

Comparative Performance of Large Language Models in Answering Guideline-Based Questions on Juvenile Idiopathic Arthritis: A Blinded Expert Evaluation

Pelin Yücesan Şahin

yucesanpelin@gmail.com

İzmir Şehir Hastanesi

Rana İşgüder

İzmir Tepecik Eğitim ve Araştırma Hastanesi

Selçuk Yüksel

Canakkale Onsekiz Mart Üniversitesi Tıp Fakültesi Hastanesi

Süleyman İmamoğlu

İzmir Şehir Hastanesi

Hatice Adıgüzel Dünder

Dr. Behçet Uz Çocuk Hastalıkları Hastanesi

Belde Kasapdemir

İzmir Şehir Hastanesi

Research Article

Keywords: juvenile idiopathic arthritis, large language models, artificial intelligence, pediatric rheumatology, clinical decision support, practice guidelines, readability

Posted Date: October 18th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11231359/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Purpose

Large language models (LLMs) are increasingly used for clinical information retrieval, but their performance in juvenile idiopathic arthritis (JIA) remains poorly characterized. We compared three contemporary LLMs across diagnosis and classification, treatment and follow-up, and clinical case scenarios.

Methods

Twenty-one guideline-based JIA questions were submitted once to GPT-5.6 Sol, Gemini 3.1 Pro, and Claude Opus 5 on August 25, 2026, each in a separate conversation. Five pediatric rheumatologists independently rated anonymized responses, presented in randomized order, for accuracy and guideline concordance, clinical usability, and overall quality. Linear mixed-effects models included LLM, domain, and their interaction, with evaluator and question as crossed random intercepts. Response length, readability, inter-rater reliability, and ordinal sensitivity analyses were also assessed.

Results

The 63 responses yielded 315 ratings per outcome. GPT-5.6 Sol achieved the highest mean scores for accuracy (4.71 ± 0.47), usability (4.70 ± 0.55), and quality (6.47 ± 0.71), compared with Gemini 3.1 Pro (4.04 ± 0.95 , 4.24 ± 0.86 , 5.32 ± 1.11) and Claude Opus 5 (3.93 ± 0.89 , 3.90 ± 0.81 , 5.26 ± 1.05 ; all $p < 0.001$). LLM $\times$ domain interactions were significant for accuracy ($p = 0.005$) and usability ($p < 0.001$), but not quality ($p = 0.053$), reflecting a domain-dependent ranking of Gemini and Claude. Claude Opus 5 produced the longest responses without higher ratings, and GPT-5.6 Sol had the lowest Flesch-Kincaid Grade Level in every domain. Ordinal analyses supported the primary findings. Panel-level reliability was moderate (ICC[2,k] 0.664–0.729).

Conclusion

Contemporary LLMs differed meaningfully in guideline-based JIA performance, with differences varying by clinical task. These findings support supervised rather than autonomous use of LLMs in pediatric rheumatology and highlight the need for model-specific, periodically repeated validation.

What is Known

- Large language models are widely accessible, but their accuracy varies by model and clinical task.
- In juvenile idiopathic arthritis, evidence is limited to a single evaluation of one model.

What is New:

- In the first blinded multi-model comparison against current JIA guidelines, GPT-5.6 Sol ranked highest in every domain, while Gemini and Claude swapped ranks between diagnosis and management.
- Longer answers were not rated higher, and fluent responses sometimes blurred classification criteria and strength of evidence, supporting only specialist-supervised use.

1. INTRODUCTION

Juvenile idiopathic arthritis (JIA) is the most common chronic inflammatory arthritis of childhood and comprises a heterogeneous group of conditions with substantial variation in clinical phenotype, extra-articular manifestations, prognosis, and treatment requirements [1]. The International League of Associations for Rheumatology (ILAR) classification remains widely used, while the provisional Pediatric Rheumatology International Trials Organization (PRINTO) framework was developed to define more homogeneous disease entities and address limitations of the ILAR categories [2, 3]. Management has also become increasingly complex. Treatment now includes conventional synthetic disease-modifying antirheumatic drugs, multiple biologic agents, and targeted synthetic therapies, with choices guided by disease phenotype, activity, prognostic features, and previous treatment response [4–6]. Current recommendations additionally address medication monitoring, immunization, and infection-related considerations [7]; systemic JIA and macrophage activation syndrome require distinct strategies centered on early control of systemic inflammation and targeted cytokine inhibition [8]; and uveitis screening and treatment depend on ocular and JIA-related risk factors [9]. These parallel developments in classification and management make accurate interpretation of current recommendations particularly important.

Large language models (LLMs) are general-purpose models trained on large bodies of text that can generate fluent, well-organized answers to clinical questions [10], and their performance on medical licensing-style questions drew early attention to their potential role in medicine [11]. In rheumatology, LLM use is emerging, although routine clinical adoption remains limited; rheumatologists have identified diagnostic support as one of the areas in which AI may be particularly useful [12]. A recent German survey found that 57.6% of patients with rheumatic diseases would welcome the use of AI-based clinical support by their rheumatologists [13]. Linguistic plausibility, however, does not guarantee factual accuracy, guideline concordance, or clinical safety. Evaluation of LLMs in health care has predominantly focused on question answering and medical knowledge benchmarks, while only a small proportion of studies have used real patient-care data [14]. When ten models were tested on guideline-based pediatric nephrology cases, accuracy varied widely and hallucinations occurred across all evaluated models [15]. This is especially relevant in fields where recommendations change over time, because commercially deployed models are themselves updated and may differ in how they represent recent guidance. The Chatbot Assessment Reporting Tool (CHART) was developed in response to heterogeneity in how such

studies are conducted and reported and emphasizes transparent reporting of model identifiers, access conditions, prompting strategy, and evaluation procedures [16].

Evidence specific to rheumatology remains limited but is expanding. In a 2024 study using rheumatology board-level questions, ChatGPT-4o achieved higher accuracy than Gemini (86.9% vs 60.2%) [17]. More recently, La Bella et al. evaluated ChatGPT-4o responses to ten guideline-derived questions on oligoarticular JIA across five countries. Depending on the evaluator, 52–84% of responses were rated as mostly or fully adherent to ACR recommendations; readability was consistently difficult, and inter-rater agreement was generally weak [18]. A 2026 systematic review of LLM applications in inflammatory arthritis identified 18 studies spanning patient education, guideline adherence, and clinical reasoning, but only one focused specifically on JIA, and most evaluated GPT-based models rather than directly comparing several contemporary systems [19]. The review also reported lower performance for case-based clinical scenarios than for frequently asked questions and structured guideline questions, suggesting that factual retrieval alone may not capture performance in clinically contextualized tasks [19].

Against this background, comparative evidence on contemporary LLM performance across the broader spectrum of JIA-specific clinical tasks remains limited. We therefore aimed to compare GPT-5.6 Sol, Gemini 3.1 Pro, and Claude Opus 5 in answering guideline-based questions covering diagnosis and classification, treatment and follow-up, and clinical case scenarios in JIA. Responses were evaluated independently by pediatric rheumatologists against a prespecified multi-source reference standard. We additionally examined response length and readability, inter-rater reliability, and whether relative model performance varied across clinical domains. By combining guideline-based questions with clinically contextualized scenarios, the study was designed to evaluate not only the accuracy of retrieved information but also its translation into clinically usable responses.

2. METHODS

2.1. Study Design and Reference Standard

This cross-sectional comparative study evaluated three LLMs using guideline-based questions on juvenile idiopathic arthritis (JIA). The study was designed and reported in accordance with the CHART statement [16]. The prespecified reference standard comprised recommendations available at the time of response generation: the ILAR [2] and provisional PRINTO [3] classification criteria for diagnosis and classification, and the 2026 American College of Rheumatology (ACR) JIA guideline summary [5] for treatment, monitoring, and uveitis. Topics not covered by this summary were evaluated using the 2021 ACR guideline for latent infection screening [7] and the 2019 ACR/Arthritis Foundation uveitis guideline for uveitis risk stratification and anti-TNF selection [9]; the EULAR/PReS Still's disease recommendations were additionally considered for systemic JIA and macrophage activation syndrome [8]. Because no JIA-specific guideline addresses biologic therapy during serious infection, this question was evaluated against product labeling and the consensus of the two pediatric rheumatologists who developed the

questions. Where sources differed, the most recent applicable recommendation was prioritized, and responses consistent with an alternative guideline-supported recommendation were not considered incorrect.

Formal ethics committee approval was not obtained. The study involved no patient participation, identifiable patient information, biological material, access to medical records, or clinical interventions; it consisted of expert assessment of AI-generated textual responses to predefined guideline-based questions and hypothetical clinical scenarios. The study protocol was not prospectively registered.

2.2. Question Development and LLM Response Generation

The methodological framework of this study was informed by a previous expert-based comparative LLM evaluation study [20]. The present study extended this framework to JIA using a prespecified multi-source clinical reference standard, randomized response presentation, domain-specific analyses, and mixed-effects modelling to account for the crossed structure of expert ratings. A total of 21 questions were formulated by two pediatric rheumatologists on the basis of current international pediatric rheumatology classification frameworks and clinical guidelines defined in the prespecified reference standard. The questions were grouped into three domains: diagnosis and classification (Questions 1–6), treatment and follow-up (Questions 7–16), and clinical case scenarios (Questions 17–21). The first domain addressed core diagnostic and classification concepts; the second focused on therapeutic selection, treatment escalation or switching, biologic therapy, infection screening and management, uveitis surveillance, and drug-toxicity monitoring; and the case-based domain assessed the ability of the models to integrate multiple clinical variables and formulate clinically appropriate management recommendations. All questions were written in English, and identical wording was used for all three LLMs.

Three commercially available LLMs were evaluated: GPT-5.6 Sol (OpenAI), Gemini 3.1 Pro (Google), and Claude Opus 5 (Anthropic). All three were proprietary, closed-source models. All responses were generated on August 25, 2026, through the respective consumer web interfaces using active paid subscriptions, with all queries performed in İzmir, Türkiye. All models were used with their default settings; no optional features, such as extended reasoning or thinking modes, web search, file uploads, custom instructions, or memory and personalization functions, were enabled. The same configuration was maintained for all questions within each platform. Because commercially deployed LLMs are continuously updated, the exact date of response generation was recorded to define the temporal version of each evaluated system and enhance reproducibility.

Each question was submitted once to each LLM, yielding 63 model-generated responses. To minimize carryover effects from preceding study prompts or within-conversation context, every question was entered in a separate, newly initiated conversation for each model. No follow-up, corrective, or iterative prompting was performed. The first complete response generated by each model was recorded and retained for evaluation. The textual content of all responses was preserved verbatim, with no editing, summarization, rewriting, truncation, or correction. To reduce the recognizability of individual models,

formatting elements such as headings, bullet points, bold text, and tables were largely removed, and all responses were presented to evaluators as plain text in a uniform document format. Response length was not restricted or standardized, thereby preserving the native output characteristics of each model and more closely reflecting real-world use. The full text of all 63 responses is provided in Supplementary File 1.

2.3. Expert Evaluation, Response Length, and Readability

The anonymized responses were independently evaluated by five pediatric rheumatologists, comprising two professors, two associate professors, and one specialist. Before scoring, evaluators received a standardized written assessment framework defining each evaluation domain and its Likert-scale anchors and reviewed the reference standards used in the study. To minimize assessment bias, model identities were concealed using neutral labels (Model X, Model Y, and Model Z), and the presentation order of the three model responses was independently randomized for each question so that no model occupied a consistent position across the questionnaire. Evaluators scored all responses independently and were unaware of one another's ratings. No patients or members of the public were involved in question development or response evaluation.

Each response was rated across three predefined evaluation dimensions: accuracy and guideline concordance, clinical usability, and overall quality. Accuracy and guideline concordance were assessed on a 5-point Likert scale ranging from 1 (completely incorrect) to 5 (completely correct and guideline-concordant). Clinical usability was assessed on a 5-point Likert scale ranging from 1 (clinically unusable) to 5 (directly usable in clinical practice), taking into account clarity, applicability, clinical relevance, and the absence of misleading recommendations. Overall quality was assessed on a 7-point Likert scale ranging from 1 (very poor) to 7 (excellent), incorporating completeness, clarity, appropriate clinical prioritization, avoidance of unnecessary detail, inclusion of essential information, and the extent to which the response constituted a satisfactory expert-level answer. A 7-point scale was used for this broader construct to allow finer discrimination between responses of similar accuracy, consistent with a previous expert-based LLM evaluation study [20]. Higher scores indicated better performance in all domains.

Response length and readability were evaluated independently of expert-rated performance. For each LLM, the original verbatim responses within each prespecified clinical domain were combined, excluding question text and model-identification labels, and analyzed using the Readable.com platform. Total word count was recorded for each model-domain combination and was treated as a descriptive measure of verbosity rather than an intrinsic measure of response quality. Readability was assessed using the Flesch-Kincaid Grade Level (FKGL), Gunning Fog Index (GFI), and Flesch Reading Ease (FRE) score for each model-domain combination.

2.4. Statistical Analysis

Descriptive statistics were calculated for each LLM overall and within each prespecified question domain. Expert-rated outcomes were summarized as mean $\pm$ standard deviation (SD). Word count and

readability indices were reported descriptively for each model-domain combination.

Because every question was answered by all three LLMs and each response was scored by the same five evaluators, the data had a crossed, repeated-measures structure. The individual rating (question $\times$ model $\times$ evaluator) was therefore used as the unit of analysis. Accuracy, clinical usability, and overall quality were compared using linear mixed-effects models with LLM, question domain, and the LLM $\times$ domain interaction specified as fixed effects and evaluator and question included as crossed random intercepts. Fixed effects were tested using F tests with Satterthwaite approximation for the denominator degrees of freedom. Given the limited number of questions within each domain, the LLM $\times$ domain interaction was considered exploratory. When a significant overall LLM effect or LLM $\times$ domain interaction was identified, prespecified pairwise contrasts of estimated marginal means were performed as appropriate, with Holm adjustment for multiple comparisons. Given the ordinal nature of the rating scales, cumulative logit mixed models were additionally fitted as sensitivity analyses using LLM, question domain, and the LLM $\times$ domain interaction as fixed effects, with question included as a random intercept and evaluator as a fixed blocking factor.

Inter-rater reliability among the five evaluators was assessed using intraclass correlation coefficients (ICCs) with 95% confidence intervals. The primary reliability measure was ICC(2,k), based on a two-way random-effects, absolute-agreement, average-measures model, reflecting the reliability of the five-expert panel. ICC(2,1), based on the corresponding single-rater absolute-agreement model, was additionally calculated to characterize individual-rater reliability. ICC(3,k), based on a two-way mixed-effects, consistency, average-measures model, was also reported to facilitate comparison with a previous expert-based LLM study [20]. ICC values < 0.50 , $0.50\text{--}0.75$, $0.75\text{--}0.90$, and > 0.90 were interpreted as indicating poor, moderate, good, and excellent reliability, respectively [21]. No formal sample size calculation was performed because the study used a purposively constructed, guideline-based question set rather than a sampled clinical population. The 21 questions were selected to provide prespecified coverage of the three clinical domains of interest. No imputation was planned for missing ratings; analyses were based on available observations, and the frequency of missing data, if any, was reported. All statistical tests were two-sided, and a p value < 0.05 was considered statistically significant. Statistical analyses were performed using IBM SPSS Statistics version 28.0 (IBM Corp., Armonk, NY, USA) and R version 4.6.1 (R Foundation for Statistical Computing, Vienna, Austria) with the lme4, lmerTest, ordinal, emmeans, and irr packages. Figures were generated using the ggplot2 package in R.

3. RESULTS

3.1. Response Set and Overall Model Performance

A total of 21 questions were evaluated across three prespecified JIA domains: diagnosis and classification ($n = 6$), treatment and follow-up ($n = 10$), and clinical case scenarios ($n = 5$) (Table 1). Each question was answered by all three LLMs, yielding 63 model-generated responses. Each response was

independently evaluated by five pediatric rheumatologists, resulting in 315 individual ratings for each evaluation dimension. No ratings were missing.

Table 1
Questions submitted to the large language models

No.	Question
A. Diagnosis and Classification	
Q1	What are the minimum entry criteria and required symptom duration for establishing a diagnosis of Juvenile Idiopathic Arthritis?
Q2	How is Early-Onset ANA-Positive JIA defined in the current PRINTO classification system, and why was this category introduced?
Q3	What clinical and laboratory criteria are required for the classification of Rheumatoid Factor-Positive JIA?
Q4	What clinical features and laboratory findings support the diagnosis and classification of Enthesitis-Related Arthritis?
Q5	How is Psoriatic JIA classified according to current criteria, and which features help distinguish it from other JIA subtypes?
Q6	What are the major differences between the PRINTO and ILAR classification systems for Juvenile Idiopathic Arthritis?
B. Treatment and Follow-up	
Q7	What is the recommended next-line therapeutic approach for a patient with oligoarthritis who has an inadequate response to initial NSAIDs and intra-articular glucocorticoid injections?
Q8	What is the preferred first-line systemic treatment option for a patient newly diagnosed with active Systemic JIA without Macrophage Activation Syndrome?
Q9	Which specific biologic mechanisms or interleukin inhibitors are preferred as first-line therapy when Systemic JIA is complicated by Macrophage Activation Syndrome?
Q10	What is the recommended switching strategy when a JIA patient exhibits an inadequate clinical response to their first TNF inhibitor?
Q11	What is the recommended approach to tapering or discontinuing biologic therapy in patients with Systemic JIA who achieve sustained inactive disease?
Q12	What is the standard required frequency of screening for chronic anterior uveitis in a high-risk JIA patient?
Q13	Which mandatory screening protocol must be completed for latent infections prior to initiating any biologic DMARD therapy in a child with JIA?
Q14	How should biologic therapy generally be managed in a patient with JIA who develops a serious acute infection?
Q15	Which specific anti-TNF agents are preferred, and which one should be avoided, for a JIA patient who has active chronic anterior uveitis?
<i>Abbreviations: JIA, juvenile idiopathic arthritis; PRINTO, Pediatric Rheumatology International Trials Organization; ILAR, International League of Associations for Rheumatology; MAS, macrophage activation syndrome; NSAIDs, nonsteroidal anti-inflammatory drugs; TNF, tumor necrosis factor; DMARD, disease-modifying antirheumatic drug; RF, rheumatoid factor; MCP, metacarpophalangeal.</i>	

No.	Question
A. Diagnosis and Classification	
Q16	Which specific laboratory parameters must be monitored, and at what standard intervals, to assess drug toxicity during long-term Methotrexate therapy?
C. Clinical Case Scenarios	
Q17	A 13-year-old girl with RF-negative Polyarticular Juvenile Idiopathic Arthritis (JIA) continues to have active arthritis involving both wrists, several metacarpophalangeal (MCP) joints, and one ankle despite 8 months of Methotrexate therapy and 6 months of Etanercept treatment. Inflammatory markers remain elevated, although there is no evidence of structural joint damage on imaging. How should her treatment strategy be modified?
Q18	A 3-year-old girl with Oligoarticular JIA involving the knee and wrist shows persistent active synovitis despite maximum-dose NSAIDs and repeated intra-articular corticosteroid injections. What is the next recommended step in her therapeutic management?
Q19	A 12-year-old female with RF-positive Polyarticular JIA experiences progressive severe pain and limited motion in her hip and jaw despite 9 months of Methotrexate and Adalimumab therapy. How should her biologic treatment framework be modified?
Q20	A 15-year-old female diagnosed with Polyarticular JIA has achieved sustained and complete clinical remission for the past 12 months on a combination of Methotrexate and a TNF inhibitor. What should be the next step in her therapeutic management?
Q21	A 10-year-old boy with Systemic JIA develops a second episode of MAS within one year despite previous successful treatment with pulse corticosteroids and Anakinra. He is currently asymptomatic between episodes. What long-term therapeutic strategy would be most appropriate to reduce the risk of future MAS episodes and disease-related complications?
<i>Abbreviations: JIA, juvenile idiopathic arthritis; PRINTO, Pediatric Rheumatology International Trials Organization; ILAR, International League of Associations for Rheumatology; MAS, macrophage activation syndrome; NSAIDs, nonsteroidal anti-inflammatory drugs; TNF, tumor necrosis factor; DMARD, disease-modifying antirheumatic drug; RF, rheumatoid factor; MCP, metacarpophalangeal.</i>	

Across all 21 questions, significant differences among the three LLMs were observed for all three outcomes (all $p < 0.001$; Table 3), and GPT-5.6 Sol achieved the highest mean scores for each. In Holm-adjusted pairwise comparisons, GPT-5.6 Sol scored significantly higher than both other models for all outcomes, whereas Gemini 3.1 Pro and Claude Opus 5 differed significantly only for clinical usability, with Gemini ranking second. The distributions of individual expert ratings across models and clinical domains are shown in Fig. 1.

Table 2

Comparative performance of large language models across juvenile idiopathic arthritis task domains

Outcome	GPT-5.6 Sol	Gemini 3.1 Pro	Claude Opus 5	P value*
Diagnosis and Classification				
Accuracy and guideline concordance	4.63 ± 0.49 ^a	3.60 ± 1.10 ^c	4.07 ± 0.78 ^b	< 0.001
Clinical usability	4.53 ± 0.57 ^a	3.87 ± 0.97 ^b	4.23 ± 0.73 ^{ab}	0.001
Overall quality	6.37 ± 0.72 ^a	5.03 ± 1.25 ^c	5.57 ± 0.86 ^b	< 0.001
Word count	1,289	1,213	1,211	
Readability				
FKGL	15.6	18.9	16.9	
GFI	17.7	18.0	17.0	
FRE	9.0	-6.4	1.6	
Treatment and Follow-up				
Accuracy and guideline concordance	4.82 ± 0.39 ^a	4.24 ± 0.77 ^b	4.00 ± 0.90 ^b	< 0.001
Clinical usability	4.80 ± 0.45 ^a	4.46 ± 0.73 ^b	3.88 ± 0.69 ^c	< 0.001
Overall quality	6.56 ± 0.64 ^a	5.46 ± 0.97 ^b	5.20 ± 1.07 ^b	< 0.001
Word count	1,791	637	2,708	
Readability				
FKGL	15.2	20.8	17.2	
GFI	18.3	23.8	17.7	
FRE	12.8	-4.4	0.3	
Clinical Case Scenarios				
Values for expert-rated outcomes are mean ± SD. Diagnosis and classification included 30 ratings per model (6 questions × 5 evaluators), treatment and follow-up included 50 ratings per model (10 × 5), and clinical case scenarios included 25 ratings per model (5 × 5). *P values represent the within-domain omnibus LLM effect from linear mixed-effects models including LLM, question domain, and the LLM × domain interaction as fixed effects, with evaluator and question included as crossed random intercepts. Superscript letters denote Holm-adjusted pairwise comparisons within each outcome and domain; values sharing at least one letter were not significantly different. Word count and readability metrics were calculated from pooled responses within each model-domain combination using Readable.com and are descriptive. FKGL, Flesch-Kincaid Grade Level; GFI, Gunning Fog Index; FRE, Flesch Reading Ease.				

Outcome	GPT-5.6 Sol	Gemini 3.1 Pro	Claude Opus 5	P value*
Diagnosis and Classification				
Accuracy and guideline concordance	4.60 ± 0.58 ^a	4.16 ± 0.94 ^b	3.64 ± 0.95 ^c	< 0.001
Clinical usability	4.72 ± 0.68 ^a	4.24 ± 0.83 ^b	3.56 ± 1.00 ^c	< 0.001
Overall quality	6.40 ± 0.82 ^a	5.40 ± 1.19 ^b	5.00 ± 1.15 ^b	< 0.001
Word count	977	636	2,677	
Readability				
FKGL	15.3	19.3	18.8	
GFI	18.9	22.3	22.2	
FRE	16.0	-0.1	-2.1	
Values for expert-rated outcomes are mean ± SD. Diagnosis and classification included 30 ratings per model (6 questions × 5 evaluators), treatment and follow-up included 50 ratings per model (10 × 5), and clinical case scenarios included 25 ratings per model (5 × 5). *P values represent the within-domain omnibus LLM effect from linear mixed-effects models including LLM, question domain, and the LLM × domain interaction as fixed effects, with evaluator and question included as crossed random intercepts. Superscript letters denote Holm-adjusted pairwise comparisons within each outcome and domain; values sharing at least one letter were not significantly different. Word count and readability metrics were calculated from pooled responses within each model-domain combination using Readable.com and are descriptive. FKGL, Flesch-Kincaid Grade Level; GFI, Gunning Fog Index; FRE, Flesch Reading Ease.				

Table 3
Overall performance of the large language models across all 21 JIA questions

Outcome	GPT-5.6 Sol	Gemini 3.1 Pro	Claude Opus 5	P value*
Accuracy and guideline concordance	4.71 ± 0.47 ^a	4.04 ± 0.95 ^b	3.93 ± 0.89 ^b	< 0.001
Clinical usability	4.70 ± 0.55 ^a	4.24 ± 0.86 ^b	3.90 ± 0.81 ^c	< 0.001
Overall quality	6.47 ± 0.71 ^a	5.32 ± 1.11 ^b	5.26 ± 1.05 ^b	< 0.001
Values are presented as mean ± SD based on 105 expert ratings per model (21 questions × 5 evaluators).				
*P values represent the overall LLM effect from linear mixed-effects models with evaluator and question included as crossed random intercepts. Superscript letters indicate Holm-adjusted pairwise comparisons; values sharing at least one letter were not significantly different.				

3.2. Domain-Specific Performance

Model performance differed significantly within each of the three clinical domains (Table 2). In the diagnosis and classification domain, GPT-5.6 Sol had the highest mean accuracy and guideline concordance score (4.63 ± 0.49), compared with Claude Opus 5 (4.07 ± 0.78) and Gemini 3.1 Pro (3.60 ± 1.10 ; $p < 0.001$). All three pairwise comparisons were statistically significant. Overall quality showed the same pattern, with mean scores of 6.37 ± 0.72 , 5.57 ± 0.86 , and 5.03 ± 1.25 for GPT-5.6 Sol, Claude Opus 5, and Gemini 3.1 Pro, respectively ($p < 0.001$), with significant differences between each pair of models. Clinical usability also differed among models ($p = 0.001$); GPT-5.6 Sol scored higher than Gemini 3.1 Pro (4.53 ± 0.57 vs 3.87 ± 0.97), whereas Claude Opus 5 (4.23 ± 0.73) did not differ significantly from either model after Holm adjustment.

In the treatment and follow-up domain, GPT-5.6 Sol again achieved the highest scores for accuracy and guideline concordance (4.82 ± 0.39), clinical usability (4.80 ± 0.45), and overall quality (6.56 ± 0.64), with significant overall differences among models for all three outcomes (all $p < 0.001$). For accuracy, GPT-5.6 Sol scored significantly higher than Gemini 3.1 Pro (4.24 ± 0.77) and Claude Opus 5 (4.00 ± 0.90), whereas the latter two models did not differ significantly. For clinical usability, all pairwise comparisons were significant, with scores decreasing from GPT-5.6 Sol to Gemini 3.1 Pro (4.46 ± 0.73) and Claude Opus 5 (3.88 ± 0.69). For overall quality, GPT-5.6 Sol scored significantly higher than both Gemini 3.1 Pro (5.46 ± 0.97) and Claude Opus 5 (5.20 ± 1.07), whereas Gemini and Claude did not differ significantly.

A similar pattern was observed for the clinical case scenarios. GPT-5.6 Sol achieved mean scores of 4.60 ± 0.58 for accuracy and guideline concordance, 4.72 ± 0.68 for clinical usability, and 6.40 ± 0.82 for overall quality. Corresponding scores were 4.16 ± 0.94 , 4.24 ± 0.83 , and 5.40 ± 1.19 for Gemini 3.1 Pro and 3.64 ± 0.95 , 3.56 ± 1.00 , and 5.00 ± 1.15 for Claude Opus 5 (all omnibus $p < 0.001$). Accuracy and clinical usability differed significantly between all three models after Holm adjustment. For overall quality, GPT-5.6 Sol scored significantly higher than both competing models, whereas the difference between Gemini 3.1 Pro and Claude Opus 5 was not statistically significant.

The exploratory LLM $\times$ domain interaction was significant for accuracy and guideline concordance ($F[4,284] = 3.86$, $p = 0.005$) and clinical usability ($F[4,284] = 5.82$, $p < 0.001$), indicating that the magnitude of differences between models varied across clinical domains. The interaction for overall quality did not reach statistical significance ($F[4,284] = 2.37$, $p = 0.053$).

3.3. Response Length and Readability

Response length varied substantially across models (Table 2 and Fig. 2). Claude Opus 5 generated the most words in total (6,596), followed by GPT-5.6 Sol (4,057) and Gemini 3.1 Pro (2,486). Word counts were similar across models in diagnosis and classification, whereas Claude Opus 5 produced substantially longer responses in treatment and follow-up and in clinical case scenarios. Because the domains contained different numbers of questions, domain-specific word counts were interpreted descriptively.

Readability indices also differed descriptively across models (Table 2). GPT-5.6 Sol had the lowest FKGL (15.2–15.6) and the highest FRE scores (9.0–16.0) in all three domains, whereas Gemini 3.1 Pro had the highest FKGL (18.9–20.8) and negative FRE values in every domain. Claude Opus 5 had the lowest Gunning Fog Index in two of the three domains. Overall, the readability measures indicated high linguistic complexity across all three models.

3.4. Inter-Rater Reliability and Sensitivity Analyses

The reliability of the five-expert panel was moderate for all three overall evaluation dimensions (Supplementary Table S1). The primary average-measures absolute-agreement ICC [ICC(2,k)] was 0.729 (95% CI, 0.603–0.823) for accuracy and guideline concordance, 0.664 (95% CI, 0.514–0.779) for clinical usability, and 0.706 (95% CI, 0.574–0.806) for overall quality. Domain-specific ICC(2,k) values ranged from 0.587 to 0.841. Reliability was good for accuracy and guideline concordance (ICC = 0.793, 95% CI 0.594–0.912) and overall quality (ICC = 0.841, 95% CI 0.687–0.932) within the diagnosis and classification domain, while the remaining domain-specific estimates were in the moderate range. Single-rater absolute-agreement reliability was lower, with overall ICC(2,1) values of 0.350 for accuracy and guideline concordance, 0.284 for clinical usability, and 0.324 for overall quality.

Sensitivity analyses using cumulative logit mixed models were broadly consistent with the primary linear mixed-effects analyses, with the direction of all pairwise comparisons preserved. Two pairwise significance decisions differed after Holm adjustment: in the diagnosis and classification domain, the Claude Opus 5 versus Gemini 3.1 Pro comparison for accuracy was significant in the primary analysis but not in the ordinal sensitivity analysis (adjusted $p = 0.055$), whereas in the clinical case scenario domain, the Claude Opus 5 versus Gemini 3.1 Pro comparison for overall quality became significant in the ordinal analysis (adjusted $p = 0.037$). These differences did not alter the overall pattern of findings.

4. DISCUSSION

The three LLMs showed clear and generally consistent differences in performance across the JIA question set. GPT-5.6 Sol had the highest mean scores for accuracy and guideline concordance, clinical usability, and overall quality in each prespecified domain. The significant LLM $\times$ domain interactions for accuracy and clinical usability indicate that the magnitude of between-model differences depended partly on the clinical task, whereas the interaction for overall quality did not reach statistical significance. Response length followed a different pattern: Claude Opus 5 generated the greatest volume of text, while GPT-5.6 Sol received the highest expert ratings. Thus, within this dataset, longer responses did not necessarily correspond to higher expert-rated performance.

The overall pattern is consistent with evidence that relative LLM performance varies across models, specialties, and evaluation tasks. In rheumatology board-style questions, Is and Menekseoglu reported higher accuracy for ChatGPT-4o than for Gemini (86.9% vs 60.2%) [17]. In pediatric nephrology, Niel et al. evaluated ten LLMs and found overall accuracy ranging from 50.8% to 86.9%; the authors also reported hallucinations in every model and potentially life-threatening decisions in some models [15]. In a recent

pediatric study, Gemini 3 Pro outperformed ChatGPT 5.2 in quality, reliability, and readability when providing feeding and oral health guidance for children with cleft lip and palate [22], in contrast to the superiority of the GPT model observed in our JIA question set. Using an expert-rated design conceptually similar to ours, Baş Aksu et al. compared ChatGPT, Gemini, and Claude in adrenal incidentaloma management and found broadly similar performance for routine questions, with clearer differences in more complex scenarios [20]. These findings argue against assuming a fixed hierarchy among commercial model families across specialties or versions. A recent systematic review of LLMs in inflammatory arthritis identified 18 studies but only one JIA-specific study, which evaluated ChatGPT-4o rather than directly comparing several contemporary models [18, 19]. The present study therefore adds a direct multi-model evaluation across classification, treatment, follow-up, and case-based JIA tasks. Our results reflect the model versions accessed on August 25, 2026, and may not apply to later versions.

The significant interactions for accuracy and clinical usability largely reflected a reversal in the relative ranking of the other two models: GPT-5.6 Sol had the highest mean scores in every domain, whereas Claude Opus 5 scored above Gemini 3.1 Pro in diagnosis and classification but below Gemini in treatment and follow-up and in clinical case scenarios. Diagnosis and classification require distinction between the established ILAR framework and the provisional PRINTO criteria, which reorganize several traditional JIA categories and introduce early-onset ANA-positive JIA as a distinct entity [2, 3]. Inspection of the verbatim responses provides an illustrative example. In one response, “asymmetric arthritis” was incorporated into the defining criteria for early-onset ANA-positive JIA. Although asymmetry is characteristic of the phenotype, it is not itself an operational PRINTO classification criterion; the provisional definition instead specifies arthritis duration, age at onset, repeated ANA positivity, and relevant exclusions [3]. This shows how an LLM may merge characteristic disease features with formal classification rules.

Treatment and follow-up questions posed a different challenge because recommendations depend on previous therapy, disease phenotype, associated uveitis, systemic features, MAS risk, and guideline updates [5, 7–9]. A second illustrative example was the inclusion of rituximab among post-TNF options for RF-positive polyarticular JIA. We do not regard this response as incorrect under the prespecified reference framework, but it shows how an LLM may present a clinically conceivable option without conveying the relative strength of supporting evidence. The 2019 ACR guideline for polyarthritis preferentially recommended abatacept or tocilizumab after failure of a first TNF inhibitor and placed rituximab later in the treatment sequence [6], whereas the 2026 APLAR recommendations considered the evidence for rituximab insufficient [23]. Neither guideline was part of our scoring standard, and both examples are illustrative rather than the result of a formal error analysis.

The clinical case scenarios required integration of phenotype, previous treatment exposure, treatment response, remission status, and risk. All three models were significantly separated for accuracy and clinical usability in this domain. Similar separation was seen in other domains, so this mainly shows that the differences persisted when recommendations had to be applied to individual patients. Roemer et al. illustrated the potential consequence of an incorrect but plausible suggestion in a randomized

rheumatology study: although LLM assistance improved medical students' overall diagnostic reasoning, exposure to an erroneous JIA suggestion in one rheumatoid arthritis vignette increased selection of JIA as the leading diagnosis [24]. Future evaluations should therefore include clinical cases, not only factual questions.

Claude Opus 5 produced 6,596 words across the complete question set, compared with 4,057 for GPT-5.6 Sol and 2,486 for Gemini 3.1 Pro, yet this ordering did not mirror the expert ratings. Because word count was analyzed descriptively, we cannot link length to quality statistically; however, longer answers were clearly not needed to obtain higher ratings. Readability was demanding across all three models, with FKGL values above 15 in every model-domain combination. La Bella et al. similarly reported difficult readability for ChatGPT-4o responses to guideline-derived oligoarticular JIA questions, with a median Flesch Reading Ease score of 30 [18]. Advanced reading levels have also been reported in rheumatology patient-education and pediatric dialysis LLM studies [25, 26]. Those studies mostly examined patient materials, whereas our questions were written for clinicians. This level of complexity is acceptable for specialists, but the same answers would need simplifying before being shared with families.

Panel-level reliability was moderate, with ICC(2,k) values of 0.729 for accuracy and guideline concordance, 0.664 for clinical usability, and 0.706 for overall quality [21]. Single-rater ICCs were considerably lower, indicating variation in how individual specialists judged the same responses. La Bella et al. likewise reported generally weak inter-rater agreement when experts assessed adherence of JIA-related LLM responses [18], whereas higher panel reliability has been reported in other specialty-specific evaluations, including Baş Aksu et al. [20]. Differences in scoring instruments, clinical content, number of raters, and ICC specification preclude direct numerical comparison. In practice, a single expert's rating would not have been a reliable benchmark for these models.

In clinical terms, these models can support pediatric rheumatologists but cannot replace them. The models often synthesized guideline-based information into clinically useful answers, but the two examples show that a fluent answer can blur the line between a classification criterion and a typical feature, or between a possible option and a recommended one. Expert verification therefore remains necessary, particularly when recommendations depend on recent guideline updates or individual context. Madrid-García et al. evaluated a retrieval-augmented framework incorporating EULAR and ACR recommendations and reported improvements in accuracy, safety, and completeness compared with the underlying LLM without retrieval augmentation [27]. Such approaches may be useful in pediatric rheumatology, where relevant evidence is distributed across several guidance documents, but they require prospective validation before routine clinical use.

Several features strengthen the present study. Three current commercial models were tested using identical questions; each prompt was submitted in a separate conversation; and only the first complete response was retained without editing or corrective prompting. Responses were anonymized, their order was randomized for each question, and five pediatric rheumatologists independently evaluated every response. The prespecified reference framework incorporated multiple JIA sources, and the mixed-

effects models accounted for evaluator- and question-level variation. Ordinal sensitivity analyses preserved the direction of all pairwise comparisons, supporting the main pattern of findings.

The study also has limitations. The 21 questions were purposively constructed and cannot represent the full spectrum of JIA practice, and domain-specific estimates were based on relatively few questions, particularly the five clinical scenarios; accordingly, the LLM × domain analyses should be considered exploratory. Each question was submitted only once, so stochastic variation and reproducibility across repeated generations were not assessed. The study evaluated English-language responses from consumer web interfaces at a single time point, limiting generalizability to later versions, APIs, other languages, or retrieval-enabled systems. Although model identities were concealed and formatting was standardized, evaluators may still have recognized characteristic writing styles, introducing residual bias. The 2026 ACR recommendations were available as a board-approved guideline summary rather than final peer-reviewed guideline manuscripts at the time of evaluation [5]. In addition, no JIA-specific guideline directly addresses management of biologic therapy during a serious acute infection; this item therefore required interpretation using product labeling and expert consensus. Word count and readability were pooled at model-domain level and were descriptive, and readability formulas do not directly measure clinical comprehension or reasoning quality. Finally, hypothetical questions cannot establish effects on real clinical decisions or patient outcomes.

In conclusion, the three contemporary LLMs differed meaningfully in their responses to guideline-based JIA questions, with GPT-5.6 Sol receiving the highest expert ratings across all evaluated domains. The magnitude of between-model differences varied according to the clinical task, and greater response length did not consistently accompany higher scores. These models may be useful as supervised tools in pediatric rheumatology, but not for independent diagnostic or treatment decisions in JIA.

Declarations

Conflict of interest:

The authors declare no conflicts of interest.

Ethics approval:

Not applicable; the study involved no patients, patient data, or human participants other than the expert evaluators, as described in the Methods.

Funding:

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Author Contribution

P.Y.Ş. and B.K.D. conceived and designed the study. P.Y.Ş. and B.K.D. developed the questions and defined the reference standard. P.Y.Ş. generated and anonymized the LLM responses, performed the statistical analysis, and wrote the first draft of the manuscript. R.İ., S.Y., S.İ., H.A.D., and B.K.D. evaluated the model responses. B.K.D. supervised the study. R.İ., S.Y., S.İ., H.A.D., and B.K.D. critically revised the manuscript for important intellectual content. All authors read and approved the final manuscript.

Acknowledgement

During the preparation of this manuscript, the authors used artificial intelligence (AI) tools to assist with the translation of the original Turkish draft into English and to improve language, grammar, and readability. These tools were not used to generate scientific content, perform data analysis, or interpret the study findings. All AI-assisted text was subsequently reviewed and edited by the authors, who take full responsibility for the accuracy and integrity of the published content. The large language models evaluated in this study were used solely as study subjects, as described in the Methods section.

Data Availability

The full verbatim texts of all model-generated responses are provided in Supplementary File 1. The anonymized expert ratings and analysis code are available from the corresponding author upon reasonable request.

References

1. Sandborg CI, Schulert GS, Kimura Y (2025) Juvenile idiopathic arthritis. *N Engl J Med* 393(2):162–174
2. Petty RE et al (2004) International League of Associations for Rheumatology classification of juvenile idiopathic arthritis: second revision, Edmonton, 2001. *J Rheumatol* 31(2):390–392
3. Martini A et al (2019) Toward new classification criteria for juvenile idiopathic arthritis: first steps, pediatric rheumatology international trials organization international consensus. *J Rheumatol* 46(2):190–197
4. Shenoi S et al (2024) Treatment of non-systemic juvenile idiopathic arthritis. *Nat Rev Rheumatol* 20(3):170–181
5. American College of Rheumatology (2026) *American College of Rheumatology (ACR) Juvenile Idiopathic Arthritis Guideline Summary*. 2026 May 13 [cited 2026 Sep 24]; Available from: <https://rheumatology.org/juvenile-idiopathic-arthritis-guideline>
6. Ringold S et al (2019) *American College of Rheumatology/Arthritis Foundation guideline for the treatment of juvenile idiopathic arthritis: therapeutic approaches for non-systemic polyarthritis*,

- sacroiliitis, and enthesitis*. Arthritis care & research, 2019. 71(6): pp. 717–734
7. Onel KB et al (2021) *American College of Rheumatology guideline for the treatment of juvenile idiopathic arthritis: recommendations for nonpharmacologic therapies, medication monitoring, immunizations, and imaging*. Arthritis care & research, 2022. 74(4): pp. 505–520
 8. Fautrel B et al (2024) EULAR/PReS recommendations for the diagnosis and management of Still's disease, comprising systemic juvenile idiopathic arthritis and adult-onset Still's disease. Ann Rheum Dis 83(12):1614–1627
 9. Angeles-Han ST et al (2019) *American College of Rheumatology/Arthritis Foundation guideline for the screening, monitoring, and treatment of juvenile idiopathic arthritis–associated uveitis*. Arthritis care & research, 2019. 71(6): pp. 703–716
 10. Thirunavukarasu AJ et al (2023) Large language models in medicine. Nat Med 29(8):1930–1940
 11. Singhal K et al (2023) Large language models encode clinical knowledge. Nature 620(7972):172–180
 12. Holzer M-T et al (2024) Artificial intelligence in rheumatology: status quo and quo vadis—results of a national survey among German rheumatologists. Therapeutic Adv Musculoskelet Disease 16:1759720X241275818
 13. Labinsky H et al (2025) Patient experiences, attitudes, and profiles regarding artificial intelligence in rheumatology: a German national cross-sectional survey study. Rheumatol Int 45(12):269
 14. Bedi S et al (2025) Testing and evaluation of health care applications of large language models: a systematic review. JAMA 333(4):319–328
 15. Niel O, Dookhun D, Caliment A (2025) Performance evaluation of large language models in pediatric nephrology clinical decision support: a comprehensive assessment. Pediatr Nephrol 40(10):3211–3218
 16. Collaborative CHART (2025) Reporting guideline for chatbot health advice studies: Chatbot Assessment Reporting Tool (CHART) statement. Annals Family Med 23(5):389–398
 17. Is EE, Menekseoglu AK (2024) Comparative performance of artificial intelligence models in rheumatology board-level questions: evaluating Google Gemini and ChatGPT-4o. Clin Rheumatol 43(11):3507–3513
 18. La Bella S et al (2025) Global variations in artificial intelligence-generated information on juvenile idiopathic arthritis. Rheumatology 64(11):5687–5697
 19. Adiniaev Y et al (2026) ChatGPT and Other Large Language Models in Inflammatory Arthritis: A Systematic Review Across Clinical Tasks. The Journal of Rheumatology
 20. Baş Aksu Ö et al (2026) Artificial intelligence in endocrine practice: comparing ChatGPT, Gemini, and Claude for adrenal incidentaloma care. J Endocrinol Investig 49(1):69–79
 21. Koo TK, Li MY (2016) A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 15(2):155–163

22. Beşparmak A, Beşparmak TM (2026) Performance of artificial intelligence chatbots in providing feeding management and oral health guidance for children with cleft lip and palate: ChatGPT 5.2 vs Gemini 3 Pro. *Eur J Pediatrics* 185(7):498
23. Arkachaisri T et al (2026) The Asia-Pacific League of Associations for Rheumatology Consensus Recommendations on the Management of Juvenile Idiopathic Arthritis (JIA): Polyarticular Course JIA, Temporomandibular Joint Arthritis, Imaging, and Non-Pharmacologic Therapies. *Int J Rheum Dis* 29(4):e70640
24. Roemer A et al (2026) Large language models enhance diagnostic reasoning of medical students in rheumatology: a randomized controlled trial. *BMC Med Educ* 26(1):579
25. Erden Y, Temel MH, Bağcıer F (2025) Evaluating ChatGPT-4 for rheumatology patient education: a comparative analysis of readability, reliability, and similarity to the American College of Rheumatology's fact sheets. *Reumatologia* 63(5):313–320
26. Ensari E, Onder EN, Ertan P (2025) A comparative assessment of large language models in pediatric dialysis: reliability, quality and readability. *Therapeutic Apheresis Dialysis* 29(5):739–746
27. Madrid-García A et al (2025) Optimising the clinical application of rheumatology guidelines using large language models: a retrieval-augmented generation framework integrating EULAR and ACR recommendations. *EULAR Rheumatol Open* 1(3):228–236

Figures

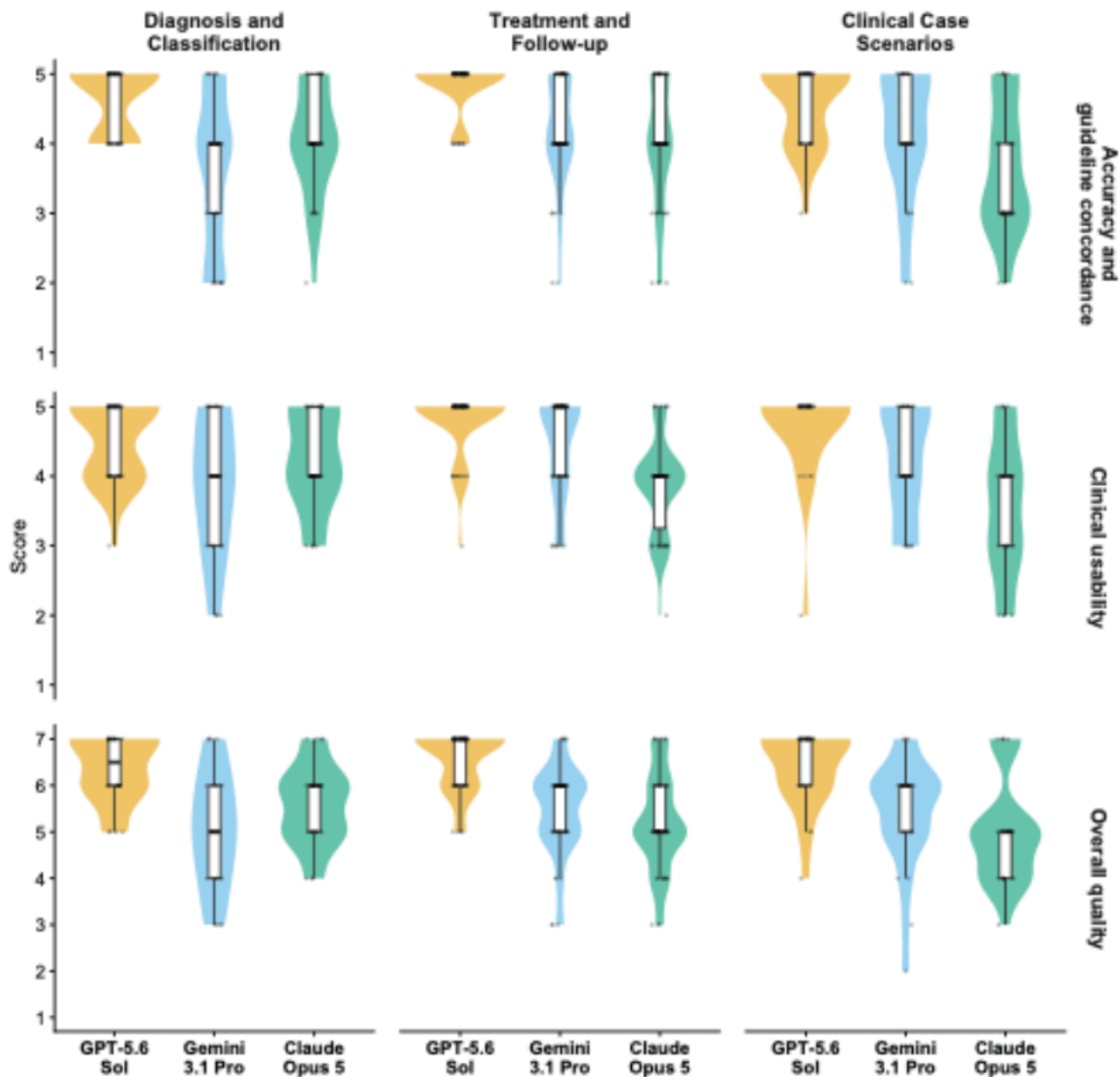


Figure 1

Distribution of expert ratings across large language models and juvenile idiopathic arthritis task domains. Violin plots show the distribution of individual expert ratings for accuracy and guideline concordance, clinical usability, and overall quality across the diagnosis and classification, treatment and follow-up, and clinical case scenario domains. Accuracy and guideline concordance and clinical usability were rated on 5-point Likert scales, and overall quality was rated on a 7-point Likert scale; y-axes display the full range of each scale. Internal box plots indicate the median and interquartile range, with whiskers extending to 1.5 times the interquartile range. Individual points represent ratings provided by the five

pediatric rheumatology experts (30, 50, and 25 ratings per model in the three domains, respectively). Results of the statistical comparisons are presented in Table 2.

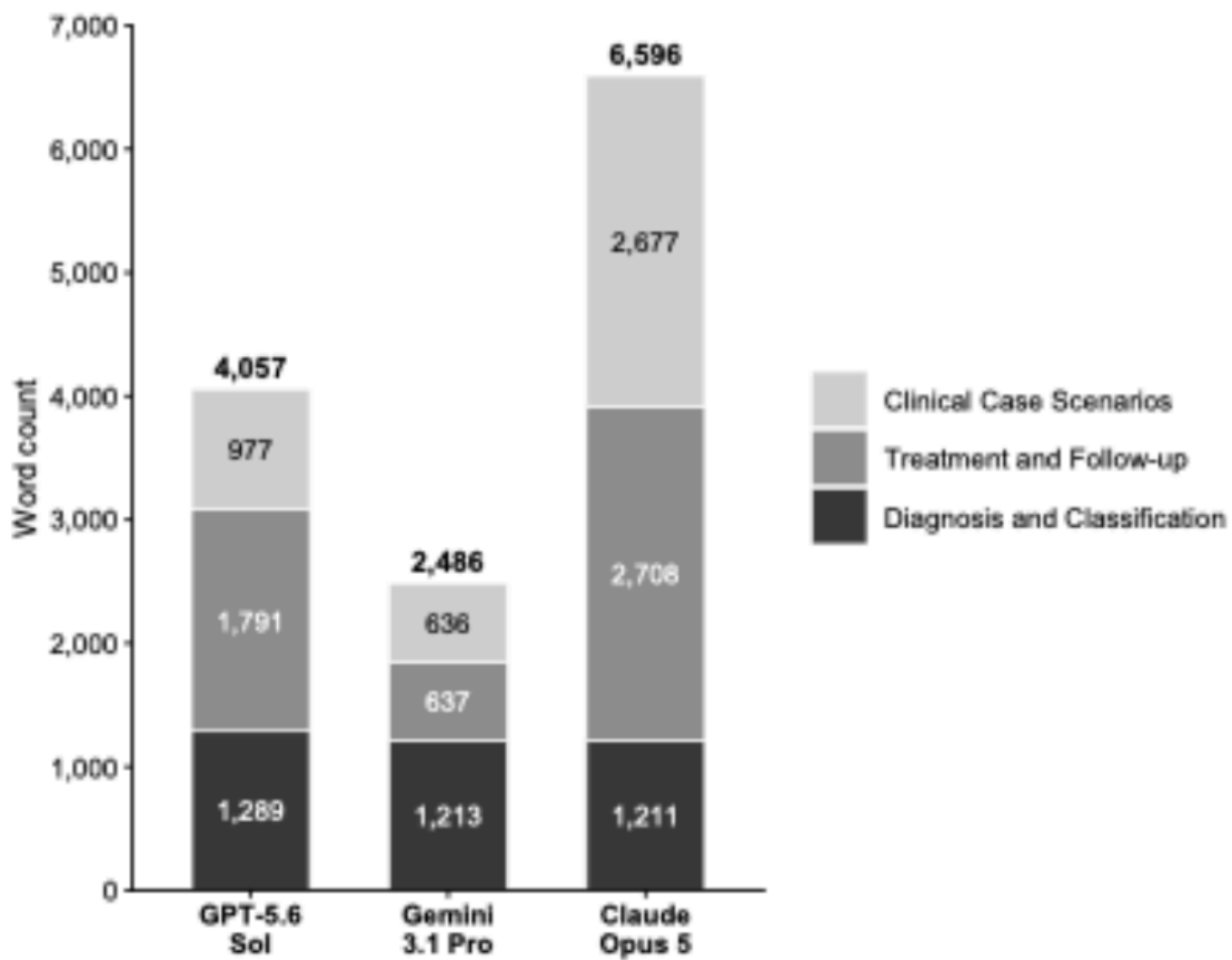


Figure 2

Total word count generated by each large language model across the three prespecified juvenile idiopathic arthritis task domains. Word counts were calculated using Readable.com from the pooled verbatim responses within each model-domain combination, excluding question text and model identification labels. Stacked bars represent the cumulative word count, values within each segment indicate domain-specific word counts, and values above each bar indicate the total word count for each model. The diagnosis and classification, treatment and follow-up, and clinical case scenario domains comprised 6, 10, and 5 questions, respectively; segment sizes therefore reflect both response length and the number of questions per domain. Word count is presented as a descriptive measure of verbosity and not as an indicator of response quality.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [supplementtables.docx](#)
- [CHART.docx](#)
- [supplementaryfile1.docx](#)