

Comparison Of The Performance Of Chatgpt-3.5, Chatgpt-4o And Gemini 2.5 In The Management Of Rheumatoid Arthritis: Analysis Of EULAR 2022 Recommendations

Ozlem Kuculmez^{1*}, Ahmet Usen²

¹ Department of Physical Medicine and Rehabilitation, Baskent University Alanya Hospital, Antalya, Turkey

² Department of Physical Medicine and Rehabilitation, University/Institution Name, City, Country
akanozlem07@gmail.com; ahmetusen1@hotmail.com

¹ <https://orcid.org/0000-0000-0000-0000>; ² <https://orcid.org/0000-0000-0000-0000>

*Corresponding Author: akanozlem07@gmail.com

Abstract

Background: Clinical guidelines are essential for diagnosing and managing rheumatic conditions like rheumatoid arthritis, which can result in deformities and impairments. Large language models such as ChatGPT and Gemini have become widely popular due to their fast responses, user-friendly interfaces, and ability to improve through human feedback and reinforcement learning. Prior studies have examined how accurate and reliable these models are when answering common public health questions.

Objective: The aim of this research is to assess and contrast the answers provided by different large language models to open-ended questions based on the 2022 EULAR guideline.

Materials and Methods: This was a comparative, cross-sectional study. Five overarching principles and eleven recommendations from the 2022 EULAR guidelines were rephrased into open-ended clinical questions. These questions were presented to ChatGPT-3.5, ChatGPT-4o, and Gemini 2.5 Flash. The responses were assessed using a 7-point Likert scale for usability and reliability, readability scores (Flesch-Kincaid Reading Ease (FKRE) and Flesch-Kincaid Grade Level (FKGL)), and similarity metrics (Universal Sentence Encoder (USE) and ROUGE-L). Different models were compared statistically and $p < 0.05$ was revealed as statistically significant.

Results: Gemini 2.5 produced responses with a significantly higher FKGL than both ChatGPT-3.5 and ChatGPT-4o ($p = 0.003$ and $p = 0.004$). FKRE was significantly higher for ChatGPT-3.5 than Gemini 2.5 ($p = 0.011$), whereas the difference between ChatGPT-4o and Gemini 2.5 was not significant. Reliability did not differ between models ($p > 0.05$), but ChatGPT-4o had a significantly higher usefulness score than Gemini 2.5 ($p < 0.05$). ChatGPT-4o and Gemini 2.5 showed significantly higher semantic similarity than ChatGPT-3.5 ($p < 0.05$). ROUGE-L F1 scores showed no significant differences ($p > 0.05$). FKRE and FKGL were strongly negatively correlated, while reliability and usefulness were strongly positively correlated ($p < 0.05$).

Conclusions: It was determined that ChatGPT-3.5 produced more readable responses than Gemini 2.5, while ChatGPT-4o demonstrated greater usefulness in answers to open-ended questions derived from the EULAR 2022 guideline; however, no difference in reliability was observed.

Keywords: Arthritis, artificial intelligence, guideline, rheumatic disease, rheumatoid arthritis.

1. INTRODUCTION

Rheumatoid arthritis (RA) is a chronic inflammatory disease that typically starts slowly, affecting small joints in the hands, wrists, feet, and ankles, and sometimes larger joints.^[1] While the clinical diagnosis is often made at a relatively young age, research indicates that the underlying immune processes may start up to a decade before clear symptoms appear.^[2] It is well-established that early detection and proper management of rheumatoid arthritis are essential for enhancing patients' quality of life and preventing

permanent joint damage. [3] As such, adopting effective treatment approaches is of considerable clinical significance.

The European League Against Rheumatism (EULAR) is crucial in shaping the decision-making process for both patients and healthcare providers. Its guidelines provide evidence-based recommendations for the most effective treatment approaches. These guidelines help ensure proper disease management and improve patients' quality of life. EULAR also provides protocols for disease monitoring, regular follow-ups, and treatment options, creating a standardized framework for managing the disease and minimizing negative outcomes. [4, 5]

Chat Generative Pretrained Transformer (ChatGPT) and Gemini are two of the most popular large language models (LLMs), both of which improve their performance through human feedback and reinforcement learning. Their ability to provide quick responses and their user-friendly interfaces have played a significant role in their widespread use and increasing popularity. Previous research has evaluated the accuracy and dependability of LLM-generated answers to frequently asked public health questions. [6] Wang et al. highlighted AI's potential in diagnosing rheumatoid arthritis by processing extensive clinical, lab, and imaging data. They also underlined its usefulness in early detection, disease monitoring, and treatment guidance. [7] Similarly, Benavent et al. noted AI's role in evaluating treatment effectiveness for both rheumatoid arthritis and spondyloarthropathies. [8] Building on these studies, the current research aims to assess the adequacy and accuracy of responses from various LLMs, including ChatGPT-3.5, ChatGPT-4o Plus, and Gemini, to questions related to key aspects of managing rheumatoid arthritis, in accordance with current evidence-based guidelines. Specifically, this study intends to compare these models in terms of usability, reliability, and readability, as well as both semantic and surface-level similarity, using open-ended questions based on the 2022 EULAR guidelines for RA management. Furthermore, the research sought to qualitatively assess how effectively these models provide reliable advice to patients, as well as the thoroughness and precision of the details they present.

2. MATERIALS AND METHODS:

This study was conducted as a cross-sectional, comparative analysis. The XXX University Ethics Committee approved the study (E-10840098-2024.02-450, date: 04.06.2025), and all procedures were carried out in accordance with the principles outlined in the Declaration of Helsinki.

Development of Open-Ended Questions

Five overarching principles and eleven recommendations from the 2022 EULAR guidelines were carefully converted into open-ended clinical questions, following a methodology similar to that used in the study by Liu et al. [9] This task was carried out by two specialists in physical medicine and rehabilitation, each with over ten years of clinical experience in rheumatoid arthritis. Special care was taken to maintain the original clinical intent of each recommendation while formulating the questions. Each question was designed to capture the essence of the corresponding guideline statement, and the expected answers were aligned with the original content of the EULAR recommendations. The questions covered a variety of clinically significant topics, such as treatment approaches, monitoring disease activity, and assessing treatment responses. To aid in the analysis, the questions were grouped into three thematic categories, following the classification method used by Sciberras et al. [10]: (1) general management and monitoring strategies, (2) pharmacological treatment, and (3) treatment response and adaptation strategies. The complete set of open-ended questions based on the 2022 EULAR guidelines provided is in Table 1.

Large Language Models Utilized and Response Collection Procedure

In this study, three large language models were used: ChatGPT-3.5 (OpenAI, America, California), ChatGPT-4o (Plus; OpenAI, California), and Gemini 2.5 Flash (Google LLC, Alphabet Inc., America, California). These models are trained on extensive text datasets, enabling them to generate contextually relevant and coherent responses that closely mimic those of an experienced human respondent.^[11] ChatGPT-3.5 and ChatGPT-4o are recently developed, widely adopted LLMs by OpenAI. Gemini 2.5 Flash is a comparable model created by Google.^[12] Each open-ended question was submitted to the models in separate, newly initiated sessions to eliminate any prior conversational context. A standardized prompt was applied consistently across all models: "Please answer without considering the previous conversation." The responses were recorded in text form and anonymized to ensure an unbiased evaluation during the subsequent analysis.

Evaluation Process of Responses of Large Language Models

The responses produced by each large language model for the open-ended questions were independently reviewed by two experienced psychiatrists (AU and OK) using a 7-point Likert scale developed by Uz et al.^[13], which evaluated both the reliability and practical usefulness of each answer. In this evaluation system, a score of 1 signifies complete unreliability, 2 indicates very low reliability, 3 represents moderate reliability, 4 corresponds to reliable, 5 denotes fairly high reliability, 6 indicates high reliability, and 7 signifies absolute reliability. Additionally, in the usefulness evaluation, a score of 1 indicated not useful at all, 2 signified very limited usefulness, 3 represented somewhat useful, 4 and 5 corresponded to moderately useful, 6 denoted highly useful, and 7 indicated extremely useful. In instances where the two evaluators assigned different scores, efforts were made to reach a consensus. If consensus could not be achieved, a third expert psychiatrist (AA), also specialized in the field, acted as an adjudicator.

To assess semantic similarity between the LLM-generated responses and the original ASAS-EULAR 2022 guideline statements, the Universal Sentence Encoder (USE) was employed. This Transformer-based model encodes text into 512-dimensional vectors, this specific dimensionality is chosen based on empirical optimization to achieve an effective balance between capturing semantic meaning and maintaining computational efficiency. This enabled cosine similarity to be calculated between each response and its corresponding guideline reference.

Scores range from 0 to 1, with values closer to 1 indicating stronger semantic correspondence. All computations were carried out in Python (v3.10) using the tensorflow_hub and scikit-learn libraries.^[14]

Surface-level similarity was measured using the ROUGE-L metric, which identifies the longest common subsequence (LCS) to evaluate both word sequence and lexical overlap between each model-generated response and its corresponding reference text. ROUGE-L F1 scores were calculated using the rouge-score package in Python, with higher scores indicating greater textual similarity.^[15, 16]

To evaluate readability, two well-established indices were used: the Flesch-Kincaid Reading Ease (FKRE) and the Flesch-Kincaid Grade Level (FKGL). FKRE provides an estimate of how easily a passage can be read, with higher scores signifying greater readability. FKGL, on the other hand, translates text complexity into a corresponding U.S. school grade level needed for comprehension. FKRE is derived from sentence length and syllables per word, while FKGL interprets this complexity into a numeric grade level.^[17] Both metrics were calculated using the textstat Python library. The process of designing the study is delineated in Figure 1.

3. STATISTICAL ANALYSIS:

All statistical procedures were conducted utilizing IBM SPSS Statistics for Windows, Version 21.0 (IBM Corp., Armonk, NY, USA). The Shapiro-Wilk test was employed to

evaluate the normality of continuous variables. Data exhibiting normal distribution were expressed as mean $\pm$ standard deviation (SD), whereas those deviating from normality were summarized by median and range (minimum–maximum). Comparative analyses among the three large language models (ChatGPT-3.5, ChatGPT-4o, and Gemini 2.5) were performed using the Paired Samples t-test for normally distributed variables or the Wilcoxon Signed-Rank Test when data were non-normally distributed, contingent upon the distribution characteristics. Associations between numerical variables were examined via Spearman's rank-order correlation coefficient. Statistical significance was established at a two-tailed p-value threshold of less than 0.05, consistent with the prevailing biomedical research convention to balance Type I error control and statistical sensitivity.

4. RESULTS:

Descriptive statistics for each large language model are presented in Table 2. In terms of FKGL and FKRE values, statistically significant differences were observed across models. For FKGL, Gemini 2.5 [15.96 (14.03–20.9)] scored significantly higher than both ChatGPT-3.5 [15.29 (13.27–19.39)] and ChatGPT-4o [14.94 (12.1–18.14)] ($p = 0.003$ and $p = 0.004$, respectively). Regarding FKRE, ChatGPT-3.5 [19.31 (–4.3–28.83)] yielded significantly higher readability scores than Gemini 2.5 [14.04 (–5.27–28.56)]. ($p = 0.011$), although the difference with ChatGPT-4o [17.52 (–5.44–28.14)] was not statistically significant ($p = 0.13$).

Although reliability scores were higher for both ChatGPT-4o [6.5 (3–7)] and Gemini 2.5 [6.5 (3–7)] compared to ChatGPT-3.5 [5 (3–7)], these differences were not statistically significant ($p = 0.114$ and $p = 0.066$, respectively). Usefulness scores were significantly higher for ChatGPT-4o [6.5 (3–7)] compared to Gemini 2.5 [5 (3–7)] ($p = 0.009$). No statistically significant difference was found between ChatGPT-3.5 and either model ($p = 0.089$ and $p = 0.549$, respectively).

Regarding semantic similarity, both ChatGPT-4o [62.48 (38.9–75.59)] and Gemini 2.5 [63.83 (29.84–77.4)] achieved significantly higher scores than ChatGPT-3.5 [60.8 (30.07–70.91)] ($p = 0.017$ and $p = 0.013$, respectively), whereas no significant difference was found between ChatGPT-4o and Gemini 2.5 ($p = 0.623$). Although Gemini 2.5 [13.21 (8.49–17.82)] yielded lower ROUGE-L F1 scores compared to ChatGPT-3.5 [14.05 (8.65–18.18)], this difference did not reach statistical significance ($p = 0.085$), while no significant differences were observed between ChatGPT-4o [14.8 (8.78–19.17)] and either model ($p = 0.739$ and $p = 0.327$, respectively). (Table 3)

Spearman correlation analyses for each model are presented in Table 4. A strong and statistically significant negative correlation was observed between FKGL and FKRE across all models: $r = -0.904$ ($p < 0.001$) for ChatGPT-3.5, $r = -0.871$ ($p < 0.001$) for ChatGPT-4o, and $r = -0.868$ ($p < 0.001$) for Gemini 2.5. Additionally, a strong positive correlation between reliability and usefulness was found in all three models: $r = 1.000$ ($p < 0.001$) for ChatGPT-3.5, $r = 0.991$ ($p < 0.001$) for ChatGPT-4o, and $r = 0.779$ ($p < 0.001$) for Gemini 2.5. No other statistically significant correlations were observed among the remaining variable pairs (all $p > 0.05$).

Interrater reliability was evaluated using Cohen's kappa coefficient. Agreement across all domains was high, with unweighted κ values ranging from 0.81 to 1.00 and weighted κ values from 0.97 to 1.00, indicating near-perfect agreement according to the criteria proposed by Landis and Koch.

5. DISCUSSION:

In the current study, Gemini 2.5 generated outputs with a higher FKGL than both ChatGPT-3.5 and ChatGPT-4o, indicating greater textual complexity. FKRE was higher for

ChatGPT-3.5 than Gemini 2.5, while ChatGPT-4o and Gemini 2.5 did not differ significantly in FKRE. While no statistically significant differences in reliability were observed among the models, the ChatGPT-4o model demonstrated a significantly greater perceived usefulness relative to Gemini 2.5; this distinction was not evident for ChatGPT-3.5. In terms of semantic similarity, both ChatGPT-4o and Google Gemini 2.5 produced substantially higher scores than ChatGPT-3.5. ROUGE-L F1 scores did not differ significantly across the models. Additionally, a strong inverse relationship was found between FKGL and FKRE scores, while reliability and usefulness ratings exhibited a strong positive correlation, while no significant correlations were identified among the other parameters.

Rheumatoid arthritis (RA) is a common inflammatory disease affecting people worldwide, mostly between ages 20 and 54, and is about three times more common in women than men.^[18] RA usually starts gradually, with joint pain, stiffness, and swelling developing over weeks to months. Early symptoms can include fatigue, muscle pain, low fever, and weight loss. Inflammation may also affect tissues around joints, causing tenderness and warmth.^[19] The disease involves complex immune processes, such as abnormal antigen presentation and activation of immune cells like T and B lymphocytes.^[20]

In recent years, the definition of early rheumatoid arthritis has evolved, and the treat-to-target (T2T) strategy has improved early detection, disease control, and prevention of joint damage.^[20] Advances in medications and non-drug therapies have enhanced inflammation management and reduced long-term joint damage. Evidence-based guidelines from groups like EULAR have been key in standardizing and guiding treatment.^[4, 5]

Moreover, with the widespread adoption of artificial intelligence (AI) across various sectors, its use in rheumatology has become increasingly essential.^[20] Both patients and healthcare providers have started leveraging AI-driven tools to investigate possible diagnoses by entering clinical symptoms.^[21] In certain cases, laboratory test results have also been fed into these systems to assess potential diagnostic outcomes. Progress has continued with the integration of imaging technologies; radiological images are now uploaded to AI platforms, which are then used to generate and prioritize differential diagnoses based on the provided information.^[22] By incorporating radiological images into these models, AI not only aids in making diagnoses but also plays a significant role in the classification process.^[23] Treatment response and flare-up periods were determined with the help of these models.^[24] Additionally, AI-enhanced game-based programs have been introduced in rheumatologic rehabilitation.^[25]

Artificial intelligence platforms like ChatGPT and Gemini, created by OpenAI and Google, respectively, serve as leading examples of large language models (LLMs). These systems are designed to generate contextually appropriate responses by utilizing vast training datasets and, in certain instances, by accessing external databases. Moreover, through reinforcement learning from user interactions, these models can continuously improve their outputs, producing more precise and customized responses over time. This ability to adapt has played a key role in their increasing popularity and widespread use. Successive versions of these models have been introduced, offering enhanced capabilities and superior performance.^[26] However, the rapid growth of their application has raised concerns about the accuracy, safety, and scientific dependability of the information they provide.^[27] To address these issues, an expanding body of research has been dedicated to investigating the role of artificial intelligence in clinical practice, particularly in the realm of rheumatology.^[28]

In their review exploring the role of artificial intelligence in rheumatoid arthritis, Wang et al. emphasized that these models may be valuable in diagnosis due to their ability to

analyze comprehensive clinical, laboratory, and radiological data. Furthermore, they noted that AI is beneficial not only in the early diagnosis of RA but also in monitoring disease activity and guiding treatment protocols. Benavent et al. stated that these models may be utilized to assess treatment response in both patients with rheumatoid arthritis and those diagnosed with spondyloarthropathies. [8] Gilvaz et al. described artificial intelligence as an overarching concept encompassing both machine learning and deep learning. They emphasized that deep learning is particularly effective due to its capacity to process complex data types such as images, audio, and natural language. Consequently, it was highlighted that these models may offer valuable support not only in diagnosis and treatment but also in assessing disease activity, detecting flare-ups, and predicting prognosis. [29]

Celik et al. administered a 100-item questionnaire developed based on EULAR recommendations, diagnostic criteria, and clinical case scenarios to ChatGPT. The responses were independently evaluated by two rheumatologists, who found that the model achieved an accuracy rate of 84–86%. [30] In their study, Bai et al. input data from 291 patients with rheumatoid arthritis, including age, sex, rheumatoid factor, anti-citrullinated peptide antibody, 14-3-3 η protein, and anti-carbamylated protein antibodies, into an artificial intelligence system. Their findings indicated that the use of AI enhanced diagnostic accuracy. They emphasized that the anti-CCP antibody emerged as the most influential factor in establishing the diagnosis. [31] In their study involving 38 patients with rheumatoid arthritis, Valdivieso et al. proposed that the likelihood of responding to anti-TNF therapies could be predicted by inputting data on C-reactive protein, interleukin 2, interleukin 6, and the disease activity score into an artificial intelligence model. [32] When various models were compared in terms of their performance, Kurshe et al. input patient data into the ChatGPT-4o system and discovered that its preliminary diagnostic predictions were more accurate and cost-effective compared to those of rheumatologists. [33] In a cross-sectional study by Efe et al., ChatGPT-4o and Google Gemini were tested on 420 questions from the Board Vitals question bank. ChatGPT achieved an accuracy rate of 86.9%, while Gemini scored 60.2%. After receiving feedback, the accuracy rates increased to 86.7% and 60.5%, respectively. [34] In a study by Coskun et al., 23 questions related to methotrexate (MTX) from a previous study were directed to both ChatGPT-3.5 and ChatGPT-4 programs. The accuracy of GPT-4 was 100%, whereas GPT-3.5 achieved an accuracy rate of 86.96%. The study concluded that ChatGPT-4 provided more precise information regarding methotrexate use. [35] In the current study, ChatGPT-3.5 and ChatGPT-4 models were revealed as more readable and useful than Google Gemini 2.5; however, there was no difference in reliability.

An illustrative example of a study methodology employing questions formulated from guideline recommendations in the literature is the investigation conducted by Tong et al. Their research focused on glucocorticoid-induced osteoporosis and, in contrast to our approach, incorporated a secondary analysis subsequent to the provision of feedback to the models. They asked 34 frequent questions about glucocorticoid-induced osteoporosis to ChatGPT-3.5, ChatGPT-4, and Google Gemini models and after giving feedback they reevaluated the responses. They revealed the answers of ChatGPT-4, and Google Gemini better than ChatGPT-3.5. [13] Notably, our study did not implement any such feedback mechanism. However in the current study the answers of ChatGPT-4 and ChatGPT-3.5 were more readable and usefull than Google Gemini 2.5, while there was no difference in reliability.

The primary advantages of these models lie in their deep learning capabilities, the ability to integrate clinical, laboratory, and radiological data, as well as their potential to guide diagnosis, treatment, and prognosis, all while saving both time and resources. [36] However, the key drawbacks include the risk of relying on inaccurate data, challenges in

incorporating these models into existing healthcare systems, concerns regarding patient data confidentiality, and the potential for both health professionals and patients to harbor doubts about the reliability of the information as time progresses.^[37] Garcia-Agundez et al., in their review, highlighted the potential utility of artificial intelligence in assessing treatment responses in rheumatoid arthritis (RA). Nonetheless, they noted that these models often lack generalizability across diverse patient populations. Another critical limitation is the restricted context window of current models; when clinical narratives exceed this limit, the models may fail to accurately capture word-level interactions. Furthermore, concerns were raised regarding biases, misinformation, and the insufficient comprehensibility associated with lower-quality models. While ROUGE analysis may evaluate semantic similarity, it may fall short in effectively assessing the clinical validity of generated outputs.^[38] In the current study, ROUGE-F1 scoring was used to evaluate the surface-level similarity between the model-generated responses and the overarching principles and recommendations of the guideline. ROUGE-F1 is a commonly used metric in large language model performance analysis, particularly for assessing the structural resemblance of responses to open-ended questions. It has been conducted that in the medical domain, higher ROUGE scores were associated with more comprehensive and meaningful answers.^[15, 16] In the current study, no statistically significant differences were found between ChatGPT-4.0, ChatGPT-3.5, and Gemini 2.5 in terms of ROUGE-F1 scores ($p > 0.05$). This suggests that the structural alignment of model responses with the guideline recommendations was comparable across the models. Nevertheless, these findings should not be interpreted in isolation; clinical reliability, usefulness, and semantic accuracy must be evaluated together for a comprehensive understanding of model performance.

While these programs are capable of delivering accurate information, educating patients, and offering treatment suggestions to healthcare providers, the primary concern remains the potential for disseminating incorrect information. Additionally, ethical challenges and the risk of misuse are other significant issues that have been raised.^[25] During the evaluation of overarching principles, in item B, the ChatGPT-4o model was penalized for entirely omitting any reference to joint damage. In item C, the 3.5 model failed to acknowledge any medical specialties beyond rheumatology. In item D, the ChatGPT 3.5 model failed to address the objectives of targeted therapy, including the goal of achieving a 50% reduction in disease activity within three months. In item E, points were deducted due to the ChatGPT 3.5 model's omission of any reference to biosimilar drugs. Upon evaluation of the recommendations, all models received low scores for the second question, as they failed to incorporate key terms such as bDMARD, tsDMARD, patient global assessment, and the necessity of targeting low disease activity within 3 to 6 months. In the third question, points were deducted due to the omission of the term "swollen joint." In the fifth question, the deduction resulted from the inappropriate recommendation of additional hydroxychloroquine and azathioprine. For the sixth question, all models were penalized for suggesting combinations of glucocorticoids with DMARDs, bDMARDs, and tsDMARDs. In the seventh question, all models were penalized for recommending the triple combination therapy and prioritizing it over biological treatments. In the ninth question, none of the models emphasized that methotrexate (MTX) should be co-administered with biologic agents at low doses, such as 10 mg. The score was reduced for the tenth question due to the omission of sarilumab and for the eleventh question due to the failure to present the correct sequence for treatment de-escalation. Overall, the readability of Google Gemini was assessed as more challenging compared to the ChatGPT program.

The most important strength of the study is the assessment of the responses to open-ended questions blindly, unaware of the source program. Other strengths of this study lie

in its comparison of the most recent versions of the programs, as well as its thorough and objective evaluations regarding reliability, usability, answer length, comprehensibility, and both superficial and semantic similarity.

Although the study evaluated semantic quality, the main limitation is that clinical quality was not assessed, and no second analysis was conducted by providing feedback on the answers received. Additionally, the lack of feedback in this context could have further influenced the accuracy and relevance of the responses. Another key limitation was that the perceived usefulness and reliability varied greatly based on the evaluator's clinical expertise. To address this, the study involved two physiotherapy specialists, each with over a decade of experience in their field working in rheumatology area. Despite this limitation, the study entitled "Management of Rheumatoid Arthritis: Consensus," conducted in Türkiye, asserts that physiatrists may also contribute to patient management in settings where access to rheumatologists is constrained.^[39] An additional limitation is that the study only 16 open-ended questions, which were developed based on the EULAR 2022 guidelines. In this study, all eleven recommendations from the 2022 EULAR guideline for managing rheumatoid arthritis were directly converted into open-ended clinical questions, using the guideline text as the basis. Each question was assessed both individually by seasoned physiatrists and objectively through software-based evaluation metrics, including ROUGE, USE, FKRE, and FKGL. Consequently, this study encompasses the entire guideline rather than a selective portion, thereby reflecting its comprehensive structure both methodologically and substantively. Although it would have been feasible to generate multiple questions from each recommendation, potentially exceeding thirty questions, doing so would have fragmented the core clinical message of each recommendation. This fragmentation could undermine the coherence of the original content and compromise the objectivity of the evaluation process. Nevertheless, in order to reduce this limitation, overarching principles were also evaluated along with recommendations. In addition to this, since LLMs such as ChatGPT and Gemini receive frequent updates, reproducibility might change over time, potentially influencing the statistical outcomes. Arguably the most critical yet frequently overlooked limitation of these models is their propensity for generating hallucinated content. While they are capable of producing extensive and seemingly authoritative information, they may mislead users when prompted to provide supporting evidence citing incorrect sources, journals, authors, titles, or publication dates. The omission of an analysis of these hallucinatory tendencies in the present study represents a significant shortcoming.^[40] To address this issue adequately, more rigorous and comprehensive investigations are warranted. While the statistical differences reached significance, caution is needed when considering their clinical relevance. For instance, improved reliability and readability might help patients comprehend information better and aid clinical decisions, but these results require confirmation through practical clinical trials before firm conclusions can be made. While incorrect answers are discussed in detail qualitatively, there is a need for new, objective parameters to more clearly assess these risks.

6. CONCLUSION:

When asked open-ended questions based on EULAR 2022 recommendations, ChatGPT-4o and ChatGPT-3.5 gave clearer and more useful answers than Gemini 2.5, though all models were similarly reliable. Despite some risk of false info, these AI tools could help educate patients and assist clinicians with treatment planning. Future research should expand questions, add structured feedback, and create interactive dialogues.

7. ABBREVIATIONS:

AI: Artificial intelligence

RA: Rheumatoid arthritis
EULAR: European League Against Rheumatism
ChatGPT: Chat Generative Pre-trained Transformer
LLM: Large language models
LCS: Longest common subsequence
USE: Universal Sentence Encoder
FKRE: Flesch-Kincaid Reading Ease
FKGL: Flesch-Kincaid Grade Level
MTX: Methotrexate
csDMARD: Conventional disease-modifying antirheumatic drugs
bDMARD: Biologic disease-modifying antirheumatic drugs
tsDMARD: Target specific disease-modifying antirheumatic drugs
T2T: Treat-to-target

ETHICS APPROVAL: The study approved by XXX University Non-Interventional Clinical Research Ethics Committee (E-10840098-202.4.02-450, date: 04.06.2025) All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

CONSENT FOR PUBLICATION: All of the authors have contributed significantly to this manuscript. Additionally, all authors have seen and agree with the content of this manuscript, disclose no financial interests, and take full responsibility for the data and give consent for publication.

DECLARATION OF CONFLICT OF INTEREST: The authors declare that they have no conflict of interest.

FUNDING: The researchers did not take any funding for the study.

ACKNOWLEDGEMENT: We are grateful to Ahmet Akyol for acting as a referee in cases of disagreement during the 7-point Likert scale evaluations of the AI-generated answers to open-ended questions.

DATA AVAILABILITY STATEMENT: The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

REFERENCES

- [1] Di Matteo A, Bathon JM, Emery P. Rheumatoid arthritis. Lancet. 2023 Nov 25;402(10416):2019-2033. doi: 10.1016/S0140-6736(23)01525-8.
- [2] Jang S, Kwon EJ, Lee JJ. Rheumatoid Arthritis: Pathogenic Roles of Diverse Immune Cells. Int J Mol Sci. 2022 Jan 14;23(2):905. doi: 10.3390/ijms23020905.
- [3] Akgöz A, Yakut Y, Tüfekçi O, Aktaş BE, Sari EO, Barlak A et al. Biopsychosocial model-based exercise improves muscle strength, proprioception, pain, function, and quality of life in rheumatoid arthritis patients with knee involvement: a randomized controlled clinical trial. Rheumatol Int. 2025 Sep 6;45(9):221. doi: 10.1007/s00296-025-05976-3.
- [4] Smolen JS, Landewé RBM, Bergstra SA, Kerschbaumer A, Sepriano A, Aletaha D et al. EULAR recommendations for the management of rheumatoid arthritis with synthetic and biological disease-modifying antirheumatic drugs: 2022 update. Ann Rheum Dis. 2023 Jan;82(1):3-18. doi: 10.1136/ard-2022-223356.

- [5] Nagy G, Roodenrijs NMT, Welsing PMJ, Kedves M, Hamar A, van der Goes MC et al. EULAR points to consider for the management of difficult-to-treat rheumatoid arthritis. *Ann Rheum Dis*. 2022 Jan;81(1):20-33. doi: 10.1136/annrheumdis-2021-220973.
- [6] Park YJ, Pillai A, Deng J, Guo E, Gupta M, Paget M et al. Assessing the research landscape and clinical utility of large language models: a scoping review. *BMC Med Inf Decis Mak*. 2024;24(1):72. <https://doi.org/10.1186/s12911-024-02459-6>
- [7] Wang J, Tian Y, Zhou T, Tong D, Ma J, Li J. A survey of artificial intelligence in rheumatoid arthritis. *Rheumatol Immunol Res*. 2023 Jul 22;4(2):69-77. doi: 10.2478/rir-2023-0011.
- [8] Benavent D, Carmona L, García Llorente JF, Montoro M, Ramirez S, Otón T et al. Artificial intelligence to predict treatment response in rheumatoid arthritis and spondyloarthritis: a scoping review. *Rheumatol Int*. 2025 Apr 7;45(4):91. doi: 10.1007/s00296-025-05825-3.
- [9] Liu S, Wright AP, Patterson BL, Wanderer JP, Turer RW, Nelson SD et al. Using AI-generated suggestions from ChatGPT to optimize clinical decision support. *J Am Med Inform Assoc*. 2023;30(7):1237-1245. doi: 10.1093/jamia/ocad072.
- [10] Sciberras M, Farrugia Y, Gordon H, Furfaro F, Allocca M, Torres J et al. Accuracy of Information given by ChatGPT for Patients with Inflammatory Bowel Disease in Relation to ECCO Guidelines. *J Crohns Colitis*. 2024;18(8):1215-1221. doi: 10.1093/ecco-jcc/jjae040.
- [11] Liu M, Okuhara T, Chang X, Shirabe R, Nishiie Y, Okada H et al. Performance of ChatGPT Across Different Versions in Medical Licensing Examinations Worldwide: Systematic Review and Meta-Analysis. *J Med Internet Res*. 2024 Jul 25;26:e60807. doi: 10.2196/60807.
- [12] Tong L, Zhang C, Liu R, Yang J, Sun Z. Comparative performance analysis of large language models: ChatGPT-3.5, ChatGPT-4 and Google Gemini in glucocorticoid-induced osteoporosis. *J Orthop Surg Res*. 2024;19(1):574. doi: 10.1186/s13018-024-04996-2.
- [13] Uz C, Umay E. "Dr ChatGPT": Is it a reliable and useful source for common rheumatic diseases? *Int J Rheum Dis*. 2023;26(7):1343-1349. doi: 10.1111/1756-185X.14749.
- [14] Izzidien A, Fitz S, Romero P, Loe BS, Stillwell D. Developing a sentence level fairness metric using word embeddings. *Int J Digit Humanit*. 2022;1-36. doi: 10.1007/s42803-022-00049-4.
- [15] Tang L, Sun Z, Idnay B, Nestor JG, Soroush A, Elias PA et al. Evaluating Large Language Models on Medical Evidence Summarization. *medRxiv [Preprint]*. 2023 Apr 24:2023.04.22.23288967. doi: 10.1101/2023.04.22.23288967. Update in: *NPJ Digit Med*. 2023;6(1):158. doi: 10.1038/s41746-023-00896-7.
- [16] Gulden C, Kirchner M, Schüttler C, Hinderer M, Kampf M, Prokosch HU et al. Extractive summarization of clinical trial descriptions. *Int J Med Inform*. 2019;129:114-121. doi: 10.1016/j.ijmedinf.2019.05.019.
- [17] Kılıcoglu V.S, Yurdakul O.V, Aydın T. Evaluation of Chat Generative Pre-trained Transformer's responses to frequently asked questions about psoriatic arthritis: A study on quality and readability. *Ann Med Res* 2025;32(2):79-84
- [18] Zhang Z, Gao X, Liu S, Wang Q, Wang Y, Hou S et al. Global, regional, and national epidemiology of rheumatoid arthritis among people aged 20-54 years from 1990 to 2021. *Sci Rep*. 2025 Mar 28;15(1):10736. doi: 10.1038/s41598-025-92150-1.
- [19] Haville S, Deane KD. Pre-RA: Can early diagnosis lead to prevention? *Best Pract Res Clin Rheumatol*. 2022 Mar;36(1):101737. doi: 10.1016/j.berh.2021.101737.

- [20] Hügler T. The wide range of opportunities for large language models such as ChatGPT in rheumatology. *RMD Open*. 2023;9(2):e003105. doi: 10.1136/rmdopen-2023-003105.
- [21] Chandwar K, Prasanna Misra D. What does artificial intelligence mean in rheumatology? *Arch Rheumatol*. 2024 Feb 12;39(1):1-9. doi: 10.46497/ArchRheumatol.2024.10664.
- [22] Bird A, Oakden-Rayner L, McMaster C, Smith LA, Zeng M, Wechalekar MD et al. Artificial intelligence and the future of radiographic scoring in rheumatoid arthritis: a viewpoint. *Arthritis Res Ther*. 2022 Dec 12;24(1):268. doi: 10.1186/s13075-022-02972-x.
- [23] Obayya M, Alamgeer M, S Alzahrani J, Alabdan R, N Al-Wesabi F, Mohamed A et al. Artificial Intelligence Driven Biomedical Image Classification for Robust Rheumatoid Arthritis Classification. *Biomedicines*. 2022 Oct 26;10(11):2714. doi: 10.3390/biomedicines10112714.
- [24] Alivernini S, Cañete JD, Bacardit J, Kurowska-Stolarska M. Using explainable artificial intelligence to predict and forestall flare in rheumatoid arthritis. *Nat Med*. 2024 Apr;30(4):925-926. doi: 10.1038/s41591-024-02818-w.
- [25] Varga G, Stoicu-Tivadar L, Nicola S. Serious Gaming and Artificial Intelligence in Rehabilitation of Rheumatoid Arthritis. *Stud Health Technol Inform*. 2022 Jun 29;295:562-565. doi: 10.3233/SHTI220790.
- [26] Ray PP. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet Things Cyber-Phys Syst*. 2023;3:121–154. doi:10.1016/j.iotcps.2023.07.003.
- [27] Chen SY, Kuo HY, Chang SH. Perceptions of ChatGPT in healthcare: usefulness, trust, and risk. *Front Public Health*. 2024;12:1457131. doi: 10.3389/fpubh.2024.1457131.
- [28] Daungsupawong H, Wiwanitkit V. Comparative performance of artificial intelligence models in rheumatology board-level questions: evaluating Google Gemini and ChatGPT-4o: correspondence. *Clin Rheumatol*. 2024;43(12):4015-4016. doi: 10.1007/s10067-024-07176-z.
- [29] Gilvaz VJ, Reginato AM. Artificial intelligence in rheumatoid arthritis: potential applications and future implications. *Front Med (Lausanne)*. 2023 Nov 16;10:1280312. doi: 10.3389/fmed.2023.1280312.
- [30] Çelik NÇ, Kılınç EA. Assessment of ChatGPT's adherence to EULAR diagnostic criteria and therapeutic protocols for rheumatoid arthritis at two distinct time points, 14 days apart, utilizing binary and multiple-choice inquiries. *Clin Rheumatol*. 2025 Apr 22. doi: 10.1007/s10067-025-07417-9.
- [31] Bai L, Zhang Y, Wang P, Zhu X, Xiong JW, Cui L. Improved diagnosis of rheumatoid arthritis using an artificial neural network. *Sci Rep*. 2022 Jun 13;12(1):9810. doi: 10.1038/s41598-022-13750-9.
- [32] Valdivieso Shephard JL, Alvarez Robles EJ, Cámara Hijón C, Hernandez Breijo B, Novella-Navarro M, Bogas Schay P et al. Predicting anti-TNF treatment response in rheumatoid arthritis: An artificial intelligence-driven model using cytokine profile and routine clinical practice parameters. *Heliyon*. 2023 Nov 28;10(1):e22925. doi: 10.1016/j.heliyon.2023.e22925.
- [33] Krusche M, Callhoff J, Knitza J, Ruffer N. Diagnostic accuracy of a large language model in rheumatology: comparison of physician and ChatGPT-4. *Rheumatol Int*. 2024;44(2):303-306. doi: 10.1007/s00296-023-05464-6.
- [34] Is EE, Menekseoglu AK. Comparative performance of artificial intelligence models in rheumatology board-level questions: evaluating Google Gemini and ChatGPT-4o. *Clin Rheumatol*. 2024;43(11):3507-3513. doi: 10.1007/s10067-024-07154-5.

- [35] Coskun BN, Yagiz B, Ocakoglu G, Dalkilic E, Pehlivan Y. Assessing the accuracy and completeness of artificial intelligence language models in providing information on methotrexate use. *Rheumatol Int.* 2024;44(3):509-515. doi: 10.1007/s00296-023-05473-5.
- [36] McMaster C, Bird A, Liew DFL, Buchanan RR, Owen CE, Chapman WW et al. Artificial Intelligence and Deep Learning for Rheumatologists. *Arthritis Rheumatol.* 2022 Dec;74(12):1893-1905. doi: 10.1002/art.42296.
- [37] Vlad AL, Popazu C, Lescai AM, Voinescu DC, Baltă AAŞ. The Role of Artificial Intelligence in the Diagnosis and Management of Rheumatoid Arthritis. *Medicina (Kaunas).* 2025 Apr 9;61(4):689. doi: 10.3390/medicina61040689.
- [38] Garcia-Agundez A, Schmajuk G, Yazdany J. Promises and pitfalls of artificial intelligence models in forecasting rheumatoid arthritis treatment response and outcomes. *Semin Arthritis Rheum.* 2025 Feb;70S:152584. doi: 10.1016/j.semarthrit.2024.152584.
- [39] Ataman S., Borman P., Evcik D., Aydog E., Ayhan F., Yildizlar D et al. Management of rheumatoid arthritis: consensus recommendations from the Turkish League Against Rheumatism. *Turkish Journal of Rheumatology*, 2011; 26.4.
- [40] Aljamaan F, Temsah MH, Altamimi I, Al-Eyadhy A, Jamal A, Alhasan K et al. Reference Hallucination Score for Medical Artificial Intelligence Chatbots: Development and Usability Study. *JMIR Med Inform.* 2024 Jul 31; 12: e54345.

FIGURE LEGENDS:

Figure 1: The process of designing and execution stages of the study

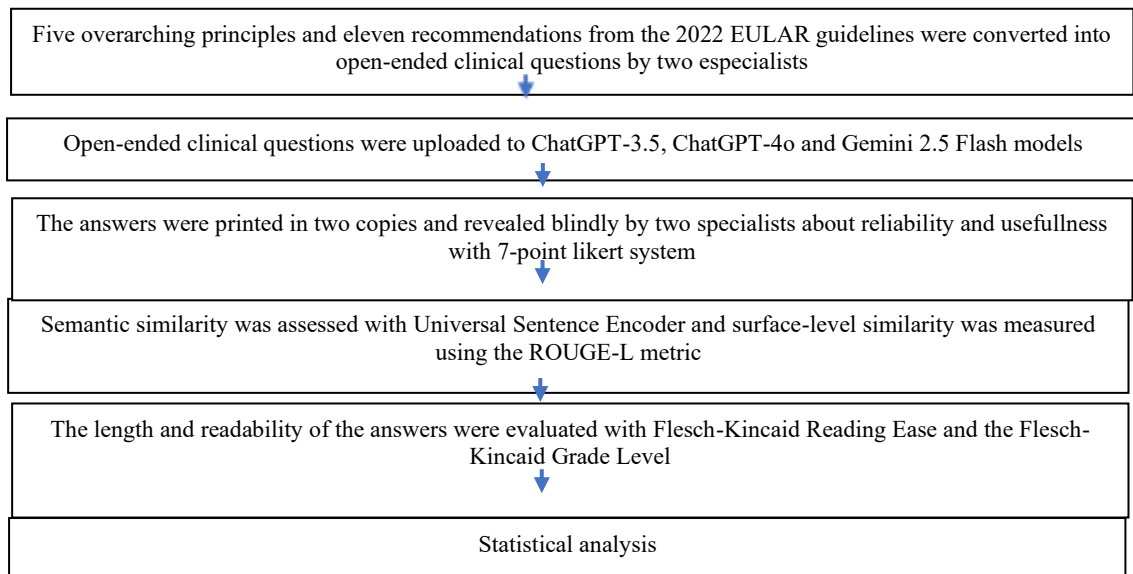


Figure 1: The process of designing and execution stages of the study

Table 1: Open-ended questions created from the EULAR 2022 guideline

Overarching Principles	
A	How should treatment decisions in rheumatoid arthritis be shared between the patient and the rheumatologist, and to what extent should patient preferences influence these decisions?

B	What are the key factors to consider when making treatment decisions in rheumatoid arthritis, and what is the role of patient-reported information in this process?
C	Which medical specialty should primarily oversee the follow-up and longitudinal management of patients with rheumatoid arthritis?
D	When and by which criteria should treatment response be assessed in patients with rheumatoid arthritis? How does this evaluation differ between early and established disease stages? Under which circumstances should treatment modification be considered, and what approach should be taken in the case of incidental flares?
E	How should cost considerations be integrated into treatment planning for rheumatoid arthritis, and what principles should guide the choice between equally effective treatment options?
Recommendations	
Q1	When and with which drug class should treatment be initiated in patients diagnosed with rheumatoid arthritis? Why is the treatment approach different in suspected RA cases, and what are the underlying reasons for this distinction?
Q2	What is the primary treatment target in patients with rheumatoid arthritis? How does this target differ between early and established RA? How is remission defined, and how might comorbid conditions such as fibromyalgia affect the achievement of treatment targets?
Q3	How frequently should disease activity be monitored in rheumatoid arthritis, and under what circumstances should the treatment be modified? Which measures should be used to guide clinical follow-up?
Q4	What should be the first-choice DMARD in the treatment of rheumatoid arthritis? Which DMARDs can be used and at what doses? What is the limited role of hydroxychloroquine, and why is it not recommended as a substitute for methotrexate? Under which circumstances are DMARD combinations appropriate, and what is the current role of triple therapy in terms of efficacy and safety?
Q5	In patients with rheumatoid arthritis who have contraindications to methotrexate or who demonstrate early intolerance, which csDMARDs should be considered as part of the treatment strategy, and what is the clinical significance of the term “early intolerance” in this context?
Q6	When initiating or switching csDMARD therapy in rheumatoid arthritis, when, at what dose, and via which route should glucocorticoids be considered? What is the appropriate duration of glucocorticoid therapy, what defines long-term use, and what are the associated risks? How does persistent glucocorticoid requirement reflect treatment inefficacy? Under which circumstances is glucocorticoid use appropriate alongside csDMARD, bDMARD, or tsDMARD therapy?
Q7	In patients with rheumatoid arthritis who fail to achieve the treatment target with the first csDMARD strategy but do not exhibit poor prognostic factors, what is the appropriate next step? When should csDMARD combinations, particularly triple therapy, be considered? What is the current role and rationale for triple csDMARD therapy?
Q8	In patients with rheumatoid arthritis who fail to achieve the treatment target with initial csDMARD therapy and present with poor prognostic factors, what are the next-line treatment options? What is the role of JAK inhibitors at this stage, and which risk factors should be considered in their use?

Q9	In patients with rheumatoid arthritis for whom treatment with a bDMARD or tsDMARD is planned, what are the justifications for continuing csDMARD co-therapy (particularly methotrexate)? Under which circumstances might monotherapy be appropriate? How should the methotrexate dose be adjusted in bDMARD–MTX combination therapy?
Q10	In patients with rheumatoid arthritis who experience treatment failure with a bDMARD or tsDMARD, what should be the subsequent treatment approach? When is it appropriate to switch to another agent within the same mechanism of action, and when should a change in mechanism be considered? How is difficult-to-treat rheumatoid arthritis defined, and what are the principles for managing such patients?
Q11	In patients with rheumatoid arthritis who are in sustained remission and have discontinued glucocorticoid therapy, when and how should DMARD dose reduction be considered? Which tapering strategies are preferred, and is complete treatment withdrawal advisable? What are the principles or sequences to consider when tapering therapy?

EULAR: European League Against Rheumatism RA: Rheumatoid arthritis csDMARD: Conventional disease-modifying antirheumatic drugs bDMARD: Biologic disease-modifying antirheumatic drugs tsDMARD: Target specific disease-modifying antirheumatic drugs

Table 2. Median and range (Min–Max) values of evaluation metrics for large language models

	ChatGPT-3.5	ChatGPT-4o	Gemini 2.5
ROUGE-L F1	14.05 (8.65 - 18.18)	14.8 (8.78 - 19.17)	13.21 (8.49 - 17.82)
FKGL	15.29 (13.27 - 19.39)	14.94 (12.1 - 18.14)	15.96 (14.03 - 20.9)
FKRE	19.31 (-4.3 - 28.83)	17.52 (-5.44 - 28.14)	14.04 (-5.27 - 28.56)
Reliability	5 (3 - 7)	6.5 (3 - 7)	6.5 (3 - 7)
Usefulness	5 (3 - 7)	6.5 (3 - 7)	5 (3 - 7)
Semantic Similarity	60.8 (30.07 - 70.91)	62.48 (38.9 - 75.59)	63.83 (29.84 - 77.4)

ROUGE-L F1: ROUGE-L F1 Score FKGL: Flesch-Kincaid Grade Level FKRE: Flesch-Kincaid Reading Ease Reliability: Reliability Score Usefulness: Usefulness Score Semantic Similarity: USE-based Semantic Similarity Score

Table 3. Pairwise comparisons of Flesch-Kincaid Grade Level, Flesch-Kincaid Reading Ease, Reliability Score, Usefulness Score, and USE-based Semantic Similarity large language models across evaluation metrics (*p*-values)

	1 vs 2	1 vs 3	2 vs 3
ROUGE-L F1	0.739*	0.085*	0.327*
FKGL	0.314*	0.003*	0.004*
FKRE	0.13*	0.011*	0.354*

Reliability	0.114 μ	0.066 μ	1 μ
Usefulness	0.089 μ	0.549*	0.009 μ
Semantic Similarity	0.017*	0.013 μ	0.623*

ROUGE-L F1: ROUGE-L F1 Score FKGL: Flesch-Kincaid Grade Level FKRE: Flesch-Kincaid Reading Ease Reliability: Reliability Score Usefulness: Usefulness Score Semantic Similarity: USE-based Semantic Similarity Score Model identifiers: 1 = ChatGPT-3.5, 2 = ChatGPT-4o, 3 = Gemini 2.5 μ Wilcoxon Signed-Rank Test *Paired Samples t-test $p < 0.05$ statistically significant values were shown in dark colour.

Table 4. Spearman correlation coefficients (r) and p -values between evaluation metrics for each large language model

			Semantic Similarity	FKGL	FKRE	Reliability	Usefulness
ROUGE-L F1 μ	ChatGPT-3.5	r	0.362	0.456	-0.288	-0.141	-0.141
		p	0.169	0.076	0.279	0.603	0.603
	ChatGPT-4o	r	0.169	0.206	-0.221	0.2	0.182
		p	0.125	0.444	0.412	0.458	0.501
	Gemini 2.5	r	0.497	-0.059	-0.026	-0.013	0.036
		p	0.05	0.829	0.922	0.963	0.895
Semantic Similarity μ	ChatGPT-3.5	r	—	0.209	0.399	-0.306	-0.306
		p		0.438	0.126	0.25	0.25
	ChatGPT-4o	r		-0.435	0.385	-0.078	-0.049
		p		0.092	0.141	0.775	0.857
	Gemini 2.5	r		-0.379	0.418	-0.379	-0.21
		p		0.147	0.107	0.148	0.435
FKGL μ	ChatGPT-3.5	r	—	—	-0.904	-0.213	-0.213
		p			<0.001*	0.427	0.427
	ChatGPT-4o	r			-0.871	0.203	0.161
		p			<0.001*	0.451	0.551
	Gemini 2.5	r			-0.868	0.124	-0.142
		p			<0.001*	0.648	0.601
FKRE μ	ChatGPT-3.5	r	—	—	—	0.09	0.09
		p				0.74	0.74
	ChatGPT-4o	r				-0.239	-0.193
		p				0.372	0.475
	Gemini 2.5	r				-0.187	0.06
		p				0.488	0.825
Reliability μ	ChatGPT-3.5	r	—	—	—	—	1
		p					<0.001*
	ChatGPT-4o	r					0.991
		p					<0.001*

	Gemini 2.5	r					0.779
		p					<0.001*

ROUGE-L F1: ROUGE-L F1 Score FKGL: Flesch-Kincaid Grade Level FKRE: Flesch-Kincaid Reading Ease Reliability: Reliability Score Usefulness: Usefulness Score Semantic Similarity: USE-based Semantic Similarity Score ^uSpearman's rank-order correlation *p < 0.05 statistically significant.