Paging Dr. AI: Which Chatbot Handles Pediatric Orthopedic Referrals Best?

Don Le[1], Jacob Siahaan, Surya Mundluru[2]
[1]Orthopedic Surgery, UTHealth Houston, [2]University of Texas Health Sciences Center

INTRODUCTION: The healthcare landscape is currently witnessing a surge in the utilization of Artificial Intelligence (AI) and Language Learning Models (LLM) to augment clinical decision-making processes. One critical area of interest is the referral process from primary care providers (PCP) to pediatric orthopaedic surgeons (POS), where traumatic and congenital orthopaedic conditions often necessitate specialized patient care from an early age. Publicly available AI chatbots present a promising solution by leveraging natural language processing (NLP) algorithms to analyze patient histories and provide triage recommendations for referral based on their access to knowledge databases, social media, and open data sources. However, the knowledge about how these chatbots compare to one another and their accuracy when compared to orthopedic specialist assessments remains unexplored. The purpose of this study was to assess and compare the efficacy of widely used artificial intelligence (AI) chatbots (ChatGPT, Microsoft Copilot, and Google Gemini) in enhancing the urgency, accuracy, and readability of PCP referrals to pediatric orthopaedists.

METHODS: A predetermined set of 5 patient histories, many of which were designed to parallel a typical pediatric orthopaedic patient interaction in a PCP clinic, served as standardized cases for evaluation. The three most widely used publicly available AI chatbots (ChatGPT, Copilot, and Google Gemini) were prompted with each patient history to ascertain whether a referral to an orthopaedic specialist is warranted and determine the urgency level of the referral on a scale of 1 to 5 (1 being no referral and 5 being urgent referral). Concurrently, four POS blindly and independently assessed the same patient histories and provided their recommendations for referral and urgency level. Differences in referral recommendations and urgencies were assessed and compared between the AI chatbots and POS to determine which AI chatbots most closely align with the clinical experience of POS. The Flesch-Kincaid Grade Level and Flesch Reading Ease indices were utilized to determine the readability and clarity of the AI chatbot responses. Statistical analyses comparing the AI chatbot responses with the POS responses to each of the five standardized cases were conducted using paired t-tests.

RESULTS: The results indicated no statistically significant differences between the averages urgency levels of each respective search engine and the average urgency levels of the POS. AI chatbot prompts and urgency levels are shown in Figure 1). Google Gemini failed to respond to prompt five, stating that it does not have the capacity to help. While not statistically significant, ChatGPT 3.5 shows the highest average difference between the AI Chatbot and POS urgency levels (p=0.27) (Figure 2). Microsoft CoPilot provided the highest word count per response with an average of 188.8 words while ChatGPT 3.5 provided the lowest word count with an average of 68.2 words. On average, ChatGPT 3.5, scored the second highest on the Flesch-Kincaid Grade Level (12.82), and lowest on the Flesch Ease Score (27.28). These scores both indicated that, while short in word length, ChatGPT 3.5's responses were among the most difficult to read of the tested AI chatbots (Figure 3).

DISCUSSION AND CONCLUSION: The findings of this pilot study indicate that the differences between AI-generated urgency levels and POS responses to standardized pediatric orthopaedic patient histories are insignificant. However, ChatGPT 3.5 showed the most difference in POS and AI chatbot urgency levels for the prompts, suggesting a decrease in accuracy when compared to ChatGPT 4, Microsoft Co-Pilot, and Google Gemini. In addition, ChatGPT 3.5 consistently generated the shortest and most difficult to read responses when compared to the other AI chatbots.

Figure 1. Standardized Prompts and Urgency Level Responses

Prompt	ChatGPT 4	ChatGPT 3.5	Microsoft Co-Pilot (More Balanced Non-Pro)	Google Gemini (Non-Advanced)	Surgeon Average
I am a pediatrician. A 10-year-old male child has been **"pigeon toeing ever since they could walk"**, and their parents are concerned. The patient is comfortable and in no pain. Should I refer to orthopedics? Answer with yes or no, explain your reasoning and rank the urgency level on a scale of 1-5 (1 being no referral and 5 being urgent referral).	2	3	3	2	1.75
I am a pediatrician. A 10-year-old female child has **"one leg more bowed than the other"** and their parents are concerned. The patient is comfortable and in no pain. Should I refer to orthopedics? Answer with yes or no, explain your reasoning and rank the urgency level on a scale of 1-5 (1 being no referral and 5 being urgent referral).	3	3	3	3	3.5
I am a pediatrician. A 10-year-old male child has **"fallen on their outstretched hand while playing on the playground."** There is no visible deformity, and their parents are concerned. The patient is comfortable and in no pain. Should I refer to orthopedics? Answer with yes or no, explain your reasoning and rank the urgency level on a scale of 1-5 (1 being no referral and 5 being urgent referral).	2	3	1	2	1.5
I am a pediatrician. A 10-year-old female child has **"a shoulder-level height difference and a family history of scoliosis"** and their parents are concerned. The patient is comfortable and in no pain. Should I refer to orthopedics? Answer with yes or no, explain your reasoning, and rank the urgency level on a scale of 1-5 (1 being no referral and 5 being urgent referral).	3	3	5	3	3.25
I am a pediatrician. A 10-year-old male child has **"a 1-month history of weight loss, leg pain, and leg swelling"** and their parents are concerned. Should I refer to orthopedics? Answer with yes or no, explain your reasoning, and rank the urgency level on a scale of 1-5 (1 being no referral and 5 being urgent referral).	4	5	4	N/A	4.5

N/A = Chatbot failure to respond, stating no capacity to help.

Figure 2. Differences Between AI Chatbot and Surgeon Average Urgency Levels

Artificial Intelligence Chatbots	Average Urgency Levels	Surgeon Urgency Level Average Difference	P-value
ChatGPT 4	2.8	0.1	0.6
ChatGPT 3.5	3.4	0.5	0.3
Microsoft Co-Pilot (More Balanced Non-Pro)	3.2	0.3	0.6
Google Gemini (Non-Advanced)	2.5	0.4	1

Figure 3. Readability Characteristics Among AI Chatbots

Artificial Intelligence Chatbots	Flesch-Kincaid Grade Level (0-18)	Flesch Reading Ease (100-0)	Text Length Average (words)
Microsoft Co-Pilot (More Balanced Non-Pro)	10.82	38.98	188.8
Google Gemini (Non-Advanced)	11.225	41.35	151*
ChatGPT 3.5	12.82	27.28	68.2
ChatGPT 4	13.78	28	134.8

* = One outlier removed due to chatbot failure to respond, stating no capacity to help.