Retrieval-Augmented Large Language Models for Adolescent Idiopathic Scoliosis Patients in Shared Decision-Making

Wenqi Shi
wshi83@gatech.edu
Georgia Tech
Atlanta, United States

Yuchen Zhuang
yzhuang43@gatech.edu
Georgia Tech
Atlanta, United States

Yuanda Zhu
yzhu94@gatech.edu
Georgia Tech
Atlanta, United States

Henry J. Iwinski
hiwinski@shrinenet.org
Shriners Children's
Lexington, United States

J. Michael Wattenbarger
mwattenbarger@shrinenet.org
Shriners Children's
Greenville, United States

May D. Wang
maywang@gatech.edu
Georgia Tech
Atlanta, United States

ABSTRACT

As health-related decision-making evolves, patients increasingly seek help from additional online resources such as "Dr. Google" and ChatGPT. Despite their potential, these tools encounter limitations, including the risk of potentially inaccurate information, a lack of specialized medical knowledge, the risk of generating unrealistic outputs (hallucinations), and significant computational demands. In this study, we develop and validate an innovative shared decision-making (SDM) tool, Chat-Orthopedist, for adolescent idiopathic scoliosis (AIS) patients and families to prepare a meaningful discussion with clinicians based on retrieval-augmented large language models. Firstly, we establish an external knowledge base with information on AIS disease and treatment options. Secondly, we develop a retrieval-augmented ChatGPT to feed LLMs with AIS domain knowledge, providing accurate and comprehensible responses to patient inquiries. In addition, we perform a cyclical process of human-in-the-loop evaluations for system validation and improvement. Chat-Orthopedist may optimize SDM workflow by enabling better interactive learning experiences, more effective clinical visits, and better-informed treatment decision-making.

CCS CONCEPTS

• **Information systems** → **Document representation**; • **Computing methodologies** → **Information extraction**.

KEYWORDS

large language models, information retrieval, pediatric healthcare, shared decision-making, adolescent idiopathic scoliosis

ACM Reference Format:
Wenqi Shi, Yuchen Zhuang, Yuanda Zhu, Henry J. Iwinski, J. Michael Wattenbarger, and May D. Wang. 2023. Retrieval-Augmented Large Language Models for Adolescent Idiopathic Scoliosis Patients in Shared Decision-Making.

ACM ISBN 979-8-4007-0126-9/23/09.
https://doi.org/10.1145/3584371.3612956

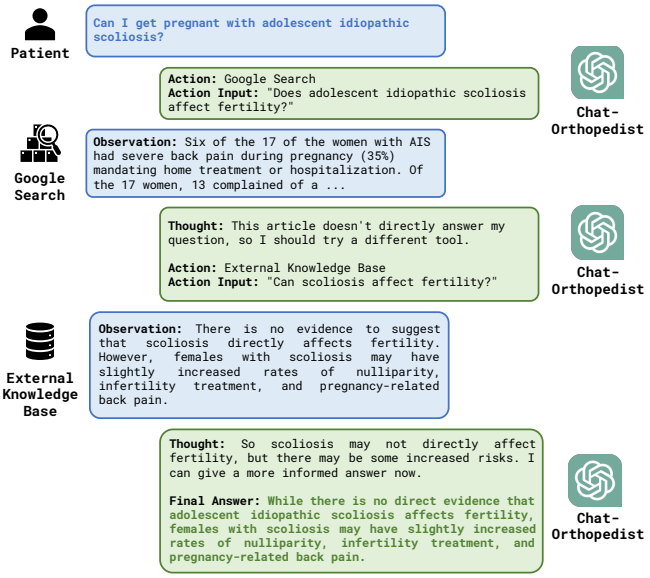


Figure 1: Example of Chat-Orthopedist, the proposed retrieval-augmented LLMs, for answering AIS-related patient questions during SDM. Specifically, we improve model transparency with reasoning (e.g., chain-of-thoughts prompting) and acting (e.g., action plan generation).

In *14th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '23), September 3–6, 2023, Houston, TX, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3584371.3612956

1 INTRODUCTION

Adolescent idiopathic scoliosis (AIS) is a structural, lateral, rotated curvature of the spine that impacts 1–4% of children within the at-risk age group of 10–16 years [34]. If left untreated, scoliosis may lead to altered spinal mechanics and degenerative changes, resulting in pain, loss of spinal mobility, possible function loss or disability, and decreased quality-of-life [33]. Common treatment options for scoliosis include observation, bracing, and spinal fusion surgery [17]. The decision to perform interventions on AIS patients depends on multiple factors, including patient maturity, curve characteristics, curve magnitude, location of the curve, and the possibility of progression [22]. For pediatric patients during

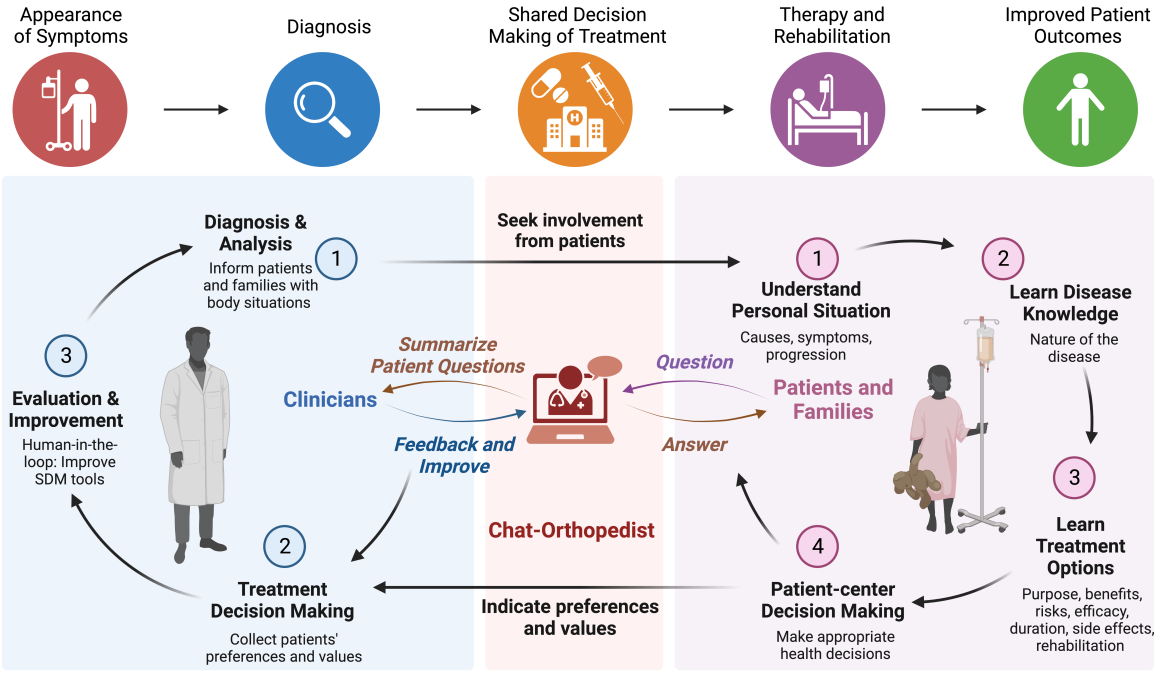


Figure 2: Clinical workflow and usability of the proposed SDM tool with LLM-enabled dialogue system. The proposed SDM tool serves as a chatbot to answer questions from AIS patients and help them to prepare knowledge about diseases and treatment for the coming clinical visit. Besides, clinicians could review patient questions and responses to understand questions from patients, prepare for clinical visits, and improve the quality of responses. Created with BioRender.com.

growth periods, the primary goals of interventions are to limit or halt the progression of the deformity, restore trunk balance, and prevent long-term consequences [15]. Non-surgical treatments, such as bracing and physiotherapy, aim to reduce the need for operations by preventing curve progression, but their effectiveness has not been rigorously assessed despite their widespread use [14]. On the other hand, all types of spinal surgery carry significant risks, both short-term (approximately a mean infection rate of 3.6%) and long-term (re-operation risk of 8.3% with a mean follow-up of 14.9 years) [19]. Consequently, there is an important clinical need for adolescent patients diagnosed with moderate to severe idiopathic scoliosis, who often face a critical decision between observation and intervention, to have access to validated tools that can assist in their treatment decision-making.

For AIS patients and families, the decision to have surgery or not can be overwhelming [26, 35]. Shared decision-making (SDM) is a collaborative approach, which enables patients, families, and physicians jointly participate in the medical decision-making process, reaching a consensus on treatment plans [1]. Specifically, clinicians possess knowledge about diseases, tests, and treatments, while patients and families are familiar with their bodies, daily life circumstances, and healthcare expectations. Healthcare providers explain treatments and alternatives to patients, assisting them in choosing the option that best aligns with their values and preferences (e.g., treatment benefits, surgical complications, pain, rehabilitation, and cost) and achieving the ideal of evidence-based and patient-centered medicine [5]. Thus, SDM mechanisms play a crucial role in empowering patients, families, and clinicians to collaboratively identify treatment solutions that are intellectually, practically, and emotionally appropriate [11]. However, current SDM tools (e.g. pre-recorded video playing, reading materials) often lack effective knowledge sharing with patients, resulting in less efficient clinical visits and conversations. Existing SDM approaches [3, 11, 16] usually rely on educational materials and passive knowledge sharing from clinicians to patients during clinical visits, which may limit patients' motivation to actively seek information and prepare for effective consultations.

The rapid advancement of conversational and chat-based language models has led to remarkable progress in artificial general intelligence. Large language models (LLMs) have demonstrated remarkable capabilities because of pre-training on a vast corpus with reinforcement learning from human feedback [4, 24]. By utilizing LLMs, we can convert a conventional passive knowledge sharing in SDM (e.g., pre-recorded videos or reading materials) to an active knowledge inquiry, such as patient-clinician question-and-answer (Q&A) (Figure 1). However, adapting LLMs to biomedicine has been rarely explored due to the lack of medical domain knowledge [6, 30, 42]. Moreover, generative models are susceptible to producing hallucinated information and often struggle with logical reasoning in the context of complex inferences. In addition, other concerns, such as computational costs and model transparency, further impede the adoption of LLMs in real-world clinical settings [7, 12, 28, 36, 39].

To address these challenges, we propose an innovative SDM tool for AIS patients and families, leveraging a retrieval-augmented ChatGPT to equip LLMs with AIS disease and treatment knowledge

(Figure 2). With a retriever and an external knowledge base, the proposed SDM tool could augment ChatGPT and any LLMs by leveraging external resources (e.g., search engines, medical papers, treatment guidelines, etc.) to answer queries related to clinical concepts and treatment recommendations. This framework mitigates the need for time-intensive and costly fine-tuning and facilitates timely updates without the necessity of re-training the entire model. Notably, we perform a human-in-the-loop assessment for validation and improvement with a diverse group of targeted users and domain experts. Furthermore, the SDM tool provides a positive clinical impact by minimizing human biases in treatment recommendations enabled by LLMs based on large objective domain knowledge.

The main contribution of our work is four-fold:

- We develop an innovative SDM tool for AIS patients and families to facilitate comprehensive pre-operative consultations, thereby improving patient treatment outcomes and the efficiency of clinical visits.
- We validate the proposed tool with targeted user groups and domain experts to quantitatively and qualitatively demonstrate the feasibility of adopting LLMs in clinical settings.
- We employ a retrieval-based framework to augment LLMs with the most recent domain-specific knowledge, thereby significantly improving the computation efficiency.
- We effectively mitigate the risk of hallucination with an external knowledge base. In addition, it enhances model transparency by identifying source information and enabling human-in-the-loop validation.

2 RELATED WORKS

Recent advancements in general-domain LLMs [4, 24] have demonstrated exceptional capabilities in following instructions and generating responses that closely mimic human conversation. However, few LLMs have been adapted or fine-tuned for the biomedical domain [6, 23]. As a result, when generating responses related to domain-specific topics, standard LLMs often suffer from a deficiency in providing sound medical advice. Due to challenges such as insufficient domain knowledge and computational costs, only a few LLMs [10, 40] have been adapted for biomedicine by fine-tuning open-source LLMs (typically LLMs with 6.5B-13B parameters) for medical consultation. For example, ChatDoctor [40] has fine-tuned LLaMA [31] (with 7B parameters) to answer clinical questions based on 100k real-world patient-physician conversations from an online medical consultation site. Similarly, MedAlpaca [10] has also fine-tuned LLaMA [31] with publicly available medical datasets, such as Anki Medical Curriculum flashcards, for biomedical Q&A tasks.

However, several significant challenges exist when attempting to implement LLMs in practical clinical settings (Table 1). First, models specific to the medical domain often utilize comparatively smaller-scale LLMs (e.g., LLaMA [31] compared to ChatGPT with 175B parameters), which may resent less accurate and robust representations [23]. Second, the fine-tuning of even these smaller LLMs, typically comprised of 7 to 13 billion parameters, is both computationally demanding and cost-intensive [40]. Furthermore, the introduction of new knowledge necessitates the complete retraining of the model, imposing additional burdens on developers. Third, in general, LLMs are susceptible to hallucination and struggle to represent the comprehensive long tail of knowledge from the training corpus [2, 8, 18, 20].

To solve these challenges, we propose to leverage retrieval-augmented language models [13, 27, 38] to access medical knowledge from an external database for enabling domain expertise, reducing computational costs, minimizing hallucination, and enhancing coverage. Our goal is to improve and accelerate LLMs for clinical use cases by incorporating patient interaction and clinical prompting into dialogue systems. Comparing existing applications of LLMs in healthcare, the proposed Chat-Orthopedist aims to answer patient questions during SDM, using an external training corpus in conjunction with an off-the-shelf retrieval model. This approach allows for asynchronous updates with new knowledge, eliminating the need to retrain the entire model. To the best of our knowledge, our proposed work represents one of the first innovative attempts to leverage the advantages of retrieval-augmented LLMs (>175B model parameters) for SDM in clinical research and practice.

3 METHODOLOGY

Given the rapid advancement in AI, it is feasible to facilitate online clinical conversations to provide necessary knowledge support for patients in SDM. To equip LLMs with domain-specific knowledge related to scoliosis, we introduce Chat-Orthopedist, a retrieval-augmented ChatGPT, for AIS patient Q&A during pre-operative SDM. The proposed Chat-Orthopedist is comprised of three key components: an external AIS knowledge base, a retriever, and an LLM. With user query as input, the retriever seeks out the most relevant content from an external knowledge base, which contains additional information that is not typically stored within the LLM's parameters. Once a subset of the most relevant content has been identified, this information is seamlessly reintegrated into the prompts, thereby augmenting the inherent capabilities of the original LLM. Paired with the user's query, this enriched context is then conveyed to the LLM for an improved response with optimized domain knowledge.

3.1 Knowledge Base Establishment

In Chat-Orthopedist, we collect external knowledge based on multiple evidence-based and physician-authored clinical knowledge data resources (Figure 3). As the sizes of paragraphs are various from different sources (or even from the same source), we apply chunking to segment the supportive materials into manageable units. Each corpus consists of 2000 tokens, thereby standardizing the information input irrespective of its original source. The utilization of an external knowledge base in our approach offers the flexibility to readily update existing materials in alignment with the latest advancements and clinical recommendations. Specifically, it is achieved without necessitating the retraining of the entire model, thereby optimizing the computation efficiency while maintaining its relevance in an ever-evolving field.

3.2 Retrieval Process

To fully capture the semantic information from the external knowledge base, the proposed retriever in Chat-Orthopedist follows a dense retrieval manner. In the retriever, we use a dense encoder $E_P(\cdot)$ to map all the text passages into d-dimensional real-valued

Table 1: Comparison of different online dialogue systems or search engines for healthcare SDM, including conventional searching engine (e.g., Google search), fine-tuning LLMs (e.g., LLaMA with 6.5B parameters), instruction tuning LLMs (ChatGPT with 175B parameters), and our proposed retrieval-augmented LLMs, Chat-Orthopedist.

Methods	# Parameters	Instruction Type	Human-Like Dialogue	AIS Knowledge	Knowledge Update	No Hallucination	Source Transparency	Multi-Source Reasoning
Google Search	-	-	✗	✔	✔	-	✔	✔
LLaMA [31]	7-65B	-	✔	✗	✗	✗	✗	✗
ChatGPT [25]	>175B	-	✔	✗	✗	✗	✗	✗
ChatDoctor [40]	13B	Tuning	✔	✔	✗	✔	✗	✗
MedAlpaca [10]	7-13B	Tuning	✔	✔	✗	✔	✗	✗
Chat-Orthopedist	>175B	Prompting	✔	✔	✔	✔	✔	✔

vectors. We then build an index for all the M passages that we will use for retrieval. During the retrieval process, to ensure congruity in encoding between the user's query and the corresponding external knowledge, we adopt a uniform dense encoder $E_Q(\cdot) = E_P(\cdot) = E(\cdot)$ to map the user queries into d-dimensional vectors, where $E_Q(\cdot)$ is the user query encoder. Furthermore, it facilitates the retrieval of the k passages whose vector representations are the closest in proximity to the vector corresponding to the question, thereby aligning the user's inquiry with the most relevant knowledge extracts. We use dot product between the high dimensional vectors to define the similarities between passages and user queries:

$$sim(q, p) = E_Q(q)^T E_P(p) = E(q)^T E(p).$$

Regarding the encoders, we adopt an off-the-shelf sentence transformer model, MP-Net [29], for both query and passage encoding purposes. Specifically, we take the representation at `[CLS]` token as the output, with the high dimensionality d of embeddings set as 768. This retrieval strategy effectively enables the optimal coherence between queries and relevant passages.

3.3 Augmented LLM

To augment the LLMs with domain-specific knowledge, it is important to include the retrieved materials in the context for model reference, as shown in Figure 4. However, the data originating from various sources in the knowledge base play different roles in SDM. The appropriate source of information can vary depending on the specific situation and the nature of the knowledge sought. For example, if a patient's question primarily pertains to the diagnosis, priority is given to retrieving information from guidelines related to physician diagnosis. If the patient is asking about complications related to treatment, we recommend seeking information from clinical trial-related papers. If the model cannot answer the question with information from both sources, it then seeks Google Search for additional assistance. Consequently, to model the logical reasoning underpinning the knowledge from these disparate resources, we regard the retrievers from these various knowledge sources as distinct tools. We then apply the architecture of ReAct [38] to synergize reasoning and acting to leverage LLMs' reasoning ability to induce and modify the actions.

3.3.1 Reasoning Over Steps. To enable the model to understand which data resource to retrieve knowledge from, we enable Chat-Orthopedist to retrieve from a single source of data for each step and learn to reason over these steps to combine the knowledge from different sources. Consider Chat-Orthopedist as an agent that interacts with the environment with different retrieval tools and obtains information from different sources. At time step t, the agent needs to decide which action to take $a_t \in \mathcal{A}$, where $\mathcal{A}$ is the action space, containing retrieval operations from the pre-defined and created external knowledge bases. To learn how to make the decision, the target of the agent is to follow a policy $\pi(a_t|c_t)$, where $c_t = (o_1, a_1, o_2, a_2, \cdots, o_{t-1}, a_{t-1}, o_t)$ is the context to the agent and o_t is the observation obtained from the environment at the current step. For example, in Figure 1, when $t = 2$, Chat-Orthopedist obtains the information from the past records o_1, a_1, o_2 that the previous action "`Google Search`' did not receive enough information to answer the questions and determined the next action, a_t, is to seek help from the external knowledge base.

3.3.2 Prompt Engineering. We concatenate information from the following dimensions to augment LLMs with domain-expert knowledge and reasoning ability to organize the information from different resources:

- **Data Source Descriptions:** We offer a brief description covering the introduction of what the source contains and when the model needs to seek this source for information. More specifically, for the Google Search source, we leverage a description as:"Google Search is a portal to access public web pages. When you think you cannot answer the questions correctly only with information from the knowledge base, you may seek information in this source.";
- **Few-Shot Exemplar:** To enable the model to effectively retrieve information from these varied resources in the appropriate manner (format of call functions), we present three different exemplars to briefly guide the generation;
- **Historical Reasoning Records:** As reasoning and action are synergized step by step, we incorporate all historical reasoning records to enable the model to comprehend the historical states and the most recent environmental feedback. Then the model can be aware of the most suitable next action in line with the current context accordingly.

3.4 Statistical Analysis

We performed multiple quantitative and qualitative examinations to comprehensively assess the feasibility, acceptability, and effectiveness of adopting LLMs in SDM tool development.

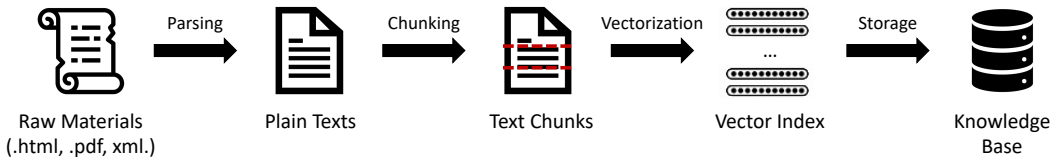


Figure 3: Automated pipeline for large-scale external domain-specific knowledge database establishment using multiple formats of raw materials, including document parsing, chunking, vectorization, and storage.

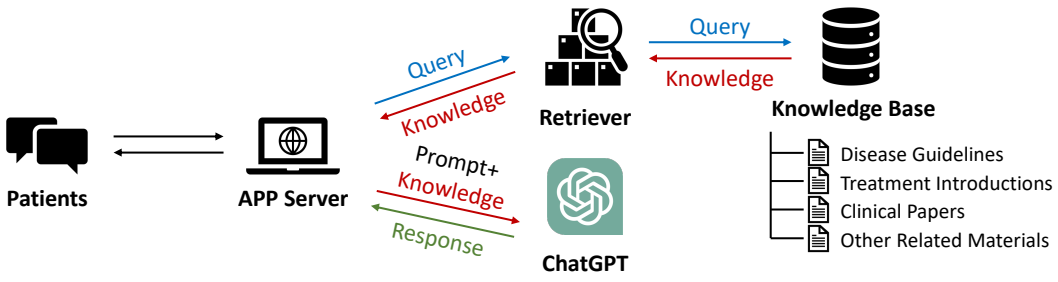


Figure 4: Overview of developing an SDM tool via retrieval augmented ChatGPT. First, we generate embedding for each document in the knowledge base with a text embedding model. For each query, we generate the embedding of the query and search the indexes of the top-*K* most relevant documents in the embedding space. Next, we use the indexes to retrieve the corresponding documents. Lastly, we use the retrieved relevant documents as context with the prompt and question, and send them to ChatGPT to generate the response.

3.4.1 Usability test. Following System Usability Scale (SUS) standard [9], we first conducted a usability study of our SDM tool to evaluate the effectiveness and usability of targeted users (i.e., parents). The SUS (user form) is a 4-item questionnaire designed to assess information accuracy, response clarity, answer relevance, and ease of understanding. The evaluation utilizes a scale ranging from 1 to 5, with 5 representing the most positive response.

3.4.2 Knowledge test. We performed two parallel 6-item, multiple-choice knowledge tests on two distinct user groups (those with and without access to our SDM tool) to determine the intervention's effects on knowledge of the relevant AIS disease condition and associated treatment options. A univariate analysis was subsequently conducted to compare the level of patient knowledge between the control group and the SDM group. The difference in scores from the knowledge test between these two groups was analyzed using Mann-Whitney U tests, with statistical significance defined as a p-value < 0.05.

3.4.3 Human-in-the-loop. We recruited a multidisciplinary team of orthopedic surgeons and researchers from multiple sites to conduct an iterative human-in-the-loop analysis and clinical validation. Frequently asked questions (FAQs) with generated answers, along with retrieved source information from the knowledge base, were reviewed by the expert team for a comprehensive evaluation and further improvement.

3.4.4 Tool comparison. We conducted an extensive comparative analysis of the responses generated by different tools with the expert team, including conventional search engines (Google), LLMs (ChatGPT), and retrieval-augmented LLMs (Chat-Orthopedist). Specifically, we integrated the Google search and ChatGPT API into our SDM tools to ensure a single-blinded experimental design. Beyond the original 4-item SUS, we designed an additional questionnaire (expert form) to provide a more comprehensive comparison of the three tools employed in SDM settings. This expanded questionnaire further encompasses factors such as completeness, fluency, credibility, verifiability, level of aggressiveness, and ethical concerns.

4 RESULTS AND DISCUSSIONS

Our central objective was to evaluate the feasibility and acceptability of the proposed SDM tool as a knowledge-sharing instrument in clinical settings. Consequently, our analyses of collected data were primarily descriptive and qualitative, supplemented by several inferential statistical analyses. We refrain from attributing statistical significance to purely descriptive results. All statically analytical procedures were carried out using IBM-SPSS version 28.0 and Prism version 9. Specifically, the following hypotheses were tested:

- **Hypothesis 1**: The mean scores on the usability test with the SUS questionnaire will indicate positive satisfaction (mean item scores >3.0) with the developed SDM tool.
- **Hypothesis 2**: The mean scores from the knowledge tests for parents with access to the SDM tool (the SDM group) will be significantly higher in comparison to the mean scores of parents without access (the control group).
- **Hypothesis 3**: The mean score of each item on the SUS questionnaire (expert form) will indicate positive satisfaction (mean item scores >3.0) with the LLMs-enabled tools.
- **Hypothesis 4**: As an exploratory hypothesis, the mean scores of the SUS questionnaire (expert form) of the proposed SDM tool will significantly exceed the mean scores

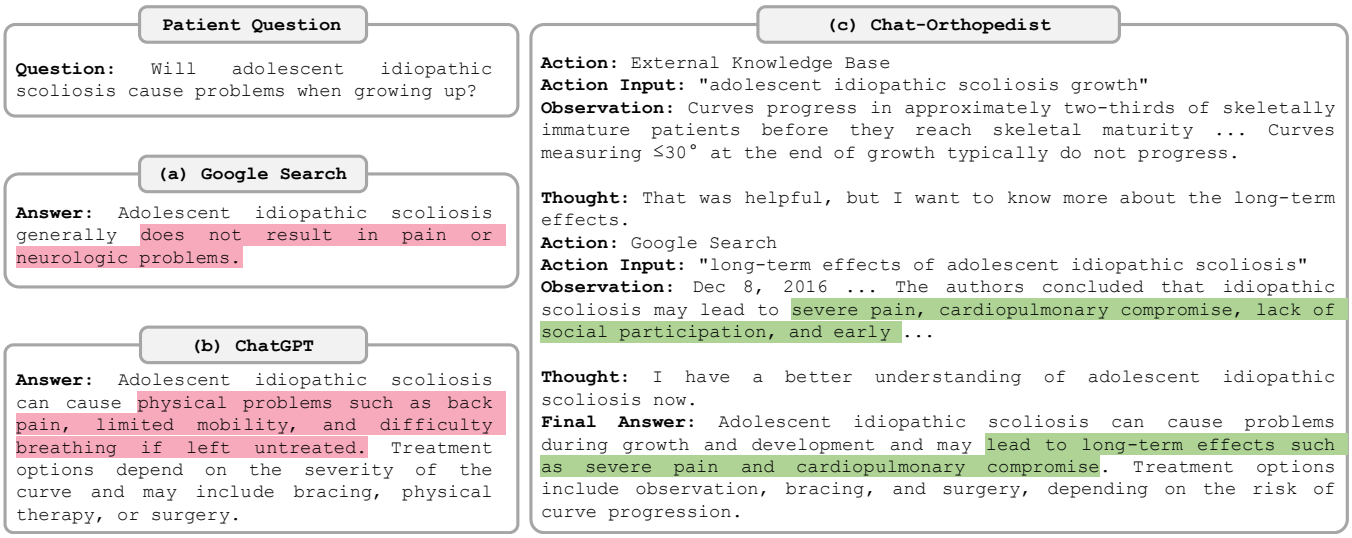


Figure 5: Comparison of three potential SDM tools in answering AIS disease and treatment-related questions. Actions and observations shrouded in pink are incorrect conclusions or imperfect inferences, whereas those in green are correct.

for both conventional search engines (Google) and LLMs (ChatGPT).

The analyses for Hypotheses 1 and 3 were primarily descriptive, based on the examination of the distributions of item scores and total scores across various measures. In addition, the analyses for Hypotheses 2 and 4 employed Mann-Whitney U tests, t-tests, and ANOVA tests to perform comparative statistical analyses, with a focus on establishing statistical significance (i.e., p-value< 0.05).

4.1 Case Studies

We presented an example of patient Q&A outcomes using Chat-Orthopedist and other potential SDM tools, such as Google and ChatGPT, in Figure 5. Specifically, the patient is asking a question about the long-term effect of AIS. Google search provides low-quality source information with wrong answers since untreated AIS will have a severe influence on both physical and mental health of patients [34]. Similarly, ChatGPT only focuses on the short-term effect on physical condition, which fails to provide a comprehensive answer. Chat-Orthopedist is able to successfully understand the requirement long-term effect in the query. With the external knowledge base and Google search engine, Chat-Orthopedist provides a more accurate and comprehensive answer regarding the potential long-term physical effect and self-reported outcomes, by reasoning from multiple sources (e.g., Google and external knowledge base). This also demonstrates that the established knowledge base is able to provide domain-specific information to facilitate SDM for AIS patients and parents. In Chat-Orthopedist, the external knowledge base contains multiple evidence-based and physician-authored clinical knowledge data resources, including PubMed clinical papers (e.g., meta-analysis, case studies), Scoliosis Research Society's (SRS) practice guidelines[1], UpToDate[2], and Dynamed[3]. Specifically, the proposed framework can be readily generalized to incorporate other interventions and other diseases by updating our existing knowledge base to encompass the new knowledge domain.

4.2 Usability Tests (Hypothesis 1)

A total of 128 targeted users (i.e., parents) were included in the usability test and final analysis after quality control. The demographic characteristics of participants represent a diverse national user group of parents from all genders, ages, regions, and income levels. Figure 6 provides descriptive statistics of summarized outcomes of all questions in the SUS (user form) from multiple perspectives. The summary plot reveals that the proposed SDM tool has attained average scores and standard deviations (mean±std) of 3.57±0.84, 3.83±0.73, 3.82±0.76, and 3.84±0.77 for accuracy, clarity, relevance, and simplicity (i.e., ease to understand), respectively. Given that all mean item scores surpass 3.5, the usability test outcome reflects a broad positive satisfaction among the targeted user groups. From the detailed distribution, we can observe a relatively consistent and robust mean item score among different examples, with accuracy ranging from 3.45 to 3.69, clarity from 3.69 to 3.98, relevance from 3.77 to 3.90, and simplicity from 3.56 to 4.09.

4.3 Knowledge Tests (Hypothesis 2)

Eighty-three intervention group users (with access to the proposed SDM tool) and 70 control group users were included in the final analysis. We selected the top 6 questions as knowledge tests from FAQs of AIS patients and families summarized by surgeons. Table 2 provides descriptive statistics and Mann-Whitney U Test outcomes for two groups in the knowledge test. We also present the difference between the two groups for each question in estimation plots (Figure 7) for more direct visualization. The assumption of normality for the scores of the knowledge tests was confirmed, thus negating the necessity for any data transformations. The SDM group obtained a significantly higher average score compared to the control group from each question (Q1-Q6), as indicated by the Mann-Whitney U Test results ($p < 0.0001$). The average score for the control group

[1]SRS: https://www.srs.org
[2]UpToDate: http://uptodate.com
[3]Dynamed: https://www.dynamed.com

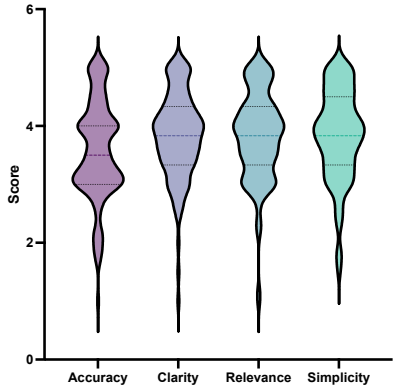


Figure 6: Descriptive statistics for average results of all questions in usability test from accuracy, clarity, relevance, and simplicity perspectives.

across all questions was 0.2667 out of 1, whereas the SDM group yielded a significantly higher average score of 0.6044. The difference between these group means was 0.3378 (95% CI: 0.2532 - 0.4223), which was statistically significant ($p < 0.0001$). This indicates a consistent and significant improvement in the knowledge test with the SDM approach compared to the control, suggesting the potential effectiveness of the proposed tool for knowledge enhancement.

4.4 Human-in-the-Loop (Hypothesis 3)

We recruited a multidisciplinary team with two orthopedic surgeons and 15 researchers from multiple sites to conduct an iterative human-in-the-loop analysis and clinical validation. The expert team reviewed the generated answers from the most popular patient questions identified by the clinical team, along with retrieved source information from the knowledge base. We reported the following SUS (expert form) evaluation in Figure 8. For the quality of generated answers, we can observe an overall positive evaluation on accuracy, clarity, completeness, relevance, fluency, and simplicity, with mean item scores of 4.46±0.73, 4.61±0.61, 3.89±1.07, 4.50±0.77, 4.63±0.63, and 4.40±0.90, respectively. Given that the majority of mean item scores are higher than 4.0, the SUS outcomes convey an overall positive assessment of the quality of the generated answers. Moreover, similar results on ChatGPT-generated responses further demonstrate the feasibility of adopting LLMs in clinical settings.

In addition, we also collected narrative comments from participants and the expert team during user feedback, clinical panels, and usability sessions to further evaluate the acceptability and effectiveness of the proposed SDM tools in the clinical workflow. Based on the results of the panel evaluation, we can conclude that the proposed SDM tool effectively delivers relatively accurate and unbiased information regarding the risks and benefits of various treatment options for AIS patients and their families. Moreover, after reviewing the source information of generated responses in our established knowledge database, we can observe that Chat-Orthopedist could properly infer or cite evidence-based conclusions in published treatment guidelines, scholarly reviews, and clinical papers. Successfully retrieving informative knowledge via search is critical in ensuring the quality of generated response. Through expert evaluation of source information, we can timely update the knowledge base by eliminating low-quality resources and incorporating more relevant and recent materials. Compared with the original ChatGPT or other LLMs, the retrieval-augmented framework could potentially solve the hallucination and black-box nature of LLMs, thereby promoting the adoption of LLMs in clinical settings.

4.5 SDM Tool Comparisons (Hypothesis 4)

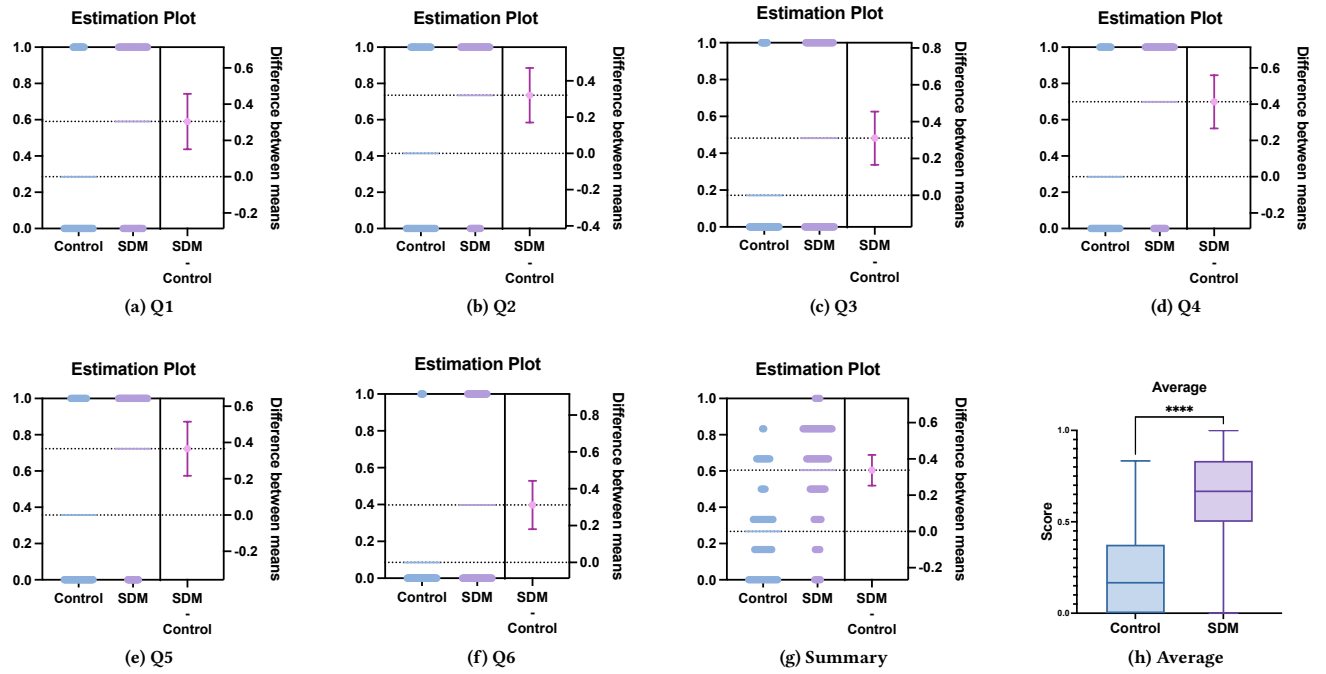
We compared Google, ChatGPT, and our proposed Chat-Orthopedist using the same SUS (expert form) evaluation as SDM tools for AIS patient care. The expert team reviewed the generated answers to the same question set by Google and ChatGPT. Specifically, we only provided original source information from Google, due to the lack of such information in ChatGPT. As an exploratory hypothesis test, we collected 150 evaluation records for each tool from a single evaluation perspective. Through the descriptive statistics, the mean item scores of three tools are presented in Figure 8. In terms of the quality of generated responses, it can be observed that the Chat-Orthopedist outperforms in the areas of accuracy, clarity, and relevance, whereas ChatGPT excels in completeness, fluency, and simplicity. Responses from the search engine (Google) were generally found to be of relatively low quality in most aspects of answer quality. However, when assessing the credibility and verifiability of source information, ChatGPT receives the lowest evaluation results. This could be attributed to the 'black-box' nature of LLMs, which hinders the provision of clear source information. Additionally, all three tools obtain similar performance with respect to low levels of aggressiveness and minimal ethical concerns. The overall scores of Google, ChatGPT, and our proposed Chat-Orthopedist are 4.22, 4.25, and 4.46, respectively. In addition, Table 3 presents the inferential statistical analysis via t-tests and ANOVA test with Greenhouse-Geisser correction. In summary, both the descriptive and inferential analyses demonstrate that LLM-enabled tools significantly outperform conventional search engines in the SDM process for AIS knowledge sharing. Specifically, Chat-Orthopedist illustrates its feasibility, acceptability, and effectiveness by augmenting ChatGPT with more accurate and relevant domain-specific knowledge. The clear identification of source information further improves credibility and enables validation through a human-in-the-loop approach.

4.6 Limitations and Future Works

Our design process for Chat-Orthopedist incorporated in-depth user feedback into model development. We also summarized potential limitations and action items for promoting LLMs-enabled SDM tool adoption in pediatric healthcare. Firstly, the responses generated must be easily comprehensible. Given the target audience, which includes AIS patients and their parents, it is crucial to consider the simplicity of the generated answers. This is particularly important in light of the evolving role of the growing patient in decision-making. It is also important to ensure that users of various ages, cultural backgrounds, and educational levels can derive benefit from this tool. Secondly, our current knowledge base is constructed through a systematic screening of existing materials and papers based on titles or keywords. An in-depth review of this knowledge base could prove instrumental in further enhancing the success rate of information retrieval. Thirdly, further demonstrations, including clinical trials assessing the efficiency of

Table 2: Descriptive statistics for knowledge test results of each question from the control and SDM groups.

ID	Mean (Control)	Mean (SDM)	Difference between Means	95% Confidence Interval	Mann-Whitney U Test (p-value)
Q1	0.2857	0.5904	0.3046 ± 0.07741	0.1517 - 0.4576	0.0002***
Q2	0.4143	0.7349	0.3207 ± 0.07604	0.1704 - 0.4709	<0.0001****
Q3	0.1714	0.4819	0.3105 ± 0.07313	0.1660 - 0.4550	<0.0001****
Q4	0.2857	0.6988	0.4131 ± 0.07442	0.2660 - 0.5601	<0.0001****
Q5	0.3571	0.7229	0.3657 ± 0.07551	0.2165 - 0.5149	<0.0001****
Q6	0.08571	0.3976	0.3119 ± 0.06651	0.1805 - 0.4433	<0.0001****
Avg.	0.2667	0.6044	0.3378 ± 0.04280	0.2532 - 0.4223	<0.0001****



(a) Q1 (b) Q2 (c) Q3 (d) Q4

(e) Q5 (f) Q6 (g) Summary (h) Average

Figure 7: Estimation plots (a)-(f) with descriptive statistics to show the difference in knowledge test results between the SDM group and the control group. In each plot, the right side shows the 95% confidence interval. The mean scores from the knowledge tests for the SDM group are significantly higher in comparison to the mean scores of the control group in (g) and (h).

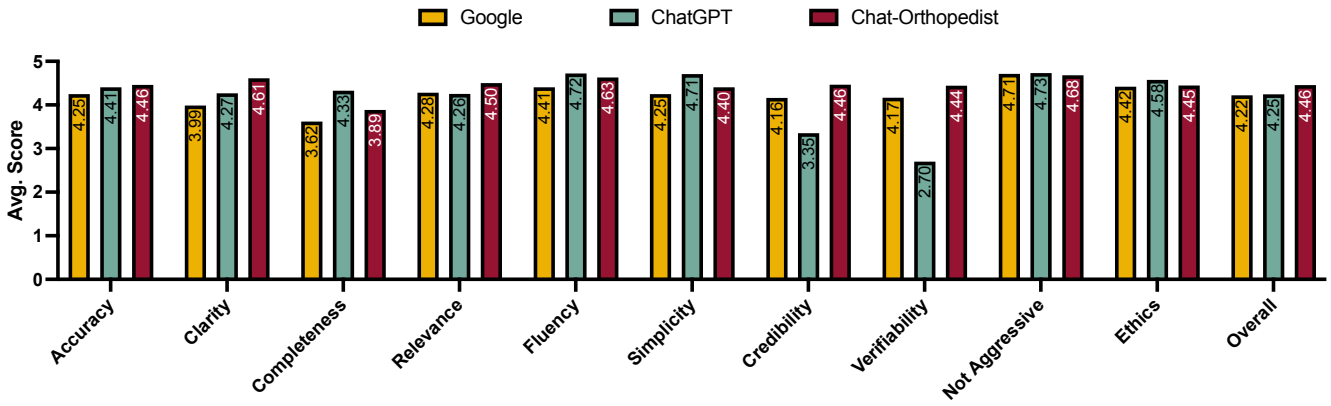


Figure 8: Descriptive statistics for expert evaluation results for comparison of Google, ChatGPT, and Chat-Orthopedist.

Table 3: Inferential statistical analyses for comparison of different Q&A tools, including (1) t-test between Google and ChatGPT (Group 1), (2) t-test between Google and Chat-Orthopedist (Group 2); (3) t-test between ChatGPT and Chat-Orthopedist (Group 3); and (4) ANOVA test among Google, ChatGPT, and Chat-Orthopedist.

	T-test (*p-value<0.05)			ANOVA
	Group 1	Group 2	Group 3	
Accuracy	0.0442*	0.0081*	0.3040	0.0238*
Clarity	0.0039*	<0.0001*	<0.0001*	<0.0001*
Completeness	<0.0001*	0.0127*	<0.0001*	<0.0001*
Relevance	0.7100	0.0139*	0.0365*	0.0083*
Fluency	<0.0001*	0.0020*	0.0429*	<0.0001*
Simplicity	<0.0001*	0.0784	<0.0001*	<0.0001*
Credibility	<0.0001*	0.0001*	<0.0001*	<0.0001*
Verifiability	<0.0001*	0.0020*	<0.0001*	<0.0001*
Not Aggressive	0.6489	0.4346	0.2242	0.4684
Ethics	0.0050*	0.5810	0.0682	0.0244*

clinical visits and improvement in post-operative patient-reported outcomes, are essential for real-world clinical applications. Lastly, SDM tools could benefit from not only patient Q&A but also the incorporation of multi-modal materials. These could include clinical illustration figures, diagnosis, and treatment decision trees, among other resources, thereby enriching the Chat-Orthopedist capacity to provide comprehensive and illustrative guidance. To accomplish the clinical mission of providing the highest quality of care to children with idiopathic scoliosis, our long-term goal is to develop validated decision support systems that can facilitate surgeons and adolescent patients to make well-informed decisions among various treatment options. Given the decision complexity between fusion surgery and observation, the rationale is that the proposed SDM could promote patient knowledge acquisition and optimize clinical workflow, leading to more efficient patient care and improved patient treatment outcomes.

5 CONSIDERATIONS OF LLMS FOR HEALTH

During human-in-the-loop iterative sessions for model validations and improvements, we focused on design preferences to improve the usefulness and usability of LLMs tools that provide AI-generated feedback on patient-provider communication during SDM and clinical encounters. Drawing insights from clinical feedback and extensive literature reviews, we distilled several major considerations with potential solutions for widely adopting LLMs in real-world healthcare for future studies.

Firstly, we have observed that LLMs encounter difficulties with less prevalent factual knowledge, which may lead to hallucinated or less reliable generations [21]. To enhance the credibility of LLMs, we have leveraged a 'knowledge brain' grounded in Google search and authoritative databases in Chat-Orthopedist. Moreover, this retrieval-augmented framework can access timely updated domain-specific information to address patients' inquiries based on a trustworthy knowledge source, which is important for clinical settings with minimal tolerance for errors or hallucinations [40].

Secondly, LLMs with over 100 billion parameters, such as GPT-3.5, are usually commercially restricted and not open-sourced [4, 25]. On the other hand, even open-source LMs like LLaMA-13B or LLaMA-65B [31] require significant computational resources for local fine-tuning. For example, fine-tuning a BLOOM-176B requires 72 A100 GPUs, each with 80GB memory and costing $15k apiece [27]. This substantial resource requirement makes LLMs largely inaccessible for researchers and developers with limited resources. Consequently, the common approach for downstream AI applications is transitioning from fine-tuning specialized models towards prompting generalist models (e.g., in-context learning) [41].

Thirdly, most of the current medical LLMs [10, 40] lack adequate security measures to assure accurate medical diagnoses and recommendations. Considering the real-world clinical practice requirements, responsible AI is a critical prerequisite for adopting LLMs in healthcare. As LLMs, such as ChatGPT, are only accessible through black-box APIs, where users can submit queries and receive responses, a trade-off emerges between model capability and model transparency. In Chat-Orthopedist, we improve model transparency by presenting source information with reasoning (e.g., chain-of-thoughts) and acting (e.g., action plan generation).

Fourthly, simply scaling the model size has not proven sufficient for achieving high performance on complicated tasks such as reasoning in intricate clinical scenarios [32]. One potential challenge in pediatric healthcare stems from the necessity of appropriate understandability, given the changes in education level during growth. Another challenge that lies in the potential applications for few-shot or zero-shot learning scenarios is the diagnosis of rare diseases. To address weaknesses in model reasoning, potential solutions like chain-of-thought [32] or tree-of-thought [37] could be readily implemented with prompts, serving as intermediate steps towards problem-solving for improved reasoning and easy understanding. It enables models to decompose multi-step problems into intermediate stages with an interpretable insight into the model behaviors.

6 CONCLUSION AND BROADER IMPACT

In this study, we developed and validated an innovative SDM tool, Chat-Orthopedist, to prepare AIS patients and families for a meaningful discussion with clinicians. The usability tests with human-in-the-loop demonstrate the effectiveness of the proposed LLM-enabled SDM tool in delivering accurate and unbiased information on disease knowledge and treatment options. In addition, we discussed several critical considerations and potential solutions for widely adopting ChatGPT-like LLMs to facilitate clinical practice. From a clinical perspective, successfully implementing the proposed SDM tool could assist AIS patients by increasing access to required clinical knowledge for effective medical consultations. In addition, it could potentially facilitate clinicians to enhance efficiency in clinical visits and in reducing workload. From a technical perspective, this work may promote the adoption of LLMs in real-world clinical applications by solving challenges associated with generalizing AI at scale for domain-specific tasks. This study may serve as a pilot and feasibility examination to support the feasibility, acceptability, and effectiveness of LLMs-enabled SDM in pediatrics. We expect the cross-discipline collaboration between LLMs and SDMs will

ultimately improve the treatment outcomes of children with AIS in a family-centered and collaborative environment.

ACKNOWLEDGMENTS

This research has been supported by Accelerate Foundation Models Academic Research Initiative from Microsoft Research. It has been also supported by Shriners Children's Hospital and Georgia Institute of Technology in Greenville-Lexington Shriner Multi-site AI-enabled Rehabilitation Technology for Scoliosis Patients Care (GL-SMART) project. In addition, this work has been supported by a Wallace H. Coulter Distinguished Faculty Fellowship, a Petit Institute Faculty Fellowship, and research funding from Amazon and Microsoft Research to Professor May D. Wang.

REFERENCES

[1] Stig Aaro and CARL Ohlund. 1984. Scoliosis and pulmonary function. *Spine* 9, 2 (1984), 220–222.

[2] John W Ayers, Adam Poliak, Mark Dredze, Eric C Leas, Zechariah Zhu, et al. 2023. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Internal Medicine* (2023).

[3] Carolina Bejarano, Lindsay Fuzzell, Catharine Clay, Sharon Leonard, Eric Shirley, et al. 2015. Shared decision making in pediatrics: A pilot and feasibility project. *Clinical Practice in Pediatric Psychology* 3, 1 (2015), 25. Publisher: Educational Publishing Foundation.

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[5] Ashley M Butler, Sara Elkins, Marc Kowalkowski, and Jean L Raphael. 2015. Shared decision making among parents of children with mental health conditions compared to children with chronic physical conditions. *Maternal and child health journal* 19 (2015), 410–418. Publisher: Springer.

[6] Aidan Gilson, Conrad W Safranek, Thomas Huang, Vimig Socrates, Ling Chi, et al. 2023. How does CHATGPT perform on the United States Medical Licensing Examination? the implications of large language models for medical education and knowledge assessment. *JMIR Medical Education* 9, 1 (2023), e45312.

[7] Felipe Giuste, Wenqi Shi, Yuanda Zhu, Tarun Naren, Monica Isgut, et al. 2023. Explainable Artificial Intelligence Methods in Combating Pandemics: A Systematic Review. *IEEE Reviews in Biomedical Engineering* 16 (2023), 5–21. https://doi.org/10.1109/RBME.2022.3185953

[8] Olga Golovneva, Moya Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, et al. 2022. ROSCOE: A Suite of Metrics for Scoring Step-by-Step Reasoning. *arXiv preprint arXiv:2212.07919* (2022).

[9] Rebecca A. Grier, Aaron Bangor, Philip T. Kortum, and S. Camille Peres. 2013. The System Usability Scale. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 57 (2013), 187 – 191.

[10] Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, et al. 2023. MedAlpaca–An Open-Source Collection of Medical Conversational AI Models and Training Data. *arXiv preprint arXiv:2304.08247* (2023).

[11] Oluwatomilona I Ifelayo, Juan P Brito, Ian G Hargraves, and A Noelle Larson. 2021. Development of a shared decision-making tool for adolescents with scoliosis to decide between observation versus fusion surgery. *Journal of Pediatric Orthopaedics* 41 (2021), S70–S74. Publisher: LWW.

[12] Monica Isgut, Logan Gloster, Katherine Choi, Janani Venugopalan, and May D. Wang. 2023. Systematic Review of Advanced AI Methods for Improving Healthcare Data Quality in Post COVID-19 Era. *IEEE Reviews in Biomedical Engineering* 16 (2023), 53–69. https://doi.org/10.1109/RBME.2022.3216531

[13] Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, et al. 2022. Atlas: Few-shot learning with retrieval augmented language models. (Aug. 2022). _eprint: 2208.03299.

[14] Joseph A Janicki and Benjamin Alman. 2007. Scoliosis: Review of diagnosis and treatment. *Paediatrics & child health* 12, 9 (2007), 771–776. Publisher: Oxford University Press.

[15] Lori A Karol. 2019. The natural history of early-onset scoliosis. *Journal of Pediatric Orthopaedics* 39 (2019), S38–S43. Publisher: LWW.

[16] Robert K Lark, Elizabeth Ellie H Garman, Mary Jackson, and Katherine S Garman. 2022. Shared Decision-Making for Juvenile Scoliosis. *Pediatrics* 149, 4 (2022). Publisher: American Academy of Pediatrics.

[17] Christopher S Lee, Soroush Merchant, and Vidya Chidambaran. 2020. Postoperative pain management in pediatric spinal fusion surgery for idiopathic scoliosis. *Pediatric Drugs* 22 (2020), 575–601. Publisher: Springer.

[18] Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. CAMEL: Communicative Agents for" Mind" Exploration of Large Scale Language Model Society. *arXiv preprint arXiv:2303.17760* (2023).

[19] Marios G Lykissas, Viral V Jain, Senthil T Nathan, Varun Pawar, Emily A Eismann, et al. 2013. Mid-to long-term outcomes in adolescent idiopathic scoliosis after instrumented posterior spinal fusion: a meta-analysis. *Spine* 38, 2 (2013), E113–E119. Publisher: LWW.

[20] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, et al. 2023. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651* (2023).

[21] Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Hannaneh Hajishirzi, et al. 2022. When Not to Trust Language Models: Investigating Effectiveness and Limitations of Parametric and Non-Parametric Memories. *arXiv preprint arXiv:2212.10511* (2022).

[22] Peter O Newton, Frances D Faro, Lawrence G Lenke, Randal R Betz, David H Clements, et al. 2003. Factors involved in the decision to perform a selective versus nonselective fusion of Lenke 1B and 1C (King-Moe II) curves in adolescent idiopathic scoliosis. *Spine* 28, 20S (2003), S217–S223. Publisher: LWW.

[23] 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* (2023).

[24] OpenAI. 2023. GPT-4 Technical Report. arXiv:cs.CL/2303.08774

[25] OpenAI. 2023. Introducing ChatGPT. https://openai.com/blog/chatgpt

[26] Wenqi Shi, Felipe O. Giuste, Yuanda Zhu, Ashley M. Carpenter, Henry J. Iwinski, et al. 2021. A FHIR-compliant Application for Multi-Site and Multi-Modality Pediatric Scoliosis Patient Rehabilitation. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 1524–1527. https://doi.org/10.1109/BIBM52615.2021.9669649

[27] Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, et al. 2023. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652* (2023).

[28] Wenqi Shi, Li Tong, Yuanda Zhu, and May D Wang. 2021. COVID-19 automatic diagnosis with radiographic imaging: Explainable attention transfer deep neural networks. *IEEE Journal of Biomedical and Health Informatics* 25, 7 (2021), 2376–2387.

[29] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems* 33 (2020), 16857–16867.

[30] Haotian Sun, Yuchen Zhuang, Lingkai Kong, Bo Dai, and Chao Zhang. 2023. AdaPlanner: Adaptive Planning from Feedback with Language Models. arXiv:cs.CL/2305.16653

[31] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[32] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, et al. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903* (2022).

[33] Stuart L Weinstein. 1989. Adolescent idiopathic scoliosis: prevalence and natural history. *Instructional course lectures* 38 (1989), 115–128.

[34] Stuart L Weinstein. 2019. The natural history of adolescent idiopathic scoliosis. *Journal of Pediatric Orthopaedics* 39 (2019), S44–S46. Publisher: LWW.

[35] Stuart L Weinstein, Lori A Dolan, Jack CY Cheng, Aina Danielsson, and Jose A Morcuende. 2008. Adolescent idiopathic scoliosis. *The lancet* 371, 9623 (2008), 1527–1537. Publisher: Elsevier.

[36] Po-Yen Wu, Chih-Wen Cheng, Chanchala D. Kaddi, Janani Venugopalan, Ryan Hoffman, et al. 2017. –Omic and Electronic Health Record Big Data Analytics for Precision Medicine. *IEEE Transactions on Biomedical Engineering* 64, 2 (2017), 263–273. https://doi.org/10.1109/TBME.2016.2573285

[37] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, et al. 2023. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. *arXiv preprint arXiv:2305.10601* (2023).

[38] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, et al. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629* (2022).

[39] Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander Ratner, et al. 2023. Large Language Model as Attributed Training Data Generator: A Tale of Diversity and Bias. arXiv:cs.CL/2306.15895

[40] Li Yunxiang, Li Zihan, Zhang Kai, Dan Ruilong, and Zhang You. 2023. Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge. *arXiv preprint arXiv:2303.14070* (2023).

[41] Wangchunshu Zhou, Yuchen Eleanor Jiang, Ryan Cotterell, and Mrinmaya Sachan. 2023. Efficient Prompting via Dynamic In-Context Learning. *arXiv preprint arXiv:2305.11170* (2023).

[42] Yuchen Zhuang, Yue Yu, Kuan Wang, Haotian Sun, and Chao Zhang. 2023. ToolQA: A Dataset for LLM Question Answering with External Tools. arXiv:cs.CL/2306.13304