Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

RheumBench: A Specialist-Validated Rubric-Based Benchmark for Evaluating Large Language Models in Rheumatology

Fabian Lechner
University Marburg, Institute for Digital Medicine
Jonathan Bamberger
University Marburg, Institute for Digital Medicine
Phillip Kremer
Division of Rheumatology and systemic inflammatory disorders, University Medical Center Hamburg-Eppendorf
Jutta Richter
Clinic for Rheumatology, University Hospital Düsseldorf, Medical Faculty of Heinrich-Heine-University
Matthias Schneider
Clinic for Rheumatology, University Hospital Düsseldorf, Medical Faculty of Heinrich-Heine-University
Uta Kiltz
Ruhr-University Bochum
Sebastian Kuhn
University Marburg, Institute for Digital Medicine
Lucie Flek
Bonn-Aachen International Center for Information Technology, University of Bonn
Philipp Klemm
Department of Rheumatology, Clinical Immunology, Osteology and Physical Medicine, Campus Kerckhoff, Justus-Liebig-University Gießen
Axel J Hueber
Division of Rheumatology, Klinikum Nürnberg, Paracelsus Medical University
Martin Krusche
Division of Rheumatology and systemic inflammatory disorders, University Medical Center Hamburg-Eppendorf
Hannah Labinsky
Department of Medicine II, Rheumatology and Clinical Immunology, University Hospital Würzburg
Johannes Knitza
knitza@uni-marburg.de

University Marburg, Institute for Digital Medicine

Research Article

Keywords: Large language models (LLMs), Rheumatology, Clinical decision support, Artificial intelligence, Patient safety

Posted Date: September 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10825548/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: Competing interest reported. Johannes Knitza, Jutta Richter, Matthias Schneider, Uta Kiltz, Martin Krusche, and Hannah Labinsky are members of a project group involved in the development and improvement of the RHELA system. The other authors declare no competing interests.

Abstract

Large language models (LLMs) are increasingly used for clinical decision support, yet scalable and specialty-specific methods for evaluating their performance and safety remain limited. We developed RheumBench, the first rubric-based LLM benchmark in rheumatology, providing a specialist-validated framework for open-ended evaluation. We evaluated ten systems, comprising three frontier proprietary general-purpose LLMs, one frontier open-weight general-purpose LLM, and six clinical retrieval-augmented generation (RAG) systems, including two certified medical products and one rheumatology-specific RAG system. RheumBench comprised 100 clinical cases (70 diagnostic and 30 diverse clinical management) and 1,020 case-specific rubric criteria independently reviewed and weighted for clinical appropriateness and potential harm by two board-certified rheumatologists. Systems were evaluated for answer quality, prompt sensitivity, potential harm, diagnostic accuracy, and system confidence. Responses were scored by a three LLM judge panel, yielding 50,553 rubric-level LLM evaluations, with high agreement with rheumatologist consensus (Cohen's $\kappa = 0.91$). Overall system performance (0-100%) varied substantially, with a maximum RheumBench score of 54.0% (Claude Fable 5), and frontier general-purpose LLMs generally outperformed clinical RAG systems, including certified Prof. Valmed and DR.INFO. Management case performance varied widely, with thorough prompting improving mean management scores in 9 of 10 systems and yielding a peak score of 72.9% with GPT-5.6 Sol. Potentially severe errors occurred with all systems and were predominantly omission errors. Diagnostic top-1 accuracy reached 57.1% (GPT-5.6 Sol), while top-3 accuracy reached 67.1% (Claude Fable 5). Self-reported system confidence was generally high and diagnostic calibration analyses showed systematic overconfidence for all systems, with the lowest Brier score of 0.203 (Claude Fable 5). RheumBench identified substantial differences that were not explained by clinical specialization or medical product certification. These findings highlight limitations of current LLM-based clinical decision support, and underscore the need for human oversight. RheumBench will be continuously expanded and updated to enable ongoing safety analyses and support system improvement.

Introduction

Large language models (LLMs) are increasingly used by healthcare professionals and patients for clinical support and health-related decision-making [1–5]. Their adoption is facilitated by convenient and rapid access to information [6, 7], while successive model generations have shown improving performance on medical tasks [7, 8]. Rheumatology is a complex specialty facing a growing mismatch between workforce capacity and demand from patients and referring clinicians [9], which may further encourage the use of LLM-based support [10]. Early studies have shown promising results in diagnostic reasoning [7, 11, 12], patient communication [13, 14], treatment recommendations [15], and AI-assisted clinical documentation [16]. In some settings, LLMs approached the performance of board-certified rheumatologists [12] and improved the performance of less experienced healthcare professionals [17]. Patient surveys also suggest broad acceptance of physician-supervised AI, with majorities supporting its use in diagnostic and therapeutic decision-making [3]. However, inaccurate, incomplete, or overconfident

recommendations may cause patient harm and promote automation bias, particularly among less experienced users [17, 18].

Not only are LLMs increasingly used for medical purposes, but the LLM ecosystem itself is rapidly expanding to include proprietary frontier models, open-source models, clinical retrieval-augmented generation (RAG) systems, and even certified LLM-based medical products [7, 11]. This trend underscores the need for clinically meaningful evaluation frameworks that allow patients, healthcare professionals, developers, and regulators, to compare systems, guide procurement, identify opportunities for model improvement, and support safe clinical implementation.

Despite the clear need for rigorous evaluation, assessment methods struggle to keep pace with rapid LLM development [19, 20]. Randomized controlled trials are costly and time-consuming, while multiple-choice benchmarks are increasingly saturated and capture clinical reasoning only incompletely [20]. Open-ended clinical questions provide a more realistic assessment, but expert review of free-text responses remains labor-intensive, subjective, and difficult to scale. As a result, models may already be outdated by the time an evaluation is completed. Scalable, representative, and standardized evaluation frameworks are therefore needed [21].

Existing medical benchmarks have advanced the evaluation of open-ended clinical responses and potential harms, but they provide limited insight into rheumatology-specific performance. HealthBench [22] did not include rheumatology among its 26 designated medical specialties, while NOHARM [23] evaluated ten other specialties without rheumatology. Neither benchmark systematically assessed the relationship between self-reported confidence and performance, despite evidence that LLMs may be highly confident in incorrect answers and growing recommendations that calibration be incorporated into medical LLM evaluations [7, 24]. Furthermore, to our knowledge, no previous LLM benchmark has compared multiple certified medical products with general-purpose and clinical RAG systems to determine whether certification is associated with higher performance, better calibration, or greater safety.

To address these gaps, we developed RheumBench, the first rheumatology-specific, specialist-validated benchmark for LLMs and applied it to ten systems. We assessed overall answer quality, sensitivity to prompt design, potential harm, diagnostic accuracy, and calibration.

Methods

The overall methodology (Fig. 1) was largely informed by HealthBench [22] and NOHARM [23]. A formal inquiry was submitted to the Ethics Committee of Philipps University Marburg, Germany (26–266 ANZ). Given the anonymous and non-interventional nature of the study, the committee confirmed that formal institutional review board approval was not required.

Benchmark case curation

A panel of board-certified rheumatologists, including members of the committee for digital rheumatology of the German society for rheumatology and clinical immunology

and the digital rheumatology network, compiled the benchmark cases. Given the central importance of accurate diagnosis in rheumatology, the panel placed a strong emphasis on diagnostic cases and incorporated a previously curated set of 60 published case reports covering rare rheumatic diseases and clinically important differential diagnoses [7]. To reduce the potential impact of model contamination from previously published material, 10 additional de-identified cases from the University Hospital Hamburg were included. This resulted in 70 diagnostic cases, of which 42 (60.0%) represented inflammatory rheumatic diseases and 28 (40.0%) represented clinically relevant non-inflammatory differential diagnoses. Two rheumatologists independently reviewed the cases for completeness and quality of the clinical information. Each case contained core demographic and clinical information, including age, sex, symptoms, and available laboratory, physical examination, and imaging findings. The panel additionally developed clinically relevant management cases. These ranged from detailed, de-identified clinical scenarios with broader management questions to short, focused questions on topics such as treatment selection and dosing. Panel members contributed real-world questions based on consultations from colleagues, questions raised by rheumatologists, their own clinical practice, and physician rheumatology examinations. Four panel members independently rated the relevance of each candidate case on a scale from 0 to 10, and the 30 highest-priority cases were selected. The panel leads (HL and JK) subsequently reviewed the final set to remove thematic overlap and ensure clinical relevance and question quality.

Rubric creation

For each of the 30 management cases, an ensemble of frontier LLMs from different developers, including OpenAI GPT-5.6 Sol (high thinking), Anthropic Claude Fable 5, and Google Gemini 3.1 Pro, independently generated draft answers (Supplementary Tables S1-S2). GPT-5.6 Sol was then used as the designated consolidator model to synthesize these outputs, reconcile areas of agreement and disagreement, and produce a single reference-answer draft for each case. Two board-certified rheumatologists subsequently reviewed and refined each draft. For the diagnostic cases, the reference answer consisted of the correct diagnosis, which was reviewed for clinical plausibility by two board-certified rheumatologists.

Based on the reference answers, the same LLMs generated candidate rubric criteria (see supplementary table S3 for details) and the consolidator model then merged overlapping items and removed duplicates to produce a preliminary fixed rubric. Management-case rubrics contained between 7 and 37 criteria and covered the domains previously defined in NOHARM [23]: diagnostic evaluation, medication, counseling, follow-up, and procedural management. Most criteria represented appropriate clinical actions, although potentially inappropriate actions were also included. Consistent with NOHARM [23], a rubric criterion, “do nothing”, representing the option of taking no clinical action was included where appropriate.

Diagnostic-case rubrics contained five fixed criteria: identification of the correct diagnosis as the top-ranked diagnosis, inclusion of the correct diagnosis among the top three suggestions, and inclusion of each of three clinically important differential diagnoses among the top three suggestions. The three differential diagnoses were initially proposed by an LLM and subsequently reviewed by the rheumatologists.

All proposed rubric criteria were independently reviewed by two board-certified rheumatologists, who could add, revise, or remove LLM-generated items. The final criteria were agreed upon through discussion. Once the rubrics had been finalized, both rheumatologists independently assigned each criterion a weight reflecting its clinical appropriateness and importance. Based on the HealthBench approach [22], weights ranged from − 10 to + 10, with negative values indicating inappropriate actions and positive values indicating appropriate actions. To improve consistency, ratings were restricted to −10, −7, − 4, −1, + 1, +4, + 7, and + 10. The rheumatologists also independently rated the potential harm associated with each criterion as severe, moderate, mild, or uncertain, following an approach adapted from NOHARM [23]. When ratings differed by only one category, the lower rating was selected automatically, for example, + 7 rather than + 10 or mild rather than moderate. Larger disagreements were resolved through discussion. For diagnostic cases, raters were advised to assign + 10 points for ranking the correct diagnosis first, + 7 points for including it among the top three, and generally + 4 points for each plausible differential diagnosis, while retaining discretion to adjust weights according to the individual case. This process resulted in a total of 1020 case-specific rubric criteria. The rubric comprised 877 (86.0%) clinically appropriate criteria with positive weights and 143 (14.0%) clinically inappropriate criteria with negative weights. The rubric criteria covered five clinical domains, including diagnostic evaluation, 569 (55.8%), medication, 262 (25.7%), procedures, 47 (4.6%), counseling, 83 (8.1%), and follow-up, 59 (5.8%). Agreement between rheumatologists was high, with exact agreement for 92.9% of rubric weights and 92.9% of harm ratings and weighted Cohen's κ values of 0.94 and 0.95, respectively. Larger disagreements requiring consensus discussion occurred for 0.4% of weight ratings and 0.4% of harm ratings.

System evaluation

Characteristics of the ten systems assessed are summarized in Table 1, comprising three proprietary frontier models, one open-weight model, and six clinical RAG systems, including two certified medical devices and one rheumatology-specific system currently under development and not publicly available.

Table 1
Characteristics of evaluated systems

System family	System name	System type	Medical product status	Specialty focus
General-purpose LLMs	GPT-5.6 Sol	Proprietary general-purpose LLM	No	General
	Claude Fable 5	Proprietary general-purpose LLM	No	General
	Gemini 3.1 Pro	Proprietary general-purpose LLM	No	General
	Kimi K2.6	Open-weight general-purpose LLM	No	General
Clinical RAG systems	OpenEvidence	Clinical RAG system	No	General medicine
	AMBOSS AI	Clinical RAG system	No	General medicine
	Prof. Valmed	Clinical RAG system	CE-marked Class I medical device	General medicine
	DR.INFO	Clinical RAG system	CE-marked Class IIb medical device	General medicine
	OpenLeitlinien	Clinical RAG system	No	General medicine
	RHELA	Clinical RAG system	No	Rheumatology

RAG, retrieval-augmented generation; LLM, large language model.

Each system was evaluated on the same 100 cases. Diagnostic cases were assessed using a single concise prompt, whereas management cases were evaluated using both a concise prompt and a thorough prompt adapted from NOHARM (see supplementary table S4). All prompts additionally asked systems to report their confidence in the generated answer on a scale from 0% to 100%.

Each system response was evaluated against the frozen rubric by a three LLM judging panel, (see supplementary table S5). Clinical RAG systems were evaluated by a fixed panel comprising GPT-5.6 Sol, Claude Fable 5, and Gemini 3.1 Pro. Responses from general-purpose LLMs were evaluated by a judge pool, consisting of GPT-5.6 Sol, Claude Fable 5, Kimi K2.6, and Gemini 3.1 Pro. The judging panel rotated so that no model evaluated its own output, using a leave-one-out approach. This evaluation yielded 50,553 rubric-level LLM evaluations across the 10 systems. For each rubric criterion, the assigned weight and harm rating were counted only when at least two of the three judges agreed, using a majority-vote best-of-three approach. Refusals and empty responses received a score of zero without judge

evaluation. Consensus criterion-level scores were summed for each case and converted to a percentage of the maximum achievable score.

To validate the LLM-as-a-judge, two rheumatologists independently evaluated a representative, sample of 100 cases and paired system answers from different systems against the same 1020 rubric criteria, see Fig. 2. Initial inter-rater agreement between the two rheumatologists was almost perfect (Cohen's κ = 0.96; exact agreement 98.0%). Following adjudication, agreement between the final rheumatologist consensus and the three-LLM majority consensus was 95.6% (Cohen's κ = 0.91), supporting the validity of the LLM-as-a-judge approach. Pairwise Cohen's κ among the three LLM judges ranged from 0.84 to 0.85, with exact agreement ranging from 92.4% to 92.8%.

Evaluation metrics

RheumBench assessed answer quality, prompt sensitivity, potential harm, diagnostic accuracy, and system confidence. Overall RheumBench scores were calculated as achieved rubric points divided by the maximum achievable points per case and expressed as percentages. Across all 100 cases, diagnostic and concise-management case scores were combined using case weighting. For management cases, scores were additionally calculated by clinical domain and compared between concise and thorough prompts. Potential harm was assessed from rubric criteria predefined as potentially harmful and summarized by affected cases, triggered harmful criteria, severity, and omission versus commission errors. Omission errors reflected failures to take a necessary action, whereas commission errors reflected actions taken that could cause harm. A hypothetical "do nothing" criterion served as a reference scenario. Diagnostic accuracy was evaluated using top-1 and top-3 accuracy. Diagnostic calibration was assessed using Brier scores and an over-underconfidence index. The observed Brier score was calculated as the mean squared difference between the stated diagnostic confidence and the binary outcome of diagnostic correctness, with lower values indicating better probabilistic accuracy. As a benchmark, a reference Brier score was calculated as (p(1-p)), where (p) represents the system's overall top-1 diagnostic accuracy. An observed Brier score below the reference value indicates that case-specific confidence estimates provide predictive information beyond overall diagnostic accuracy. The over-underconfidence index is calculated as mean self-reported confidence minus mean top-1 accuracy, with positive values indicating overconfidence.

Response acquisition

Model responses were obtained from the provider-native REST endpoints (API) over urllib. Each LLM was called individually over the entire project to not interfere with previous chats, resulting in 55,162 API calls and 780 RAG system answers in total. The six clinical RAG systems provided no public API, so their 780 responses were collected through their web interfaces under institutional accounts, driven by browser automation with per-system recipes for prompt entry, submission, and extraction. A new conversation was started for every case so that no response was conditioned on preceding queries. Refusals, empty

responses, and interface errors were recorded as such rather than retried silently, and each was verified manually.

Statistical analysis

Stratified case-bootstrap 95% confidence intervals for overall RheumBench scores retained the original distribution of 70 diagnostic and 30 management cases and used 20,000 percentile resamples with a fixed random seed. For management cases, prompt-effect confidence intervals were estimated using paired case-level bootstrap resampling with 20,000 percentile resamples and a fixed random seed. Spearman's rank correlation coefficient (ρ) was used to assess case-level associations between RheumBench scores and system self-reported confidence. Interrater agreement was quantified using Cohen's κ. Differences between concise and thorough prompting were assessed using two-sided paired sign-flip randomization tests with 1,000,000 Monte Carlo permutations per system. P values were adjusted using the Holm method across the 10 system-specific management-case comparisons. A two-sided $P < 0.05$ was considered statistically significant. Analyses were performed in Python 3.9.6 using openpyxl 3.1.5, matplotlib 3.9.4, NumPy 2.0.2, SciPy 1.13.1, and formulas 1.3.4. Paired sign-flip randomization tests used NumPy 2.3.5. Figures were generated using Python 3.12.13, with Sharp 0.34.5 used for rasterization.

Results

Overall system performance using concise prompts across all 100 cases (70 diagnostic and 30 management), yielding 1,000 system responses, is summarized in Fig. 3. These responses were evaluated against the 1,020 RheumBench rubric criteria by a three-LLM judge panel, resulting in 30,489 LLM ratings. Overall performance varied substantially across systems. Claude Fable 5 attained the highest overall score (54.0%). In contrast RHELA demonstrated the lowest score (17.8%).

Management case performance, prompt effects and confidence-score associations

Figure 4 illustrates system performance across 30 management cases evaluated using both concise and thorough prompts (also see Supplementary Tables S6-S9). GPT-5.6 Sol and Claude Fable 5 achieved the highest mean management scores with the thorough prompt (72.9% and 71.9%, respectively). With the concise prompt, Claude Fable 5 achieved the highest score (68.3%), followed by GPT-5.6 Sol (55.4%). Among the clinical RAG systems, OpenEvidence achieved the highest score with the thorough prompt (57.0%), whereas AMBOSS AI performed best with the concise prompt (51.3%). RHELA consistently demonstrated the lowest performance across both prompting conditions.

Performance varied substantially across management domains, with GPT-5.6 Sol and Claude Fable 5 achieving the highest scores in most domains. Under the concise prompt, Claude Fable 5 led in three of five domains, while AMBOSS AI achieved the highest score in the follow-up domain; under the thorough

prompt, Claude Fable 5 led in three domains and GPT-5.6 Sol in the remaining two. Procedural advice was the weakest domain for the majority of systems.

The thorough prompt produced higher overall management scores for all systems except DR.INFO, with improvements observed across most management domains. The mean paired differences and bootstrapped 95% confidence intervals for each system are shown in Fig. 4b. DR.INFO achieved higher scores with the concise prompt, possibly because, unlike the thorough prompt, it did not impose a word limit and therefore generated longer responses (mean length: 634 vs 215 words). After Holm adjustment, prompt effects were statistically significant for GPT-5.6 Sol, Gemini 3.1 Pro, Kimi K2.6, OpenEvidence, OpenLeitlinien, DR.INFO, and RHELA (all adjusted $p < 0.05$). Across the fixed set of evaluated systems, the mean paired difference was 7.4 percentage points (95% CI 4.6 to 10.4; $p < 0.001$).

Self-reported confidence was high across systems and prompts despite substantial variation in answer quality, as measured by mean RheumBench management scores (Fig. 4c). Mean scores ranged from 16.8% to 72.9%, whereas reported confidence ranged from 83.4% to 96.5%. This mismatch was particularly evident for Prof. Valmed, which reported confidence of 94.7% and 95.0% despite mean scores of 28.1% and 38.2% under the concise and thorough prompts, respectively. Similarly, Gemini 3.1 Pro reported the highest confidence under both prompts despite not achieving the highest RheumBench scores. At the case level, confidence was weakly to moderately associated with answer quality under the concise prompt ($\rho = -0.87$–0.51). Under the thorough prompt, associations were weaker and ranged from − 0.33 to + 0.35, see Supplementary Tables S6-S7.

Harm potential

Severe omission or commission errors occurred under both prompting conditions (Fig. 5 and Supplementary Tables S10-S11). Severe omission and severe commission errors were mutually exclusive; no management case exhibited both error types. Across systems, model-case responses containing at least one potentially severely harmful error ranged from 20.0% (AMBOSS AI) to 36.7% (Kimi K2.6) with the concise prompt and from 6.7% (GPT-5.6 Sol) to 33.3% (DR.INFO) with the thorough prompt. To assess the potential consequences of taking no clinical action, we applied a hypothetical “do nothing” rubric criterion, which was associated with severe potential harm in 70.0% of cases. The thorough prompt reduced the total number of severe errors from 135 to 101. Severe omission errors were substantially more frequent than severe commission errors (52.0% vs 0.6% with the concise prompt and 38.8% vs 0.5% with the thorough prompt), despite there being considerably more rubric-defined opportunities for severe commission than omission errors (780 vs 250 criteria, respectively).

Diagnostic accuracy and calibration of systems

Diagnostic performance and calibration varied substantially across systems (Fig. 6 and supplementary Tables S12-S13. GPT-5.6 Sol achieved the highest top-1 accuracy (57.1%), whereas Claude Fable 5 achieved the highest top-3 accuracy (67.1%), see Fig. 6a. Self-reported mean confidence always

exceeded mean top-1 accuracy (Fig. 6b), indicating systematic overconfidence (positive O-U index). Claude Fable 5 had the lowest O–U index (+ 0.16), closely followed by GPT-5.6 Sol (+ 0.16), whereas the highest values were observed for RHELA (+ 0.70), Prof. Valmed (+ 0.59), and DR.INFO (+ 0.57). Claude Fable 5 achieved the lowest observed Brier score (0.203), followed by GPT-5.6 Sol (0.218), indicating the best overall agreement between stated diagnostic confidence and actual top-1 correctness (Fig. 6c). For both systems, the observed Brier score was lower than the corresponding reference Brier score, indicating that their case-specific confidence estimates provided meaningful predictive information beyond simply assigning each case a constant confidence equal to the system's overall accuracy. RHELA demonstrated the lowest diagnostic accuracy (top-1, 14.3%; top-3, 18.6%) and worst Brier score (0.611), indicating the poorest overall diagnostic performance.

Discussion

RheumBench provides the first dedicated, specialist-validated, rubric-based framework for evaluating LLMs in rheumatology. Performance varied substantially across systems, management domains, and prompting conditions. Importantly, no system achieved an overall score above 54.0%, indicating substantial room for system improvement and an absence of apparent score saturation.

The leading frontier general-purpose LLMs, GPT-5.6 Sol and Claude Fable 5, achieved the strongest overall performance and led most individual analyses, including management performance, diagnostic accuracy, potential harm, and calibration. They outperformed all clinical RAG systems, including the two certified medical products and the rheumatology-specific system currently under development. Notably, the open-weight model Kimi K2.6 also outperformed both certified medical products in most analyses. These findings indicate that access to medical retrieval sources, clinical specialization, or medical product certification alone does not guarantee higher-quality clinical decision support. These results are consistent with Vishwanath et al. [25], who found that frontier models outperformed clinical RAG systems across several medical evaluations without a clear safety advantage for specialized systems. Similarly, Freudenberg et al. [26] reported superior treatment-planning performance of ChatGPT-5 Thinking compared with OpenEvidence and Prof. Valmed. Our findings also extend previous rheumatology-specific research. Madrid-García et al. and Felde et al demonstrated that rheumatology guideline retrieval improved the performance of LLMs [27, 28]. RheumBench, however, evaluates a broader set of specialist-curated diagnostic and management cases using case-specific, clinically weighted rubric criteria. Although RAG may improve an individual model, our results suggest that retrieval alone does not necessarily compensate for differences in underlying model capability.

Consistent with NOHARM, omission errors predominated, and all systems produced potentially severe errors [29]. Unlike NOHARM, clinical RAG systems did not consistently outperform general-purpose LLMs. The thorough prompt reduced or did not change severe errors for all systems except DR.INFO. Prompt design had a substantial effect on performance. The thorough prompt improved management scores for nearly all systems across most domains, whereas DR.INFO performed better with the concise prompt, potentially because only the thorough prompt imposed a word limit. As clinicians and patients

may commonly use short prompts without dedicated training [3, 5, 30–32], these findings demonstrate the potential of relatively simple prompt guidance to improve LLM-generated clinical responses. This is supported by Rebitschek et al., who found that guidance on more informed prompting improved the evidence-based quality of AI-generated health information among lay users [32]. Self-reported confidence was high despite substantial differences in RheumBench scores and diagnostic accuracy, and diagnostic calibration analyses demonstrated systematic overconfidence across all systems. These findings are consistent with Naderi et al. [33], who similarly reported poor self-confidence calibration across LLMs in gastroenterology. Although the leading frontier models demonstrated comparatively better calibration, confidence frequently remained poorly aligned with actual performance. Our results underscore the need for oversight when LLMs are used for clinical decision support. Potentially severe errors occurred across all systems, including certified medical products, showing that certification does not eliminate clinically relevant risk. This concern may be amplified by automation bias [34, 35], when highly confident but incorrect recommendations may be accepted without sufficient scrutiny by clinicians and potentially patients.

RheumBench has several strengths. It evaluates realistic free-text responses across diagnostic and management tasks and provides both an overall performance measure and clinically interpretable assessments of management domains, potential harm, confidence, and calibration. Its case-specific weighted rubrics may also improve discrimination between high-performing systems when conventional diagnostic accuracy approaches ceiling effects. The benchmark is based on clinically relevant cases and questions, with rubric content, clinical appropriateness, and potential harmfulness independently assessed by board-certified rheumatologists. Furthermore, the high agreement between LLM judges and rheumatologists supports the use of LLM-as-a-judge approaches for more scalable evaluation. Several limitations should be considered. RheumBench currently included 100 cases and therefore does not capture the full spectrum of rheumatology. In addition, 60 of the 70 diagnostic cases had been previously published [7], creating a potential risk of data contamination. Future versions will expand the case set, incorporate additional real-world prompts, and include multimodal tasks. Although benchmark content and harm ratings underwent expert review, these assessments remain partly subjective, and automated judges may lack clinical calibration or rarely abstain on ambiguous criteria [36]. Prompt comparisons may also have been affected by response-length differences because only the thorough prompt imposed a word limit. We did not evaluate within-model response variability through repeated sampling, and performance may change as models, retrieval sources, and interfaces are updated. Confidence analyses were limited by missing or invariant values for some systems, and management-case correlations reflect associations between confidence and graded answer quality rather than absolute probabilistic calibration. We also did not assess the quality of references provided by the systems, which has been suggested previously [37] and may become increasingly relevant as retrieval sources and commercial incentives evolve [38]. Finally, RheumBench currently requires substantial rheumatologist input for rubric development and validation. More automated approaches to rubric generation may improve scalability, as suggested by recent work demonstrating agreement between automated and physician-developed clinical rubrics [39, 40]. Most importantly, RheumBench assesses

responses to standardized cases rather than effects on clinician behavior or patient outcomes. Prospective studies are therefore required to determine whether benchmark performance translates into safer and more effective clinical decision support.

Conclusions

RheumBench provides the first dedicated, specialist-validated, rubric-based benchmark for evaluating LLMs in rheumatology. Frontier general-purpose LLMs achieved the strongest overall performance and generally outperformed clinical RAG systems, including certified medical products and a rheumatology-specific system. However, substantial performance differences, potentially severe errors, and systematic overconfidence were observed across systems, highlighting the limitations and clinically relevant risks of current LLM-based decision support. These findings indicate that medical specialization or certification alone does not ensure superior performance or safety and reinforce the need for meaningful human oversight, particularly given the potential for automation bias. Prompt design also substantially influenced performance, demonstrating its potential as a simple and scalable means of improving LLM-generated clinical responses while emphasizing the importance of evaluating systems under realistic use conditions. RheumBench will be continuously expanded and updated to support ongoing safety assessment, system improvement, and informed stakeholder decision-making.

Abbreviations

AI	Artificial Intelligence
API	Application Programming Interface
CE	European Conformity
CRP	C-Reactive Protein
ESR	Erythrocyte Sedimentation Rate
GCA	Giant Cell Arteritis
GPT	Generative Pre-trained Transformer
LLM	Large Language Model
NOHARM	First, do NOHARM benchmark
O-U	Over-underconfidence index
RAG	Retrieval-Augmented Generation
SD	Standard Deviation

Declarations

Acknowledgements

The authors thank all contributors involved in the development and evaluation of RheumBench.

Author Contributions

Fabian Lechner: data curation, formal analysis, methodology, software, visualization. Jonathan Bamberger: formal analysis, visualization, writing, review and editing. Phillip Kremer, Jutta Richter, Matthias Schneider, Uta Kiltz, Philipp Klemm, Axel J. Hueber, and Martin Krusche: methodology, clinical expertise, interpretation, and writing, review and editing. Sebastian Kuhn: methodology, interpretation, and writing, review and editing. Lucie Flek: supervision, review, editing. Hannah Labinsky: conceptualization, methodology, clinical expertise, supervision, interpretation, and writing, review and editing. Johannes Knitza: first manuscript draft, conceptualization, methodology, supervision, project administration, data curation, interpretation, and writing, review and editing. All authors made substantial contributions to the work, critically revised the manuscript for important intellectual content, approved the final version, and agree to be accountable for the integrity and accuracy of the work.

Funding

No specific funding was received for this study. Johannes Knitza was supported in part by the European Union’s Horizon Europe program under grant agreement No. 101080711 (SPIDERR).

Data Availability

The raw data supporting the conclusions of this article will be made available by the authors upon reasonable request.

Ethics Approval and Consent to Participate

The Philipps-University Marburg Ethics Committee confirmed that no ethical approval was required (reference: 26-266 ANZ) as the analysis was based on anonymous data.

Competing Interests

Johannes Knitza, Jutta Richter, Matthias Schneider, Uta Kiltz, Martin Krusche, and Hannah Labinsky are members of a project group involved in the development and improvement of the RHELA system. The other authors declare no competing interests.

Consent for Publication

Not applicable.

Clinical Trial Number

Not applicable.

Generative AI and AI-Assisted Technologies

Generative AI use as part of the study methods is stated in the methods section. During the preparation of the manuscript, the authors used GPT-5.5/5.6 (OpenAI) to support language editing, manuscript formatting, and wording refinement. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

1. Blease CR, Locher C, Gaab J, et al (2024) Generative artificial intelligence in primary care: an online survey of UK general practitioners. BMJ Health Care Inform 31:. https://doi.org/10.1136/bmjhci-2024-101102
2. Ayo-Ajibola O, Davis RJ, Lin ME, et al (2024) Characterizing the Adoption and Experiences of Users of Artificial Intelligence–Generated Health Information in the United States: Cross-Sectional Questionnaire Study. Journal of Medical Internet Research 26:e55138. https://doi.org/10.2196/55138
3. Labinsky H, Klemm P, Graalmann L, et al (2025) Patient experiences, attitudes, and profiles regarding artificial intelligence in rheumatology: a German national cross-sectional survey study. Rheumatol Int 45:269. https://doi.org/10.1007/s00296-025-06023-x
4. Holzer M-T, Meinecke A, Müller F, et al (2024) Artificial intelligence in rheumatology: status quo and quo vadis-results of a national survey among German rheumatologists. Ther Adv Musculoskelet Dis 16:1759720X241275818. https://doi.org/10.1177/1759720X241275818
5. La Bella S, Rebollo-Giménez AI, Aouad K, et al (2026) Artificial intelligence in rheumatology and paediatric rheumatology: insights from an international survey by EMEUNET. EULAR Rheumatol Open 2:100153. https://doi.org/10.1016/j.ero.2026.03.001
6. Matthes J, Reinhardt A, Hodzic S, et al (2026) Predicting the Intention to Use Generative Artificial Intelligence for Health Information: Comparative Survey Study. Journal of Medical Internet Research 28:e75648. https://doi.org/10.2196/75648
7. Kremer P, Schiebisch H, Lechner F, et al (2025) Comparative analysis of large language models and traditional diagnostic decision support systems for rare rheumatic disease identification. EULAR Rheumatology Open 1:51–59. https://doi.org/10.1016/j.ero.2025.04.007
8. Griewing S, Knitza J, Boekhoff J, et al (2024) Evolution of publicly available large language models for complex decision-making in breast cancer care. Arch Gynecol Obstet 310:537–550. https://doi.org/10.1007/s00404-024-07565-4
9. Battafarano DF, Ditmyer M, Bolster MB, et al (2018) 2015 American College of Rheumatology Workforce Study: Supply and Demand Projections of Adult Rheumatology Workforce, 2015–2030. Arthritis Care Res (Hoboken) 70:617–626. https://doi.org/10.1002/acr.23518

10. Knitza J, Gupta L, Hügle T (2024) Rheumatology in the digital health era: status quo and quo vadis? Nat Rev Rheumatol 20:747–759. https://doi.org/10.1038/s41584-024-01177-7
11. Kremer P, Langballe E, Haase I, et al (2026) Diagnostic performance of Prof. Valmed, ChatGPT-5 Thinking, and OpenEvidence in rheumatology: A comparative evaluation. Rheumatol Int 46:31. https://doi.org/10.1007/s00296-025-06068-y
12. Krusche M, Callhoff J, Knitza J, Ruffer N (2023) Diagnostic accuracy of a large language model in rheumatology: comparison of physician and ChatGPT-4. Rheumatol Int. https://doi.org/10.1007/s00296-023-05464-6
13. Haase I, Xiong T, Rissmann A, et al (2024) ChatSLE: consulting ChatGPT-4 for 100 frequently asked lupus questions. Lancet Rheumatol 6:e196–e199. https://doi.org/10.1016/S2665-9913(24)00056-0
14. Kremer P, Knitza J, Reichard A, et al (2026) A patient-derived benchmark for evaluating large language models in connective tissue diseases: blinded multi-stakeholder assessment and guideline comparison. Rheumatol Int 46:210. https://doi.org/10.1007/s00296-026-06178-1
15. Labinsky H, Nagler L-K, Krusche M, et al (2024) Vignette-based comparative analysis of ChatGPT and specialist treatment decisions for rheumatic patients: results of the Rheum2Guide study. Rheumatol Int. https://doi.org/10.1007/s00296-024-05675-5
16. Knitza J, Aries P (2026) Ambient AI scribes in rheumatology: early real-world clinician and patient experience. EULAR Rheumatol Open 2:100101. https://doi.org/10.1016/j.ero.2025.12.002
17. Roemer A, Schlicker N, Kernder A, et al (2026) Large language models enhance diagnostic reasoning of medical students in rheumatology: a randomized controlled trial. BMC Med Educ. https://doi.org/10.1186/s12909-026-09079-w
18. Gaube S, Suresh H, Raue M, et al (2021) Do as AI say: susceptibility in deployment of clinical decision-aids. npj Digit Med 4:31. https://doi.org/10.1038/s41746-021-00385-9
19. Andersen ES, Birk-Korch JB, Hansen RS, et al (2024) Monitoring performance of clinical artificial intelligence in health care: a scoping review. JBI Evid Synth 22:2423–2446. https://doi.org/10.11124/JBIES-24-00042
20. Goh E, Wu D, Walton C, et al (2026) Toward a test of medical AI superintelligence. Nat Med. https://doi.org/10.1038/s41591-026-04539-8
21. Williams G, Rutunda S, Nzabakira F, Mateen BA (2026) Human evaluators vs. LLM-as-a-Judge: toward scalable evaluation of GenAI in global health. npj Digit Med. https://doi.org/10.1038/s41746-026-02992-w
22. Arora RK, Wei J, Hicks RS, et al (2025) HealthBench: Evaluating Large Language Models Towards Improved Human Health
23. Wu D, Haredasht FN, Maharaj SK, et al (2025) First, do NOHARM: towards clinically safe large language models
24. Naderi N, Safavi-Naini SAA, Savage T, et al (2026) Across generations, sizes, and types, large language models poorly report self-confidence in gastroenterology clinical reasoning tasks. npj Gut Liver 3:6. https://doi.org/10.1038/s44355-026-00053-3

25. Vishwanath K, Alyakin A, Ghosh M, et al (2026) General-purpose large language models outperform specialized clinical AI tools on medical benchmarks. Nat Med 32:2405–2409. https://doi.org/10.1038/s41591-026-04431-5
26. Freudenberg J, Knitza J, Gremke N, et al (2026) AI-enabled clinical decision support in breast cancer care: a blinded multicenter benchmarking study comparing medically specialized with a general-purpose system. J Med Syst 50:107. https://doi.org/10.1007/s10916-026-02434-w
27. Felde S, Buchkremer R, Chehab G, et al Performance and Practical Considerations of Large and Small Language Models in Clinical Decision Support in Rheumatology
28. Madrid-García A, Benavent D, Plasencia-Rodríguez C, 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 Rheumatology Open 1:228–236. https://doi.org/10.1016/j.ero.2025.08.001
29. Wu D, Haredasht FN, Maharaj SK, et al (2025) First, do NOHARM: towards clinically safe large language models
30. Kharko A, Garcia Sanchez C, Hagström J, et al (2025) General practitioners' opinions of generative artificial intelligence in the UK: An online survey. Digit Health 11:20552076251360863. https://doi.org/10.1177/20552076251360863
31. Siden R, Kerman H, Gallo RJ, et al (2025) A typology of physician input approaches to using AI chatbots for clinical decision-making. npj Digit Med. https://doi.org/10.1038/s41746-025-02184-y
32. Rebitschek FG, Carella A, Kohlrausch-Pazin S, et al (2025) Evaluating evidence-based health information from generative AI using a cross-sectional study with laypeople seeking screening information. npj Digit Med 8:343. https://doi.org/10.1038/s41746-025-01752-6
33. Naderi N, Safavi-Naini SAA, Savage T, et al (2026) Across generations, sizes, and types, large language models poorly report self-confidence in gastroenterology clinical reasoning tasks. npj Gut Liver 3:6. https://doi.org/10.1038/s44355-026-00053-3
34. Qazi IA, Ali A, Khawaja AU, et al (2026) Automation Bias in Large Language Model–Assisted Diagnostic Reasoning among Physicians Trained in AI Literacy — A Randomized Clinical Trial. NEJM AI 3:AIoa2501001. https://doi.org/10.1056/AIoa2501001
35. Knitza J, Tascilar K, Gruber E, et al (2021) Accuracy and usability of a diagnostic decision support system in the diagnosis of three representative rheumatic diseases: a randomized controlled trial among medical students. Arthritis Res Ther 23:233,. https://doi.org/10.1186/s13075-021-02616-6.
36. Philipp W, Fassbender F, Fister D, et al (2026) Clinician-Level Agreement Without Clinical Caution: LLM Evaluator Limits in Medical AI Benchmarking
37. Kirschner P, Epping PZ, Meuth SG, et al (2026) Evaluating the First CE -Marked LLM -Based Chatbot for Evidence-Based Neurology. Euro J of Neurology 33:e70736. https://doi.org/10.1111/ene.70736
38. Brender TD, Detsky AS, Yourman L (2026) Artificial Intelligence and the Financialization of Medical Knowledge. JAMA Intern Med. https://doi.org/10.1001/jamainternmed.2026.2961

39. Shah A, Hines A, Downs A, et al (2026) Case-Specific Rubrics for Clinical AI Evaluation: Methodology, Validation, and LLM-Clinician Agreement Across 823 Encounters
40. Sylolypavan A, Sleeman D, Wu H, Sim M (2023) The impact of inconsistent human annotations on AI driven clinical decision making. npj Digit Med 6:26. https://doi.org/10.1038/s41746-023-00773-3

Figures

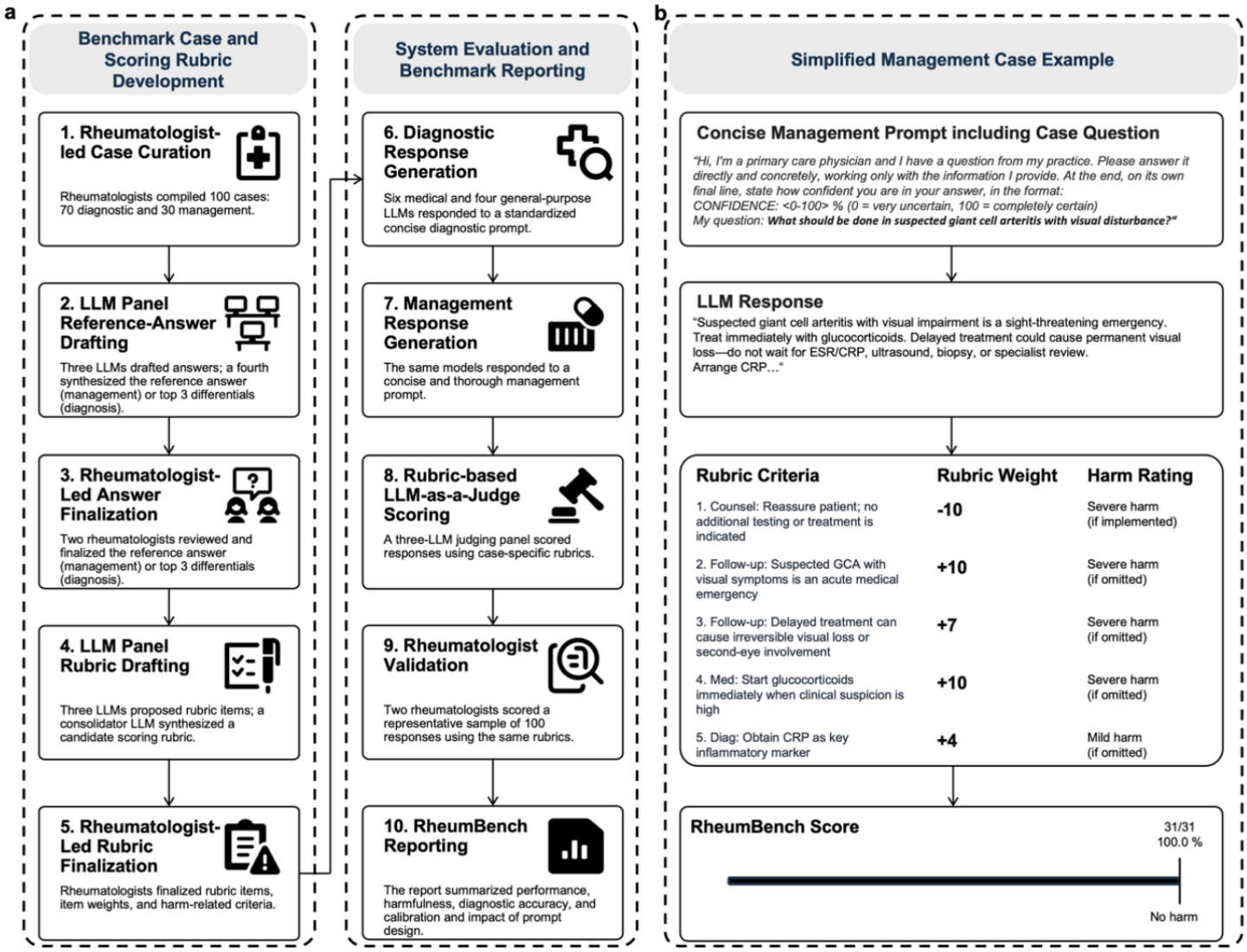


Figure 1

Overview of (**a**) RheumBench development, model evaluation, and (**b**) rubric-based scoring and simplified management case example including showing the prompt, model response, weighted rubric criteria, and potential harm ratings. Positive rubric weights indicate appropriate recommendations, whereas negative weights indicate inappropriate recommendations. Harm ratings reflect the potential consequences of omitting an appropriate criterion or implementing an inappropriate criterion. LLM: large language model, GCA: giant cell arteritis, ESR: erythrocyte sedimentation rate, CRP: C-reactive protein.

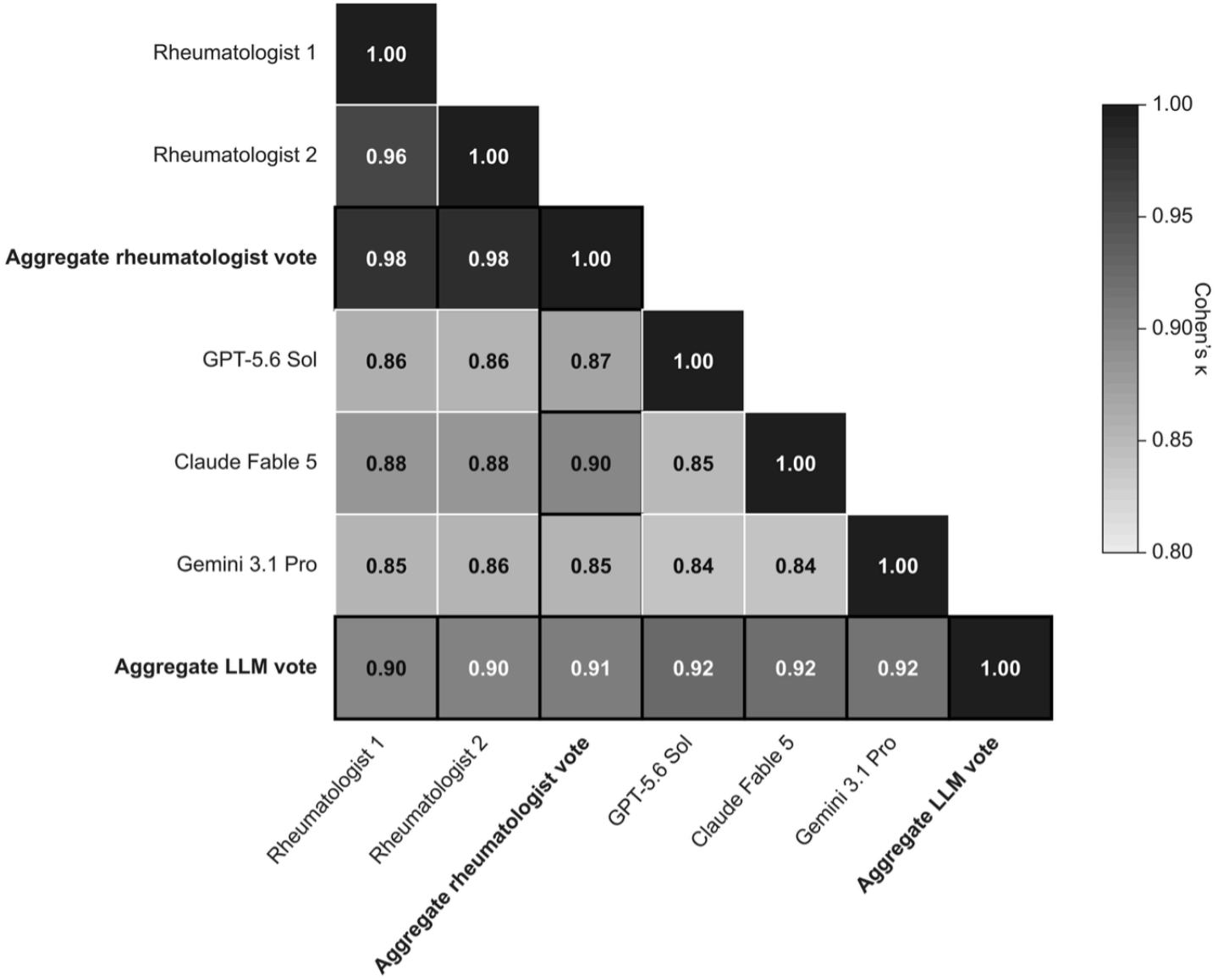


Figure 2

Pairwise agreement between rheumatologists and large language models. Heatmap showing Cohen's κ for pairwise agreement between two board-certified rheumatologists, their aggregate vote, three LLMs (GPT-5.6 Sol, Claude Fable 5, and Gemini 3.1 Pro), and the aggregate LLM vote. Higher κ values indicate greater agreement. Diagonal values represent perfect self-agreement (κ = 1.00).

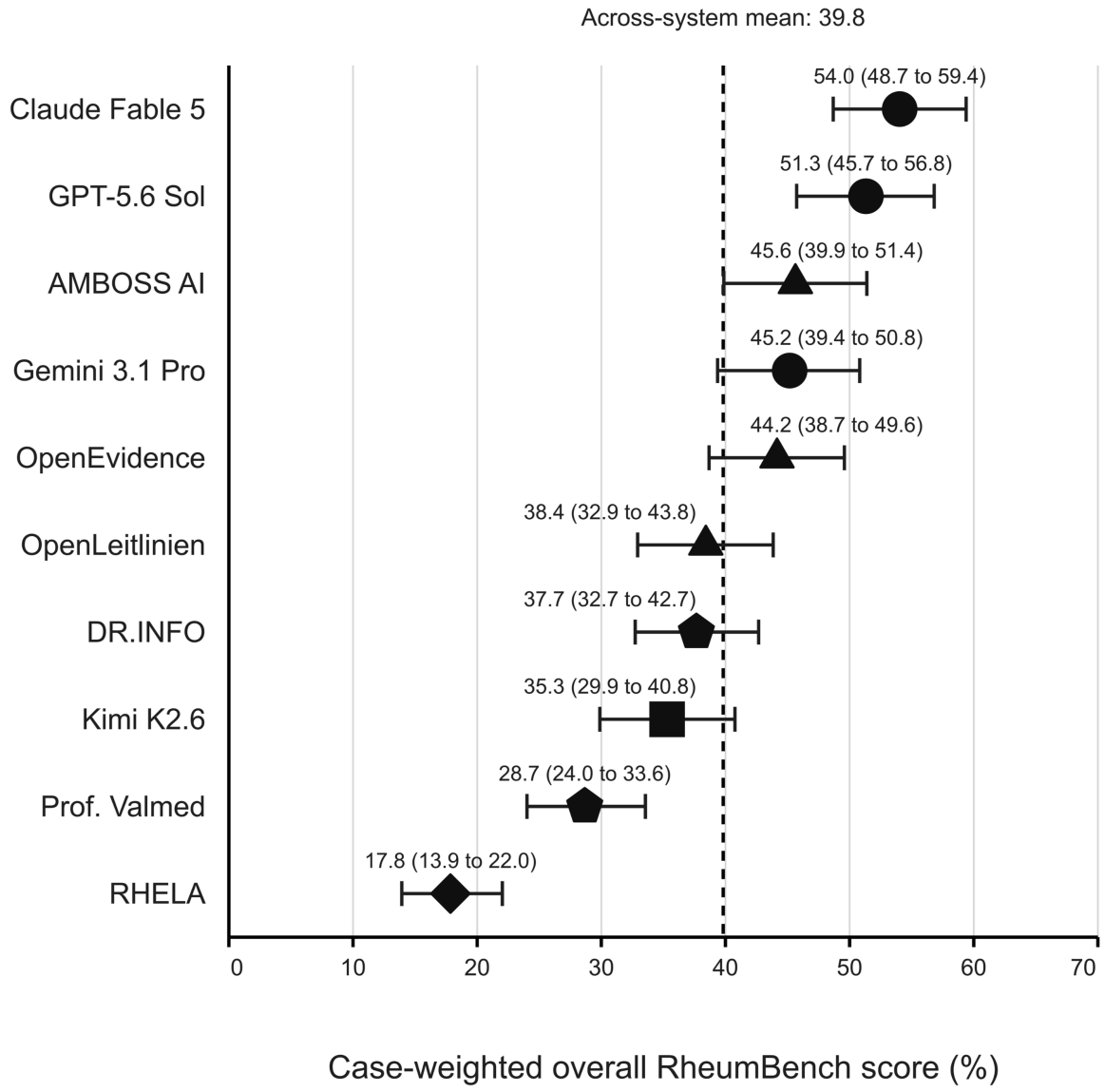


Figure 3

Case-weighted overall RheumBench scores with bootstrapped 95% confidence intervals. The overall score combines 70 diagnostic and 30 concise-management cases using case weighting. Markers indicate the overall score for each system, horizontal error bars show the bootstrapped 95% confidence interval. Circles denote proprietary general-purpose LLMs, the square denotes an open-weight general-purpose LLM, upward triangles denote non-certified general medicine clinical RAG systems, filled pentagons denote general medicine clinical RAG systems certified as medical devices, and the diamond denotes a rheumatology-specific non-certified clinical RAG system. The dashed vertical line marks the across-system mean of 39.8. Numeric labels report each system's overall score and 95% confidence interval.

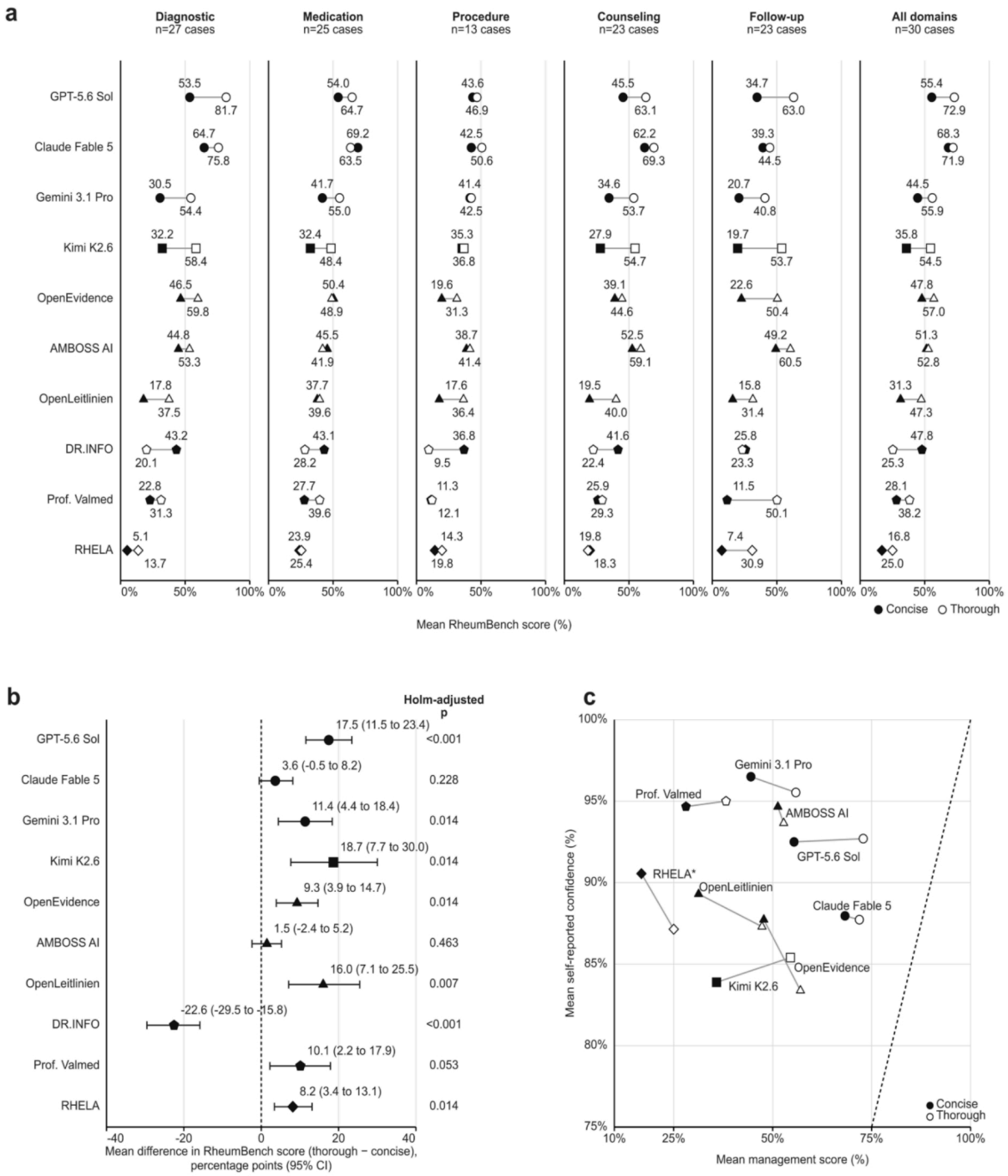


Figure 4

Management performance across clinical domains using different prompts and its relationship with self-reported confidence. **(a)** Mean RheumBench scores across five management domains and across all domains for each system. Paired markers indicate scores obtained with the concise (filled) and thorough prompts (empty), with connecting lines illustrating the effect of prompt design. Domain sample sizes reflect the number of cases containing rubric criteria from the respective domain. (**b**) Effect of prompt

design on RheumBench management performance across the 30 management cases. Points show the mean within-system paired difference in RheumBench score between thorough and concise prompting (thorough minus concise), expressed in percentage points (pp); horizontal error bars indicate 95% confidence intervals estimated by paired case-level bootstrap resampling. Positive values favor the thorough prompt, whereas negative values favor the concise prompt. The vertical dashed line at zero indicates no difference between prompting conditions. Two-sided Holm-adjusted *P*values were obtained from paired sign-flip randomization tests. **(c)** Mean self-reported confidence plotted against mean RheumBench management score for each system using the concise prompt (filled markers) and thorough prompt (empty markers). The dashed line indicates equality between confidence and RheumBench score. Circles denote proprietary general-purpose LLMs, the square denotes an open-weight general-purpose LLM, upward triangles denote non-certified general medicine clinical RAG systems, pentagons denote general medicine clinical RAG systems certified as medical devices, and the diamond denotes a rheumatology-specific non-certified clinical RAG system. DR.INFO is not shown because self-reported confidence was not available. The asterisk indicates incomplete confidence reporting for RHELA (n=9 concise; n=7 thorough).

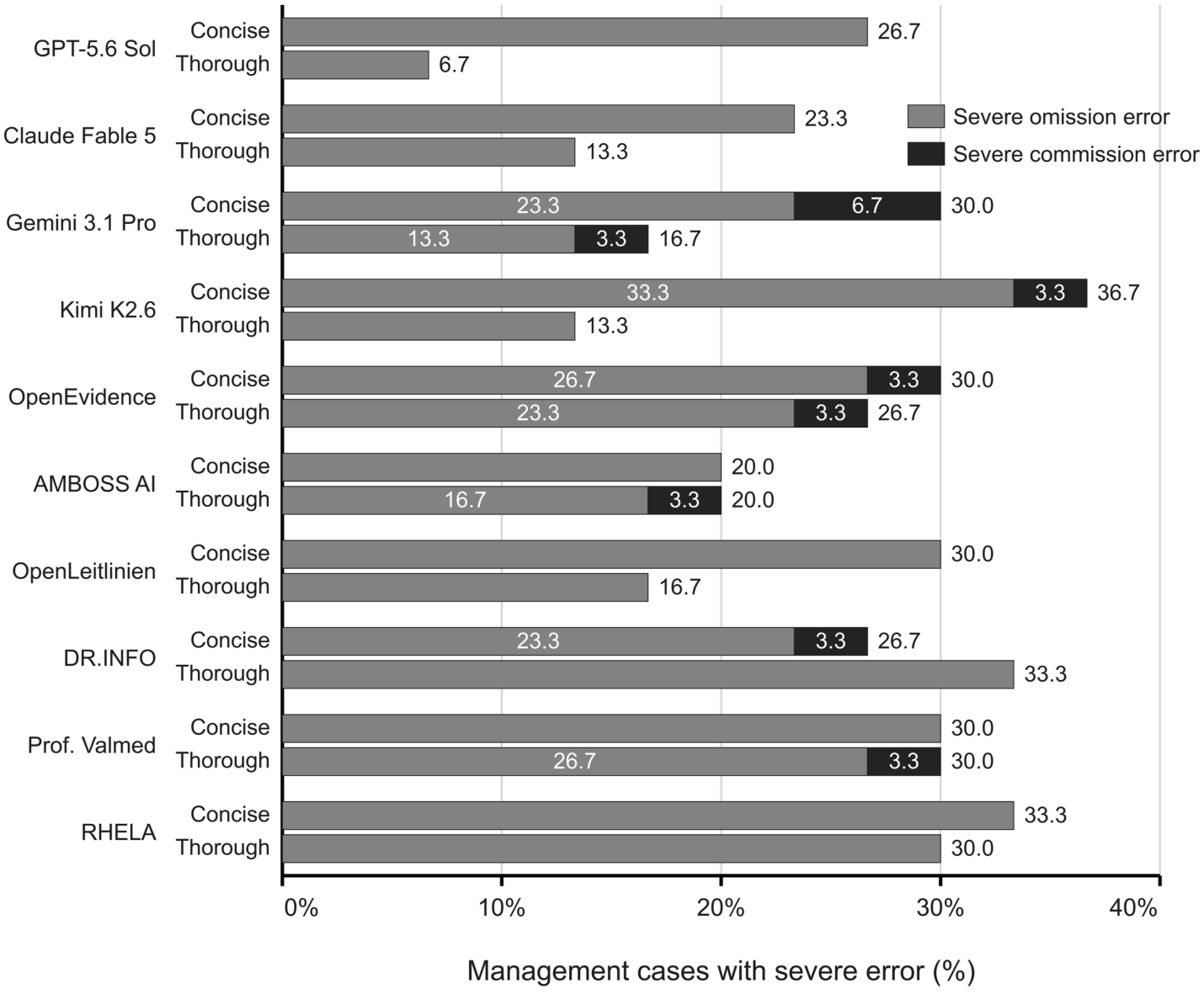


Figure 5

Percentage of the 30 management cases containing at least one severe error, shown separately for concise and thorough prompts. Stacked segments distinguish severe omission errors from severe commission errors. Numbers within segments report component percentages, and labels at bar ends report total percentages.

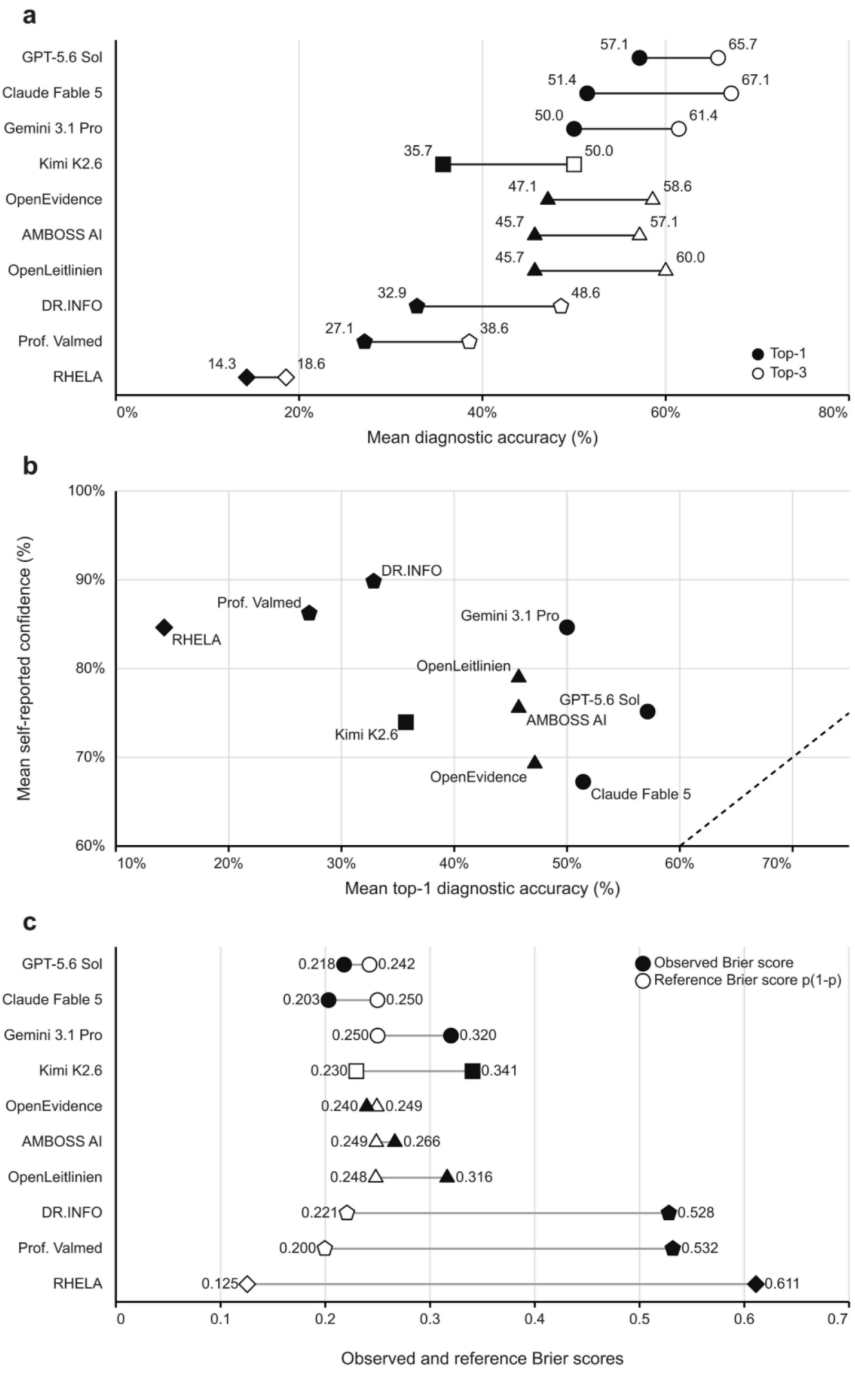


Figure 6

Diagnostic accuracy and confidence calibration across evaluated systems. **(a)** Top-1 (filled markers) and top-3 (empty markers) diagnostic accuracy, defined as identification of the correct diagnosis as the highest-ranked diagnosis or within the three highest-ranked diagnoses, respectively. **(b)**Mean self-reported confidence plotted against top-1 diagnostic accuracy. The dashed diagonal indicates perfect agreement between mean confidence and accuracy, with values above the line indicating

overconfidence. **(c)** Observed Brier scores (filled markers) are compared with system-specific reference Brier scores p(1−p) (open markers), calculated on the same cases; lower scores indicate better calibration. Circles denote proprietary general-purpose LLMs, the square denotes an open-weight general-purpose LLM, upward triangles denote non-certified general medicine clinical RAG systems, pentagons denote general medicine clinical RAG systems certified as medical devices, and the diamond denotes a rheumatology-specific non-certified clinical RAG system.

Supplementary Files

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

- Supplementaryfile1.docx