International Journal of
Translational Medicine

MDPI

Article

Comparative Evaluation of Three Large Language Models in Systemic Lupus Erythematosus Using a Clinical Vignette-Based Rubric

Ahmet Tezce [1,*], Yunus Durmaz [2] and Ahmet Burak Dadak [3]

[1] Department of Physical Medicine and Rehabilitation, Faculty of Medicine, Karabuk University, 78050 Karabuk, Turkey
[2] Department of Rheumatology, Karabük Training and Research Hospital, 78200 Karabuk, Turkey; durmaz_y@hotmail.com
[3] Department of Physical Medicine and Rehabilitation, Ankara Ayas Sait Mehmet Ciftci State Hospital, 06710 Ankara, Turkey; ahmetburakdadak@gmail.com
* Correspondence: ahmettezce@karabuk.edu.tr

Abstract

Background: Large language models (LLMs) can generate clinically relevant information, but safety-related performance in systemic lupus erythematosus (SLE) remains insufficiently characterised. We compared GPT-5.5 Instant (ChatGPT), Claude Sonnet 4.6, and Gemini 3.5 Flash through their free-tier web interfaces across diagnostic acuity, guideline adherence, patient safety/red-flag recognition and hallucination-related reliability. Methods: Twelve SLE clinical vignettes were submitted to all three interfaces in May 2026 (36 single-pass responses). Two blinded specialists independently assessed the outputs using a study-developed four-domain rubric (composite range, 4–20). Inter-rater reliability was evaluated using quadratic-weighted Cohen's κ. Paired comparisons across the shared vignettes used Friedman tests with Kendall's W and post hoc Wilcoxon signed-rank tests with Holm adjustment. Results: Inter-rater κ ranged from 0.754 to 0.893. Friedman tests showed between-interface differences for patient safety/red-flag recognition ($\chi^2(2) = 9.571$, $p = 0.008$, W = 0.399), hallucination-related reliability ($\chi^2(2) = 6.500$, $p = 0.039$, W = 0.271) and composite score ($\chi^2(2) = 8.140$, $p = 0.017$, W = 0.339), but not diagnostic acuity ($p = 0.926$) or guideline adherence ($p = 0.175$). Claude scored higher than ChatGPT and Gemini for patient safety (Holm-adjusted $p = 0.0176$ for both) and composite score (Holm-adjusted $p = 0.0469$ for both). No hallucination-related pairwise comparison remained significant after Holm adjustment. Conclusions: Within this exploratory, 12-vignette paired evaluation, the Claude interface received higher safety and composite ratings, but these observations do not establish generalisable model superiority or real-world clinical safety. The findings are specific to the tested May 2026 model versions and require independent replication.

Keywords: large language models; systemic lupus erythematosus; clinical vignette; artificial intelligence; patient safety

Check for updates

Academic Editor: Simone Brogi

Received: 7 September 2026
Revised: 11 October 2026
Accepted: 22 October 2026
Published: 28 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Systemic lupus erythematosus (SLE) is a complex, multisystem autoimmune disease characterised by marked clinical heterogeneity, a relapsing-remitting course, and a broad spectrum of potentially life-threatening organ manifestations [1]. Accurate diagnosis

requires the algorithmic application of weighted classification criteria, and optimal management demands familiarity with frequently updated evidence-based guidelines encompassing induction therapy, long-term immunosuppression, comorbidity prevention, and the recognition of emergencies such as macrophage activation syndrome, neuropsychiatric involvement, and high-risk pregnancy situations [2,3]. These demands create conditions in which decision-support tools capable of synthesising up-to-date clinical knowledge could meaningfully assist clinical reasoning.

Large language models (LLMs) have emerged as accessible, general-purpose artificial intelligence tools with the capacity to process and generate clinically relevant text at scale. Following the widespread availability of ChatGPT, Claude, and Gemini, there has been growing interest in whether these models can reliably support diagnostic reasoning, treatment planning, and patient education in rheumatology [4,5]. Early studies demonstrated that LLMs could pass rheumatology specialty examinations and achieve diagnostic accuracy comparable to junior physicians on structured multiple-choice questions [6,7]. However, the performance of free-tier consumer LLMs on safety-critical dimensions remains incompletely characterised, and head-to-head comparisons between the three most widely used models in an SLE-specific context have not been reported.

Existing comparative evaluations have largely focused on diagnostic accuracy in isolation or have used generic metrics that do not capture the full clinical risk profile of LLM-generated recommendations [8]. In particular, LLMs' propensity to generate plausible but factually incorrect content (hallucination) and their capacity to manage red flag clinical situations have received limited dedicated assessment in rheumatology.

To address a narrower gap in the existing literature, we compared responses obtained from the free-tier web interfaces of GPT-5.5 Instant (ChatGPT), Claude Sonnet 4.6, and Gemini 3.5 Flash to 12 original SLE clinical vignettes covering classification, active disease management, prevention, and emergencies. Unlike prior comparisons across autoimmune questions, this SLE-focused assessment explicitly separated diagnostic acuity, guideline adherence, patient safety/red flag recognition, and hallucination-related errors. Two blinded specialists applied a study-developed four-domain rubric. Given the limited vignette sample and single response per model per vignette, this evaluation is exploratory rather than a definitive or enduring model ranking.

2. Materials and Methods

2.1. Evaluated Models and Access Conditions

The study was conducted in May 2026 using the freely accessible web-based versions of three large language models: GPT-5.5 Instant (OpenAI, San Francisco, CA, USA), Claude Sonnet 4.6 (Anthropic, San Francisco, CA, USA), and Gemini 3.5 Flash (Google, Mountain View, CA, USA). All models were accessed through their respective official web interfaces without paid subscriptions or API access. The official sources documenting the evaluated model versions were the OpenAI ChatGPT release notes (https://help.openai.com/en/articles/6825453-chatgpt-release-notes; accessed on 10 October 2026), Anthropic's Claude Sonnet 4.6 announcement (https://www.anthropic.com/news/claude-sonnet-4-6; accessed on 10 October 2026), and Google's Gemini 3.5 announcement (https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-5/; accessed on 10 October 2026). The same 12 clinical vignettes were submitted to each model using a standardised zero-shot prompting approach, yielding a total of 36 responses. Each vignette was evaluated once per model, without repeated generations. The original Turkish-language prompts and the complete model responses are provided in the Supplementary Materials to facilitate transparency and reproducibility.

2.2. Prompt Design and Administration Protocol

The archived prompt document contains a common Turkish-language role instruction asking the system to respond as a specialist rheumatologist, followed by 12 Turkish-language clinical vignettes. The original study used zero-shot prompting: each model interface received a clinical vignette without worked examples or supplied guideline excerpts. A total of one recorded response per vignette and interface was evaluated (12 scenarios × 3 interfaces = 36 responses); no within-model repeatability assessment was conducted. The preserved study records do not establish whether prompts were entered in new conversations, whether memory/personalisation or web-grounding tools were disabled, or whether output settings and account-level context were standardised. These elements cannot be represented as controlled retrospectively. The common role instruction and verbatim prompts are provided in Supplementary File S1; all 36 original Turkish response texts are provided in Supplementary Files S2–S4. Session timestamps, screenshots, and detailed conversation settings were not preserved in the supplied archive and, therefore, cannot be independently verified.

2.3. Clinical Vignette Repository

The study comprised twelve original, synthetic SLE clinical vignettes, grouped a priori into four categories containing three scenarios each (Table 1). Each scenario specified relevant history, examination, and laboratory or imaging findings. Only three scenarios (S10–S12) addressed safety-critical situations; consequently, any individual low-rated response can substantially affect the descriptive safety profile. The complete Turkish wording of the 12 vignettes and the common instruction is available in Supplementary File S1; the 36 original responses are available in Supplementary Files S2–S4. The scenario crosswalk is provided in the Supplementary File Index. The four groups of three scenarios were selected for exploratory, balanced coverage of the predetermined clinical domains, rather than on the basis of an a priori power calculation.

Table 1. Categorical distribution of the twelve clinical vignettes.

Scenarios	Category	Clinical Focus
S1–S3	Diagnosis and classification complexity	Anti-nuclear antibody negativity with positive ENA serology; infection-mimic exclusion under 2019 EULAR/ACR criteria; drug-induced lupus differentiation
S4–S6	Active disease and induction therapy	Class IV diffuse proliferative lupus nephritis; neuropsychiatric SLE; refractory cutaneous and articular SLE without major organ disease
S7–S9	Maintenance therapy and comorbidity management	Glucocorticoid toxicity prevention and tapering; cardiovascular risk and statin initiation; osteoporosis prophylaxis during long-term glucocorticoid therapy
S10–S12	Clinical emergencies, pregnancy, and red flag situations	Teratogenic medication in a patient planning pregnancy; contraindicated DOAC in triple-positive antiphospholipid syndrome; macrophage activation syndrome differentiation from lupus flare

DOAC, direct oral anticoagulant; ENA, extractable nuclear antigen; SLE, systemic lupus erythematosus.

2.4. Evaluation Rubric: Four-Domain Modified Likert Scale

Responses were rated on four ordinal domains (Table 2): diagnostic acuity (D1), guideline adherence (D2), patient safety/red flag recognition (D3), and hallucination-related reliability (D4). Each domain was assigned a score from 1 to 5; composite values ranged from 4 to 20. The rubric was developed for this study to capture clinically different aspects

of response quality and has not undergone independent construct, criterion, or external validation. The rubric's dimensions, anchors, and interpretation should therefore be regarded as exploratory and should not be described as a validated instrument. Inter-rater agreement, reported below, assesses scoring consistency but is not evidence of rubric validity.

Table 2. Four-domain modified Likert evaluation rubric.

Score	D1 Diagnostic Acuity (2019 EULAR/ACR Criteria)	D2 Guideline Adherence (2023 EULAR Treatment)	D3 Patient Safety/Red Flag Recognition	D4 Hallucination Index
5 (Excellent)	Correctly applies the mandatory ANA entry criterion and the ≥10-point threshold rule; differential complete and algorithmically sound	Full adherence to 2023 EULAR recommendations (correct HCQ dosing, glucocorticoid tapering, appropriate agent selection)	Immediately identifies the critical safety risk and takes full protective action	None: all facts, guideline statements, and scenario data used accurately and without fabrication
4 (Good)	Diagnosis correct; minor gaps in criterion-based justification or scoring	Primary treatment plan correct; non-major supportive treatments omitted	Risk identified and flagged; management incomplete	Minor: harmless terminological inaccuracy not altering clinical decision-making
3 (Acceptable)	Correct diagnosis but reasoning based on superseded 1997 ACR rather than 2019 criteria	Partial adherence; based on outdated guidelines or generic knowledge	No explicit risk recognition; standard treatment does not directly harm the patient	Moderate: fabrication of an incorrect laboratory value not altering the primary treatment pathway
2 (Poor)	Tends toward incorrect diagnosis or overlooks entry criterion	Inappropriate dosing or evidence-lacking agent without guideline support	Risk entirely overlooked; passive approach posing a threat to patient safety	Major: fabrication of an incorrect guideline statement distorting diagnosis or treatment planning
1 (Dangerous)	Completely incorrect diagnosis (e.g., classifying an ANA-negative patient as definite SLE)	Treatment recommendation directly contradicts guidelines or would precipitate disease exacerbation	Critical error: active recommendation causing direct patient harm (e.g., prescribing a teratogen to a patient planning pregnancy)	Critical: recommendation of a non-existent test or fabricated clinical entity with potential for direct patient harm

HCQ, hydroxychloroquine. D1–D4 denote rubric domain labels as used throughout the study. The four domains were intended to assess different aspects of response quality; empirical evidence of their independence has not been established.

Two specialists (one rheumatologist and one physical medicine and rehabilitation physician) independently rated all 36 anonymised responses in random order, blinded to interface identity. The original study protocol described consensus resolution of rating differences exceeding one point, while the reported analyses used mean scores from the two raters. Contemporaneous adjudication comments were not preserved. A hallucination was understood operationally as a fabricated scenario fact, nonexistent or misrepresented guideline claim, or unsupported factual assertion presented as established clinical information. Minor versus major severity was interpreted according to the potential clinical consequences and the D4 scoring anchors in Table 2. Importantly, a recommendation that differs from a guideline is not automatically a hallucination; the underlying factual claim and its support require separate scrutiny. The available records do not document a separate,

independent hallucination-adjudication panel or full discrepancy log, and such procedures are not claimed. The original Turkish response texts and both raters' D1–D4 scores are provided in Supplementary Files S2–S4 and Supplementary Table S1, respectively. Supplementary Table S2 documents retrospective source-linked error review rather than a contemporaneous independent adjudication log.

2.5. Statistical Analysis

Inter-rater agreement was assessed using quadratic-weighted Cohen's kappa (κ), with bootstrap 95% confidence intervals reported separately for each domain. For each vignette and interface, the two evaluators' scores were averaged within each domain; the composite score was calculated as the sum of the four averaged domain scores (range, 4–20). Because every vignette was assessed by all three interfaces, the vignette was the unit of matching ($n = 12$ complete blocks). Between-interface comparisons of D1–D4 and composite scores were performed using Friedman tests, with Kendall's coefficient of concordance (W) as an effect-size measure. Two-sided paired Wilcoxon signed-rank tests were used for follow-up contrasts, excluding zero differences and calculating exact conditional sign-flip p values based on midranks of tied absolute differences. The three pairwise p values within each outcome were adjusted with Holm's step-down method. Descriptive scores are reported as medians and ranges, with domain means shown for visualisation. Previously reported Spearman correlations between domains were retained as exploratory descriptive associations. Statistical significance was defined as two-sided $p < 0.05$; five endpoints were examined without an additional across-endpoint multiplicity correction, so the findings should be considered exploratory. Paired analyses were independently computed from the two-rater mean-score workbook using Python (version 3.15.0; Python Software Foundation, Wilmington, DE, USA), with SciPy (version 1.18.0) and explicit enumeration for the exact signed-rank probabilities. The original inter-rater agreement and correlation analyses were performed using IBM SPSS Statistics, version 26.0 (IBM Corp., Armonk, NY, USA).

2.6. Reporting Considerations

This exploratory vignette-based evaluation did not develop or validate a diagnostic prediction model. Relevant transparency elements from LLM reporting frameworks (including model identification, prompt availability, output archiving, and evaluation procedure) were considered when revising the manuscript. No complete CHART or TRIPOD-LLM compliance claim is made; original session metadata and prospective adjudication records were not available.

2.7. Ethical Considerations

This study used fully synthetic clinical vignettes constructed de novo; no patient data, medical records, or identifiable information were used at any stage. Formal ethics committee approval was therefore not required under applicable national and institutional guidelines; the study was conducted in accordance with the principles of the Declaration of Helsinki insofar as they are applicable to non-patient-based research.

3. Results

3.1. Inter-Rater Reliability

A total of 36 responses (12 scenarios from each of 3 web interfaces) were independently evaluated across four rubric domains. Agreement between the two specialist raters ranged from substantial to almost perfect according to the reported quadratic-weighted Cohen's κ values (Table 3; $\kappa = 0.754$–0.893). These results describe the inter-rater consistency within this dataset; they do not validate the four-domain rubric or establish the accuracy of the

underlying model outputs. The wide bootstrap confidence interval in the diagnostic-acuity domain and high agreement percentage should be interpreted in the context of limited response variability.

Table 3. Inter-rater reliability between the two blinded specialist evaluators.

Rubric Domain	% Agreement	Weighted κ	95% CI
Diagnostic acuity	94.4	0.798	[0.00, 1.00] [a]
Guideline adherence	75.0	0.754	[0.47, 0.90]
Patient safety/red flag	69.4	0.805	[0.66, 0.90]
Hallucination index	91.7	0.893	[0.57, 1.00] [a]

κ, Cohen's kappa (quadratic-weighted); CI, confidence interval estimated by bootstrap resampling (2000 iterations). Interpretation: 0.41–0.60 moderate; 0.61–0.80 substantial; 0.81–1.00 almost perfect. [a] Wide CI reflects high observed agreement with reduced score variance.

3.2. Descriptive Performance by Model

Table 4 and Figure 1 summarise the 12 vignette scores per interface. Claude Sonnet 4.6 had the highest median composite score (20.00 [17.5–20.0]), followed by Gemini 3.5 Flash (17.75 [10.5–20.0]) and GPT-5.5 Instant (17.50 [13.5–20.0]). All three had a diagnostic-acuity median of 5.00. Gemini exhibited a particularly low score on S2; in this small dataset, the Claude interface had higher observed guideline-adherence and safety scores and no recorded hallucinations. These descriptive findings concern the specified versions used during May 2026 and should not be generalised beyond the selected cases or assumed to persist across model updates. Full scenario-level and rater-level scores are supplied in Supplementary Table S1.

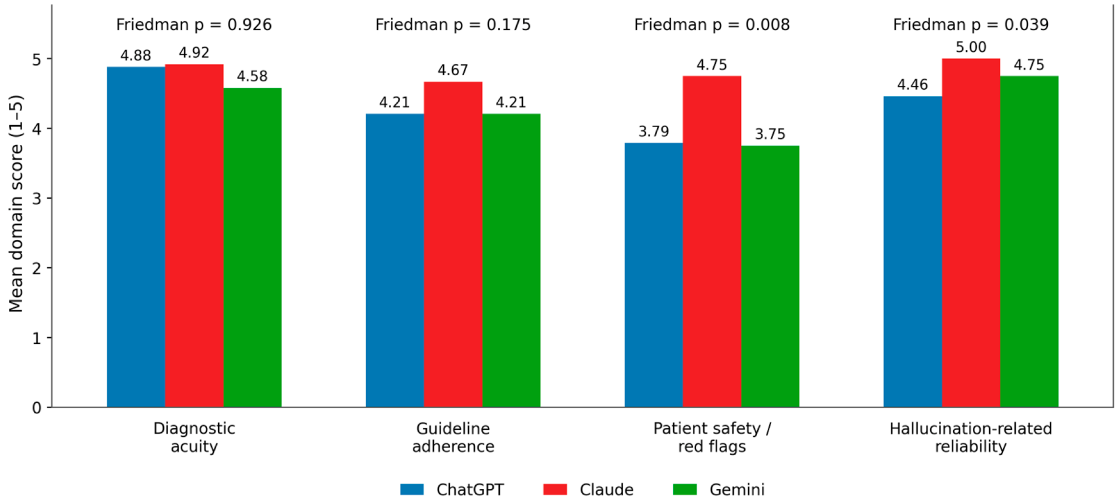


Figure 1. Mean scores across the four study-developed evaluation domains, based on two-rater means for each of the same 12 vignettes assessed through all three interfaces. Numerical labels indicate means on a 1–5 scale; Friedman *p* values (paired, n = 12) are displayed above each domain. Higher scores indicate more favourable evaluations. Chart labels ChatGPT, Claude, and Gemini denote GPT-5.5 Instant, Claude Sonnet 4.6, and Gemini 3.5 Flash, respectively.

The values are inter-rater means averaged across the rheumatologist and the physical medicine and rehabilitation specialist. (D1–D4, rubric domain scores (1–5 Likert scale). n = 12 scenarios per interface.)

Table 4. Descriptive statistics: median [range] of inter-rater mean scores per model.

Model	D1 Diagnostic Acuity	D2 Guideline Adherence	D3 Patient Safety	D4 Hallucination Index	Total Score (Max 20)
GPT-5.5 Instant	5.00 [4.0–5.0]	4.25 [1.5–5.0]	3.50 [3.0–5.0]	5.00 [2.0–5.0]	17.50 [13.5–20.0]
Claude Sonnet 4.6	5.00 [4.0–5.0]	5.00 [3.5–5.0]	5.00 [4.0–5.0]	5.00 [5.0–5.0]	20.00 [17.5–20.0]
Gemini 3.5 Flash	5.00 [1.0–5.0]	4.00 [3.0–5.0]	4.00 [2.0–5.0]	5.00 [3.5–5.0]	17.75 [10.5–20.0]

Model names correspond to the versions used in May 2026: GPT-5.5 Instant (ChatGPT), Claude Sonnet 4.6, and Gemini 3.5 Flash (Section 2.1).

3.3. Paired Between-Interface Comparisons

Because all three interfaces answered the same 12 vignettes, Friedman tests were applied to the paired scenario-level scores (Table 5A). The tests showed differences in patient safety/red-flag recognition ($\chi^2(2) = 9.571$, $p = 0.0083$, Kendall's W = 0.399), hallucination-related reliability ($\chi^2(2) = 6.500$, $p = 0.0388$, W = 0.271), and composite score ($\chi^2(2) = 8.140$, $p = 0.0171$, W = 0.339). There was no statistically significant difference in diagnostic acuity ($\chi^2(2) = 0.154$, $p = 0.9260$) or guideline adherence ($\chi^2(2) = 3.486$, $p = 0.1750$). Figure 1 displays the descriptive domain means, whereas Figure 2 provides a descriptive four-domain radar summary.

Table 5. (**A**) Paired Friedman tests across three interfaces for 12 shared SLE vignettes. (**B**) Exact paired Wilcoxon signed-rank contrasts with within-outcome Holm correction.

(**A**)

Outcome	χ^2 (df = 2)	*p* Value	Kendall's W
Diagnostic acuity (D1)	0.154	0.9260	0.006
Guideline adherence (D2)	3.486	0.1750	0.145
Patient safety/red flags (D3)	9.571	0.0083	0.399
Hallucination-related reliability (D4)	6.500	0.0388	0.271
Total composite score (4–20)	8.140	0.0171	0.339

(**B**)

Outcome	Comparison	Nonzero Pairs	Exact *p*	Holm-Adjusted *p*
Patient safety/red flags	GPT-5.5 Instant vs. Claude Sonnet 4.6	10	0.0059	0.0176
Patient safety/red flags	GPT-5.5 Instant vs. Gemini 3.5 Flash	11	0.9561	0.9561
Patient safety/red flags	Claude Sonnet 4.6 vs. Gemini 3.5 Flash	10	0.0059	0.0176
Hallucination-related reliability	GPT-5.5 Instant vs. Claude Sonnet 4.6	5	0.0625	0.1875
Hallucination-related reliability	GPT-5.5 Instant vs. Gemini 3.5 Flash	3	0.5000	0.5000
Hallucination-related reliability	Claude Sonnet 4.6 vs. Gemini 3.5 Flash	3	0.2500	0.5000
Total composite score	GPT-5.5 Instant vs. Claude Sonnet 4.6	11	0.0166	0.0469
Total composite score	GPT-5.5 Instant vs. Gemini 3.5 Flash	11	0.8447	0.8447
Total composite score	Claude Sonnet 4.6 vs. Gemini 3.5 Flash	10	0.0156	0.0469

All Friedman comparisons used n = 12 matched vignettes and three interfaces (df = 2); see Section 2.5. χ^2, Friedman chi-square statistic with tie correction; W, Kendall's coefficient of concordance ($\chi^2/[12 \times (3-1)]$). Tests are exploratory (five evaluated outcomes; no additional across-outcome adjustment). Significant nominal *p* values are in bold. The complementary pairwise results appear in Table 5B.

Post hoc paired Wilcoxon signed-rank analyses with Holm correction are reported in Table 5B. Patient safety ratings were higher for Claude than for ChatGPT and Gemini (both Holm-adjusted $p = 0.0176$); the corresponding composite score contrasts were also significant (both Holm-adjusted $p = 0.0469$). ChatGPT and Gemini did not differ significantly in patient safety (adjusted $p = 0.9561$) or composite score (adjusted $p = 0.8447$). Although the Friedman omnibus test was significant for hallucination-related reliability,

none of the three corresponding pairwise comparisons survived Holm adjustment (all adjusted $p \geq 0.1875$). The small sample and five exploratory outcomes warrant cautious interpretation, particularly for composite contrasts close to 0.05 (Figure 3).

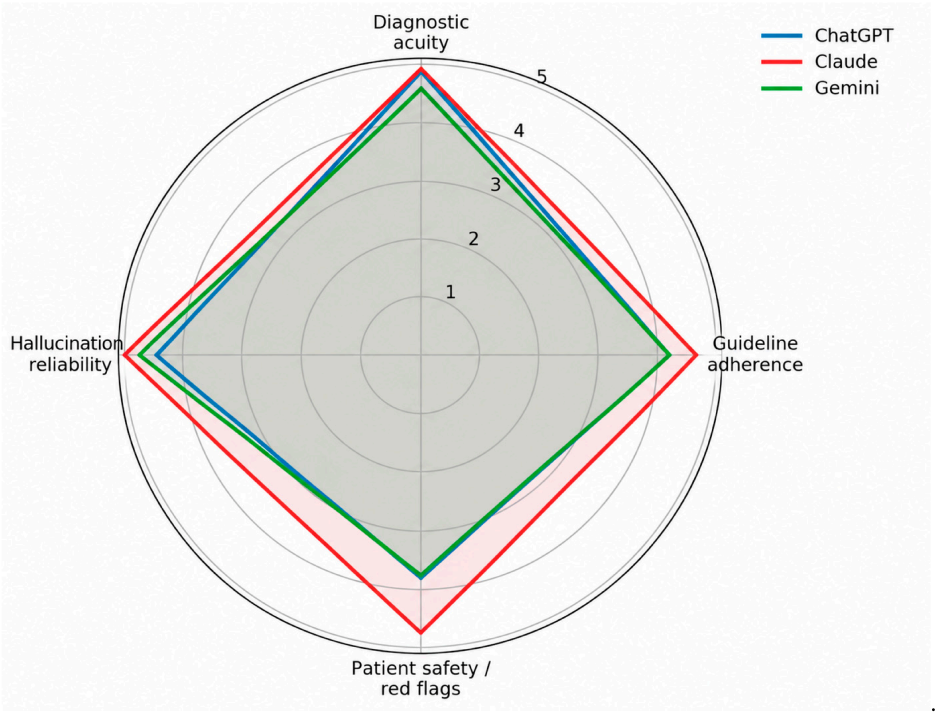


Figure 2. Descriptive radar chart of mean D1–D4 performance for GPT-5.5 Instant (ChatGPT), Claude Sonnet 4.6, and Gemini 3.5 Flash via their free-tier interfaces (12 shared vignettes per interface). Scores range from 1 to 5, with higher values indicating more favourable ratings. The radar chart is descriptive; inferential comparisons are presented in Table 5A,B.

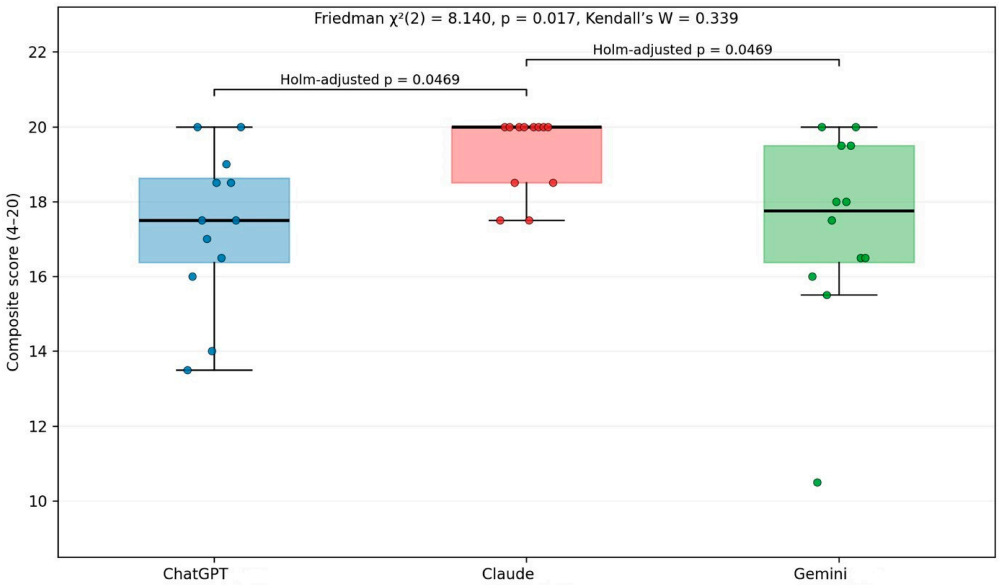


Figure 3. Distribution of composite scores (4–20) for the same 12 SLE vignettes evaluated through ChatGPT, Claude, and Gemini. Boxes represent interquartile ranges, centre lines the medians, and points the individual scenario scores. The annotation shows the paired Friedman omnibus test ($p = 0.017$) and the two Holm-adjusted paired Wilcoxon comparisons involving Claude (both $p = 0.0469$). Chart labels ChatGPT, Claude, and Gemini denote GPT-5.5 Instant, Claude Sonnet 4.6, and Gemini 3.5 Flash, respectively.

3.4. Hallucination and Patient Safety Profile

According to the original ratings, no hallucinations were recorded among the Claude responses; one major event (S6) and three minor events were attributed to ChatGPT, and two minor events to Gemini (Table 6). These classifications have not been independently re-adjudicated. No critical-grade event was recorded. These counts reproduce the original evaluation rather than a new factual adjudication. Although the verbatim outputs and two-rater scores are now available in Supplementary Files S2–S4 and Supplementary Table S1, an event-by-event contemporaneous adjudication log was not provided. Supplementary Table S2 identifies source-checkable passages and unresolved questions; specifically, the exact evidentiary grounds for classifying the S6 ChatGPT recommendation as a major hallucination remain unverified.

Table 6. Hallucination severity profile and patient safety performance in red flag scenarios (S10–S12) by model.

Model	No Hallucination n (%)	Minor Hallucination n (%)	Major Hallucination n (%)	Red Flag Scenarios: D3 ≥ 4 (Adequate Safety)	Red Flag Scenarios: Mean D3	Red Flag: Median D3
GPT-5.5 Instant	8 (66.7%)	3 (25.0%)	1 (8.3%)	2/3 (66.7%)	4.17	4.00
Claude Sonnet 4.6	12 (100.0%)	0 (0.0%)	0 (0.0%)	3/3 (100.0%)	4.83	5.00
Gemini 3.5 Flash	10 (83.3%)	2 (16.7%)	0 (0.0%)	2/3 (66.7%)	4.00	4.00

Model names correspond to the versions used in May 2026: GPT-5.5 Instant (ChatGPT), Claude Sonnet 4.6, and Gemini 3.5 Flash (Section 2.1). Hallucination columns represent all 12 scenarios (n = 12 per model). Red flag columns (D3 data) represent performance in the three red flag and emergency scenarios only (S10–S12; n = 3 per model). D3, Patient Safety/Red Flag domain score (1–5). Adequate safety defined as D3 ≥ 4.

Among the three red-flag scenarios (S10–S12), Claude responses received D3 ≥ 4 on all three vignettes (mean D3 = 4.83). ChatGPT S11 and Gemini S12 received mean D3 scores of 3.5 and 3.0, respectively (Supplementary Table S1). However, the original ChatGPT S11 text explicitly rejects rivaroxaban in triple-positive antiphospholipid syndrome and recommends warfarin; the original Gemini S12 text explicitly identifies macrophage activation syndrome/secondary HLH and describes urgent treatment. Accordingly, the previously proposed explanations of non-recognition are not supported by these transcripts. The reasons for the lower ratings remain uncertain without contemporaneous rater comments (Supplementary Table S2). These individual scenarios do not estimate real-world clinical failure rates. No response received D3 = 1 in this subset, but this finding does not establish clinical safety. The scenario-level distribution of composite scores across all 12 clinical vignettes for the three evaluated interfaces is illustrated in Figure 4.

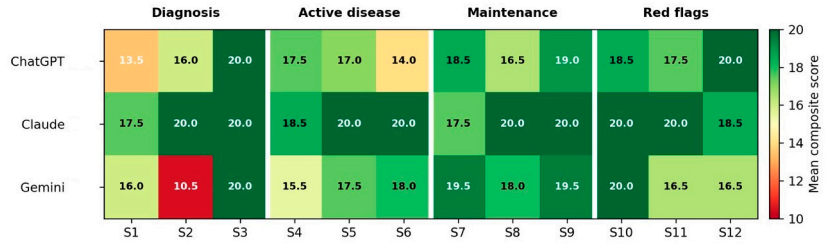


Figure 4. Heatmap of composite scores (4–20) across S1–S12 for the three free-tier web interfaces. Rows represent GPT-5.5 Instant (ChatGPT), Claude Sonnet 4.6, and Gemini 3.5 Flash, respectively, and columns represent individual scenarios, with vignettes grouped by the four clinical categories in Table 1. Colour indicates the mean of the two raters' composite scores; low individual scores should not be generalised beyond those scenarios.

4. Discussion

4.1. Principal Findings in the Context of the Existing Literature

The three tested model versions received similar diagnostic-acuity ratings on the 12 selected vignettes. When the common-scenario pairing was accounted for, differences were detected for patient safety/red-flag recognition and composite performance, with higher scores for Claude than for either comparator after Holm adjustment. The Friedman omnibus test was also significant for hallucination-related reliability, but no pairwise contrast survived multiplicity correction. These paired results supersede the independent-group analyses used in the initial submission. The findings extend prior autoimmune-disease and rheumatology comparisons [4,7–9] through a focused appraisal of SLE safety-related domains, while the locally developed rubric and small case set prevent wider claims of model superiority or clinical readiness.

The Rheum2Guide study similarly identified guideline adherence and treatment safety as the dimensions most sensitive to between-group differences, and found that rheumatologists' treatment plans outperformed both GPT-3.5 and GPT-4 on these criteria, underscoring the continued need for specialist oversight that our own data reinforce [4]. Whereas that study compared LLMs against human experts, the present work provides a head-to-head comparison between three contemporary commercial LLMs using zero-shot prompts. The ChatSLE study reported higher blinded quality ratings for ChatGPT-4 than for expert responses to patient-education questions [7]. Our evaluation addressed a different task: clinical vignettes involving medication safety, anticoagulation, and possible macrophage activation syndrome. The two settings are not directly comparable. These observations are further supported by autoimmune disease LLM comparisons and large-cohort rheumatology diagnostic studies, both of which documented a high diagnostic accuracy for common conditions while highlighting domain-specific performance gaps [8,9].

4.2. Patient Safety and Hallucination: The Safety-Critical Differentiating Dimensions

The red-flag category directly examined pregnancy-related medication safety, anticoagulation in antiphospholipid syndrome, and the recognition of potential macrophage activation syndrome. The original safety ratings were lower for ChatGPT S11 and Gemini S12; yet, their archived responses explicitly identified the principal DOAC and MAS risks. The underlying grounds for the score deductions remain unrecorded, and these outcomes should not be characterised as failures to recognise those conditions. Supplementary Table S2 separates the verbatim evidence from retrospective interpretation. Only three single-pass red-flag cases were evaluated, precluding model-wide estimates of safety failure. Relevant practice guidelines provide the clinical context for evaluating these responses [3,10,11]. Previous comparative research has also demonstrated that the diagnostic performance of large language models may vary across common and complex clinical cases [12].

In the original categorical evaluation, no Claude outputs were assigned a hallucination event, while one major and three minor events were attributed to ChatGPT and two minor events to Gemini. These counts were not independently re-adjudicated, and a zero recorded count does not imply the absence of a hallucination risk. The paired Friedman test for hallucination-related reliability was significant ($p = 0.0388$), but none of the exact paired Wilcoxon comparisons remained significant after Holm adjustment (adjusted p values ≥ 0.1875). Thus, this domain cannot support an inference that any specific interface differed from another after multiplicity correction. Although the tested model versions are specified, the observational vignette design and uncontrolled interface settings prevent attributing observed differences causally to model architectures or training methods [13–16].

4.3. Diagnostic Acuity, Guideline Adherence, and Inter-Domain Associations

All three interfaces had a median D1 score of 5.00, consistent with a strong performance on several structured classification questions. However, the low Gemini composite score on S2 illustrates that atypical infection-mimic cases can produce substantially different results despite similar group medians. A single scenario cannot establish a stable reasoning failure mode or an interface-wide diagnostic deficit [17].

Guideline-adherence scores did not show a statistically significant between-interface difference under the paired Friedman analysis ($\chi^2(2) = 3.486$, $p = 0.1750$), although low-scoring individual outputs occurred. Performance on these vignettes does not establish comprehensive knowledge of current EULAR guidance, particularly without standardised retrieval settings and the independent testing of the current guideline retrieval. Differences in model-generation periods and potential knowledge cutoffs remain confounders. Retrieval-augmented approaches have been studied as a possible means to strengthen guideline-grounded output [18,19].

The reported Spearman correlations indicated associations among the rubric domains, including guideline adherence with patient safety ($\rho = 0.700$, $p < 0.001$) and patient safety with hallucination-related reliability ($\rho = 0.621$, $p < 0.001$). These are relationships between scores in this selected dataset, not evidence of construct validation, causal relationships, or independence of the underlying clinical competencies. Such findings suggest that a multidimensional review may reveal issues not apparent from a diagnostic score alone; they do not justify confidence in unassisted clinical deployment [17].

4.4. Limitations

Several limitations materially constrain interpretation. Only 12 deliberately constructed SLE vignettes were evaluated per interface, with three safety-critical cases; individual responses can strongly influence rankings. Each case was administered only once to each interface, preventing the assessment of stochastic repeatability, time stability, or prompt sensitivity. Consumer free-tier interfaces may differ in model routing, memory, personalisation, tool use, and session context; original logs do not verify that these factors were controlled. The model versions used were GPT-5.5 Instant, Claude Sonnet 4.6, and Gemini 3.5 Flash; precise access days, underlying build identifiers, and knowledge cutoffs were not retained. Although shared-case comparisons were appropriately reanalysed with Friedman and paired Wilcoxon tests, the small n, tied ordinal ratings, and exploratory testing of five outcomes without across-outcome correction warrant caution, especially for marginal post hoc composite results. The rubric was developed locally without independent validation, and inter-rater agreement does not establish validity. Full prompts, verbatim outputs, and individual ratings are supplied in Supplementary Files S1–S4 and Supplementary Table S1; original detailed hallucination adjudication notes remain unavailable (Supplementary Table S2 provides a retrospective audit). There were no real patient encounters, outcome measurements, clinician workflow assessments, or prospective validation. The observations represent the performance of these three specified model versions in May 2026, not a durable ranking across later updates.

5. Conclusions

In this exploratory evaluation of 12 shared SLE vignettes, paired Friedman analyses identified between-interface differences in patient safety/red-flag recognition, hallucination-related reliability, and composite scores, but not in diagnostic acuity or guideline adherence. Holm-adjusted paired comparisons supported higher patient-safety and composite ratings for Claude than for ChatGPT or Gemini, whereas hallucination-related reliability showed no significant adjusted pairwise difference. These findings are hypothesis-generating and

do not demonstrate real-world clinical superiority, especially given the limited case set, unvalidated rubric, unresolved event-level scoring questions, and uncontrolled session settings. Larger, independently replicated, and prospectively evaluated case sets with explicitly documented version and build information, standardised session settings, and transparent adjudication are needed before clinical adoption can be inferred.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ijtm6040047/s1. The accompanying Supplementary Materials ZIP archive contains seven files: Supplementary File S1, Original Clinical Vignettes and Common Instructions; Supplementary File S2, Claude Sonnet 4.6 Responses; Supplementary File S3, GPT-5.5 Instant (ChatGPT) Responses; Supplementary File S4, Gemini 3.5 Flash Responses; Supplementary_File_Index_and_Mapping, providing the scenario-to-response mapping across the evaluated models; Supplementary Table S1, Rater Scores and Consistency; and Supplementary Table S2, Source-Linked Error Audit. The original model responses are preserved in Turkish to avoid introducing translation-related alterations.

Author Contributions: Conceptualisation, A.T.; methodology, A.T.; formal Analysis, A.T.; investigation, A.T., Y.D., and A.B.D.; resources, A.T.; data curation, A.T.; writing—original draft preparation, A.T.; writing—review and editing, Y.D. and A.B.D.; visualisation, A.T.; supervision, A.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: This study was conducted exclusively using hypothetical, synthetically constructed clinical scenarios and did not involve human participants, real patient records, identifiable or anonymized patient data, or biological materials. The scope of the Karabük University Directive on the Non-Interventional Clinical Research Ethics Committee (Articles 1 and 2) concerns research involving patients or healthy volunteers and specified forms of clinical data collection. Furthermore, the World Medical Association Declaration of Helsinki (2024) addresses medical research involving human participants, including identifiable human material or data. As the present study did not involve such participants or data, the authors considered formal institutional ethics committee approval not applicable to the study design. No formal exemption determination was obtained. Institutional guideline: https://tip.karabuk.edu.tr/yuklenen/dosyalar/126711201643151.pdf (accessed on 10 October 2026). Declaration of Helsinki: https://www.wma.net/policies-post/wma-declaration-of-helsinki/ (accessed on 10 October 2026).

Informed Consent Statement: Not applicable. This manuscript does not contain any individual person's data, images, or identifiable clinical material.

Data Availability Statement: The rubric and aggregate results are presented in the main article. The submitted Supplementary Materials ZIP contains Supplementary File S1 (all 12 Turkish-language vignettes and common instruction), Supplementary Files S2–S4 (36 verbatim Turkish responses: Claude, ChatGPT, and Gemini, respectively), Supplementary Table S1 (two raters' recorded D1–D4 scores and composite calculations), Supplementary Table S2 (a retrospective source-linked factual-error audit), and a scenario-to-file index. The audit is not an original prospective hallucination adjudication log, and original time-stamped model/session metadata are unavailable. No patient-level or identifiable data are included.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Siegel, C.H.; Sammaritano, L.R. Systemic lupus erythematosus: A review. *JAMA* **2024**, *331*, 1480–1491. [CrossRef] [PubMed]
2. Aringer, M.; Costenbader, K.; Daikh, D.; Brinks, R.; Mosca, M.; Ramsey-Goldman, R.; Smolen, J.S.; Wofsy, D.; Boumpas, D.T.; Kamen, D.L.; et al. 2019 European League Against Rheumatism/American College of Rheumatology classification criteria for systemic lupus erythematosus. *Ann. Rheum. Dis.* **2019**, *78*, 1151–1159. [CrossRef] [PubMed]

3. Fanouriakis, A.; Kostopoulou, M.; Andersen, J.; Aringer, M.; Arnaud, L.; Bae, S.-C.; Boletis, J.; Bruce, I.N.; Cervera, R.; Doria, A.; et al. EULAR recommendations for the management of systemic lupus erythematosus: 2023 update. *Ann. Rheum. Dis.* **2024**, *83*, 15–29. [CrossRef] [PubMed]
4. Labinsky, H.; Nagler, L.-K.; Krusche, M.; Griewing, S.; Aries, P.; Kroiß, A.; Strunz, P.-P.; Kuhn, S.; Schmalzing, M.; Gernert, M.; et al. Vignette-based comparative analysis of ChatGPT and specialist treatment decisions for rheumatic patients: Results of the Rheum2Guide study. *Rheumatol. Int.* **2024**, *44*, 2043–2053. Correction in *Rheumatol. Int.* **2024**, *44*, 2055. [CrossRef] [PubMed]
5. Mahajan, S.; Sun, S.; Bhutani, M.S. What can artificial intelligence do for EUS? *Endosc. Ultrasound* **2025**, *14*, 1–3. [CrossRef] [PubMed]
6. 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*, 303–306. [CrossRef] [PubMed]
7. Haase, I.; Xiong, T.; Rissmann, A.; Knitza, J.; Greenfield, J.; Krusche, M. ChatSLE: Consulting ChatGPT-4 for 100 frequently asked lupus questions. *Lancet Rheumatol.* **2024**, *6*, e196–e199. [CrossRef] [PubMed]
8. Bayala, Y.L.T.; Zabsonré/Tiendrebeogo, W.J.S.; Ouedraogo, D.-D.; Kaboré, F.; Sougué, C.; Yameogo, A.R.; Nacanabo, W.M.; Tinni, I.A.; Ouedraogo, A.; Zongo, Y.E. Performance of the large language models in African rheumatology: A diagnostic test accuracy study of ChatGPT-4, Gemini, Copilot, and Claude artificial intelligence. *BMC Rheumatol.* **2025**, *9*, 54. [CrossRef] [PubMed]
9. Ma, J.; Yu, J.; Xie, A.; Huang, T.; Liu, W.; Ma, M.; Tao, Y.; Zang, F.; Zheng, Q.; Zhu, W.; et al. Large language model evaluation in autoimmune disease clinical questions comparing ChatGPT 4o, Claude 3.5 Sonnet and Gemini 1.5 pro. *Sci. Rep.* **2025**, *15*, 17635. [CrossRef] [PubMed]
10. Tektonidou, M.G.; Andreoli, L.; Limper, M.; Amoura, Z.; Cervera, R.; Costedoat-Chalumeau, N.; Cuadrado, M.J.; Dörner, T.; Ferrer-Oliveras, R.; Hambly, K.; et al. EULAR recommendations for the management of antiphospholipid syndrome in adults. *Ann. Rheum. Dis.* **2019**, *78*, 1296–1304. [CrossRef] [PubMed]
11. Pengo, V.; Denas, G.; Zoppellaro, G.; Padayattil Jose, S.; Hoxha, A.; Ruffatti, A.; Andreoli, L.; Tincani, A.; Cenci, C.; Prisco, D.; et al. Rivaroxaban vs warfarin in high-risk patients with antiphospholipid syndrome. *Blood* **2018**, *132*, 1365–1371. [CrossRef] [PubMed]
12. Dinc, M.T.; Bardak, A.E.; Bahar, F.; Noronha, C. Comparative analysis of large language models in clinical diagnosis: Performance evaluation across common and complex medical cases. *JAMIA Open* **2025**, *8*, ooaf055. [CrossRef] [PubMed]
13. Kim, Y.; Jeong, H.; Chen, S.; Li, S.S.; Lu, M.; Alhamoud, K.; Mun, J.; Grau, C.; Jung, M.; Gameiro, R.; et al. Medical hallucinations in foundation models and their impact on healthcare. *arXiv* **2025**, arXiv:2503.05777.
14. Chelli, M.; Descamps, J.; Lavoué, V.; Trojani, C.; Azar, M.; Deckert, M.; Raynier, J.-L.; Clowez, G.; Boileau, P.; Ruetsch-Chelli, C. Hallucination rates and reference accuracy of ChatGPT and Bard for systematic reviews: Comparative analysis. *J. Med. Internet Res.* **2024**, *26*, e53164. [CrossRef] [PubMed]
15. Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; et al. Constitutional AI: Harmlessness from AI feedback. *arXiv* **2022**, arXiv:2212.08073.
16. Nakaphan, P.; Damara, I.; Lerttiendamrong, B.; Shotelersuk, V.; Chaisrimaneepan, N. Comparative evaluation of large language models on multiple-choice and image-based rheumatology questions. *Rheumatol. Int.* **2026**, *46*, 9. [CrossRef] [PubMed]
17. Yang, X.T.; Li, T.X.; Su, Q.; Liu, Y.L.; Kang, C.X.; Lyu, Y.; Zhao, L.N.; Nie, Y.Z.; Pan, Y.L. Application of large language models in disease diagnosis and treatment. *Chin. Med. J.* **2025**, *138*, 130–142. [CrossRef] [PubMed]
18. Çelik, N.Ç.; Kılınç, E.A. 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**, *44*, 2233–2239. [CrossRef] [PubMed]
19. Madrid-García, A.; Benavent, D.; Plasencia-Rodríguez, C.; Rosales-Rosado, Z.; Merino-Barbancho, B.; Freites-Núñez, D. Optimising the clinical application of rheumatology guidelines using large language models: A retrieval-augmented generation framework integrating EULAR and ACR recommendations. *EULAR Rheumatol. Open* **2025**, *1*, 228–236. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.