployment. While we expanded to two disease settings (prostate cancer and type 2 diabetes) with 20 human participants, generalizability to other conditions and populations remains untested. Our within-subject design limits the number of system variants each participant can compare. Simulated patients are constructed from Reddit posts, which may not represent all patient populations. The safety evaluation covers a limited set of outputs; the one mismatched case highlights the need for eligibility screening.

To facilitate safe and effective real-world application, several improvements are needed:

1. **Enhanced Safety Controls:** More robust safeguards to detect and mitigate harmful or inaccurate outputs, flag uncertain information, and redirect users to verified medical resources.

2. **Transparent Risk Communication:** Explicit disclosure of AI hallucination risks and possibility of incorrect or incomplete information.

3. **Clinical Integration:** Generate summary cards for patients to bring to appointments, alerting physicians to AI-assisted preparation and enabling them to address misconceptions or gaps.

4. **Improved Interview Proactiveness:** Enhance the interview to more actively probe concerns, identify knowledge gaps, and adapt questioning based on responses.

Finally, randomized controlled trial (RCT) evaluation is essential to establish true efficacy. Such studies should compare AI-assisted preparation against human medical assistants or traditional educational resources, measuring: (1) time to visit-readiness, (2) patient confidence levels, (3) question quality and clinical appropriateness, and (4) physician assessment of patient preparedness during visits. While logistically challenging, this validation is necessary to justify integration into cancer care workflows.

References

Mohammad Abbasinia, Fazlollah Ahmadi, and Anoshirvan Kazemnejad. 2020. Patient advocacy in nursing: a concept analysis. *Nursing ethics*, 27(1):141–151.

Amogh Ananda Rao, Pei Lun Chen, Sydney Pugh, and Kevin B Johnson. 2025. Developing an llm-based conversational agent for primary-care pre-visit planning (pcp-bot).

John W Ayers, Adam Poliak, Mark Dredze, Eric C Leas, Zechariah Zhu, Jessica B Kelley, Dennis J Faix, Aaron M Goodman, Christopher A Longhurst, Michael Hogarth, and 1 others. 2023. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA internal medicine*, 183(6):589–596.

Xinyue Chen, Kunlin Ruan, Kexin Phyllis Ju, Nathan Yap, and Xu Wang. 2025. More ai assistance reduces cognitive engagement: Examining the ai assistance dilemma in ai-supported note-taking. *Proceedings of the ACM on Human-Computer Interaction*, 9(7):1–29.

Glyn Elwyn, Padhraig Ryan, Daniel Blumkin, and William B Weeks. 2024. Meet generative ai… your new shared decision-making assistant. *BMJ evidence-based medicine*, 29(5):292–295.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Sandra G Hart. 2006. Nasa-task load index (nasa-tlx); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 50, pages 904–908. Sage publications Sage CA: Los Angeles, CA.

Yu Lun Hsu, Yun-Rung Chou, Chiao-Ju Chang, Yu-Cheng Chang, Zer-Wei Lee, Rokas Gipiškis, Rachel Li, Chih-Yuan Shih, Jen-Kuei Peng, Hsien-Liang Huang, and 1 others. 2025. Precare: Designing ai assistants for advance care planning (acp) to enhance personal value exploration, patient knowledge, and decisional confidence. *arXiv preprint arXiv:2505.09115*.

Eunkyung Jo, Daniel A Epstein, Hyunhoon Jung, and Young-Ho Kim. 2023. Understanding the benefits and challenges of deploying conversational ai leveraging large language models for public health intervention. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pages 1–16.

Melina Klepsch, Florian Schmitz, and Tina Seufert. 2017. Development and validation of two instruments measuring intrinsic, extraneous, and germane cognitive load. *Frontiers in psychology*, 8:1997.

Nesrine Lajmi, Mehul Patel, Gareth Obery, Archana Dorge, Ashish Sharma, Jack Halligan, and Ernest Lo. 2026. Simulation-based evaluation of a large language model–enabled clinical decision support platform in oncology. *JCO Clinical Cancer Informatics*, (10).

Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan Ilgen, Emma Pierson, Pang Wei W Koh, and Yulia Tsvetkov. 2024. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. *Advances in Neural Information Processing Systems*, 37:28858–28888.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers)*, pages 3469–3483.

Yixin Liu and Pengfei Liu. 2021. SimCLS: A simple framework for contrastive learning of abstractive summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1065–1072, Online. Association for Computational Linguistics.

Donald R Martin and Bryan K Tipton. 2007. Patient advocacy in the usa: Key communication role functions. *Nursing & Health Sciences*, 9(3):185–191.

Nahara Anani Martinez-Gonzalez, Andreas Plate, Stefan Markun, Oliver Senn, Thomas Rosemann, and Stefan Neuner-Jehle. 2019. Shared decision making for men facing prostate cancer treatment: a systematic review of randomized controlled trials. *Patient preference and adherence*, pages 1153–1174.

Reza Negarandeh, Fatemeh Oskouie, Fazlollah Ahmadi, Mansoure Nikravesh, and Ingalill Rahm Hallberg. 2006. Patient advocacy: barriers and facilitators. *BMC nursing*, 5(1):3.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Comfort Nsiah, Mate Siakwa, and Jerry PK Ninnoni. 2020. Barriers to practicing patient advocacy in healthcare setting. *Nursing Open*, 7(2):650–659.

Annette M. O'Connor. 2003. The ottawa decision support framework (dsf) guiding the development of decision aids for patients and practitioners considering health care intervention. `http://204.187.39.28/resources.html`. [Cited 15 December 2003].

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Prostate Cancer Foundation. 2023. *Patient Guide to Localized Prostate Cancer*. Prostate Cancer Foundation. Information on diagnosis, treatment, side effects, and more for patients with early-stage prostate cancer.

Mathieu Ravaut, Shafiq Joty, and Nancy Chen. 2022. Summareranker: A multi-task mixture-of-experts re-ranking framework for abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4504–4524.

Ananya B Sai, Akash Kumar Mohankumar, Siddhartha Arora, and Mitesh M Khapra. 2020. Improving dialog evaluation with a multi-reference adversarial dataset and large scale pretraining. *Transactions of the Association for Computational Linguistics*, 8:810–827.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, Shekoofeh Azizi, Karan Singhal, Yong Cheng, Le Hou, Albert Webson, Kavita Kulkarni, S Sara Mahdavi, Christopher Semturs, Juraj Gottweis, and 6 others. 2024. Towards conversational diagnostic ai.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and 1 others. 2024. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450.

A Per-Disease Human Study Results

Metric	Plain LLM	Ours
Overall Mental Effort ↓	4.80	**4.20**
Intrinsic Load (IL1)	3.80	4.50
Extraneous Load (EL1) ↓	1.80	**1.40**
Extraneous Load (EL2) ↓	1.40	**1.40**
Germane Load (GL1) ↑	6.10	**6.30**
Germane Load (GL2) ↑	5.30	**5.70**

Table 6: Diabetes-only results (n=10).

Metric	Plain LLM	Ours
Overall Mental Effort ↓	4.30	**3.40**
Intrinsic Load (IL1)	3.30	3.70
Extraneous Load (EL1) ↓	3.50	**3.20**
Extraneous Load (EL2) ↓	3.00	**2.70**
Germane Load (GL1) ↑	4.90	**6.10**
Germane Load (GL2) ↑	5.20	**5.60**

Table 7: Prostate cancer-only results (n=10).

B Simulated Patient Example

Below is an example data point from our test set. The simulation prompt used to generate patient responses from these materials is in Appx D.4.

Post context (used as patient profile): "Hello everyone. I'm 46 and was diagnosed with pc on October 1st. PSA jumped from 6.69 to 7.97 in 3 months. Had a biopsy and 11 out of 12 core samples were positive for cancer. I have a Gleason 7 possibly a Gleason 8, localized in the prostate with capsular abutment. I chose radiation (7 weeks) and 6 months of ADT (Lupron) was recommended. I'm on week 3 of ADT and just started experiencing hot flashes and discomfort in my chest."

Reference questions:

- **Original:** "What do y'all think of my diagnosis"
- **Refined:** "Based on my PSA rise, biopsy (11/12 cores positive), Gleason 7–8, and capsular abutment, how severe is my prostate cancer and what risk category does it likely fall into?"

Atomic facts (simulation constraints):

1. The patient is 46 years old.
2. PSA rose from 6.69 to 7.97 in three months.
3. 11 out of 12 biopsy cores were positive.
4. Gleason score of 7, possibly 8.
5. Cancer is localized with capsular abutment.
6. Chose radiation therapy (seven weeks).
7. Six months of ADT (Lupron) recommended.
8. Currently on week 3 of ADT.
9. Experiencing hot flashes and chest discomfort.

C Retrieval Sources

Prostate cancer. Patient Guide to Localized Prostate Cancer (Prostate Cancer Foundation, 2023), published by the Prostate Cancer Foundation. Pages 1–60 used for retrieval.

Type 2 diabetes. Four sources:

- NIH: Understanding Type 2 Diabetes[3]
- NIH: Type 2 Diabetes Mellitus[4]
- Mayo Clinic: Type 2 Diabetes – Diagnosis & Treatment[5]
- Mayo Clinic: Type 2 Diabetes – Symptoms & Causes[6]

D Prompts

D.1 Template-Guided Response Generation

The following prompt is used to generate each interviewer turn in the balanced design (Sec 3.4). The LLM must include the template question while adding personalization.

```
You are a medical interviewer conducting a
patient consultation. This is stage one of
the interview, focused on exploring patient
information to understand their decision-
making landscape.

Based on the conversation history below,
generate personalized responses that:

1. MUST include the following template
   question exactly as provided:
   "{template_question}"
2. You may add ONE personalized segment
   BEFORE the template question that
   responds to what the patient just said
   (acknowledging, providing useful
   information, or building on their answer)
3. You may optionally add ONE additional
   personalized question either BEFORE or
   AFTER the template question
4. The purpose is to make the response more
   responsive and personalized while
   following the clinical structure

The template question you MUST include:
{template_question}

Conversation History:
{conversation_text}

Generate the complete response including:
- Required: A personalized segment (1-3
  sentences) responding to the patient
- Optional: One additional personalized
  question (if you decide to add one)
- Required: The template question exactly
  as provided

Format your response naturally, as if
speaking to the patient.
```

[3] https://www.ncbi.nlm.nih.gov/books/NBK513253/
[4] https://www.ncbi.nlm.nih.gov/books/NBK553486/
[5] https://www.mayoclinic.org/diseases-conditions/type-2-diabetes/diagnosis-treatment/drc-20351199
[6] https://www.mayoclinic.org/diseases-conditions/type-2-diabetes/symptoms-causes/syc-20351193

D.2 Knowledge Panel Generation with RAG

The following prompt is used to generate the knowledge panel after Interview Stage 1 (Sec 3.3). The LLM receives retrieved context with page numbers and must produce structured output grounded only in the provided evidence.

```
The following is a self-description from a
patient:
{query_str}
The following is relevant information
extracted from educational sources.
Each segment is labeled with its page number
in the format: [Page X] followed by content.
Please track which page numbers you actually
use when generating the knowledge content.
---------------------
{context_str}
---------------------
We are planning a follow-up response during
an interview with the patient. The response
will follow this template:
---------------------
Thanks for sharing your story. Based on what
you've shared, we've identified the following
information that may help you make a decision
about ${decision_topic} or address your
concerns in this area.
* Knowledge
${knowledge}
* Key factors related to your concern
${factors}
* Possible options and their associated
  benefits or risks
${options}
---------------------
Task Requirements:
1. Identify the patient's decision topic
   based on their self-description.
2. Fill in ${knowledge} using accurate
   information extracted from the context.
3. Infer key influencing factors related to
   the patient's decision topic.
4. List all possible options including
   benefits and risks.
5. Use only the extracted context when
   generating ${knowledge}.
6. Track which page numbers you actually
   used. List all unique page numbers.
7. Output in JSON format with fields:
   response, decision_topic, knowledge,
   factors, options, page_numbers.
```

D.3 Visit Preparation Question Generation

The following prompt is used to generate know-them and ask-them questions after Interview Stage 2 (Sec 3.3). The LLM generates questions grounded in retrieved evidence (know-them) and questions requiring clinician input (ask-them).

```
The following is a self-description about a
patient:
{query_str}
The following is relevant information
extracted from educational sources:
---------------------
{context_str}
---------------------
A patient has shared their concerns in the
interview. To assist them in preparing for
their outpatient visit, your task is to:
1. Generate five questions directly related
   to their concerns that can be answered
   using the provided book information.
   - Append the corresponding segments
     extracted from the book after each.
2. Generate five additional questions
   directly related to their concerns that
   cannot be answered using the provided
   book information but should be addressed
   by a doctor.
   - Leave them unanswered to indicate they
     require a doctor's input, using
     first-person tone.
All questions should:
- Be specific to the patient's concerns.
- Reflect the context provided in the book
  and their situation accurately.
- Help guide their preparation and ensure a
  meaningful discussion with their doctor.
Output in JSON format with fields:
  book_questions: [{question, answer}, ...],
  doctor_questions: [{question}, ...]
```

D.4 Simulated Patient Response

The following prompt is used for simulated patient role-play. The patient responds using only atomic facts for factual questions, and may infer preferences/feelings from context.

```
You are a patient participating in an
interview to prepare for an upcoming medical
visit. You are being asked questions by an
AI assistant to help identify your concerns
and decision-making needs.

Below is a list of atomic facts about you,
use ONLY the information in this list and
answer the assistant's question.
{atomic_facts_text}

Context (for reference only):
{context_text}

Question from the assistant: {question}

IMPORTANT RULES:
- If options are provided: Respond with ONLY
  the option (number or text).
- For factual questions (dates, numbers,
  test results): Use atomic facts. If none
  match, say "I don't have that information."
- For preference/value/feeling questions:
  Infer reasonable answers from context.
  Atomic facts can serve as evidence.

Instructions:
- Use atomic facts for factual information
- Derive preferences, values, and concerns
  from context
- Convert to first person perspective
- Be authentic and realistic

Generate your answer in this format:

STATEMENTS:
[selected atomic facts, or "No matching
statements"]

FIRST PERSON:
```

```
[your response based on atomic facts
(factual) or context (preferences/feelings)]
```

Role	Content
Assistant	Hello. I am your Prostate Cancer Patient Assistant. I'm here to help you prepare for your upcoming medical visit. To do that, I'd like to understand what choices or uncertainties you're currently facing in your prostate cancer journey. This will help ensure you get the most out of your conversation with your doctor. Are you ready to get started?
Patient	Yes. Let's get started
Assistant	First, we would like to know whether you currently have any concerns, questions, or decisions related to one of the optional areas. Please select the category of concerns you're currently thinking about as part of your decision-making.
Patient	Long-Term Management and Monitoring
Assistant	It's great to hear that you're thinking about long-term management and monitoring, as it's an important part of managing prostate cancer effectively. Understanding your plan can help you feel more in control and prepared for the future. What do you understand about your follow-up care or monitoring plan? Is there any specific aspect of your long-term management that you're particularly concerned about or want more information on?
Patient	From what I understand, my follow-up care will mostly involve regular PSA tests and maybe occasional imaging...
Assistant	It sounds like you're doing a great job of staying informed about your follow-up care, which is crucial for long-term management. Balancing this with arthritis can be challenging, so prioritizing self-care and understanding warning signs is key. How do you currently manage your arthritis symptoms, and have you discussed any integrated care approaches with your doctor? Do you have any questions or concerns about how you'll manage this condition in the future? What are they?
Patient	Right now, I manage my arthritis mostly with pain medication and trying to stay active when I can, though ...
Assistant	It sounds like you're handling a lot, and it's understandable to feel concerned about managing both arthritis and prostate cancer. It's important to remember that integrating care for both conditions can help you maintain a balanced lifestyle and reduce stress. Have you ever considered or tried any non-medication strategies, such as physical therapy or lifestyle changes, that might help manage both your arthritis and prostate health?
Patient	I've thought about it, but I haven't really tried anything structured like physical therapy yet...
Assistant	It's great that you're thinking about incorporating structured approaches like physical therapy into your routine, as they can offer significant benefits for managing both arthritis and prostate health. Having a tailored plan can make it easier to stay consistent and feel more in control. Have you ever discussed with a healthcare professional the possibility of a tailored exercise program or received advice on specific lifestyle changes that could benefit both conditions?
Patient	No, I haven't really talked with a healthcare professional about a specific exercise program or lifestyle changes that ...
Assistant	**Knowledge Panel** Thanks for your sharing. Based on what you've shared, we've identified the following information that may help you make a decision about **long-term management and monitoring** or address your concerns in this area. **Knowledge:** Active surveillance is a common approach for managing low-risk prostate cancer, allowing patients to maintain quality of life while monitoring for any signs of progression. Regular PSA tests, digital rectal exams, and periodic biopsies or MRIs are typical components of this strategy. It's important to discuss with your doctor the frequency of these tests and any lifestyle changes that could support your health during this period. **Key factors:** (1) Balancing prostate cancer monitoring with arthritis management; (2) Understanding the frequency and type of tests required for active surveillance; (3) Exploring integrated care approaches for managing both prostate cancer and arthritis. **Options:** *Active Surveillance* – Benefits: Maintains quality of life, delays potential side effects. Risks: Requires regular monitoring, potential anxiety. *Surgery or Radiation Therapy* – Benefits: Potentially curative, immediate treatment. Risks: Possible side effects impacting quality of life. *Integrated Care Approach* – Benefits: Holistic management of both conditions, potentially improving overall well-being. Risks: Requires coordination among healthcare providers, may involve lifestyle adjustments. *For your reference: This information is based on the Patient Guide to Localized Prostate Cancer (Page 18, Page 29-30, Page 41, Page 45).*
Patient	I have read it.
Assistant	Having professional guidance on a tailored exercise program could make a significant difference in managing both your arthritis and prostate health, especially considering your current challenges with consistency due to pain. It may be beneficial to bring this up with your healthcare team to explore options that fit your needs. Which options do you currently see as available to you? Please list all of them out.
Patient	At the moment, the ways I'm managing both my arthritis and prostate health are mostly using my pain meds for arthritis ...
Assistant	It's commendable that you are actively managing your symptoms and considering professional guidance to improve your control over both conditions. Discussing a tailored plan with your healthcare provider could ensure you have a comprehensive strategy that accommodates your arthritis while keeping up with prostate cancer monitoring. Have you thought about setting specific goals with your doctor to address both conditions together? What parts still feel unclear?
Patient	I've thought about setting goals with my doctor, but we haven't really done that yet...

Table 8: Example conversation between the system and a simulated patient (Part 1 of 2). Template questions are highlighted in blue.

Role	Content
Assistant	It's great that you're considering setting specific goals with your doctor, as this can provide a clearer roadmap for managing both your arthritis and prostate health. Having a structured plan can help you feel more confident and in control. Which factors matter most to you in this decision? (If possible, list your top 2–3.) You can also propose your own factors. Balancing prostate cancer monitoring with arthritis management. Understanding the frequency and type of tests required for active surveillance. Exploring integrated care approaches for managing both prostate cancer and arthritis.
Patient	What matters most to me is finding a balance between keeping up with prostate cancer monitoring and managing ...
Assistant	It sounds like you're focused on achieving a balance that allows you to manage both your prostate cancer monitoring and arthritis effectively. Understanding the necessary frequency and types of tests can certainly help alleviate some stress, making it easier to commit to your health plan. Which tradeoff feels hardest to weigh right now?
Patient	The hardest part for me right now is figuring out how to stay active with my arthritis while also keeping up with ...
Assistant	Finding the right balance between staying active with arthritis and keeping up with prostate cancer monitoring is indeed a challenging task, especially when dealing with pain and stiffness. It might be helpful to discuss with your healthcare team strategies that can accommodate both needs, perhaps through modified exercises or a flexible monitoring schedule that considers your arthritis flare-ups. What makes you feel less confident or ready?
Patient	I think what makes me feel unprepared is not really knowing what I can safely do with my arthritis while still keeping ...
Assistant	It's understandable to feel uncertain about how much activity is right for you, especially when managing two conditions. Your healthcare team can provide guidance on exercises that accommodate both your arthritis and prostate cancer needs, helping you feel more confident in your routine. Have you discussed this with anyone (family, friends, clinicians, support groups)? What did they say?
Patient	Not really, I haven't talked much about it with anyone yet...
Assistant	It sounds like you're doing your best to navigate these challenges on your own, and reaching out to your family is a great first step. Joining a support group or discussing your concerns with your healthcare team might provide additional insights and encouragement. What steps do you think you could take to start involving your healthcare team in creating a plan that addresses both your arthritis and prostate cancer needs?
Patient	I think the first step would be to have a proper conversation with my doctor where I can explain all my concerns ...
Assistant	It's a great idea to bring a list of questions and track your symptoms; this can provide your doctor with a clearer picture of your needs. Collaborating with a physical therapist or specialist could offer tailored strategies that align with both your arthritis and prostate cancer management. How comfortable do you feel about initiating these discussions with your healthcare team, and is there anything specific you might need to help you prepare for that conversation?
Patient	I feel a little nervous about bringing it up because I don't want to take up too much time or forget something important ...
Assistant	Thank you for your time today. We have received all your responses. This interview has been completed.
Assistant	**Know-Them Questions:** *Educational background: The answers below are from Patient Guide to Localized Prostate Cancer. Not medical advice—please confirm with your clinician.* • **What type of testing will be required if I do active surveillance?** Men on active surveillance will usually have a PSA blood test done twice per year, an annual DRE, and repeat biopsies generally every 1 to 5 years. MRI may also be used to help monitor the cancer and guide timing and decision for repeat biopsies. • **How frequently will I be tested under active surveillance?** PSA testing is done twice per year, and MRIs are generally performed every two years to monitor any changes in the prostate. • **What are proactive changes to my health habits that I should consider?** Exercise as regularly as you can, most days of the week, and increase the intensity of your exercise as you are able. An anti-inflammatory, heart-healthy diet is also recommended. • **What is my baseline PSA number, and what number would be concerning?** Questions about the specific tests needed under active surveillance and PSA benchmarks are important to discuss with your doctor. • **How does exercise impact prostate cancer recovery and long-term health?** Exercise reduces risk of prostate cancer recurrence and improves survival. It also supports better sleep, cognition, and reduces depression and anxiety. **Ask-Them Questions:** • How can I safely integrate exercises that benefit both my arthritis and prostate health? • What specific lifestyle changes should I make to manage both prostate cancer and arthritis effectively? • Are there any specific warning signs I should watch for at home that indicate a change in my prostate cancer status? • How can I balance the need for physical activity with the limitations imposed by my arthritis? • Would involving a physical therapist be beneficial for creating a tailored exercise program for my conditions?

Table 9: Example conversation between the system and a simulated patient (Part 2 of 2). Template questions are highlighted in blue.

Role	Content
Assistant	Hello. I am your Prostate Cancer Patient Assistant. I'm here to help you prepare for your upcoming medical visit. To do that, I'd like to understand what choices or uncertainties you're currently facing in your prostate cancer journey. This will help ensure you get the most out of your conversation with your doctor. Are you ready to get started?
Patient	Yes. Let's get started
Assistant	First, we would like to know whether you currently have any concerns, questions, or decisions related to one of the optional areas. Please select the category of concerns you're currently thinking about as part of your decision-making.
Patient	Diagnosis and Screening
Assistant	Got it. Diagnosis and screening can be challenging to navigate. Let me ask a few questions to better understand your situation: 1. Diagnosis Understanding: How comfortable do you feel with the details of your diagnosis? Do you have any specific questions about your test results or what they mean? 2. Screening Tests: Have you had any recent prostate-specific antigen (PSA) tests, biopsies, or imaging studies? If so, do you need clarification on their purpose or results? 3. Next Steps: Are you aware of the recommended next steps in your screening or diagnostic process, and do you understand why they are suggested? 4. Concerns or Symptoms: Do you have any symptoms or concerns that you feel haven't been addressed yet? Your answers will help us tailor the conversation to focus on the most important aspects of your upcoming doctor visit.
Patient	I've had multiple PSA tests and biopsies and imaging studies what would be next.
Assistant	It sounds like you've already undergone a comprehensive set of tests. Here are some potential next steps and questions you might consider for your doctor visit: 1. Discussing Results: Make sure to review the results of your PSA tests, biopsies, and imaging studies with your doctor. Ask for an explanation of what the results mean for your specific situation. 2. Staging and Grading: Ask about the stage and grade of your prostate cancer, as this information is crucial for determining the appropriate treatment plan. 3. Treatment Options: Inquire about the available treatment options based on your current diagnosis. Common treatments include active surveillance, surgery, radiation therapy, hormone therapy, or a combination of these. 4. Risks and Benefits: Discuss the potential risks and benefits of each treatment option, and how they align with your personal values and lifestyle. 5. Second Opinion: Consider asking if a second opinion might be beneficial, especially if you're uncertain about the recommended treatment plan. 6. Support and Resources: Ask about resources for support, such as local support groups or counseling services, to help you through this process. These questions can guide your conversation with your doctor and help you make informed decisions about your care. How does that sound to you? Is there anything else you'd like to discuss?
Patient	The staging and grading Information i'm most focused on.
Assistant	Thank you for sharing your focus on staging and grading information. This helps clarify the aspects that are most important to you in your discussion with your doctor. Remember to ask for detailed explanations to ensure you understand the implications for your treatment. Thank you for sharing your concerns. Your focus on staging and grading is crucial for your treatment decisions. The interview is complete.
...	...

Table 10: Example interview process between plain LLM and a simulated patient. Plain LLM tends to ask many questions in a single message.