Rapid Evolution of ChatGPT's Reliability for Self-Diagnosis: A Repeated-Prompt Study on Five Common Orthopedic Conditions

Tomoyuki Kuroiwa[1], Toru Sasaki[2], Sara Sugiura[1], Sunghyun Lee[3], Tomohiko Waki[1], Yusuke Inomori[1], Takuya Ibara[2], Toshitaka Yoshii[1], Koji Fujita[4]
[1]Department of Orthopaedic and Spinal Surgery, Graduate School of Medical and Dental Sciences, Institute of Science Tokyo, Tokyo, Japan
[2]Department of Functional Joint Anatomy, Graduate School of Medical and Dental Sciences, Institute of Science Tokyo, Tokyo, Japan
[3]Center for Stem Cell and Regenerative Medicine, Institute of Science Tokyo, Tokyo, Japan
[4]Medical Design Section, Center for Medical Innovation, Institute of Science Tokyo, Tokyo, Japan

Email of Presenting Author: Kuroiwa.orth@tmd.ac.jp

Disclosures: Tomoyuki Kuroiwa (N), Toru Sasaki (N), Sara Sugiura (N), Sunghyun Lee (N), Tomohiko Waki (N), Yusuke Inomori (N), Takuya Ibara (N), Toshitaka Yoshii (N), Koji Fujita (N)

INTRODUCTION: With the rapid development of generative AI, applications in the medical field are expanding beyond image diagnosis support and pathology analysis to electronic medical record summarization, treatment plan proposals, patient education, and mental health care. Furthermore, because generative AI can now comprehensively support the process of "verbalizing symptoms → suggesting suspected diseases → advising on medical consultations," and its practical use as a self-diagnosis support tool for the general public is becoming widespread. However, AI output is probabilistic[1], leading to misinformation (hallucination) and inconsistent answers (lack of reproducibility). Particularly in the context of self-diagnosis, core requirements for safety and trust are i) diagnostic accuracy (reaching the correct disease) and (ii) reproducibility (arriving at the same conclusion for the same input). **We previously quantified these in ChatGPT 3.5 by repeatedly submitting standardized patient-style symptom prompts**, focusing on these two points, and also examined the strength of the recommendation for medical consultation in the answers[2]. The results showed a low accuracy rate, low reproducibility, and weak consultation recommendations, revealing the limitations of this tool as a self-diagnosis support tool. However, AI is evolving rapidly, and ChatGPT has released new versions every year, with its inference capabilities said to be at or above the level of doctoral students in physics and chemistry. Therefore, **we re-evaluated recent and latest versions using the same repeated-prompt design to measure improvements in accuracy, reproducibility, and referral recommendations.**

METHODS: The target diseases were five common orthopedic conditions with relatively clear symptoms: carpal tunnel syndrome (CTS), cervical spondylotic myelopathy (CSM), lumbar spinal stenosis (LSS), knee osteoarthritis (KOA), and hip osteoarthritis (HOA). **We created a canonical prompt describing typical symptoms for each disease,** always ending with the words, " Can you give me a primary diagnosis and a list of five potential differential diagnoses?." Four examiners entered the exact same prompt for five days (1 disease x 20 runs/model).

Evaluation items were **(1) Accuracy**—the proportion of runs in which the primary diagnosis matched the target disease (synonyms allowed); **(2) Reproducibility**—observed agreement and the Prevalence-Adjusted Bias-Adjusted Kappa (PABAK) for primary and differential diagnoses, assessed within-rater (intra-rater) and between-raters (inter-rater); and **(3) Recommendation for medical consultation**: the strength of the recommendation for medical consultation in the response was classified into five levels and counted. The **ChatGPT versions used were 4.1 mini, o4 mini high, and 5.** The rater was blinded to each version during the analysis.

RESULTS: Across versions 4.1 mini, o4 mini high, and 5, mean primary-diagnosis **accuracy was 97%, 98%, and 100%,** respectively (for the first two versions, accuracy was 100% for all diseases except CSM) (Table 1). Reproducibility for the primary diagnosis showed mean **intra-rater observed agreement of 0.95, 0.96, and 1.00**, and mean **inter-rater observed agreement of 0.94, 0.96, and 1.00**, respectively (Table 2). For differential diagnoses, mean **intra-rater observed agreement was 0.31, 0.28, and 0.33**, while mean **inter-rater observed agreement was 0.14, 0.21, and 0.30**, respectively. The proportion of answers categorizing recommendation for medical consultation as **"essential/mandatory" was 62%, 67%, and 81%,** respectively. (Table 3)

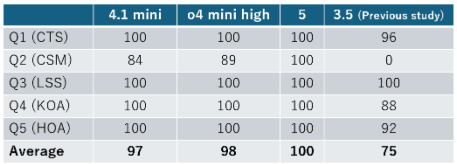
DISCUSSION: In version 3.5, the accuracy rates for CTS, KOA, and HOA were not 100%, and for CSM was extremely low at 4%. In contrast, in the recent version, **the accuracy rates for both KOA and HOA were 100%, and for CSM, they were extremely high. The latest version achieved a 100% accuracy rate even for CSM.** Furthermore, in version 3.5, the mean observed agreement for differential **diagnosis was low at 0.13 within examiners and 0.11 between examiners, but this improved in the recent and the latest versions.** Furthermore, **the recommendation for consultation was "essential" in only 3% in the answers in version 3.5, but this also improved in the recent and the latest versions.**

This study demonstrates that the diagnostic ability of LLM for typical orthopedic symptoms can improve in both accuracy and reproducibility within just 1 to 2 years. While version 3.5 demonstrated high accuracy in diagnosing CTS and LSS, which have characteristic symptoms, it exhibited remarkable: weaknesses in CSM, which has a wide differential diagnosis. However, in contrast, recent and latest versions demonstrated ceiling effects across multiple questions, suggesting improved diagnostic ability. Regarding the differences between version 4.1 mini and version o4 mini high, the **"o4" series emphasizes on inferential ability, suggesting that this inferential ability may also have a positive effect on medical diagnosis.** Furthermore, in addition to improved diagnostic ability, there has been a shift toward clarity and safety in terms of the recommendation level for medical consultation, **suggesting that the system's safety and reliability as a self-diagnosis support system have increased.** However, while this study was conducted under ideal conditions with "English, typical symptoms, and standard phrases," real-world clinical settings are complex, involving ambiguous patient descriptions, multiple coexisting diseases, atypical clinical histories, and diverse linguistic and socio-cultural backgrounds.

We plan to incorporate patients' own descriptions to more closely resemble the realities of clinical practice in the future research. Furthermore, we aim to develop AI software specifically for medical use that can integrate multimodal input, including language, images, and sensor information, to aid diagnosis.

SIGNIFICANCE/CLINICAL RELEVANCE: In this study, the latest version of ChatGPT showed notable improvements in diagnostic accuracy, reproducibility, and strength of consultation recommendations compared with the older version. **These results suggest that ChatGPT may now be a sufficiently reliable tool for self-diagnosis support, at least for simple orthopedic diseases.**

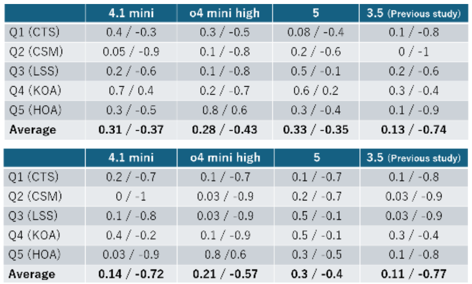
Table 1. Accuracy



	4.1 mini	o4 mini high	5	3.5 (Previous study)
Q1 (CTS)	100	100	100	96
Q2 (CSM)	84	89	100	0
Q3 (LSS)	100	100	100	100
Q4 (KOA)	100	100	100	88
Q5 (HOA)	100	100	100	92
Average	**97**	**98**	**100**	**75**

* These values are in %.

Table 2. Reproducibility

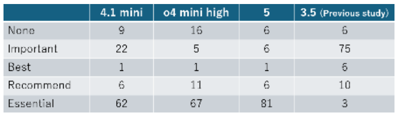


	4.1 mini	o4 mini high	5	3.5 (Previous study)
Q1 (CTS)	0.4 / -0.3	0.3 / -0.5	0.08 / -0.4	0.1 / -0.8
Q2 (CSM)	0.05 / -0.9	0.1 / -0.8	0.2 / -0.6	0 / -1
Q3 (LSS)	0.2 / -0.6	0.1 / -0.8	0.5 / -0.1	0.2 / -0.6
Q4 (KOA)	0.7 / 0.4	0.2 / -0.7	0.6 / 0.2	0.3 / -0.4
Q5 (HOA)	0.3 / -0.5	0.8 / 0.6	0.3 / -0.4	0.1 / -0.9
Average	**0.31 / -0.37**	**0.28 / -0.43**	**0.33 / -0.35**	**0.13 / -0.74**

	4.1 mini	o4 mini high	5	3.5 (Previous study)
Q1 (CTS)	0.2 / -0.7	0.1 / -0.7	0.1 / -0.7	0.1 / -0.8
Q2 (CSM)	0 / -1	0.03 / -0.9	0.2 / -0.7	0.03 / -0.9
Q3 (LSS)	0.1 / -0.8	0.03 / -0.9	0.5 / -0.1	0.03 / -0.9
Q4 (KOA)	0.4 / -0.2	0.1 / -0.9	0.5 / -0.1	0.3 / -0.4
Q5 (HOA)	0.03 / -0.9	0.8 /0.6	0.3 / -0.5	0.1 / -0.8
Average	**0.14 / -0.72**	**0.21 / -0.57**	**0.3 / -0.4**	**0.11 / -0.77**

* Values indicate Observed agreement / PABAK

Table 3. Recommendations



	4.1 mini	o4 mini high	5	3.5 (Previous study)
None	9	16	6	6
Important	22	5	6	75
Best	1	1	1	6
Recommend	6	11	6	10
Essential	62	67	81	3

* These values are in %.

References
1. Kameda et al. *N Engl J Med.* 2025
2. Kuroiwa et al. *J Med Internet Res.* 2023