Comparing AI-Generated and Clinician-Created Personalized Self-Management Guidance for Knee Osteoarthritis Patients: A Blinded Observational Study

Kai Du, Ao Li, Qi-Heng Zuo, Chen-Yu Zhang, Ren Guo, Ping Chen, Wei-Shuai Du, Shu-Ming Li

Submitted to: Journal of Medical Internet Research
on: October 22, 2024

Table of Contents

Comparing AI-Generated and Clinician-Created Personalized Self-Management Guidance for Knee Osteoarthritis Patients: A Blinded Observational Study

Kai Du[1*]; Ao Li[1*]; Qi-Heng Zuo[1]; Chen-Yu Zhang[1]; Ren Guo[2]; Ping Chen[2]; Wei-Shuai Du[2]; Shu-Ming Li[2]

[1]Beijing University of Chinese Medicine Beijing CN
[2]Beijing Hospital of Traditional Chinese Medicine Beijing CN
[*]these authors contributed equally

Corresponding Author:
Shu-Ming Li
Beijing Hospital of Traditional Chinese Medicine
23 Meishuguan Houjie, Dongcheng District, Beijing 100010, China
Beijing
CN

Abstract

Background: Personalized education is crucial for effective knee osteoarthritis (OA) management, but providing it remains challenging due to imbalanced patient-provider ratio and limited resources. This study explores the potential of GPT-4, a large language model, in generating tailored self-management guidance and compares its performance with physician-provided advice.

Objective: Objective?This study aims to evaluate the effectiveness of GPT-4 in generating personalized education materials for patients with knee osteoarthritis (OA) and compare it with experienced clinicians. Specifically, the comparison is made in terms of efficiency, readability, accuracy, personalization, comprehensiveness, and safety. By leveraging patient data from previous trials, it is evaluated whether AI can improve the quality and accuracy of patient education and evaluate its potential in improving patient care and outcomes.

Methods: A two-phase, blinded, observational study was conducted using patient data from a previous trial. In phase one, two experienced orthopedic specialists created personalized education materials. In phase two, the same data were input into GPT-4 by a physician to generate content. Materials were evaluated for efficiency (words per minute), readability (Flesch-Kincaid Grade Level, Gunning Fog Index, Coleman-Liau Index, and SMOG Index), accuracy, personality, comprehensiveness, and safety.

Results: GPT-4 demonstrated higher efficiency than clinicians (median 530.03 vs. 37.29 WPM, $P < 0.001$). GPT-4 content exhibited superior readability on the Flesch-Kincaid grade level, Gunning Fog Index, and SMOG Index ($P < 0.001$). Expert evaluations revealed that GPT-4 outperformed clinicians in accuracy (5.307 ± 1.731 vs. 4.76 ± 1.098, $P = 0.047$), personality (54.32 ± 6.212 vs. 33.2 ± 5.395, $P < 0.001$), comprehensiveness (51.74 ± 6.471 vs. 35.26 ± 6.657, $P < 0.001$), and safety (median 61 vs. 50, $P < 0.001$).

Conclusions: Conclusion: GPT-4 shows promise in generating high-quality, personalized patient education for knee OA, surpassing human experts. This study provides novel evidence for the potential of AI in enabling precise and intelligent patient education. Further research is needed to validate the findings in larger populations and assess the impact on patient outcomes.

Conclusion: GPT-4 shows promise in generating high-quality, personalized patient education for knee OA, surpassing human experts. This study provides novel evidence for the potential of AI in enabling precise and intelligent patient education. Further research is needed to validate the findings in larger populations and assess the impact on patient outcomes.

(JMIR Preprints 22/10/2024:67830)
DOI: https://doi.org/10.2196/preprints.67830

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Comparing AI-Generated and Clinician-Created Personalized Self-Management Guidance for Knee Osteoarthritis Patients: A Blinded Observational Study

Kai Du[1], Ao Li[1], Qi -Heng Zuo[1], Chen -Yu Zhang[1], Ren Guo[2], Ping Chen[2], Wei -Shuai Du[2], Shu-Ming Li[2*]

[1] Graduate School, Beijing University of Chinese Medicine, 11 Beisanhuan Donglu, Chaoyang District, Beijing 100029, China

[2] Department of Pain Medicine, Beijing Hospital of Traditional Chinese Medicine, Capital Medical University, 23 Meishuguan Houjie, Dongcheng District, Beijing 100010, China

K. Du and A. Li contributed equally to this work and should be considered co-first authors.

*** Correspondence:**

Corresponding Author: Dr. Shu-Ming Li lishuming@bjzhongyi.com

Abstract

Background: Personalized education is crucial for effective knee osteoarthritis (OA) management, but providing it remains challenging due to imbalanced patient-provider ratio and limited resources. This study explores the potential of GPT-4, a large language model, in generating tailored self-management guidance and compares its performance with physician-provided advice.

Objective This study aims to evaluate the effectiveness of GPT-4 in generating personalized

education materials for patients with knee osteoarthritis (OA) and compare it with experienced clinicians. Specifically, the comparison is made in terms of efficiency, readability, accuracy, personalization, comprehensiveness, and safety. By leveraging patient data from previous trials, it is evaluated whether AI can improve the quality and accuracy of patient education and evaluate its potential in improving patient care and outcomes.

Methods: A two-phase, blinded, observational study was conducted using patient data from a previous trial. In phase one, two experienced orthopedic specialists created personalized education materials. In phase two, the same data were input into GPT-4 by a physician to generate content. Materials were evaluated for efficiency (words per minute), readability (Flesch-Kincaid Grade Level, Gunning Fog Index, Coleman-Liau Index, and SMOG Index), accuracy, personality, comprehensiveness, and safety.

Results: GPT-4 demonstrated higher efficiency than clinicians (median 530.03 vs. 37.29 WPM, P < 0.001). GPT-4 content exhibited superior readability on the Flesch-Kincaid grade level, Gunning Fog Index, and SMOG Index (P < 0.001). Expert evaluations revealed that GPT-4 outperformed clinicians in accuracy (5.307 ± 1.731 vs. 4.76 ± 1.098, $P = 0.047$), personality (54.32 ± 6.212 vs. 33.2 ± 5.395, $P < 0.001$), comprehensiveness (51.74 ± 6.471 vs. 35.26 ± 6.657, $P < 0.001$), and safety (median 61 vs. 50, $P < 0.001$).

Conclusion: GPT-4 shows promise in generating high-quality, personalized patient education for knee OA, surpassing human experts. This study provides novel evidence for the potential of AI in enabling precise and intelligent patient education. Further research is needed to validate the findings in larger populations and assess the impact on patient outcomes.

Keywords: Artificial Intelligence in Healthcare, Large Language Models, Knee Osteoarthritis, Self-Management, Personalized Medicine, Patient Education.

1 Introduction

Knee Osteoarthritis (OA) is a prevalent chronic degenerative joint disorder that imposes a substantial burden on patients' quality of life and overall well-being [1]. Epidemiological studies have revealed that knee OA affects hundreds of millions of individuals worldwide, with its prevalence increasing with age [2]. In addition to age, other major risk factors for knee OA include obesity, history of joint injury or surgery, and excessive joint usage due to occupational or athletic activities [3]. Knee OA causes pain and functional limitations for individual patients. Moreover, it places a heavy burden on

society and the economy, leading to decreased work productivity and increased healthcare costs [4].

Although knee OA remains incurable, appropriate treatment and management can effectively alleviate symptoms, slow disease progression, and improve prognosis. Patient education and self-management guidance play a pivotal role in knee OA management[5,6]. By providing patients with comprehensive education and personalized guidance on disease knowledge, risk factors, treatment options, and self-management strategies, healthcare providers can enhance patients' understanding of their condition, strengthen their self-efficacy, and ultimately improve treatment adherence and health outcomes[7,8].

However, in real-world clinical practice, the imbalance in patient-provider ratio and limited time and resources often constrain the breadth and depth of patient education. Physicians frequently struggle to offer adequate, individualized guidance to each patient, leading to inconsistent education quality [9]. This issue is particularly prominent in knee OA management, as the disease course is lengthy, and the patient population is highly heterogeneous[10]. Knee OA patients vary considerably in terms of disease severity, symptom presentation, comorbidities, lifestyle factors, and health literacy levels, necessitating more personalized guidance[11]. Yet, in current clinical practice, the education and guidance received by knee OA patients often remain superficial and one-size-fits-all[12]. This highlights the pressing need for innovative patient education models that can enhance the efficiency and quality of guidance while reducing the burden on clinicians.

The advancement of artificial intelligence (AI) technologies, particularly the breakthroughs in natural language processing, offers a potential solution to the challenges. In recent years, large language models (LLMs) based on transformer architectures, such as ChatGPT, Gemini, and Claude, have emerged as promising tools in the medical domain [13]. These LLMs can acquire human-like language understanding and generation capabilities through training on vast amounts of text data, demonstrating remarkable performance in tasks such as medical literature analysis, clinical decision support, information extraction, disease monitoring, and diagnosis, ultimately enhancing efficiency and accuracy in healthcare processes[14–20]. The powerful information extraction, knowledge integration, and content generation abilities of LLMs hold promise for assisting physicians in delivering personalized patient education, thereby improving education quality and efficiency while reducing clinicians' workload[21,22].

Despite the promising potential of LLMs in healthcare, their application in guiding self-management for knee OA patients remains largely unexplored[23–25]. Although LLMs have demonstrated impressive capabilities in generating medical content, most studies to date have primarily focused on general applications in healthcare, such as improving medical documentation or supporting clinical decision-

making, rather than on personalized patient education for specific conditions like knee OA[18,19]. Given the unique challenges knee OA patients face in managing their condition, targeted research is urgently needed to explore how LLMs can be applied effectively in this area, identifying both their potential benefits and limitations.

This study seeks to explore the potential of LLMs, specifically GPT-4, in generating personalized self-management guidance for knee OA patients, and to compare its effectiveness with that of physician-provided guidance. By assessing key factors (efficiency, readability, accuracy, comprehensiveness, personality, and safety), this research aims to provide a novel approach to enhancing patient education in knee OA. The study's findings are expected to shed light on the feasibility of integrating AI-based models into clinical practice, potentially paving the way for more tailored and efficient patient education solutions across various chronic diseases.

2 Methods

2.1 Ethical Considerations

Participants in this study were sourced from our earlier research project, "Acupotomy for Knee Osteoarthritis: A Randomized Controlled Trial," conducted at Beijing Hospital of Traditional Chinese Medicine, Capital Medical University[26]. The original trial was registered with the Chinese Clinical Trial Registry (ChiCTR2100043005) on February 4, 2021, and received ethical approval from the hospital's Medical Ethical Review Committee (No. 2020BL02-057-02). All participants provided written informed consent, which included a provision allowing for the secondary use of their data in future research. To protect participants' privacy, all personal identifiers had been removed from the dataset prior to our analysis. All study procedures were conducted in accordance with the Declaration of Helsinki.

2.2 Study design

This retrospective study, conducted at the Pain Department of Beijing Hospital of Traditional Chinese Medicine (affiliated with Capital Medical University), employs a two-phase, blinded, observational approach to compare the effectiveness of LLMs, specifically OpenAI's ChatGPT-4, in generating personalized patient education content for knee OA patients against content manually created by clinicians. The study utilized anonymized patient data from a previous trial, which included detailed information about patients' condition, symptoms, and medical history. This study's flow is shown in Figure 1.

In the first phase, from August 1 to August 15, 2024, Doctor 1 (R. Guo) and Doctor 2 (S. M. Li),

both orthopedic specialists with a minimum of ten years' experience in knee OA management, generated personalized patient education materials. These specialists were selected based on their comparable educational background, clinical experience, and familiarity with the latest knee OA management guidelines to minimize individual variability. The anonymized patient data from the previous trial was randomly divided into 2 sets, each containing 25 unique patient profiles. These datasets were then randomly assigned to the 2 specialists, ensuring that each specialist received 25 patient profiles to work with. The specialists were provided with comprehensive information about patients' current condition, symptomatology, medical history, and knee OA-relevant lifestyle factors based on the assigned patient profiles. Each specialist was given 15 days to create personalized education content for their assigned patients. Content creation occurred within this standardized timeframe to ensure consistency across all manually generated materials, which served as the control group for subsequent analysis.

In the second phase, from August 16 to August 20, 2024, the same 50 anonymized patient datasets used in the first phase were input into GPT-4, an advanced LLM. Doctor 3 (A. Li) was responsible for assisting with the data input, ensuring consistency in the interaction with GPT-4. To maintain consistency and minimize potential bias, all GPT-4 interactions were conducted by a single researcher within a standardized timeframe, using default parameter settings. Each patient dataset was used to initiate a new GPT-4 conversation session, ensuring the independence of responses. The researcher began each session with a standardized set of instructions: 'You are an orthopedic surgeon with extensive experience treating osteoarthritis of the knee.' This was followed by the prompt: '1. Analyze the patient's condition carefully, step by step. 2. Make full use of all the data provided to create educational content and self-management guidance for patients to help them better understand and manage their health. 3. Avoid including specific patient details in the response.

To mitigate bias, both manually created and LLM-generated content underwent rigorous anonymization and randomization by Doctor 4 (Q. H. Zuo) prior to evaluation. A separate cohort of 2 clinical experts, Doctor 5 (W. S. Du) and Doctor 6 (P. Chen), each with a minimum of five years' experience in knee OA management and uninvolved in the initial content creation, assessed all educational materials. These evaluators were blinded to the content origin, ensuring objective assessment. The study encompassed 100 pieces of educational content (50 GPT-4-generated, 50 human expert-generated), corresponding to 50 unique patient profiles.

2.3 Outcomes

2.3.1 Objective Efficiency

Objective efficiency was evaluated to compare the productivity of GPT-4 and human experts in generating patient education content. This metric was quantified by calculating the words per minute (WPM) produced during the content creation process. Both GPT-4 and the human experts were timed during their respective content generation tasks. The total word count of the produced content was divided by the time taken (in minutes) to generate the WPM figure. This measure provides an objective assessment of the speed and efficiency of content production, allowing for a direct comparison between AI-generated and human-expert-created materials. However, it should be noted that WPM only reflects the speed of content generation and does not account for the quality or accuracy of the produced content.

2.3.2 Readability Evaluation

Readability is a crucial factor in determining the effectiveness of patient education materials, as it directly impacts patients' ability to understand and engage with the provided information. The readability of patient education materials was assessed using the Readable tool (https://app.readable.com/). This instrument employs four established readability metrics: the Flesch-Kincaid Grade Level, Gunning Fog Index, Coleman-Liau Index, and SMOG Index.

The Flesch-Kincaid Grade Level is a widely used readability formula that assesses the approximate reading grade level of a text, based on average sentence length and word complexity. It provides a score that corresponds to a U.S. grade level, indicating the level of education needed to understand the text. Lower scores suggest that the text is easier to read and comprehend.

The Gunning Fog Index is another readability formula that estimates the years of formal education needed to understand a piece of writing on the first reading. It considers the average sentence length and the number of complex words (defined as those with three or more syllables). A lower Gunning Fog score indicates that the text is more accessible to a wider audience.

The Coleman-Liau Index is a readability measure that relies on characters instead of syllables per word and sentence length. It calculates the approximate U.S. grade level needed to comprehend a text, with lower scores indicating greater ease of understanding.

SMOG, which stands for 'Simple Measure of Gobbledygook,' is a readability formula that estimates the years of education needed to fully understand a piece of writing. It is particularly well-suited for evaluating health-related materials, as it was developed using a sample of 30-sentence health education materials. Lower SMOG scores suggest that the text is more accessible to readers with less

education.

This comparative analysis aims to evaluate the relative accessibility of patient education materials from different sources, as higher readability may lead to better patient understanding, engagement, and ultimately, more effective knee OA management.

2.3.3 Accuracy Evaluation

A weighted consensus approach was employed to evaluate the accuracy of educational content on KOA treatment, ensuring its alignment with the most current evidence-based recommendations. This methodology synthesizes recommendations from prominent professional societies, including the American College of Rheumatology (ACR), Osteoarthritis Research Society International (OARSI), European Society for Clinical and Economic Aspects of Osteoporosis, Osteoarthritis and Musculoskeletal Diseases (ESCEO), American Academy of Orthopaedic Surgeons (AAOS), and National Institute for Health and Care Excellence (NICE).

Treatment recommendations were categorized and assigned scores based on their endorsement levels: strongly recommended (+2), conditionally recommended (+1), inconclusive (0), conditionally recommended against (-1), and strongly recommended against (-2). This categorization and scoring system allow for a nuanced assessment of the strength of each recommendation. Weights were assigned to each recommendation to reflect the level of agreement among the societies. High consensus (agreement among ≥4 societies) received a weight of 1.0, moderate consensus (3 societies) 0.75, low consensus (2 societies) 0.5, and minimal consensus (1 society) 0.25. The weighted score for each treatment recommendation was calculated by multiplying the recommendation score by the consensus weight, thereby giving more importance to recommendations with higher levels of agreement among societies.

A comprehensive table was constructed to summarize the consensus levels, weights, and weighted scores for various knee OA treatments. The overall accuracy score for the educational content was determined by summing the weighted scores for all treatments mentioned. Higher scores indicated better alignment with consensus guidelines, while lower or negative scores highlighted areas for potential revision. This approach provides a rigorous and objective methodology for evaluating the accuracy of educational content, enhancing its reliability and clinical relevance. A detailed description of the methodology and calculations can be found in Supplement 1.

2.3.4 Personality Evaluation

Personalized content is crucial for effective patient education and engagement in knee OA management. A systematic scoring method was utilized to evaluate the degree of personality in the

communication content for knee OA treatment. This approach ensured that the content was tailored to the patient's specific characteristics, symptoms, disease stage, lifestyle habits, and medical history. The personality scoring system included three key dimensions: relevance (30 points), individual relevance (40 points), and detail level (30 points).

The relevance dimension, scored from 0 to 30 points, focused on how well the content addressed the patient's symptoms and disease stage. The individual relevance dimension, scored from 0 to 40 points, evaluated the customization of the content to the patient's personal characteristics, lifestyle habits, and medical history. The detail level, scored from 0 to 30 points, assessed the depth of information provided and the practicality of the suggestions.

Each of these dimensions was scored according to specific criteria, and the total score for personality was calculated out of 100 points. This comprehensive and objective evaluation method ensures that the content is tailored to the unique needs of each patient, potentially enhancing treatment adherence and outcomes. A detailed description of the personality scoring system can be found in Supplement 2.

2.3.5 Comprehensiveness Evaluation

To ensure the comprehensiveness of patient education content generated by both clinicians and LLMs, we developed a systematic evaluation method (Supplement 3). This approach assesses whether the content adequately covers all essential topics related to patient care, providing crucial information for effective disease management. The evaluation focuses on five key categories: medication treatment, non-medication treatment, lifestyle advice, psychological support, and disease management. Each category is scored on a scale from 0 to 20 points, with specific criteria for each score range (e.g., 17-20 points for complete coverage, 13-16 points for mostly covered). The total possible score is 100 points. This scoring system guarantees that all critical aspects of patient care are thoroughly addressed, offering a well-rounded and informative framework for patient education. By ensuring the comprehensiveness of the content, this method serves as a reliable tool for supporting patient care and improving health outcomes.

2.3.6 Safety Evaluation

To ensure that the patient education content generated by both human experts and LLMs is safe, we implemented a systematic safety evaluation method (Supplement 4). This method assesses the content across five key domains: medication treatment, non-medication treatment, lifestyle advice, psychological support, and disease management. Each domain is evaluated based on the potential risks or harm associated with the recommendations, such as drug interactions, contraindications, or

unsuitable lifestyle advice. A structured scoring system is used, where each aspect is rated from 0 to 20 points (e.g., 16-20 points for very safe, 0 points for unsafe). The total safety score for each piece of content can reach a maximum of 100 points. This comprehensive method ensures that all recommendations are safe and free from significant risks, providing a reliable framework for assessing the quality and safety of patient education materials. By prioritizing patient safety, this evaluation method contributes to the development of high-quality educational content that supports optimal patient care.

2.4 Statistical Analysis

Statistical analyses were conducted by Doctor 7 (C. Y. Zhang) using SPSS software (version 27.0; IBM Corp., Armonk, NY, USA). Continuous variables with normal distribution were presented as means and standard deviations (Mean ± SD); non-normal variables were reported as medians and interquartile ranges (Median [IQR]). The means of two continuous normally distributed variables were compared by paired samples t-test, while the medians of two continuous non-normally distributed variables were compared by Wilcoxon signed-rank test. A value of $P < 0.05$ was considered significant.

3 Results

3.1 Objective Efficiency Results

Objective efficiency, assessed by WPM, revealed a disparity in content generation speed between clinicians and GPT-4 (Figure 2). Clinicians produced a median of 37.29 WPM (IQR: 3.94), whereas GPT-4 achieved a median output of 530.03 WPM (IQR: 127.03), approximately 14 times faster than clinicians. This difference in efficiency was statistically significant ($P < 0.001$).

3.2 Readability Evaluation Results

Four validated readability metrics were used to evaluate patient education materials on knee OA management generated by clinicians and GPT-4. Significant differences were observed across all metrics (Figure 3). GPT-4-generated content exhibited superior readability on the Flesch-Kincaid grade level (11.56 ± 1.08 for GPT-4 vs. 12.67 ± 0.95 for clinicians), which assesses the ease of understanding based on word and sentence length, and the Gunning Fog Index (12.47 ± 1.36 for GPT-4 vs. 14.56 ± 0.93 for clinicians), which estimates the years of formal education needed to comprehend the text. Furthermore, GPT-4 also performed well on the SMOG Index (13.33 ± 1.00 for GPT-4 vs. 13.81 ± 0.69 for clinicians), which measures the years of education required to understand

the material. Clinician-generated materials showed a slight advantage on the Coleman-Liau Index (15.15 ± 0.91 for clinicians vs. 15.90 ± 1.03 for GPT-4), which considers the number of characters per word and words per sentence. These findings underscore the nuanced strengths of both AI and human-generated materials, with GPT-4 enhancing overall accessibility and quality in patient education.

3.3 Expert Evaluation Results

GPT-4 demonstrated favorable results across all four evaluation dimensions based on expert assessments (Figure 4). For accuracy, clinician-generated materials achieved a mean score of 4.76 ± 1.10 (79.3%), while GPT-4 scored significantly higher at 5.31 ± 1.73 (88.5%) (P = 0.047). In the dimension of personality, GPT-4-generated materials demonstrated a mean score of 54.32 ± 6.21 out of 100, in stark contrast to the clinician-generated materials, which scored only 33.20 ± 5.40 out of 100 (P < 0.001). Regarding comprehensiveness, GPT-4 again outperformed clinicians, with scores of 51.74 ± 6.47 compared to 35.26 ± 6.66 out of 100 (P < 0.001). Lastly, in terms of safety, GPT-4-generated materials received a median score of 61.00 (IQR: 8.00), while clinician-generated content had a median score of 50.00 (IQR: 8.25) (P < 0.001).

4 Discussion

4.1 Main Findings

Our study reveals that GPT-4 outperformed clinicians in several key aspects of generating personalized self-management guidance for knee OA patients. In terms of efficiency, GPT-4 demonstrated a clear advantage, producing content significantly faster, which highlights its potential to alleviate clinician workload. Regarding readability, GPT-4 generally created more accessible content, though some metrics indicated a slightly higher complexity, suggesting the need for refinement to ensure all patients can easily comprehend the material. On accuracy, GPT-4 performed better, aligning more closely with current medical guidelines, ensuring that the advice provided is medically reliable. In terms of comprehensiveness, GPT-4 covered a broader range of topics, offering more detailed and thorough information than clinician-generated content. Additionally, in personality, GPT-4 excelled by tailoring guidance more effectively to individual patient profiles, considering specific symptoms and lifestyle factors. Finally, GPT-4's content was rated as safer, consistently adhering to clinical safety standards. These results indicate that GPT-4 has strong potential to improve patient education, particularly in efficiency, accuracy, comprehensiveness, and personality, though readability improvements may be necessary to optimize patient understanding.

4.2 Comparison to Prior Work

The application of LLMs in patient education has evolved through three distinct phases, reflecting increasing sophistication and clinical relevance. In the first phase, researchers primarily focused on evaluating the performance of LLMs in answering pre-set medical questions, often related to specific conditions. Studies like those on urolithiasis and Mohs surgery explored the accuracy, readability, and completeness of LLM-generated responses without direct comparison to human-created content, aiming to assess the baseline capabilities of different models [27,28]. While these studies provided valuable insights into the capabilities of LLMs, they lacked direct comparisons to human-generated content, which is essential for understanding their potential in real-world clinical settings.

In the second phase, LLM-generated outputs were compared with human-generated educational materials or institutional resources. Studies benchmarking LLM responses against traditional patient education materials, such as those related to bariatric surgery and body contouring, highlighted LLMs' ability to produce content with similar or even superior readability and accuracy [29,30]. These studies marked a significant step forward in validating the use of LLMs for patient education. However, the lack of personalization in the generated content limited their applicability to individual patient care.

Addressing this limitation, the third phase, represented by our research, has pioneered the generation of personalized patient education. By integrating patient-specific factors like medical history and symptomatology, our study on knee osteoarthritis demonstrated that GPT-4 can create tailored educational content that not only matches but often exceeds the quality of clinician-generated materials across dimensions such as accuracy, personalization, and safety. This shift towards highly customized patient education reflects healthcare's increasing emphasis on individualized care, especially in managing chronic conditions. With LLMs demonstrating the potential to produce high-quality, personalized educational content, they are poised to enhance patient engagement, self-management, and health outcomes, ultimately paving the way for broader clinical integration and transforming the delivery of patient education[31,32].

4.3 Strengths and Limitations

This study offers several strengths that distinguish it from previous research. Firstly, it leverages the advanced capabilities of GPT-4 to create highly personalized patient education content specifically for knee osteoarthritis, addressing a significant gap in tailored patient resources. GPT-4's unique ability to understand and analyze patients' medical history, symptoms, and lifestyle factors enables it to generate customized educational materials that cater to individual patient needs. Additionally, the direct comparison of AI-generated materials with those created by human experts using blinded

assessments enhances the rigor and credibility of the findings. By employing validated evaluation metrics, the study not only provides a robust assessment of content quality but also sets a new benchmark for future investigations into LLM applications in patient education. Finally, the focus on a common yet impactful condition highlights the practical relevance of the research, facilitating its potential adoption in clinical settings.

However, the small sample size may limit generalizability, necessitating validation in larger, diverse populations. The lack of assessment of patients' subjective perceptions and satisfaction hinders understanding of the impact on engagement and adherence. The study did not evaluate long-term effects on patient outcomes, which requires longitudinal investigations. Focusing on a single AI model limits conclusions about other technologies, and comparative studies are needed to identify the most effective approaches. Lastly, the study did not investigate challenges and barriers to integrating AI-generated content into clinical workflows and its acceptability among healthcare professionals.

4.4 Future Directions

Future research must rigorously validate the effects of AI-generated content on patient outcomes across diverse populations. Large-scale studies should evaluate both short- and long-term impacts of AI-driven education on patient knowledge, self-management, treatment adherence, patient-reported outcomes (PROs), and clinical outcomes[33]. Longitudinal investigations are essential to assess the sustainability and lasting efficacy of AI-assisted interventions.

Comparative studies should also focus on evaluating the performance of various AI models, ensuring they are customized for specific medical domains, rigorously validated in clinical settings, and remain aligned with current clinical standards. Continuous updates and assessments are necessary to address evolving patient needs. Additionally, it is crucial to address ethical concerns, such as ensuring transparency, preventing automation bias, avoiding the dissemination of inappropriate or inaccurate content, and protecting patient privacy and data security[34].

Further, the impact of AI-assisted tools on clinical decision-making requires careful exploration. Research should investigate how these technologies influence clinician judgment and whether they enhance or inappropriately replace human decision-making[35]. Developing frameworks that position AI as a supportive tool, rather than a substitute for clinical expertise, is critical to maintaining high standards of care.

5 Conclusion

This study demonstrates the potential of GPT-4 in generating personalized patient education content

for knee OA. The AI-generated content outperformed human-expert-created materials in objective efficiency, accuracy, comprehensiveness, personality, and safety. However, clinician-generated content showed a slight advantage in readability on the Coleman-Liau Index. These findings suggest that AI-powered tools have the potential to revolutionize the delivery of tailored health information for chronic conditions. Further research is needed to validate these results in larger populations and real-world clinical settings.

6 Conflict of Interest

A statement was made by the authors that this study was conducted without any business or financial relationships that could be perceived as a potential conflict of interest.

7 Funding

This study was supported by Capital's Funds for Health Improvement and Research (No.2020-2-2231).

8 Acknowledgments

We thank all the anonymous patients whose data made this study possible. Figure 1 was created with BioRender.com. We also acknowledge OpenAI for providing the GPT service used in this research.

9 Captions

Figure 1: Flowchart of overall study design.
Figure 2: Comparison of Objective Efficiency
Violin plot showing the distribution of WPM for content generated by doctors and GPT-4, with box plots indicating the median, interquartile range, and outliers.
Figure 3: Comparison of Readability Metrics
Violin plots comparing readability metrics (Flesch-Kincaid, Gunning Fog, Coleman-Liau, SMOG) between doctor- and GPT-4-generated content, with mean and standard deviation indicated by error bars.
Figure 4: Comparison of Accuracy, Personality, Comprehensiveness and Safety
Violin plots comparing accuracy, personality, and comprehensiveness, with mean and standard deviation shown. Safety uses a box plot to show median, interquartile range, and outliers.

10 References

1. Katz JN, Arant KR, Loeser RF. Diagnosis and treatment of hip and knee osteoarthritis: A review. *JAMA*. 2021;325(6):568-578. doi:10.1001/jama.2020.22171
2. Deyle GD, Allen CS, Allison SC, et al. Physical Therapy versus Glucocorticoid Injection for Osteoarthritis of the Knee. *N Engl J Med*. 2020;382(15):1420-1429. doi:10.1056/NEJMoa1905877

3. Perruccio AV, Young JJ, Wilfong JM, Denise Power J, Canizares M, Badley EM. Osteoarthritis year in review 2023: Epidemiology & therapy. *Osteoarthritis Cartilage*. 2024;32(2):159-165. doi:10.1016/j.joca.2023.11.012
4. Steinmetz JD, Culbreth GT, Haile LM, et al. Global, regional, and national burden of osteoarthritis, 1990–2020 and projections to 2050: a systematic analysis for the Global Burden of Disease Study 2021. *The Lancet Rheumatology*. 2023;5(9):e508-e522. doi:10.1016/S2665-9913(23)00163-7
5. Goff AJ, De Oliveira Silva D, Merolli M, Bell EC, Crossley KM, Barton CJ. Patient education improves pain and function in people with knee osteoarthritis with better effects when combined with exercise therapy: a systematic review. *Journal of Physiotherapy*. 2021;67(3):177-189. doi:10.1016/j.jphys.2021.06.011
6. Duong V, Oo WM, Ding C, Culvenor AG, Hunter DJ. Evaluation and Treatment of Knee Pain: A Review. *JAMA*. 2023;330(16):1568-1580. doi:10.1001/jama.2023.19675
7. Sharma L. Osteoarthritis of the Knee. *N Engl J Med*. 2021;384(1):51-59. doi:10.1056/NEJMcp1903768
8. Moseng T, Vliet Vlieland TPM, Battista S, et al. EULAR recommendations for the non-pharmacological core management of hip and knee osteoarthritis: 2023 update. *Ann Rheum Dis*. 2024;83(6):730-740. doi:10.1136/ard-2023-225041
9. Oosterhaven J, Pell CD, Schröder CD, et al. Health literacy and pain neuroscience education in an interdisciplinary pain management programme: a qualitative study of patient perspectives. *Pain Rep*. 2023;8(6):e1093. doi:10.1097/PR9.0000000000001093
10. Martel-Pelletier J, Barr AJ, Cicuttini FM, et al. Osteoarthritis. *Nat Rev Dis Primers*. 2016;2:16072. doi:10.1038/nrdp.2016.72
11. Hunter DJ, Bierma-Zeinstra S. Osteoarthritis. *The Lancet*. 2019;393(10182):1745-1759. doi:10.1016/S0140-6736(19)30417-9
12. Salazar-Méndez J, Cuyul-Vásquez I, Ponce-Fuentes F, et al. Pain neuroscience education for patients with chronic pain: A scoping review from teaching-learning strategies, educational level, and cultural perspective. *Patient Educ Couns*. 2024;123:108201. doi:10.1016/j.pec.2024.108201
13. Moglia A, Georgiou K, Cerveri P, Mainardi L, Satava RM, Cuschieri A. Large language models in healthcare: from a systematic review on medical examinations to a comparative analysis on fundamentals of robotic surgery online test. *Artif Intell Rev*. 2024;57(9):231. doi:10.1007/s10462-024-10849-5
14. Chen X, Xie H, Tao X, Wang FL, Leng M, Lei B. Artificial intelligence and multimodal data fusion for smart healthcare: topic modeling and bibliometrics. *Artif Intell Rev*. 2024;57(4):91. doi:10.1007/s10462-024-10712-7
15. Jayakumar P, Moore MG, Furlough KA, et al. Comparison of an Artificial Intelligence-Enabled Patient Decision Aid vs Educational Material on Decision Quality, Shared Decision-Making, Patient Experience, and Functional Outcomes in Adults With Knee Osteoarthritis: A Randomized Clinical Trial. *JAMA Netw Open*. 2021;4(2):e2037107. doi:10.1001/jamanetworkopen.2020.37107
16. Akyon SH, Akyon FC, Camyar AS, Hızlı F, Sari T, Hızlı Ş. Evaluating the Capabilities of Generative AI Tools in Understanding Medical Papers: Qualitative Study. *JMIR Med Inform*. 2024;12:e59258. doi:10.2196/59258
17. Plagwitz L, Neuhaus P, Yildirim K, Losch N, Varghese J, Büscher A. Zero-Shot LLMs for Named Entity Recognition: Targeting Cardiac Function Indicators in German Clinical Texts. *Stud Health Technol Inform*. 2024;317:228-234. doi:10.3233/SHTI240861
18. Sun VH, Heemelaar JC, Hadzic I, et al. Enhancing Precision in Detecting Severe Immune-Related Adverse Events: Comparative Analysis of Large Language Models and International Classification of Disease Codes in Patient Records. *J Clin Oncol*. Published online September 3, 2024:JCO2400326. doi:10.1200/JCO.24.00326

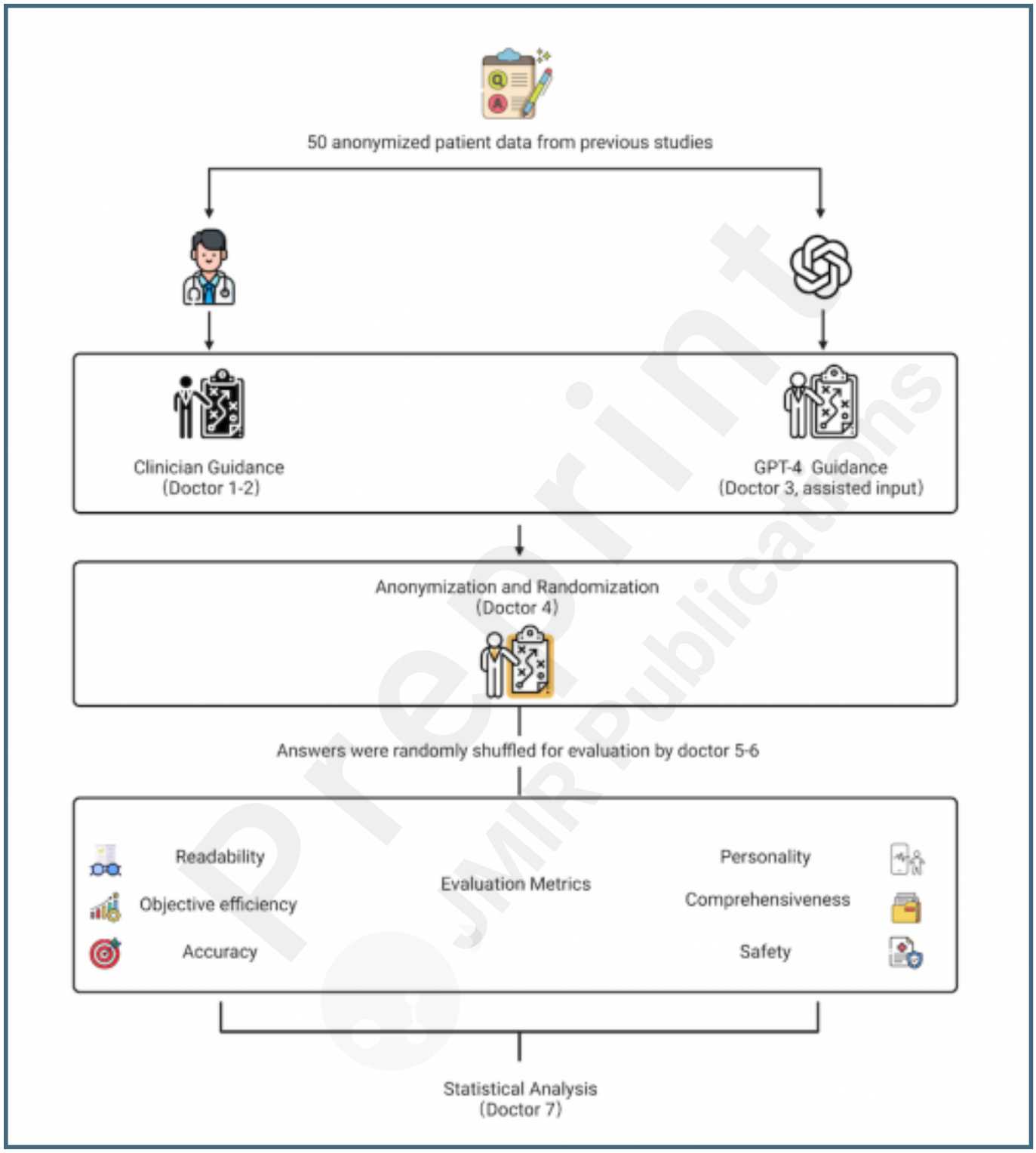
19. Liu J, Shen H, Chen K, Li X. Large language model produces high accurate diagnosis of cancer from end-motif profiles of cell-free DNA. *Brief Bioinform*. 2024;25(5):bbae430. doi:10.1093/bib/bbae430
20. Ruiz Sarrias O, Martínez Del Prado MP, Sala Gonzalez MÁ, et al. Leveraging Large Language Models for Precision Monitoring of Chemotherapy-Induced Toxicities: A Pilot Study with Expert Comparisons and Future Directions. *Cancers (Basel)*. 2024;16(16):2830. doi:10.3390/cancers16162830
21. Yeo YH, Samaan JS, Ng WH, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. *Clin Mol Hepatol*. 2023;29(3):721-732. doi:10.3350/cmh.2023.0089
22. Benary M, Wang XD, Schmidt M, et al. Leveraging Large Language Models for Decision Support in Personalized Oncology. *JAMA Netw Open*. 2023;6(11):e2343689. doi:10.1001/jamanetworkopen.2023.43689
23. Sarraju A, Bruemmer D, Van Iterson E, Cho L, Rodriguez F, Laffin L. Appropriateness of Cardiovascular Disease Prevention Recommendations Obtained From a Popular Online Chat-Based Artificial Intelligence Model. *JAMA*. 2023;329(10):842-844. doi:10.1001/jama.2023.1044
24. Romano MF, Shih LC, Paschalidis IC, Au R, Kolachalama VB. Large Language Models in Neurology Research and Future Practice. *Neurology*. 2023;101(23):1058-1067. doi:10.1212/WNL.0000000000207967
25. Beaulieu-Jones BR, Berrigan MT, Shah S, Marwaha JS, Lai SL, Brat GA. Evaluating capabilities of large language models: Performance of GPT-4 on surgical knowledge assessments. *Surgery*. 2024;175(4):936-942. doi:10.1016/j.surg.2023.12.014
26. Li SM, Li TL, Guo R, et al. Effectiveness and safety of acupotomy for knee osteoarthritis: study protocol for a randomized controlled trial. *Trials*. 2021;22(1):824. doi:10.1186/s13063-021-05786-5
27. Song H, Xia Y, Luo Z, et al. Evaluating the Performance of Different Large Language Models on Health Consultation and Patient Education in Urolithiasis. *J Med Syst*. 2023;47(1):125. doi:10.1007/s10916-023-02021-3
28. Robinson MA, Belzberg M, Thakker S, et al. Assessing the accuracy, usefulness, and readability of artificial-intelligence–generated responses to common dermatologic surgery questions for patient education: A double-blinded comparative study of ChatGPT and Google Bard. *Journal of the American Academy of Dermatology*. 2024;90(5):1078-1080. doi:10.1016/j.jaad.2024.01.037
29. Srinivasan N, Samaan JS, Rajeev ND, Kanu MU, Yeo YH, Samakar K. Large language models and bariatric surgery patient education: a comparative readability analysis of GPT-3.5, GPT-4, Bard, and online institutional resources. *Surg Endosc*. 2024;38(5):2522-2532. doi:10.1007/s00464-024-10720-2
30. Alessandri-Bonetti M, Liu HY, Palmesano M, Nguyen VT, Egro FM. Online patient education in body contouring: A comparison between Google and ChatGPT. *J Plast Reconstr Aesthet Surg*. 2023;87:390-402. doi:10.1016/j.bjps.2023.10.091
31. Wang J, Shi R, Le Q, et al. Evaluating the effectiveness of large language models in patient education for conjunctivitis. *Br J Ophthalmol*. Published online August 30, 2024:bjo-2024-325599. doi:10.1136/bjo-2024-325599
32. Dihan Q, Chauhan MZ, Eleiwa TK, et al. Large language models: a new frontier in paediatric cataract patient education. *Br J Ophthalmol*. 2024;108(10):1470-1476. doi:10.1136/bjo-2024-325252
33. Huang Z, Bianchi F, Yuksekgonul M, Montine TJ, Zou J. A visual-language foundation model for pathology image analysis using medical Twitter. *Nat Med*. 2023;29(9):2307-2316. doi:10.1038/s41591-023-02504-3
34. Li H, Moon JT, Purkayastha S, Celi LA, Trivedi H, Gichoya JW. Ethics of large language models in medicine and medical research. *Lancet Digit Health*. 2023;5(6):e333-e335.

doi:10.1016/S2589-7500(23)00083-3
35. Li R, Kumar A, Chen JH. How Chatbots and Large Language Model Artificial Intelligence Systems Will Reshape Modern Medicine: Fountain of Creativity or Pandora's Box? *JAMA Intern Med*. 2023;183(6):596-597. doi:10.1001/jamainternmed.2023.1835

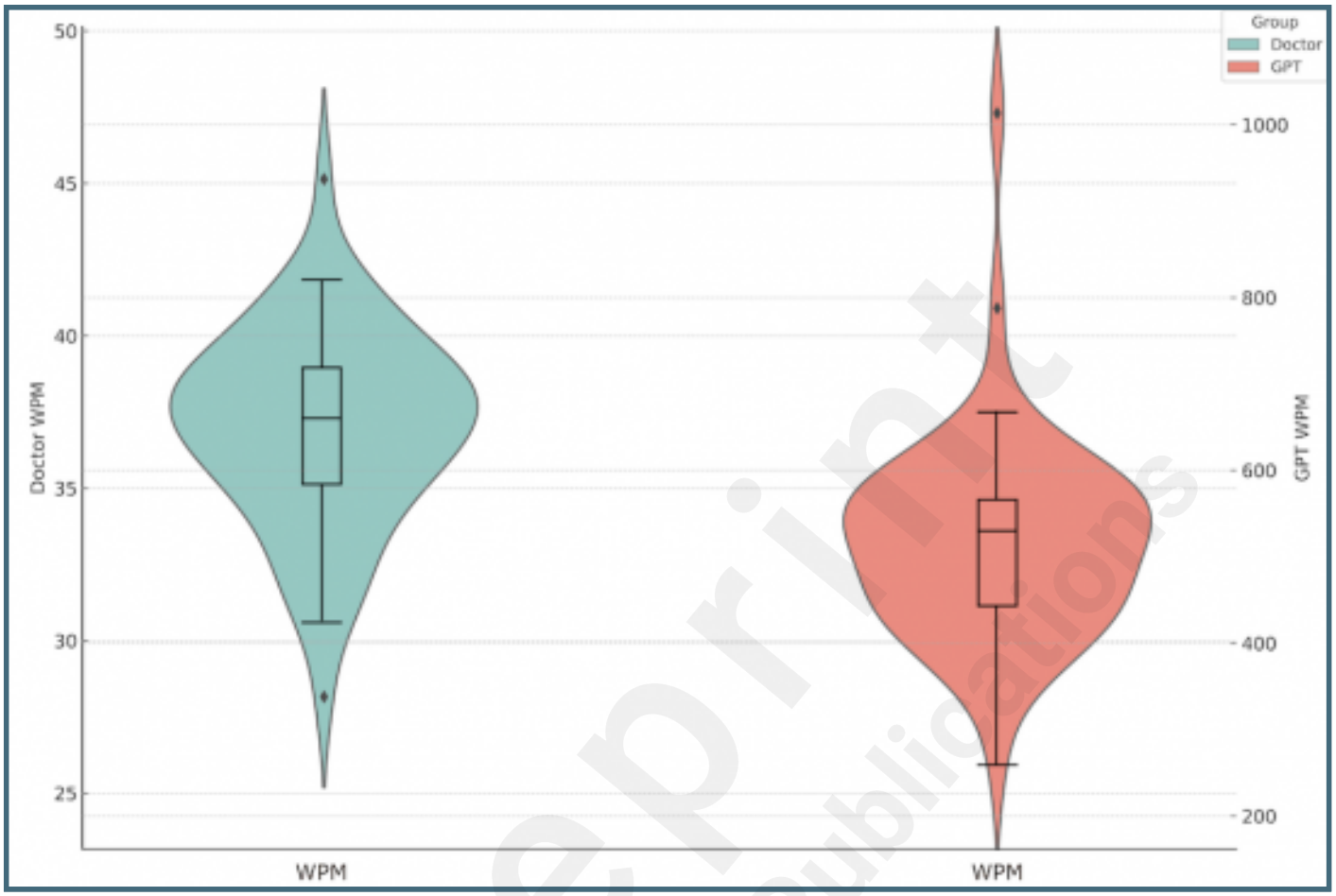
Supplementary Files

Figures

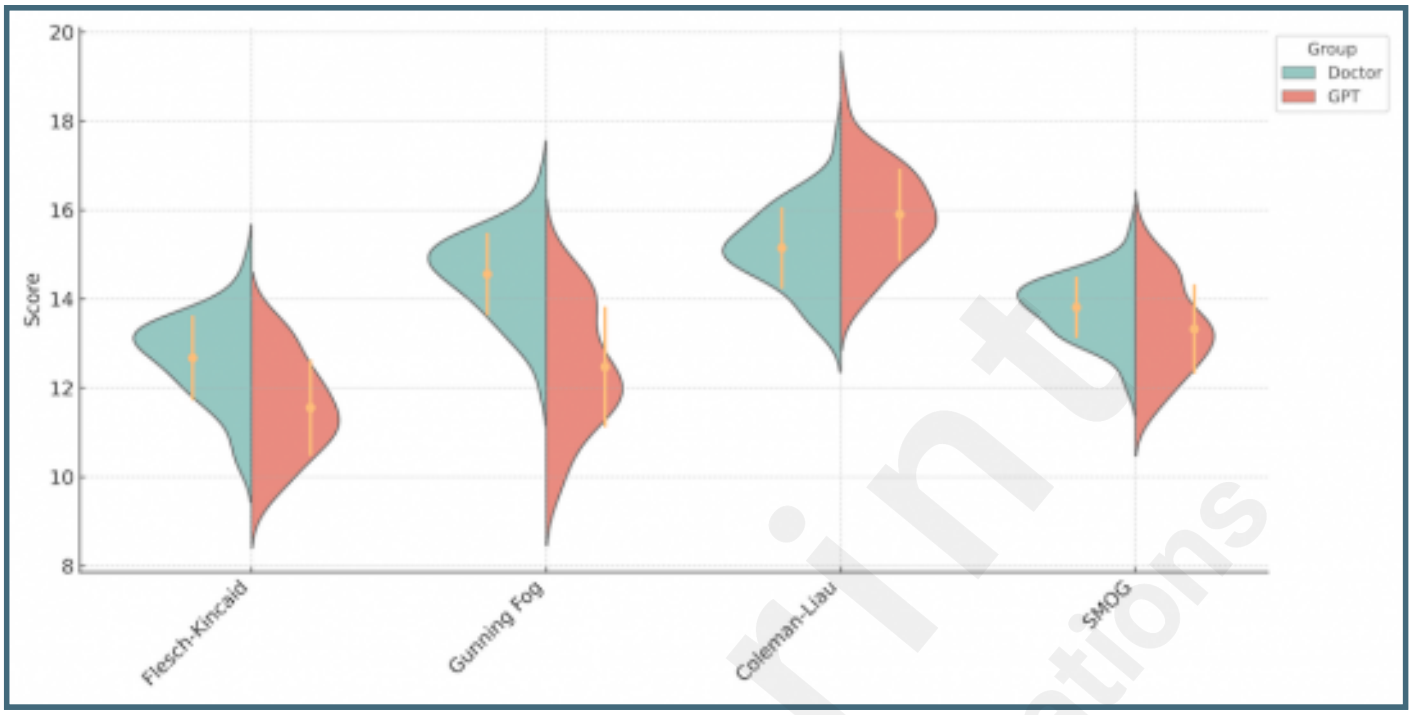
Flowchart of overall study design.



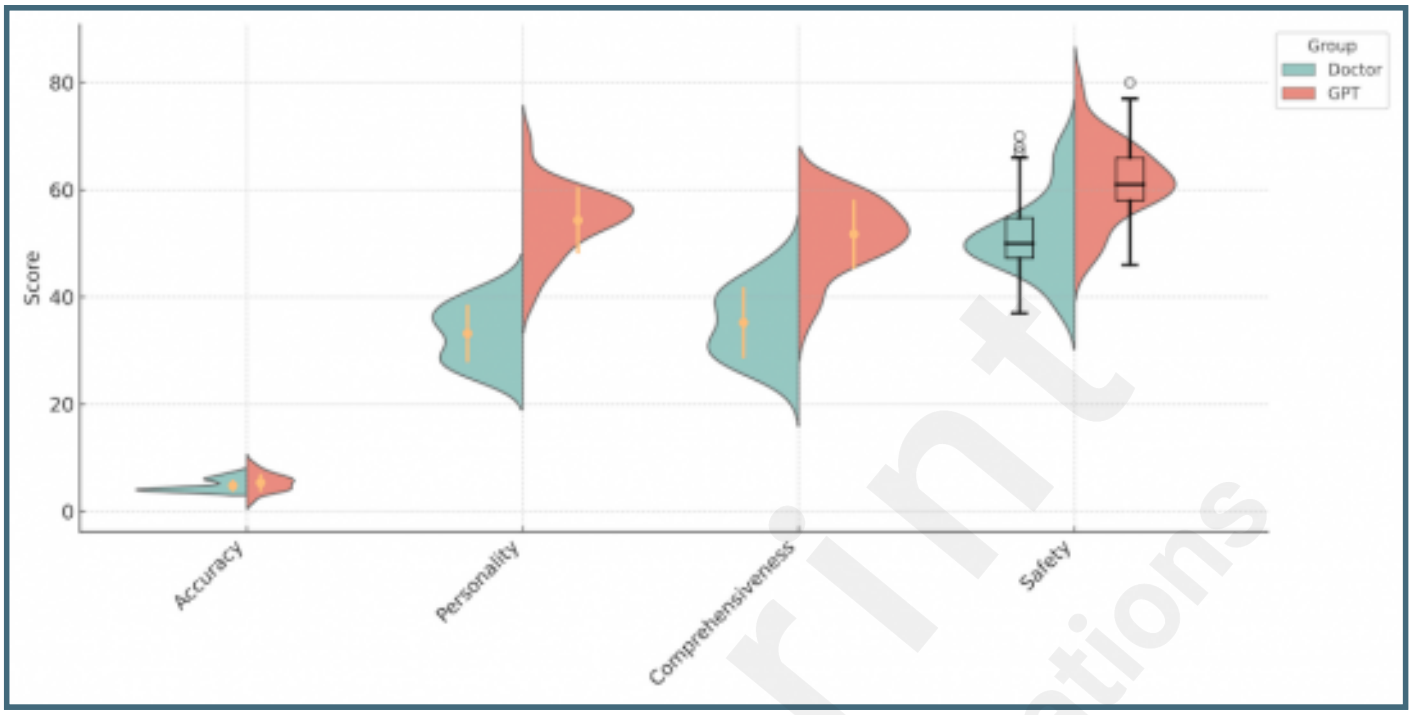
Comparison of Objective Efficiency.



Comparison of Readability Metrics.



Comparison of Accuracy, Personality, Comprehensiveness and Safety.



Related publication(s) - for reviewers eyes onlies

Detailed Calculation Methodology for Weighted Consensus Approach.
URL: http://asset.jmir.pub/assets/5ab17811f8b7be9bf9f23f513959d496.pdf

Personalization Evaluation Method.
URL: http://asset.jmir.pub/assets/b194011f317ff1c712800479f134adb5.pdf

Comprehensiveness Evaluation Method.
URL: http://asset.jmir.pub/assets/58ac5f5adb619e8f733e6f5e0689089d.pdf

Safety Evaluation Method.
URL: http://asset.jmir.pub/assets/1b1c5cc7ca637db76240fec42da7997f.pdf