

C-PATH: Conversational Patient Assistance and Triage in Healthcare System

Qi Shi[✉], Qiwei Han[✉]

School of Business and Economics
Universidade Nova de Lisboa
Carcavelos, Portugal
{qi.shi, qiwei.han}@novasbe.pt

Cláudia Soares[✉]

School of Science and Technology
Universidade Nova de Lisboa
Caparica, Portugal
claudia.soares@fct.unl.pt

Abstract—Navigating healthcare systems can be complex and overwhelming, creating barriers for patients seeking timely and appropriate medical attention. In this paper, we introduce C-PATH (Conversational Patient Assistance and Triage in Healthcare), a novel conversational AI system powered by large language models (LLMs) designed to assist patients in recognizing symptoms and recommending appropriate medical departments through natural, multi-turn dialogues. C-PATH is fine-tuned on medical knowledge, dialogue data, and clinical summaries using a multi-stage pipeline built on the LLaMA3 architecture. A core contribution of this work is a GPT-based data augmentation framework that transforms structured clinical knowledge from DDXPlus into lay-person-friendly conversations, allowing alignment with patient communication norms. We also implement a scalable conversation history management strategy to ensure long-range coherence. Evaluation with GPTScore demonstrates strong performance across dimensions such as clarity, informativeness, and recommendation accuracy. Quantitative benchmarks show that C-PATH achieves superior performance in GPT-rewritten conversational datasets, significantly outperforming domain-specific baselines. C-PATH represents a step forward in the development of user-centric, accessible, and accurate AI tools for digital health assistance and triage.

Index Terms—Large Language Models, Conversational AI, Patient Navigation, Medical Triage, Digital Health, Clinical Dialogue Systems

I. INTRODUCTION

A patient navigator in the medical domain is defined as “a barrier-focused intervention designed to help individual patients overcome obstacles to accessing and navigating the healthcare system promptly” [45]. These roles include identifying barriers at the patient level, facilitating appointments, supporting communication between patients and providers, and providing health education [29], [33]. Despite their potential, patient navigation services face significant challenges due to variability in navigator qualifications, lack of standardization, and limited certification protocols [9], [22], [40].

Recent advances in large language models (LLMs), such as ChatGPT, have prompted a growing interest in their application to medical contexts [24], [31], [39]. LLMs have shown strong performance in clinical tasks, including question answering [23], summarization [42], and information extraction [24], and have shown potential to improve clinical reasoning [36]. However, their integration in patient-facing medical conversations remains underdeveloped due to risks such as

hallucination [5], misinformation [21], and limited alignment of health literacy [6]. Additionally, the conversational fluency and trustworthiness of LLM outputs pose usability concerns, particularly in patient triage and navigation scenarios [20], [38].

In this paper, we introduce C-PATH (Conversational Patient Assistance and Triage in Healthcare), an LLM-based system designed to facilitate symptom recognition and triage patients to the appropriate medical departments through natural, multi-turn conversations. C-PATH uses patient-friendly terminology, allowing laypeople to describe symptoms effectively and understand AI responses clearly. Our approach consists of three key technical contributions:

- 1) We design a multi-stage fine-tuning pipeline using an open-source LLaMA3 model, with stages aimed at acquisition of medical knowledge, understanding of dialogue, and summarization.
- 2) We propose a novel dataset construction framework that converts structured differential diagnosis cases into realistic, multi-turn doctor-patient conversations using GPT prompting, allowing alignment with patient communication norms.
- 3) We introduce a scalable multi-turn dialogue management module that prunes history and optionally summarizes earlier turns to stay within context limits while maintaining interaction coherence.

Our experiments demonstrate that C-PATH not only produces coherent and accessible dialogues but also offers accurate department-level triage recommendations. Evaluations using GPTScore show improved understandability, informativeness, and specificity, while performance benchmarks indicate superior results compared to baselines.

In general, C-PATH offers a promising step toward improving access and navigation to digital healthcare. By automating the initial check and triage of symptoms, our system has the potential to reduce wait times, optimize primary care doctor and specialist referrals [10], [41], and improve the efficiency of hospital workflows [16]. The automatic generation of EHR-compatible summaries further integrates LLMs into clinical documentation pipelines, creating value for both patients and providers.

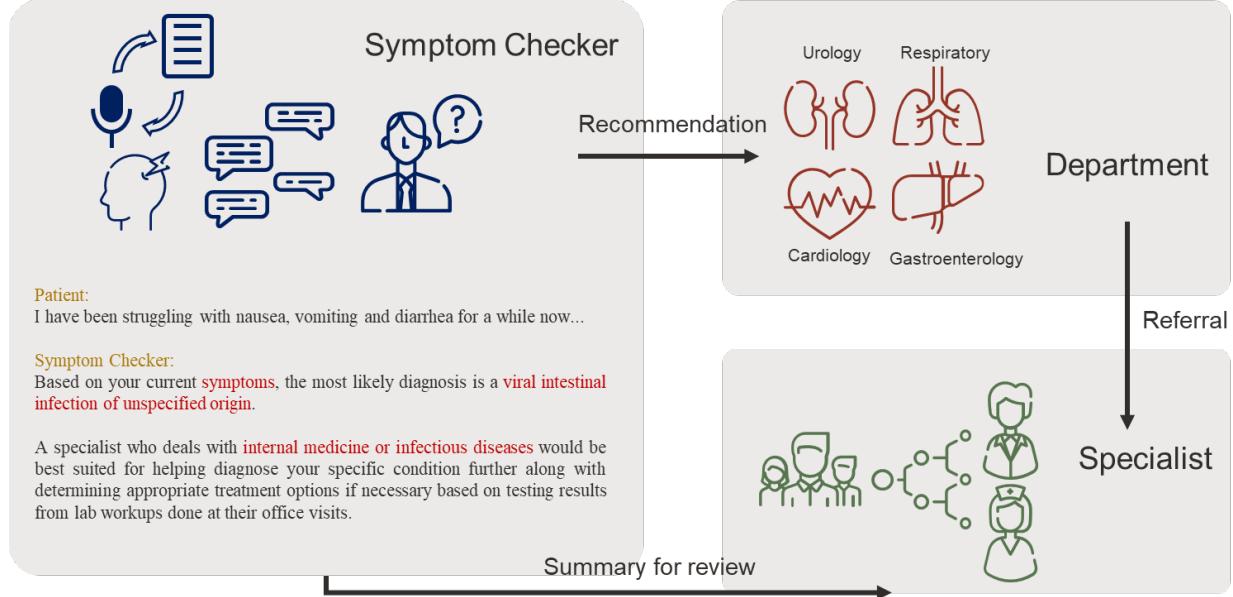


Fig. 1: Overview of the C-PATH model framework.

II. RELATED WORK

Symptom checkers for patient navigation have gained increasing attention as tools to improve the accuracy and efficiency of the diagnosis of medical conditions [43]. Their development has been particularly important in public health crises, such as the COVID-19 pandemic, where timely monitoring and rapid response are essential. For example, [30] highlighted how the symptom checkers maintained public safety by tracking and reporting infection trends. Similarly, [50] demonstrated that platforms like Symptoma provided early warnings of COVID-19 spread, offering alternative data sources for policymakers and often anticipating official reports.

Another benefit of symptom checkers lies in user satisfaction, especially when integrated with self-triage and scheduling tools. [27] found that such systems, when used during the pandemic, reduced unnecessary communication with healthcare providers by empowering users to schedule tests or consultations directly. This streamlining of patient flow has demonstrated significant operational value.

The accuracy of symptom checkers is also essential for positive health outcomes. For example, [35] evaluated Flo Health's symptom checkers to identify conditions such as endometriosis, uterine fibroids, and polycystic ovary syndrome. They reported high diagnostic accuracy, suggesting that these tools could reduce time to diagnosis and improve healthcare outcomes. In a different domain, [28] showed that integrating AI into microscopy imaging improved malaria diagnosis in low-resource settings, demonstrating a greater applicability of AI-powered diagnostic tools.

However, symptoms checkers show notable limitations in emergency care. [1] compared them to emergency physicians using Objective Structured Clinical Examinations and found that human experts significantly outperformed symp-

tom checkers in both primary and secondary diagnoses. This highlights the need to limit such tools to supportive rather than definitive roles in emergency contexts [12].

Consequently, AI-powered systems are better positioned to support healthcare workflows than to replace clinical decision-making. [15] investigated how AI improves automation in digital healthcare systems, finding benefits in operational efficiency and patient outcomes. However, [48] noted that current symptom checkers often do not meet expectations for conversational design, input flexibility, and language clarity, indicating the need for more user-friendly AI interfaces.

Recent progress in large language models (LLMs) provides a pathway to address these design shortcomings. Proprietary models such as Google's Med-PaLM 2 and MedLM [36], and open-source alternatives such as HuaTuo [44] and ChatDoctor [26], have shown promise. HuaTuo, for example, incorporates structured and unstructured medical knowledge from the Chinese Medical Knowledge Graph (CMeKG) to improve domain specificity. ChatDoctor leverages on-line medical dialogues for fine-tuning, producing more contextually relevant responses.

Despite these advancements, significant challenges remain for the deployment of LLMs in real-world healthcare applications. These include comprehensive medical knowledge [25], managing patient medical history [14], accurately recording conversations [2], and designing intuitive interfaces [37]. Overcoming these issues is vital for the effective integration of LLM-based conversational systems into hospital infrastructure, with the goal of enhancing, not replacing, professional healthcare services [46].

III. METHODOLOGY

Figure 1 provides an overview of the interaction flow and functionality of the model within the healthcare system. Initially, the component of the symptoms checker interacts with the patient to gather information about the symptoms through multi-turn conversations and perform a preliminary diagnosis. Based on this information, it recommends the most appropriate medical department for the patient’s condition. Subsequently, the patient is referred to the relevant specialist. Throughout this process, the model also generates a concise summary of the interaction as Electronic Health Records (EHR), which can be reviewed by the specialist to inform further clinical decision-making and treatment.

A. Training Workflow

1) *Knowledge Injection:* The LMFlow toolkit is used throughout the training and fine-tuning workflows. LMFlow is an open-source large language model training and fine-tuning framework designed to support mainstream open-source models [8]. It natively integrates efficient parameter fine-tuning technologies, such as LoRA, significantly reducing training costs and allowing rapid customization and optimization of large-scale models.

The model employed in this study is based on LLaMA3 and is trained in a three-stage process (see Figure 2). In the first training stage, we focus on injecting extensive medical knowledge into the model. This step involves leveraging three widely used medical question-answering datasets: PubMedQA [19], MedQA-USMLE [18], and MedMCQA [32]. These datasets encompass diverse and comprehensive content, ranging from clinical case questions to medical examination preparation materials. They are crucial to equipping the model with a broad and deep understanding of medical concepts, terminologies, symptoms, and related medical information.

2) *Instruction Tuning:* After acquiring the foundational medical knowledge, the model proceeds to the second training stage, emphasizing fine-tuning medical conversation skills. In this phase, we use a doctor-patient conversation dataset designed to simulate realistic medical interactions and consultations [11]. The objective here is to enhance the model’s capability to engage effectively in multi-turn dialogues. This fine-tuning ensures that the model can accurately interpret and respond contextually to complex, prolonged medical conversations. Such improvements are critical for delivering medically relevant, coherent, and contextually appropriate conversational responses.

To enable such multi-turn interactions in a reliable and scalable manner, we introduce a dialogue history management mechanism to maintain coherence, manage token limits, and preserve conversational flow.

3) *Multi-Turn Dialogue and History Management:* Our system is designed to support multi-turn conversations between patients and the C-PATH. This capability is essential for simulating realistic medical assistance scenarios, where symptoms may be revealed progressively and clarification questions are necessary for accurate recommendation.

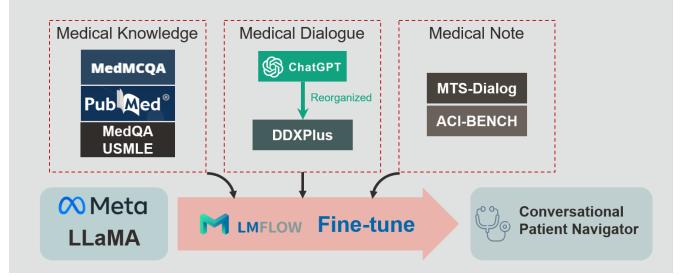


Fig. 2: Training workflow for C-PATH.

To ensure scalability and coherence, we adopt the following approach for conversation history management.

- **Context Window Pruning:** We apply a sliding window mechanism where only the most recent N dialogue turns are retained in the prompt context. This ensures that the total token length remains within the model’s maximum input length, while preserving the most relevant recent context.
- **Turn-Level Summarization:** For longer sessions, we optionally condense earlier turns into a system-level summary using a lightweight LLM-based summarizer before appending new conversational turns. This allows the assistant to retain essential background knowledge without exceeding token limits.
- **Speaker Tagging and Role Preservation:** Each conversational turn is explicitly tagged with speaker roles (e.g., [Patient], [Assistant]) to maintain turn alignment, preserve conversational roles, and prevent instruction leakage or loss of speaker identity across turns.

This history management framework is critical to enabling the model to track evolving symptom narratives, follow up with clarifying questions, and iteratively refine department recommendations over multiple turns of patient interaction. It also contributes to consistency in clinical logic and fluency in longer sessions.

4) *Summarization Tuning:* The third and final stage of the training workflow involves using summaries learned from summary notes of doctor-patient conversation scenarios [3], [47]. This step aims to transform the detailed conversations facilitated by the C-PATH into summarized clinical notes. These notes serve as references for medical specialists and are systematically integrated into Electronic Health Records (EHR), thus enhancing continuity of care and providing efficient documentation of patient interactions.

B. Model

LLaMA3: We selected LLaMA3-8B as the foundation model due to its strong performance on biomedical question-answering benchmarks such as PubMedQA. As shown in previous work [8], a fully fine-tuned LLaMA3-8B outperformed larger models like ChatGPT in domain-specific tasks. Compared to models such as GPT-J or BLOOM, LLaMA offers a better balance between performance, model size, and

computational efficiency, which is critical for future deployment in resource-constrained clinical environments, such as mobile platforms. Its open-source accessibility also provides full control over model customization, an essential feature in privacy-sensitive medical applications.

C. Data

1) *Medical Knowledge Datasets*: Three medical question-answering dataset are included in the fine-tuning process to inject medical domain knowledge into LLM as follows:

PubmedQA is a question-answering dataset based on the PubMed database, designed for biomedical applications [19]. It contains 1,000 expert-annotated QA instances, 612,000 unlabeled QA instances, and approximately 211,300 artificially generated QA instances.

MedQA-USMLE is a question-answering dataset specifically tailored for U.S. medical students and physicians preparing for the US Medical Licensing Examination [18]. It comprises 12,723 questions covering extensive medical knowledge relevant to clinical practice.

MedMCQA is a large-scale multiple-choice question-answer dataset containing approximately 194,000 records [32]. It covers over 2,400 healthcare topics across 21 distinct medical subjects, serving as a valuable resource for medical entrance exam preparation and comprehensive medical training.

2) *Medical Dialogue Datasets*: Doctor-patient dialogue scenarios typically occur in outpatient clinical settings, making these data challenging to collect. Existing open-source datasets, such as MedDialog [49], usually contain single-take conversations that focus primarily on diagnoses and treatment recommendations. This structure significantly differs from the conversational objectives of this paper, which aims to facilitate natural, multi-turn interactions between patients and medical guidance systems. To address these limitations and align more closely with the requirements, without loss of generality, we used a random sample of 5000 conversations from the DDXPlus dataset (`data_5k_ddxplus`).

DDXPlus is a comprehensive dataset comprising approximately 1.3 million patient cases, each containing differential diagnoses, confirmed pathologies, detailed descriptions of symptoms, and patient medical history [11]. Unlike traditional datasets that typically contain binary symptom indicators, DDXPlus provides categorical and multi-select symptom data organized hierarchically, which significantly facilitates logical and interactive patient conversations.

Although the DDXPlus dataset does not directly include conversational dialogues, we use its detailed symptom descriptions to construct artificial doctor-patient conversations. Initially, as Figure 3 shows, DDXPlus symptom-related questions often contained complex medical terminology, making them potentially difficult for patients to understand. To ensure clarity and promote natural human-like interactions, these queries were carefully rewritten using patient-friendly language. Medical terms were specifically avoided during the rewriting process, reflecting real clinical scenarios in which

physicians typically use layman’s terms to facilitate patient understanding. For example, the term “iliac wing(R)” was rewritten into easily comprehensible phrases such as “right side of my lower back” or “near my right hip.”

DDXPlus	
<pre>{ "name": "E_130", "code_question": "E_129", "question_fr": "De quelle couleur sont les lésions?", "question_en": "What color is the rash?", "is_antecedent": false, "default_value": "V_11", "value_meaning": { "V_11": {"fr": "NA", "en": "NA"}, "V_86": {"fr": "foncée", "en": "dark"}, "V_107": {"fr": "jaune", "en": "yellow"}, "V_138": {"fr": "pâle", "en": "pale"}, "V_156": {"fr": "rose", "en": "pink"}, "V_157": {"fr": "rouge", "en": "red"} }, "possible_values": ["V_11", "V_86", "V_107", "V_138", "V_156", "V_157"], "data_type": "C" } { "condition_name": "Myasthenia gravis", "cond-name-fr": "Myasthenie grave", "cond-name-eng": "Myasthenia gravis", "icd10-id": "G70.0", "symptoms": { "E_65": 0, "E_63": 0, "E_52": 0, "E_172": 0, "E_84": 0, "E_66": 0, "E_90": 0, "E_38": 0, "E_176": 0 }, "antecedents": { "E_28": 0, "E_204": 0 }, "severity": 3 }</pre>	<pre>{ "name": "E_130", "code_question": "E_129", "question_fr": "De quelle couleur sont les lésions?", "question_en": "What color is the rash?", "is_antecedent": false, "default_value": "V_11", "value_meaning": { "V_11": {"fr": "NA", "en": "NA"}, "V_86": {"fr": "foncée", "en": "dark"}, "V_107": {"fr": "jaune", "en": "yellow"}, "V_138": {"fr": "pâle", "en": "pale"}, "V_156": {"fr": "rose", "en": "pink"}, "V_157": {"fr": "rouge", "en": "red"} }, "possible_values": ["V_11", "V_86", "V_107", "V_138", "V_156", "V_157"], "data_type": "C" } { "condition_name": "Myasthenia gravis", "cond-name-fr": "Myasthenie grave", "cond-name-eng": "Myasthenia gravis", "icd10-id": "G70.0", "symptoms": { "E_65": 0, "E_63": 0, "E_52": 0, "E_172": 0, "E_84": 0, "E_66": 0, "E_90": 0, "E_38": 0, "E_176": 0 }, "antecedents": { "E_28": 0, "E_204": 0 }, "severity": 3 }</pre>

Fig. 3: Example of evidence and pathology description in DDXPlus.

To further enhance the realism and diversity of these conversations, each symptom-related question was reformulated into multiple variants using GPT-3.5 Turbo. For each question, as illustrated by E_183 in Figure 4, we generated four different versions to ensure variety. Questions about symptom severity, such as those illustrated by E_59, were also rewritten to provide multiple intuitive approaches to understanding severity rather than relying solely on numeric scales (1-10). Similarly, affirmative patient responses were diversified through commonly used conversational expressions generated by GPT-3.5 Turbo, such as “I think so,” “Absolutely,” “Of course,” “Definitely,” and “For sure,” rather than the simpler “Yes.” These reformulated questions and responses were compiled into an artificial dataset, termed `data_5k_artificial`. Furthermore, we used GPT-3.5 Turbo to completely rewrite the entire original DDXPlus-based conversation dataset, creating `data_5k_GPT`. This allowed us to directly compare the performance and effectiveness of GPT-generated conversations with manually revised data. While human rephrasing introduces richer linguistic variety, the GPT-3.5 rewriting consistently outperforms in downstream tasks due to its greater stylistic consistency, grammatical coherence, and normalized response structure. These qualities ensure smoother conversational flow, reduced ambiguity, and more uniform input formats, which are especially beneficial for model training and inference. Table I provides an overview of the three datasets used for evaluation.

3) *Medical Conversation Notes Datasets*: Typically, hospitals do not capture complete transcripts of doctor-patient conversations; instead, they record interactions as fragmented highlights or concise summaries in medical records. Two medical conversation notes datasets are employed to fine-tune the C-PATH for summarized doctor-patient conversations for specialist review:

E_183 Before: Do you live in a rural area?	Before: 'iliac wing(R)'
After: 1. Is your residency situated in a rural area? 2. Is your place of residence located in a rural area? 3. Is your home in a countryside setting? 4. Do you live in a rural area?	After: - 'Right side of my lower back' - 'Near my right hip' - 'Right side above the hip' - 'Right side of my waist' - 'On my right, where my waistline is'
E_59 Before: How fast did the pain appear?	
After: How fast did the pain appear? You can choose severity or score as following rules: No(0); Slow(1-3); more than 1 day; Moderate(4-6); less than 1 day; Fast(7-9); within several minutes; Immediately(10)	

Fig. 4: Example of term rewriting and question optimization.

TABLE I: Description of conversational datasets used for training and evaluation.

Dataset	Description
data_5k_ddxplus	5k multi-turn conversation samples directly composed of raw questions and answers from DDXPlus
data_5k_artificial	5k multi-turn conversation samples composed of manually rewritten questions and answers
data_5k_GPT	5k multi-turn conversation samples entirely rewritten by GPT-3.5 Turbo

MTS-Dialog dataset comprises approximately 1,700 short doctor-patient conversations accompanied by corresponding summaries [3]. Additionally, this dataset includes an augmented subset containing around 3,600 conversation-summary pairs. This augmented subset was created by back-translating the original 1,200 training pairs into French and Spanish, thereby increasing the dataset’s diversity and comprehensiveness.

ACI-BENCH dataset features complete doctor-patient conversations paired with detailed clinical notes. It incorporates data splits utilized in the MEDIQA-CHAT 2023 and MEDIQA-SUM 2023 challenges, providing robust benchmarks for clinical dialogue summarization tasks [47].

Given that these summary records frequently lack coherence, detail and completeness, to address these limitations, we further used GPT-3.5 Turbo to generate structured summaries from our designed conversations, specifically from the data set data_5k_ddxplus. Compared to traditional templated summarization methods, GPT-3.5 Turbo can more easily adapt to specific conversational contexts, producing summaries that are logical, professionally formatted, and thoroughly organized, thus significantly improving the quality of clinical documentation.

D. Fine-tuning Framework

To enhance the efficacy of generalist LLMs for medical applications, researchers have recognized the importance of training and/or fine-tuning these models on large, in-domain datasets [34]. To address this need, we integrated the LMFlow toolkit into the fine-tuning workflow [7], which simplifies the fine-tuning and inference processes of LLMs. It can incorporate instruction tuning methodologies designed to improve LLMs by explicitly training them to follow natural language

instructions and commands. The fine-tuning process for our model was performed on two Nvidia A10 graphics cards, employing the hyperparameters listed in Table II.

TABLE II: Hyperparameters used in fine-tuning the C-PATH model.

Hyperparameter	Description
<i>num_train_epochs</i> : 2	Number of epochs used for training the model on the collected dataset.
<i>learning_rate</i> : $2e - 5$	Learning rate used during model fine-tuning.
<i>block_size</i> : 128	Maximum sequence length for tokenized inputs; datasets are truncated into segments of 128 tokens for training.
<i>per_device_batch_size</i> : 6	Batch size per GPU used during fine-tuning.
<i>use_lora</i> : yes	Indicates whether Low-Rank Adaptation (LoRA) is employed.
<i>lora_r</i> : 8 bf16	Rank parameter for LoRA method. Specifies mixed precision mode as BF16 (brain floating point 16), providing enhanced numerical precision compared to FP16.
<i>dataloader_num_workers</i> : 1	Number of processes used for preprocessing during training.

Conversation Tuning is a specialized form of instruction tuning that enables LLMs to effectively handle conversational interactions, extending beyond simple text completion tasks. It demands that models not only understand context, but also manage multi-turn dialogues. Due to these specific requirements, conversation tuning requires more precise formatting compared to traditional text fine-tuning. Improper formatting can lead to performance errors even in adequately trained models. A crucial aspect distinguishing conversation tuning from general text fine-tuning is the use of special end-of-turn markers. Examples of such markers include `</s>` used in Vicuna and `<|im_end|>` employed in OpenAI’s ChatML. These markers explicitly indicate the end of each dialogue turn, enabling the model to determine when a suitable response is complete.

We adopted the conversational turning strategy to structure our conversational prompts dataset into multi-turn conversations. We used the marker `###` to indicate the end of each conversational exchange. As Figure 5 shows, the conversations were systematically decomposed into multiple individual dialogue turns, adding each turn sequentially, ensuring the complete preservation of contextual information.

E. Model Evaluation

To assess the potential performance of our model, we evaluated the quality of the DDXPlus dataset, previously introduced, as it is essential for enabling effective multi-turn dialogue interactions. To systematically assess the quality of the dialogue, we use GPTScore to assess the quality of the conversation in six dimensions (Specificity, Flexibility, Understandability, Informativeness, Patience and Accuracy) [13]. Unlike traditional NLG metrics such as BLEU or ROUGE—which are limited to surface-level lexical overlap—GPTScore allows

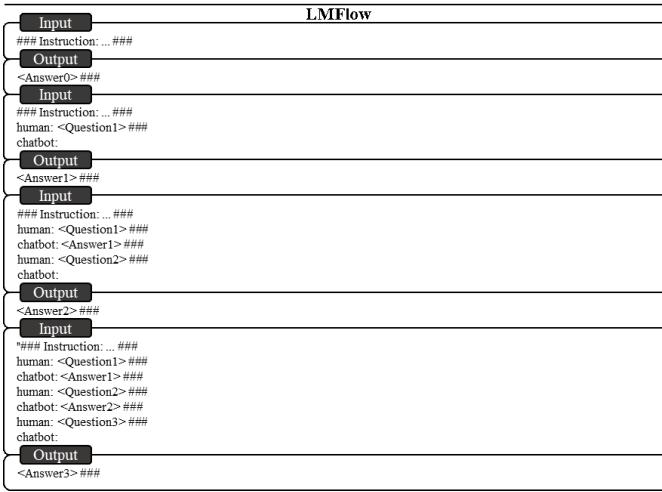


Fig. 5: Conversation formatting structure for fine-tuning with LMFlow.

for instruction-based, semantics-aware evaluation of generated text using natural language prompts. This makes it particularly well-suited for evaluating multi-turn, patient-facing medical conversations where nuance, tone, and clarity are critical. It also enables a more interpretable multidimensional assessment compared to black-box automatic metrics. Specifically, we evaluated different datasets derived from DDXPlus in six different aspects, as detailed in Table III.

Aspect	Definition
SPE	Are the responses of the Conversational Patient Navigator specific enough to the context rather than generic?
FLE	Is the Conversational Patient Navigator flexible and adaptive to individual patient interests and responses?
UND	Does the Conversational Patient Navigator clearly communicate information in a way easily understood by patients?
INF	Do the Conversational Patient Navigator's questions effectively gather sufficient information to provide accurate recommendations?
PAT	Do the questions from the Conversational Patient Navigator potentially lead to patient impatience?
ACC	Is the specialist recommended by the Conversational Patient Navigator accurately aligned with the patient's needs?

TABLE III: Definitions of evaluation aspects in GPTScore.

Since our model’s primary objective is to accurately direct patients to the appropriate medical departments, the accuracy of recommendations on the validation set is the critical metric for our evaluation. To benchmark this, we selected BERT, a widely recognized model also based on the Transformer architecture, as our baseline for comparison. In our evaluation, the entire conversation serves as input, with departmental recommendations as targets. Each dataset was partitioned into 70% training, 20% testing, and 10% validation subsets to ensure a thorough and robust evaluation process.

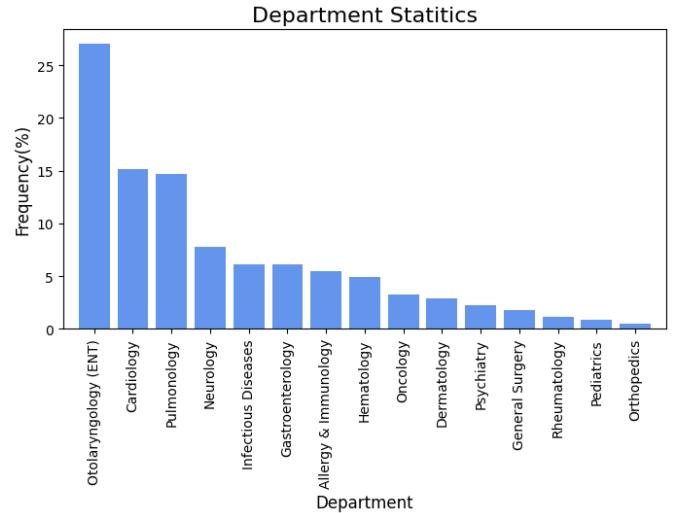


Fig. 6: Department frequency distribution in the data_5k_ddxplus dataset.

IV. RESULTS

A. Exploratory Data Analysis

Our data sampling strategy preserves the disease distribution in the original DDXPlus dataset by assigning each condition to its corresponding medical department. As shown in Figure 6, respiratory-related conditions dominate the dataset, introducing an inherent bias during model training. While this reflects real-world data availability, it highlights the importance of incorporating additional datasets in future work to broaden departmental coverage.

An essential challenge in LLM fine-tuning for multi-turn dialogue is the management of conversational history under input length constraints. Figure 7 analyzes the number of dialogue turns and token counts across the three datasets: data_5k_ddxplus, data_5k_artificial, and data_5k_GPT. While data_5k_artificial involves rewording original content, it maintains similar turn counts to the raw dataset, sometimes exceeding 40 turns per sample. In contrast, data_5k_GPT exhibits a smoother distribution with most conversations under 30 turns. This reduction likely reflects the more coherent structure introduced during GPT-based rewriting.

Token distribution results (Figures 7b and 7d) further reveal that data_5k_artificial significantly increases token counts due to its use of expanded layperson language. These longer samples risk exceeding the 1024-token input limit, complicating training and inference. Conversely, data_5k_GPT demonstrates tighter control over token lengths, reducing context overflow and improving training stability. Additionally, the GPT rewriting process normalizes departmental variation, reducing noise and enhancing consistency in conversation structures.

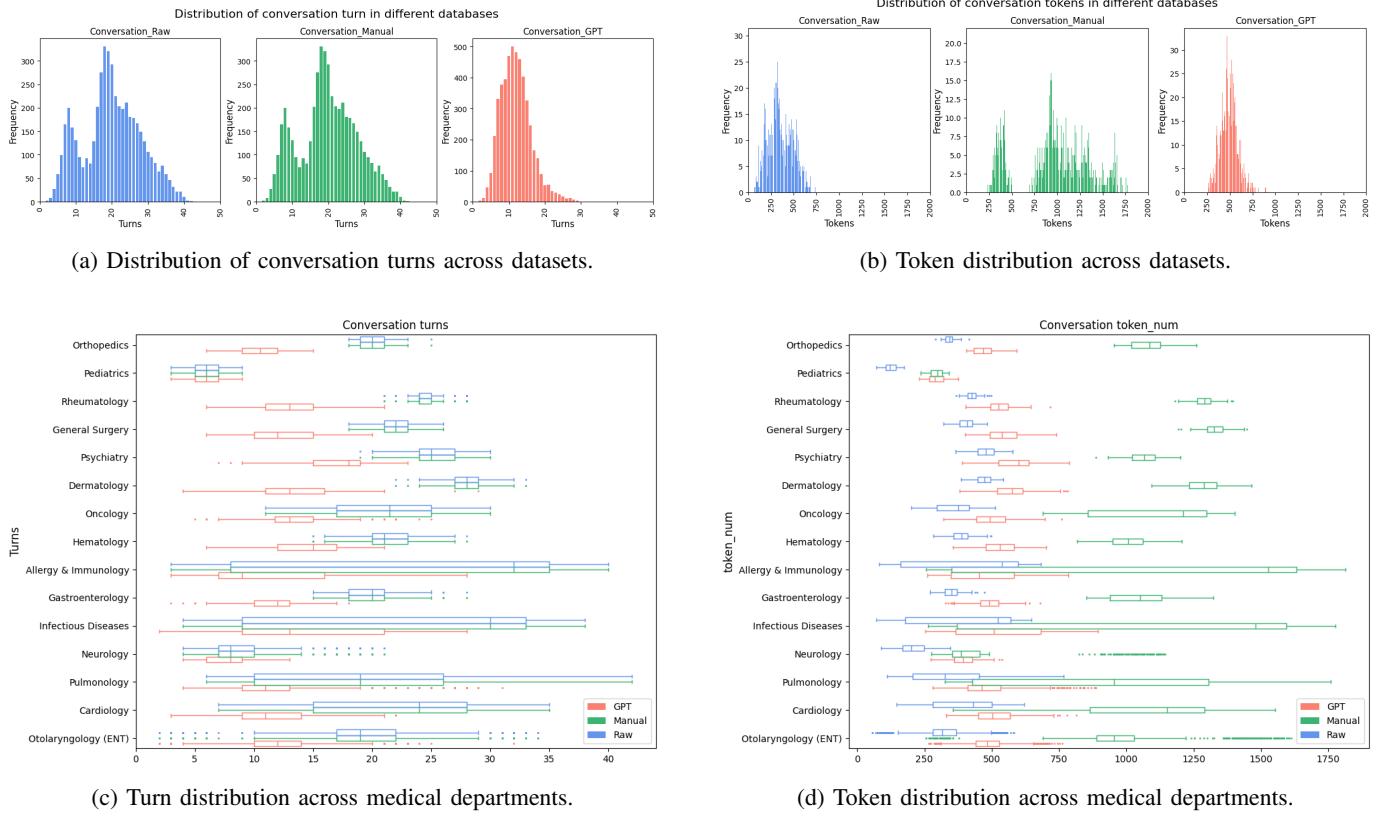


Fig. 7: Distributions of conversation turns and tokens across three datasets and departments

B. GPTScore Evaluation

We further evaluated the quality of multi-turn conversations using GPTScore, a semantic, instruction-based metric that considers multiple conversational dimensions [13]. GPTScore provides a richer assessment than surface-level lexical metrics, making it particularly suitable for evaluating patient-facing medical dialogue.

Aspect	data_5k_ddxplus	data_5k_artificial	data_5k_GPT
SPE	70.02	45.47	19.92
FLE	81.59	58.05	62.98
UND	76.26	83.70	94.57
INF	99.90	100.00	100.00
PAT	0.10	0.10	0.00
ACC	99.60	100.00	100.00

TABLE IV: GPTScore evaluation results across different datasets.

Table IV summarizes the results in six dimensions: Specificity (SPE), Flexibility (FLE), Understandability (UND), Informativeness (INF), Patience (PAT), and Accuracy (ACC). Our GPT-rewritten dataset, data_5k_GPT, achieves the highest scores in understandability and informativeness, reflecting the benefit of converting complex clinical questions into clear, accessible exchanges. The accuracy and informativeness remain consistently high across all datasets, indicating that core medical content is preserved regardless of format.

Model	data_5k_ddxplus		data_5k_artificial		data_5k_GPT	
	F1	Accuracy	F1	Accuracy	F1	Accuracy
BERT	0.773	0.930	0.998	0.998	0.996	0.996
ClinicalBERT	0.159	0.408	0.796	0.964	0.229	0.570

TABLE V: F1-Score and Accuracy of BERT and Clinical-BERT in Department Recommendation across GPT-rewritten and manually constructed dialogue datasets

Interestingly, data_5k_GPT achieves lower specificity than the raw dataset, suggesting that simplifying language may lead to more generalized expressions. However, this trade-off appears acceptable given the substantial gains in patient comprehension and coherence. The flexibility score improves over the artificial dataset, though it remains slightly behind the raw version, highlighting a potential area for prompt refinement in future work.

To further benchmark model performance in terms of classification quality, we report the F1-score and accuracy of two transformer-based models: BERT and ClinicalBERT on the department recommendation task across all three datasets. As shown in Table V, BERT achieves superior performance on data_5k_artificial and data_5k_GPT, reflecting the benefits of clearer structure and language simplification in these versions. In contrast, ClinicalBERT, despite being domain-specific, shows inconsistent results across datasets,

likely due to its training objective and limited capacity to generalize over conversational inputs. These results highlight the robustness of our fine-tuned dataset design and suggest that task-specific fine-tuning can often outperform domain-specific pretraining when applied to conversational healthcare data.

C. Discussions on Conversation Quality

Patient Navigator: Are you comfortable?

Human: I've noticed that lying down seems to aggravate my symptoms, while sitting up relieves them.

Patient Navigator: Have there been any recent instances of viral infections?

Human: I think so.

Patient Navigator: Can you recall ever being treated for pericarditis?

Human: I think yes.

Patient Navigator: Are you experiencing any pain related to your current health concern?

Human: I think yes.

Patient Navigator: Could you specify the type of your pain as following, no pain; heartbreak; haunting; tedious; sensitive; a knife stroke; tugging; burning; a cramp; heavy; a pulse; violent; sharp; sickening; scary; exhausting?

Human: Sharp.

Patient Navigator: Are there areas on your body where you feel pain? If yes, please specify the location or body part.

Human: I think it is on torso.

Patient Navigator: Is there any particular location pain you're feeling? If yes, please specify the location or body part.

Human: Torso.

Patient Navigator: Is there any particular location pain you're feeling? If yes, please specify the location or body part.

Human: It is located right torso.

Patient Navigator: Are there areas on your body where you feel pain? If yes, please specify the location or body part.

Human: I think it is on left torso.

Patient Navigator: How intense is the pain? You can describe severity or score as following rules,

No(0);

Mild(1-3), noticeable but not significantly interfere with daily activities;

Moderate(4-6), enough to cause discomfort and may require some attention or medication;

Severe(7-9), challenging to concentrate that significantly impact daily activities;

Worst(10), unbearable and causing disability to daily activities

Human: 9.

Patient Navigator: Does the pain radiate to another location? If yes, please specify the location or body part.

Human: My right back side, upper part.

Patient Navigator: How precisely is the pain located? You can describe severity or score as following rules,

Unknown(0);

Hardly described(1-3), identify a general area (like an entire limb or half of the body);

Moderately precise(4-6), identify a more specific region of the body (like the lower leg or upper arm);

Highly precise(7-9), identify a very specific area (like the inside of the right knee or the outside of the left wrist);

Extremely precise(10), identify the exact spot

Human: Moderately precise.

Patient Navigator: How fast did the pain appear? You can choose severity or score as following rules,

No(0);

Slow(1-3), more than 1 day;

Moderate(4-6), less than 1 day;

Fast(7-9), within several minutes;

Immediately(10)

Human: Fast.

Patient Navigator: Have you been experiencing palpitations or any changes in your heart's rhythm?

Human: I think so.

Patient Navigator: Have you traveled out of the country in the last 4 weeks? If yes, please specify which continent it is as follows, North Africa; West Africa; South Africa; Central America; North America; South America; Asia; South East Asia; Caraibes; Europe; Oceania.

Human: Nowhere.

Patient Navigator: Does your pain intensify upon deep inhalation?

Human: Undoubtedly.

Patient Navigator: Based on your description, I would recommend you to go to a Cardiologist for further evaluation and management of your condition. Cardiologists are medical specialists who focus on diagnosing and treating diseases and conditions related to the heart and blood vessels. If you are experiencing symptoms such as chest pain, shortness of breath, irregular heartbeats, or have risk factors for heart disease, a Cardiologist can help assess your cardiovascular health. They have expertise in conducting various heart-related tests such as electrocardiograms (ECG/EKG), echocardiograms, stress tests, and cardiac catheterization, among others, to evaluate your heart's function and identify any underlying issues. By visiting a Cardiologist, you will receive specialized care and appropriate management to maintain a healthy heart and reduce the risk of cardiovascular complications.

Fig. 8: Example conversation between a patient and the C-PATH model.

An example conversation between C-PATH and a simulated patient is shown in Figure 8. The model engages naturally, avoiding jargon while collecting essential clinical information. This interaction style improves patient accessibility and trust, particularly for users without medical training.

The model gathers relevant details such as current symptoms, medical history, and travel patterns, allowing accurate departmental recommendations. Beyond triage, C-PATH generates structured summaries suitable for Electronic Health Record (EHR) systems, facilitating efficient physician review and reducing administrative burden.

This capability also supports healthcare systems by automating patient intake processes, reducing manual triage time, and improving throughput. C-PATH can be deployed in multiple clinical contexts, including self-service triage kiosks in hospital waiting areas, mobile health applications for at-home pre-consultation, and EHR-integrated assistant systems that streamline documentation and referral decisions for physicians. In primary care settings, the system could serve as a virtual assistant to guide patients in symptom articulation before in-person visits. In telemedicine platforms, it could support pre-screening to route patients to the appropriate specialist. These applications highlight the versatility of LLM-based medical dialogue systems and indicate their potential to improve access, efficiency, and equity in healthcare delivery.

As such, the results confirm the effectiveness of our dataset construction pipeline and fine-tuning approach. GPT-based rewriting significantly improves conversation structure and readability while preserving diagnostic value, making it an ideal foundation for instruction-tuned medical dialogue systems.

V. CONCLUSIONS

In this study, we introduced C-PATH, a Conversational Patient Navigator powered by a fine-tuned Large Language Model (LLM) that assists users in symptom recognition and triage to appropriate medical departments. Using a multi-stage training pipeline that comprises acquisition of medical knowledge, conversational alignment, and summarization, we demonstrated that LLMs can be adapted to produce safe, understandable, and context-aware medical dialogues. C-PATH enables laypeople to communicate symptoms naturally, ask follow-up questions, and receive accurate referral suggestions, enhancing both patient experience and healthcare workflow efficiency.

A central innovation of this work is our GPT-based rewriting pipeline that transforms structured clinical data from DDXPlus into accessible, multi-turn conversations. This method bridges the gap between domain-specific medical knowledge and patient-friendly communication, significantly improving the clarity of dialogue and the usability of the model. Additionally, we introduced a lightweight dialogue history management strategy combining sliding window pruning and optional summarization, allowing for scalable multi-turn interactions within LLM context limits.

Our evaluations showed that C-PATH outperforms baselines in both conversation quality and triage accuracy, and the GPTScore metrics confirm improvements in understandability, informativeness, and conversational flow. These capabilities make C-PATH well suited for deployment in digital health settings such as patient self-assessment tools or intake assistants,

with the potential to reduce provider burden and accelerate access to care.

However, several limitations remain. Logical reasoning challenges inherent to LLMs, such as the “Reversal Curse” [4], restrict their ability to generalize in nuanced medical scenarios. Furthermore, hallucination remains a concern in patient-facing contexts, especially when models confidently present incorrect information due to training artifacts [17]. In addition, the training datasets derived from DDXPlus exhibit imbalances in department representation, with certain specialties such as respiratory and general medicine being overrepresented. This skew may reduce the model’s ability to generalize recommendations across underrepresented or rare medical departments. We also note the importance of reinforcement learning from human feedback (RLHF) to align model behavior with clinical standards, which is not yet integrated in this iteration of C-PATH.

Future work will explore the incorporation of RLHF with clinical expert supervision, expanding the training data to include more diverse medical domains, and conducting real-world user studies with patients and clinicians. We also aim to integrate our model into mobile health platforms, enabling greater accessibility and continuous improvement through conversational data feedback. C-PATH represents a step toward scalable, trustworthy AI assistance in healthcare and lays the foundation for future patient-centered dialogue systems.

ACKNOWLEDGEMENTS

This work was funded by Fundação para a Ciência e a Tecnologia (UIDB/00124/2020, UIDP/00124/2020 and Social Sciences DataLab - PINFRA/22209/2016), POR Lisboa and POR Norte (Social Sciences DataLab, PINFRA/22209/2016) and NOVA LINCS (grant UIDB/04516/2020) and Carnegie Mellon Portugal Program (CMU/TIC/0016/2021). This work was supported in part by Oracle Cloud credits and related resources provided by Oracle for Research. We also acknowledge the use of ChatGPT, an AI language model by OpenAI, for assistance with grammar, LaTeX formatting, and general English writing improvements throughout the manuscript.

REFERENCES

- [1] Laure Abensur Vuillaume, Julien Turpinier, Lauriane Cipolat, Thomas Dumontier, Nicolas Peschanski, Yann Kieffer, Boris Barbat, Thomas Riquier, Vincent Dinot, and Joris Galland. Exploratory study: Evaluation of a symptom checker effectiveness for providing a diagnosis and evaluating the situation emergency compared to emergency physicians using simulated and standardized patients. *Plos one*, 18(2):e0277568, 2023.
- [2] Ayo Adedeji, Sarita Joshi, and Brendan Doohan. The sound of healthcare: Improving medical transcription asr accuracy with large language models. *arXiv preprint arXiv:2402.07658*, 2024.
- [3] Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. An empirical study of clinical note generation from doctor-patient encounters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2291–2302, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.
- [4] Lukas Berglund, Meg Tong, Maximilian Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. The reversal curse: LLMs trained on “a is b” fail to learn “b is a”. In *The Twelfth International Conference on Learning Representations*, 2024.
- [5] Mikaël Chelli, Jules Descamps, Vincent Lavoué, Christophe Trojani, Michel Azar, Marcel Deckert, Jean-Luc Raynier, Gilles Clowez, Pascal Boileau, and Caroline Ruetsch-Chelli. Hallucination rates and reference accuracy of chatgpt and bard for systematic reviews: comparative analysis. *Journal of medical Internet research*, 26:e53164, 2024.
- [6] Jan Clusmann, Fiona R Kolbinger, Hannah Sophie Muti, Zunamys I Carrero, Jan-Niklas Eckardt, Narmin Ghaffari Laleh, Chiara Maria Lavinia Löffler, Sophie-Caroline Schwarzkopf, Michaela Unger, Gregory P Veldhuizen, et al. The future landscape of large language models in medicine. *Communications medicine*, 3(1):141, 2023.
- [7] Shizhe Diao, Rui Pan, Hanze Dong, Ka Shun Shum, Jipeng Zhang, Wei Xiong, and Tong Zhang. Lmflow: An extensible toolkit for finetuning and inference of large foundation models. *arXiv preprint arXiv:2302.12420*, 2023.
- [8] Shizhe Diao, Rui Pan, Hanze Dong, KaShun Shum, Jipeng Zhang, Wei Xiong, and Tong Zhang. LMFlow: An extensible toolkit for finetuning and inference of large foundation models. In Kai-Wei Chang, Annie Lee, and Nazneen Rajani, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: System Demonstrations)*, pages 116–127, Mexico City, Mexico, June 2024. Association for Computational Linguistics.
- [9] Niharika Dixit, Hope Rugo, and Nancy J Burke. Navigating a path to equity in cancer care: the role of patient navigation. In *American Society of Clinical Oncology Educational book. American Society of Clinical Oncology. Annual Meeting*, volume 41, pages 1–8, 2021.
- [10] Regina de Brito Duarte, Qiwei Han, and Claudia Soares. Dissecting medical referral mechanisms in health services: Role of physician professional networks. *arXiv preprint arXiv:2312.02387*, 2023.
- [11] Arsene Fansi Tchango, Rishab Goel, Zhi Wen, Julien Martel, and Joumana Ghosn. Ddxplus: A new dataset for automatic medical diagnosis. *Advances in Neural Information Processing Systems*, 35:31306–31318, 2022.
- [12] Hamish SF Fraser, Susan Clamp, and Carol J Wilson. Limitations of study on symptom checkers. *JAMA internal medicine*, 177(5):740–741, 2017.
- [13] Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. GPTScore: Evaluate as you desire. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576, Mexico City, Mexico, June 2024. Association for Computational Linguistics.
- [14] Katherine E Goodman, H Yi Paul, and Daniel J Morgan. Ai-generated clinical summaries require more than accuracy. *JAMA*, 331(8):637–638, 2024.
- [15] Mohammad Hammoud, Shahd Douglas, Mohamad Darmach, Sara Alawneh, Swapnendu Sanyal, Youssef Kanbour, et al. Evaluating the diagnostic performance of symptom checkers: Clinical vignette study. *JMIR AI*, 3(1):e46875, 2024.
- [16] Qiwei Han, Mengxin Ji, Inigo Martinez De Rituerto De Troya, Manas Gaur, and Leid Zejnilovic. A hybrid recommender system for patient-doctor matchmaking in primary care. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 481–490. IEEE, 2018.
- [17] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.
- [18] Di Jin, Eileen Pan, Nassim Oufattolle, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.
- [19] Qiao Jin, Bhuvan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, 2019.
- [20] Caroline Jones, James Thornton, and Jeremy C Wyatt. Artificial intelligence and clinical decision support: clinicians’ perspectives on trust, trustworthiness, and liability. *Medical law review*, 31(4):501–520, 2023.

- [21] Yubin Kim, Hyewon Jeong, Shen Chen, Shuyue Stella Li, Mingyu Lu, Kurnail Alhamoud, Jimin Mun, Cristina Grau, Minseok Jung, Rodrigo R Gameiro, et al. Medical hallucination in foundation models and their impact on healthcare. *medRxiv*, pages 2025–02, 2025.
- [22] Kristina M Kokorelias, Jessica E Shiers-Hanley, Jorge Rios, Amanda Knoepfli, and Sander L Hitzig. Factors influencing the implementation of patient navigation programs for adults with complex needs: a scoping review of the literature. *Health Services Insights*, 14:11786329211033267, 2021.
- [23] Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, et al. Performance of chatgpt on usmle: potential for ai-assisted medical education using large language models. *PLoS digital health*, 2(2):e0000198, 2023.
- [24] Siryeol Lee, Juncheol Lee, Juntae Park, Jiwoo Park, Dohoon Kim, Joohyun Lee, and Jaehoon Oh. Deep learning-based natural language processing for detecting medical symptoms and histories in emergency patient triage. *The American Journal of Emergency Medicine*, 77:29–38, 2024.
- [25] Qiyuan Li, Haijiang Liu, Caicai Guo, Deyu Chen, Meng Wang, Feng Gao, and Jinguang Gu. Merging clinical knowledge into large language models for medical research and applications: A survey. *arXiv preprint arXiv:2502.20988*, 2025.
- [26] Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6), 2023.
- [27] Andrew W Liu, Anobel Y Odisho, William Brown III, Ralph Gonzales, Aaron B Neinstein, and Timothy J Judson. Patient experience and feedback after using an electronic health record-integrated covid-19 symptom checker: Survey study. *JMIR Human Factors*, 9(3):e40064, 2022.
- [28] Carles Rubio Maturana, Allisson Dantas De Oliveira, Sergi Nadal, Besim Bilalli, Francesc Zarzuela Serrat, Mateu Espasa Soley, Elena Sulleiro Igual, Mercedes Bosch, Anna Veiga Lluch, Alberto Abelló, et al. Advances and challenges in automated malaria diagnosis using digital microscopy imaging with artificial intelligence tools: A review. *Frontiers in microbiology*, 13:1006659, 2022.
- [29] Kerry A McBrien, Noah Ivers, Lianne Barnieh, Jacob J Bailey, Diane L Lorenzetti, David Nicholas, Marcello Tonelli, Brenda Hemmelgarn, Richard Lewanczuk, Alun Edwards, et al. Patient navigators for people with chronic disease: a systematic review. *PloS one*, 13(2):e0191980, 2018.
- [30] Nicolas Munsch, Alistair Martin, Stefanie Gruarin, Jama Nataqi, Issel-mou Abdarahmane, Rafael Weingartner-Ortner, and Bernhard Knapp. Diagnostic accuracy of web-based covid-19 symptom checkers: comparison study. *Journal of medical Internet research*, 22(10):e21299, 2020.
- [31] Jesutofunmi A Omiye, Haiwen Gui, Shawheen J Rezaei, James Zou, and Roxana Daneshjou. Large language models in medicine: the potentials and pitfalls: a narrative review. *Annals of internal medicine*, 177(2):210–220, 2024.
- [32] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In Gerardo Flores, George H Chen, Tom Pollard, Joyce C Ho, and Tristan Naumann, editors, *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR, 07–08 Apr 2022.
- [33] Annette Peart, Virginia Lewis, Ted Brown, and Grant Russell. Patient navigators facilitating access to primary care: a scoping review. *BMJ Open*, 8(3):e019252, 2018.
- [34] Cheng Peng, Xi Yang, Aokun Chen, Kaleb E Smith, Nima PourNejatian, Anthony B Costa, Cheryl Martin, Mona G Flores, Ying Zhang, Tanja Magoc, et al. A study of generative large language model for medical research and healthcare. *NJP digital medicine*, 6(1):210, 2023.
- [35] Kimberly Peven, Aidan P Wickham, Octavia Wilks, Yusuf C Kaplan, Andrei Marhol, Saddif Ahmed, Ryan Bamford, Adam C Cunningham, Carley Prentice, András Meczner, et al. Assessment of a digital symptom checker tool's accuracy in suggesting reproductive health conditions: clinical vignettes study. *JMIR mHealth and uHealth*, 11(1):e46718, 2023.
- [36] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scates, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.
- [37] Xin Sun, Yunjie Liu, Jan De Wit, Jos A Bosch, and Zhuying Li. Trust by interface: How different user interfaces shape human trust in health information from large language models. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–7, 2024.
- [38] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elango-van, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature medicine*, 29(8):1930–1940, 2023.
- [39] Shubo Tian, Qiao Jin, Lana Yeganova, Po-Ting Lai, Qingqing Zhu, Xiuying Chen, Yifan Yang, Qingyu Chen, Won Kim, Donald C Comeau, et al. Opportunities and challenges for chatgpt and large language models in biomedicine and health. *Briefings in Bioinformatics*, 25(1):bbad493, 2024.
- [40] Amy E Ustjanauskas, Marissa Bredice, Sumayah Nuhaily, Lisa Kath, and Kristen J Wells. Training in patient navigation: a review of the research literature. *Health promotion practice*, 17(3):373–381, 2016.
- [41] Filipa Valdeira, Stevo Racković, Valeria Danalachi, Qiwei Han, and Cláudia Soares. Extreme multilabel classification for specialist doctor recommendation with implicit feedback and limited patient metadata. *arXiv preprint arXiv:2308.11022*, 2023.
- [42] Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Małgorzata Polacín, Eduardo Pontes Reis, Anna Seehofnerová, et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nature medicine*, 30(4):1134–1142, 2024.
- [43] William Wallace, Calvin Chan, Swathikan Chidambaram, Lydia Hanna, Fahad Mujtaba Iqbal, Amish Acharya, Pasha Normahani, Hutan Ashrafiyan, Sheraz R Markar, Viknesh Sounderajah, et al. The diagnostic and triage accuracy of digital and online symptom checker tools: a systematic review. *NPJ digital medicine*, 5(1):118, 2022.
- [44] Haochun Wang, Sendong Zhao, Zewen Qiang, Zijian Li, Chi Liu, Nuwa Xi, Yanrui Du, Bing Qin, and Ting Liu. Knowledge-tuning large language models with structured medical knowledge bases for trustworthy response generation in chinese. *ACM Transactions on Knowledge Discovery from Data*, 19(2):1–17, 2025.
- [45] Kristen J Wells, Tracy A Battaglia, Donald J Dudley, Roland Garcia, Amanda Greene, Elizabeth Calhoun, Jeanne S Mandelblatt, Electra D Paskett, Peter C Raich, and Patient Navigation Research Program. Patient navigation: state of the art or is it science? *Cancer*, 113(8):1999–2010, 2008.
- [46] Bo Wen, Raquel Norel, Julia Liu, Thaddeus Stappenbeck, Farhana Zulkernine, and Huamin Chen. Leveraging Large Language Models for Patient Engagement: The Power of Conversational AI in Digital Health. In *2024 IEEE International Conference on Digital Health (ICDH)*, pages 104–113. IEEE, 2024.
- [47] Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisen. Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Scientific data*, 10(1):586, 2023.
- [48] Yue You, Renkai Ma, and Xinning Gui. User experience of symptom checkers: a systematic review. In *AMIA Annual Symposium Proceedings*, volume 2022, page 1198, 2023.
- [49] Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, Hongchao Fang, Penghui Zhu, Shu Chen, and Pengtao Xie. MedDialog: Large-scale medical dialogue datasets. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250, Online, November 2020. Association for Computational Linguistics.
- [50] Marc Zobel, Bernhard Knapp, Jama Nataqi, and Alistair Martin. Correlating global trends in covid-19 cases with online symptom checker self-assessments. *PLoS One*, 18(2):e0281709, 2023.