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.

RULE-LLM reporting framework for large language model-as-a-judge evaluation in medicine developed from a scoping review

Junbok Lee

nanaya85@yuhs.ac

Yonsei University

Huiyeol Yun

Yonsei University

Yoon Bin Jung

Yonsei University College of Medicine

Article

Keywords: large language model, LLM-as-a-Judge, reporting framework, scoping review, artificial intelligence evaluation

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11611263/v1

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

Additional Declarations: No competing interests reported.

RULE-LLM reporting framework for large language model-as-a-judge evaluation in medicine developed from a scoping review

Junbok Lee[1,2*], Huiyeol Yun[1], Yoon Bin Jung[3*]

[1] Yonsei Institute for Digital Health, Yonsei University, Seoul, Republic of Korea

[2] Department of Human Systems Medicine, Seoul National University College of Medicine, Seoul, Republic of Korea

[3] Department of Surgery, Yonsei University College of Medicine, Seoul, Republic of Korea

Corresponding Author #1

Junbok Lee, PhD, Research Professor, Yonsei Institute for Digital Health, Yonsei University, Address: 50-1, Yonsei-ro, Seodaemun-gu, Seoul, 03722, Republic of Korea, TEL: +82-2-2228-1828, E-mail: nanaya85@yuhs.ac

Corresponding Author #2

Yoon Bin Jung, MD, PhD, Professor, Department of Surgery, Yonsei University College of Medicine, Address: 50-1, Yonsei-ro, Seodaemun-gu, Seoul, 03722, Republic of Korea, TEL: +82-2-2228-1881, E-mail: ybjung@yuhs.ac

Abstract

Large language models (LLMs) are increasingly used to evaluate other models' outputs in medicine, an approach known as LLM-as-a-Judge (LaaJ), yet no LaaJ-specific reporting checklist exists. This scoping review included 91 medical LaaJ studies: 69 from 5,571 database records and 22 from a previous review. Judges were mainly used for response evaluation (74.7%), and 76.9% of the studies used at least one proprietary LLM. Agreement with human evaluators was reported in 65.9% of the studies, with widely varying metrics. Based on these findings, this study developed RULE-LLM, a reporting framework of 21 items and eight sub-items across four domains: role and task specification, underlying data and criteria, LLM-Judge design and output, and evaluation of judge performance. Seven experts rated the initial draft (S-CVI/Ave, 0.94), and five independent raters showed substantial agreement when applying the final framework (Fleiss' κ, 0.67; Gwet's AC1, 0.73). RULE-LLM aims to improve the transparency and reproducibility of LaaJ studies.

Keywords: large language model, LLM-as-a-Judge, reporting framework, scoping review, artificial intelligence evaluation

INTRODUCTION

Owing to advances in generative artificial intelligence (AI), many studies have explored its use in healthcare[1-3]. Generative AI is applied in clinical decision support, automation of medical records, patient education, medical research, and data analysis[2]. These technologies are expected to change how healthcare professionals work, and their adoption may widen further[1,4].

Given that new healthcare technologies can directly affect patient safety, they should be thoroughly evaluated before use[5]. Discriminative AI models are typically developed by splitting data into training and test sets, and their performance is measured against a reference standard using metrics such as accuracy, sensitivity, specificity, and F1 score[6,7]. These metrics are suitable for clear and verifiable answers[7]. Conversely, generative AI produces open-ended free text for which there is often no single correct answer; therefore, these metrics cannot be applied directly[8,9]. Text similarity metrics, such as BLEU and ROUGE, can be used; however, they correlate poorly with clinical quality and human judgment[8,10,11]. Consequently, various evaluation methods have been proposed for generative AI outputs, of which evaluations by human experts remain the most widely accepted[9,12,13].

However, human evaluation requires considerable time and resources when the volume of generated output is large[14,15], and it is subject to variability between raters[9,16]. To address these limitations, an approach in which a separate large language model (LLM) evaluates the outputs of other models, referred to as “LLM-as-a-Judge” (LaaJ) or auto-evaluation, has been proposed and applied in a growing number of studies[8,14,17].

Several frameworks related to LLM evaluation have been proposed in medicine, including human evaluation frameworks (QUEST and CLEVER)[9,13], reporting guidelines for studies using LLMs (TRIPOD-LLM and MEDAI-LLM-SUMM)[18,19], and a framework for developing benchmark datasets (READY)[20]. However, these frameworks do not adequately

cover LaaJ's methodological issues, such as the selection and validation of the judge model[21,22], prompt design, and the sensitivity of results to positional bias[14,23,24]. MedJUDGE, which was recently proposed[21], is tailored to LaaJ and partly addresses these issues. However, it focuses on post-deployment governance and safety verification according to clinical risk tiers and does not specify, item by item, what should be reported about the judge system when a study is designed, conducted, or published. Without a LaaJ-specific reporting standard, LaaJ studies in medicine may be reported inconsistently and with limited transparency, making it difficult to reproduce findings or compare results across studies[12,16,21,22].

This study conducted a scoping review of LaaJ studies in medicine to identify the purposes for which LaaJ is used, its methodological characteristics, and current reporting practices[25-27]. Based on the findings, it developed the RULE-LLM, which is a reporting framework that defines the core elements to be reported in medical LaaJ studies, and evaluated it through content validity assessment by an expert panel and inter-rater reliability assessment by independent raters. The name combines "RULE," an acronym of the four domains of the framework (Role and task specification, Underlying data and criteria, LLM-Judge design and output, and Evaluation of judge performance), with "LLM," the model that serves as the evaluator. To the best of our knowledge, this is the first item-level reporting checklist for medical LaaJ studies derived from a scoping review and assessed for content validity and inter-rater reliability.

RESULTS

Study selection

This review followed the PRISMA-ScR guidelines[28] and comprised identification, screening, eligibility assessment, and inclusion (Figure 1). Searches of PubMed, Embase, IEEE Xplore, and the ACM Digital Library yielded 5,571 records, of which 4,921 remained after duplicates were removed. Among these, 892 records indexed as books, book sections, entire conference proceedings volumes, or generic entries were removed by reference type before screening, and 3,781 were excluded during the LLM-assisted title and abstract screening. Two reviewers independently assessed the full texts of the remaining 248 articles and excluded 179 articles that did not meet the inclusion criteria (Supplementary Table 2), resulting in 69 articles. Additionally, 22 journal articles identified in a previous scoping review of LaaJ[22] were included; therefore, 91 studies were included in the analysis (Supplementary Table 9)[13,15,29-117].

Characteristics of the included studies

The 91 studies were analyzed for the medical task type, input data, source of evaluation criteria, judge function and purpose, judge model type, system architecture, prompting strategy, output format, clinical validation, and agreement between the judge and human evaluators (Supplementary Tables 4–8). Given that a study could fall into more than one category, frequencies within an item may add up to more than 91 and percentages to more than 100%.

Regarding medical task type, patient communication and education were the most common (n = 31, 34.1%), followed by clinical decision support (n = 28, 30.8%) (Supplementary Table 5). Medical research assistance accounted for 19 studies (20.9%), clinical note generation accounted for 13 (14.3%), and administration and workflow accounted for two (2.2%). Two studies (2.2%) addressed tasks other than those mentioned in these

categories.

As for input data, 50 studies (54.9%) used public data, 27 (29.7%) used private data, and 29 (31.9%) used synthetic data. Synthetic data included synthetic clinical cases, simulated patient–provider interactions, synthetic patient profiles and questions, and virtual clinical scenarios. Fourteen studies (15.4%) combined public data with private or synthetic data.

Reference-based criteria, which only relied on correct answers, expert-compiled results, clinical guidelines, and other reference information, were used in 32 studies (35.2%). Self-defined criteria set by researchers according to their study objectives were used in 33 studies (36.3%). The remaining 26 studies (28.6%) combined the two approaches by applying predefined evaluation dimensions to the external reference information.

Response evaluation (assessing the quality of the generated responses) was the most common judge function (n = 68, 74.7%). The judge was used to compare the performance of multiple models or outputs in 33 studies (36.3%) and as a verifier to determine whether the outputs met specific conditions in 33 studies (36.3%). Only six studies (6.6%) used judge evaluations as feedback to refine outputs. The judge was also used in data construction for annotation in 11 studies (12.1%) and data synthesis in one study (1.1%).

Fifty studies (54.9%) only used proprietary LLMs, such as GPT, Claude, and Gemini, as judges (Supplementary Table 6). Fourteen studies (15.4%) only used open-source LLMs, and one study (1.1%) used a model trained or fine-tuned on medical data. Among the studies that combined model types, proprietary and open-source LLMs were the most often combined (n = 10, 11.0%). One study (1.1%) combined proprietary LLMs with a medically fine-tuned model, another (1.1%) combined open-source LLMs with a medically fine-tuned model, and nine studies (9.9%) used all three types. Five studies (5.5%) did not report the judge model type. Overall, 70 studies (76.9%) used at least one proprietary LLM, 34 (37.4%) used at least one open-source LLM, and 12 (13.2%) used at least one medically fine-tuned model.

Most studies used a single-LLM system, in which one LLM served as the evaluator (n = 71, 78.0%). Fourteen studies (15.4%) only used a multi-LLM system, five (5.5%) used both single- and multi-LLM systems, and one (1.1%) used a human–AI collaborative structure in which humans and LLMs shared evaluation roles.

Rubric-based prompting, which specifies the prompt's evaluation criteria or detailed evaluation items, was the most common prompting strategy (n = 73, 80.2%). Zero-shot prompting without examples was used in 49 studies (53.8%), and 39 studies (42.9%) combined zero-shot prompting with an explicit rubric. Few-shot and chain-of-thought (CoT) prompting were each used in 15 studies (16.5%).

Numerical scores or pointwise ratings were the most common output format (n = 55, 60.4%), followed by binary outputs, such as include/exclude, correct/incorrect, or met/not met (n = 36, 39.6%). Furthermore, 36 studies (39.6%) asked the judge to provide a natural language rationale with their results, and 19 (20.9%) used structured output in JSON or predefined fields. Other formats, predominantly pairwise preference, ranking of multiple outputs, and categorical labels, were used in 11 studies (12.1%).

The involvement of experts in validating the judges varied (Supplementary Table 7). In 43 studies (47.3%), clinicians reviewed the judges' results or independently evaluated the same results using the same criteria. In 17 studies (18.7%), evaluators other than physicians, such as regulatory experts, non-clinical experts, researchers, and crowd workers, participated. The remaining 31 studies (34.1%) involved no medical experts in the evaluation or validation.

Moreover, 60 studies (65.9%) reported at least one agreement metric between judges and human evaluators. Among these 60 studies, Cohen's kappa was the most frequently used single named metric (n = 12, 20.0%), followed by Pearson correlation (n = 7, 11.7%). Spearman correlation, Kendall's tau, and the intraclass correlation coefficient (ICC) were each reported in six studies (10.0%), and the mean absolute error (MAE) was reported in three studies (5.0%).

Other measures included Gwet's AC1, Krippendorff's alpha, percentage agreement, accuracy, F1 score, sensitivity, specificity, and ranking consistency. Percentages in this paragraph are based on the 60 studies that reported agreement, whereas those in Supplementary Table 7 are based on all 91 studies.

Human review and acceptance rate of AI-assisted data extraction

The LLM-based pipeline first classified 91 studies according to the analytical framework; subsequently, two reviewers checked the classifications against the original articles and corrected them where necessary. Given that the reviewers could see the AI classifications, the figures below represent the proportion of AI classifications accepted after human review rather than the inter-rater reliability between independent raters. Among 4,004 classifications across 44 items, 3,539 (88.4%; 95% confidence interval [CI], 87.4%–89.3%) were retained, and 465 (11.6%) were corrected (Supplementary Table 12). The corresponding Cohen's kappa was 0.717 (95% CI, 0.693–0.741). With the presence of a characteristic (coded as "O"; absence, "X") after human review as the positive class, the initial AI classifications had a precision of 80.7%, recall of 79.0%, F1 score of 79.8%, and specificity of 92.2%.

Development of the RULE-LLM framework

The RULE-LLM comprised 21 items and eight sub-items across four domains (Table 1; Figure 2). The role and task specification domain contained four items on whether the medical task and purpose of the judge were clearly defined: the medical task type, primary function of the judge, end users of the evaluated outputs, and clinical scope of the application. The underlying data and criteria domain contained four items regarding the information needed to reproduce the evaluation and assess its clinical validity: the input data's composition, source, and availability; the evaluated outputs and sample; research ethics approval and data protection;

and the evaluation criteria's source. The LLM-Judge design and output domain comprised six items on the judge model and how it was run: the model type, name, version, and access date; the access method and execution settings; the system architecture; the consensus method when several LLMs or humans are involved; the prompting strategy and availability of the full prompt; and the evaluation method and output format. The evaluation of judge performance domain contained seven items on how the judge was validated against experts: expert qualifications; review procedures; agreement metrics with confidence intervals; error analysis of disagreements; intra-rater reliability across repeated runs; known sources of bias; and verification of the factual accuracy of the judge's rationale, including hallucinations.

Content validity

Seven experts rated the relevance of the 24 items and nine sub-items of the initial RULE-LLM draft (Supplementary Table 13). Among the 24 items, the item-level content validity index (I-CVI) ranged from 0.71 to 1.00, and two items fell below the predefined threshold of 0.83: known sources of judge bias (item 22) and differences across patient subgroups (item 23), each with an I-CVI of 0.71. All sub-items met the threshold (I-CVI, 0.86–1.00). One expert did not rate item 24, so its I-CVI (0.83) was calculated from the six available ratings. At the scale level, the S-CVI/Ave was 0.94, which met the predefined criterion (≥ 0.90), whereas the S-CVI/UA was 0.67, which did not meet the predefined criterion (≥ 0.80).

The items were revised based on the ratings and the experts' free-text comments (Supplementary Table 14). Of the two items below the threshold, item 22 was merged into its sub-item on bias mitigation, which became a conditional item, and item 23 was removed because many LaaJ datasets do not contain the demographic information needed for subgroup analysis. Items that met the threshold were also revised when the comments indicated overlap, ambiguity, or limited applicability. Items 15 and 24, on the format and factual accuracy of judge

rationales, were merged; item 16 and sub-item 8C were removed because they overlapped with other items (Supplementary Table 14); and the items on expert validation were made conditional on an expert evaluation being conducted. Data protection measures for transmitting data to external APIs were specified as a separate sub-item (Item 7A). The final framework comprised 21 items and eight sub-items across four domains (Table 1). To illustrate the level of detail expected for each item, the framework was applied item by item to two use cases: a single-LLM judge evaluating emergency department discharge summaries and a multi-LLM judge panel with physician escalation comparing two versions of a patient-facing diabetes chatbot (Figure 3).

Inter-rater reliability

Five raters, who had not participated in the content validity assessment, independently applied the 21 items and eight sub-items (29 in total) to five LaaJ studies, rating each item as "yes," "no," or "not applicable." Across all evaluated items, responses indicating that a criterion was not met ("no") accounted for 34.1% of ratings and "not applicable" for 12.9%. Overall inter-rater agreement was substantial, with a percent agreement of 80.9%, Fleiss' kappa of 0.67 (95% CI 0.60–0.74), and Gwet's AC1 of 0.73 (95% CI 0.67–0.79), according to Landis and Koch[118].

DISCUSSION

This study reviewed 91 studies that used LaaJ in medicine and developed the RULE-LLM, a reporting framework for LaaJ studies. LaaJ was most often applied in patient communication and education (34.1%) and clinical decision support (30.8%), whereas few studies examined administration and workflow (2.2%). Reference-based (63.7%) and self-defined (64.8%) criteria were used at similar rates when mixed use was included. Most studies

so far have used proprietary models, such as GPT, Claude, and Gemini, as judges and relied on a single LLM. LaaJ research in medicine is expanding quickly; however, how judges are applied and reported varies widely across studies, emphasizing the need for reporting guidelines[21,22].

The RULE-LLM was developed in this study to fill this gap. Existing frameworks cover an adjacent area: QUEST and CLEVER address human evaluation[9,13], TRIPOD-LLM and MEDAI-LLM-SUMM address LLM studies in general[18,19], READY addresses benchmark construction[20], and MedJUDGE addresses deployment governance[21]. None of them requires item-by-item reporting on how the judge model was selected and validated, how the judge system was designed (architecture and prompting strategy), or how the reliability of its outputs was verified.

MedJUDGE is the framework closest to RULE-LLM[21] because it is specific to LaaJ; however, the two differ in purpose, format, and development method. First, MedJUDGE is a governance tool in the deployment stage. It defines three clinical risk tiers (A, B, and C) and three dimensions (validity, safety, and accountability) to help institutions decide whether to deploy a LaaJ system and how closely it should be monitored. Conversely, the RULE-LLM is a reporting checklist that specifies what an individual LaaJ study should report at publication, similar to TRIPOD or CONSORT[7,119]. Second, MedJUDGE prescribes graded actions based on quantitative thresholds (e.g., positional bias variance < 5%), whereas RULE-LLM examines whether each of its 21 items and eight sub-items is reported (yes, no, or not applicable). Third, the authors of MedJUDGE described it as a conceptual framework derived from gaps in the literature, with each requirement corresponding to a testable hypothesis, and it has not undergone expert panel or reliability assessments[21]. Conversely, the RULE-LLM was derived from a scoping review and assessed for content validity by seven experts and for inter-rater reliability by five independent raters. Therefore, the two frameworks can be considered

complementary: RULE-LLM applies when a study is reported, and MedJUDGE applies when a LaaJ system is deployed in clinical practice.

Evaluating LLMs in medicine requires attention to several dimensions beyond accuracy, such as completeness, safety, and clarity, regardless of whether the evaluator is human or an LLM[9,12,15,16]. In human evaluation, the number of judgments that experts make increases with the number of responses and dimensions; therefore, the cost and time of evaluating large output sets increase rapidly[14,15]. This scalability problem is the main reason why LaaJ has attracted attention, and its use is likely to continue to grow[21,22]. However, the agreement data identified in this review do not provide sufficient evidence to support the use of LaaJ as a replacement for human evaluation[21,22,120-123]. Although 60 studies (65.9%) reported agreement between the judge and human evaluators, the metrics used, including Cohen's kappa, Pearson and Spearman correlations, ICC, and Kendall's tau, varied so widely that the agreement could not be compared or pooled across studies; therefore, although LaaJ showed some agreement with human evaluators, the extent of this agreement remains uncertain[22,123]. Judge performance may improve as models advance; however, human evaluation remains necessary[21,22].

This study's data extraction method indirectly supports this view. The overall acceptance rate of AI classifications was 88.4% ($\kappa = 0.717$, substantial agreement by the Landis and Koch criteria[118]); therefore, 11.6% of classifications were changed during human review. Even in the relatively structured task of classifying articles into predefined categories, the initial judgment of the AI differed from the final human judgment in a non-negligible share of cases. Given that the reviewers could see the AI classifications, the true discrepancy may have been larger. These results were derived from a data extraction task and cannot be directly generalized to clinical judgment. This gap may be wider in more complex and open-ended clinical tasks, although this requires empirical testing. However, the results supported the

requirement of RULE-LLM to report whether and how experts validated the judge and suggested that LaaJ was better positioned as a complement to human evaluation than as a replacement.

Hallucinations, which refer to the generation of plausible but incorrect content, are of particular concern in medicine[124,125]. Some of the included studies used LaaJ to detect hallucinations in the LLM outputs; however, their approaches differed[29,30]. One study combined rule-based checks with a separate "mirror" LLM acting as a verifier and analyzed self-reflection logs produced during optimization to examine how hallucinations occurred and were corrected[29]. Other studies did not examine hallucinations separately but considered them as part of broader domains, such as accuracy or reliability[30]. Related terms, including "hallucination," "fabricated knowledge," "invalid generation," and "information error," were also defined differently across studies. Therefore, this study did not extract hallucinations as separate items and instead described these studies narratively.

The judge can also hallucinate. Thus, it may cite evidence that does not exist or justify a score on incorrect grounds, undermining the validity of its evaluation[21,120,124]. This is not a reason to abandon LaaJ; however, the basis for the judge's factual decisions, and whether that basis was checked, should be reported. Therefore, when the judge produces rationales, the RULE-LLM asks authors to report whether and how the factual accuracy of these rationales (e.g., the presence of hallucinations) was verified (Item 21).

This study had several limitations. First, relevant studies on LaaJ may have been overlooked. Given that titles and abstracts were screened using an LLM, more studies may have been omitted than with manual screening[126-128]. This issue was partially addressed by including 22 studies identified in a previous scoping review. Given that this study aimed to develop a reporting framework rather than estimate the prevalence of particular practices, a small number of missed studies were unlikely to have substantially changed the domains or

items of the RULE-LLM.

Second, consistent with the scoping review methodology, this study did not assess the risk of bias in individual studies. The aim was to describe reporting practices and methodological characteristics across LaaJ studies rather than to appraise study quality; therefore, it was unlikely to have affected the development of the RULE-LLM.

Third, human reviewers were not blinded at several stages, and the agreement between reviewers was not quantified. The screening prompts were refined using 30 records reviewed by a single researcher. Apart from the 100-record validation set, title and abstract screening relied solely on the LLM's majority vote. Regarding the validation set and full-text review, two reviewers independently judged eligibility and resolved disagreements by discussion; however, agreement before consensus was not calculated, which is a limitation for the full-text stage, which determined the final inclusion. Regarding data extraction, the reviewers corrected the AI classifications while being aware of them; therefore, automation bias cannot be ruled out[129]. Therefore, the acceptance rate (88.4%, $\kappa = 0.717$) should be read as the proportion of AI classifications accepted after human review, not as reliability between two independent reviewers. Future studies using LLM-assisted reviews should include a subsample that is independently and blindly double-coded.

Fourth, the RULE-LLM was not developed through formal consensus. The EQUATOR Network guidelines recommend a multi-round Delphi survey followed by a consensus meeting[130], whereas this study obtained expert input through a single round of content validity assessment in which the experts independently rated the items and did not see one another's ratings or work toward consensus. In addition, the expert panel was small and was recruited through the authors' professional networks, and inter-rater reliability was assessed in five studies. Therefore, the RULE-LLM should be regarded as an initial framework, and a Delphi study with a larger and more diverse panel, including journal editors and end users

of LaaJ evaluations, is required to refine it. Given that LaaJ methods change quickly, for example, with multimodal or agent-based judges, the RULE-LLM will need to be updated as new evidence accumulates.

In addition, the S-CVI/UA (0.67) did not meet the predefined criterion of 0.80, although the S-CVI/Ave (0.94) did. The S-CVI/UA is a stringent index because a single rating of 1 or 2 from any expert excludes an item from universal agreement, and this stringency increases with panel size. In this study, six of the eight items without universal agreement received only one such rating. The two items below the I-CVI threshold were merged or removed, but the revised framework was not re-rated by the expert panel.

With the spread of generative AI in medicine, the methods used to evaluate it require rigorous validation[5,12,18,21,22]. LaaJ offers a practical way to reduce the time and cost of human evaluations; however, without a common reporting standard, the results are difficult to compare or replicate[12,21,22]. Therefore, the RULE-LLM, developed from a scoping review of LaaJ studies in medicine, provides a basis for more transparent and reproducible reporting of LaaJ-based evaluations and may support the safe adoption of generative AI in clinical settings.

METHODS

Study design

A scoping review was conducted on LaaJ studies in medicine, and the results were used to develop the RULE-LLM, following previous studies that developed frameworks and guidelines based on scoping reviews[9,20,21]. This study analyzed the evaluated medical tasks, input data, judge models and system architectures, prompting strategies and output formats, sources of evaluation criteria, clinical validation, and methods for assessing agreement. The review question was structured using the Population, Concept, and Context (PCC) framework[27]. The population comprised studies that applied LaaJ in medicine; the concept was the design characteristics of LaaJ (e.g., judge model type, prompting strategy, evaluation criteria, and methods for assessing human–AI agreement) and reporting practices; and the context was medicine and healthcare, including studies dealing with clinical tasks and patient-related data. The protocol was retrospectively registered using the OSF Registries on October 7, 2026 (https://doi.org/10.17605/OSF.IO/GYEJA). Given that registration occurred after the study was completed, it served to improve transparency and should not be regarded as preregistration. The PRISMA-ScR checklist[28] is presented in Supplementary Table 1.

Search strategy and eligibility criteria

The search strategy combined terms for LLMs and LLM-based evaluations with terms for medicine and healthcare (Supplementary Table 3). The LLM-related terms included “large language model” and “generative artificial intelligence,” and the evaluation-related terms included “automated evaluation,” “LLM as judge,” and “automated scoring.” The domain terms included “healthcare,” “medical,” and “clinical.” PubMed and Embase were searched as the main databases for medical literature, and IEEE Xplore and ACM Digital Library were

searched to capture methodological work in LLM and computer science research. Scopus and Web of Science were not searched because they overlapped substantially with PubMed and Embase. The ACL Anthology was not searched because it predominantly covers natural language processing conference papers, overlaps little with the clinical application literature of interest, and was considered adequately covered by IEEE Xplore and the ACM Digital Library. The search covered the period from January 1, 2024, to April 30, 2026, and all databases were last searched on May 14, 2026. This start date was selected because studies applying LaaJ in medicine began to accumulate after the term and method were widely adopted [8,14]. The queries were adapted to each database's indexing and syntax. The full queries are provided in Supplementary Table 3. In addition, journal articles included in a previous scoping review of LaaJ in healthcare were reviewed against the same eligibility criteria, and 22 eligible articles were added to the studies identified through the database search.

Studies were included if they met the following criteria: (1) applied LaaJ, in which one model evaluates the output of another in a clinical, medical, or healthcare context; (2) were available in full text, including preprints that had not undergone peer review; and (3) were written in English. Studies were excluded if they (1) were research letters, letters to the editor, or opinion papers, which usually lack a full description of the methods; (2) did not apply LaaJ; (3) were written in a language other than English; or (4) were outside the medical, clinical, or healthcare domains (Supplementary Table 2).

Screening and study selection

Screening comprised title and abstract reviews, followed by full-text reviews. Owing to the large number of records, title and abstract screening was assisted by an LLM. The procedure was modeled based on a previous study that applied an LLM to systematic review

screening[126-128], in which human researchers defined the eligibility criteria and prompts, refined the prompts on a small set of records, evaluated the LLM on a separate dataset, and applied it to the remaining records.

Among the 4,029 records eligible for title and abstract screening, 30 records were selected to develop the screening prompts and system. One researcher (H.Y.Y.) reviewed these records against predefined criteria, and the instructions and prompts were adjusted until the LLM classified them consistently. Subsequently, 100 records, which were not used in the prompt development, were set aside as a validation set. Two reviewers (J.B.L. and H.Y.Y.) independently assessed these records' eligibility, compared their results, and resolved all disagreements through discussion without a third reviewer. The agreement between the reviewers before consensus was not calculated. Consensus judgments served as the reference standard against which the LLM classifications were compared.

After validation, the finalized prompts and settings were applied to the remaining records. The screening was implemented in Python using the OpenAI Responses API with GPT-5 mini (OpenAI; model ID, gpt-5-mini-2025-08-07; run on July 21–22, 2026). The screening criteria and output format were provided as model instructions, and the title and abstract of each record were supplied as separate inputs. The model was instructed to classify each record as "INCLUDE" or "EXCLUDE" using only the title and abstract. To prevent instruction-like text from affecting the decision, the model was instructed to treat such text as data rather than as commands.

Each of the remaining 3,899 records was classified three times independently with the same criteria and settings, with reasoning effort set to "none," and the final decision was made by majority vote. In total, 4,029 records were screened, and 248 were selected for full-text review. Except for the 100-record validation set, human reviewers did not repeat or verify the title and abstract screening, and the LLM's majority vote was used as the final result. The full

prompts, model settings, voting procedure, and validation performance against the human reference standard are provided in Supplementary Appendix S1 and Supplementary Tables 10 and 11.

Full-text reviews were only performed by human reviewers without an LLM. Two reviewers (J.B.L. and H.Y.Y.) independently assessed the 248 articles against the same eligibility criteria and resolved disagreements through discussion without a third reviewer. Bibliographic information, such as authors, journals, and affiliations, was not masked, and agreement between the reviewers before consensus was not calculated.

Data extraction and analytic framework

The analytic framework covered the medical application, composition, and evaluation methods of LaaJ and the validation of the judge. It comprised ten domains, which were coded as 44 binary (present/absent) items: (1) medical task type (six items), (2) input data (three), (3) source of evaluation criteria (three), (4) judge function and purpose (six), (5) judge model type (three), (6) system architecture (four), (7) prompting strategy (four), (8) output format (five), (9) clinical validation (three), and (10) judge–human agreement metrics (seven). Medical task types were classified using a clinician-validated taxonomy from MedHELM[2], which included five top-level categories: clinical decision support, clinical note generation, patient communication and education, medical research assistance, and administration and workflow; studies that fit none of these were coded as "other." The names and definitions of all 44 items are provided in Supplementary Table 4, and the number of items per domain in Supplementary Table 12. Studies that belonged to two or more categories of a domain (e.g., several task types or agreement metrics) were coded for each category.

The data were extracted using several LLMs and reviewed by the researchers. Each

full text was separately analyzed using three systems: Google NotebookLM (base model, Gemini 3.0 Pro), GPT-5.5 (OpenAI; model ID, gpt-5.5), and Claude Opus 5 (Anthropic; model ID, claude-opus-5-20260115). All extraction and reconciliation runs were performed on August 25, 2026, via the API. The temperature was set to 0 for GPT-5.5 and Claude Opus 5; NotebookLM does not allow the temperature to be configured, so its default setting was used. Each article was analyzed once, and every system received the same analysis criteria, item definitions, and decision rules (Supplementary Table 4; Supplementary Appendix S2). The systems were independently run to ensure that the output of one model could not influence that of the others. In the reconciliation step, the criteria and three result sets were given to GPT-5.5 (temperature 0), which determined the final classification based on the original text and criteria. Subsequently, two reviewers (J.B.L. and H.Y.Y.) checked the reconciled results for the 44 classification items against each original article. Given that the reviewers viewed the AI classifications, this was a human-in-the-loop correction rather than a blinded validation, and the resulting acceptance rate and Cohen's kappa approximated the rate at which AI classifications were accepted after human review, not the reliability between independent raters. The two reviewers compared their corrections and resolved disagreements by discussion without a third reviewer; agreement before consensus was not calculated.

Development of RULE-LLM

The development of the RULE-LLM was informed by the EQUATOR Network guidance for developing health research reporting guidelines[130]; however, not all of its recommended steps, such as a formal Delphi process and a consensus meeting, were conducted. The development proceeded in six steps: (1) derivation of candidate items; (2) grouping of items into domains; (3) iterative review and consensus among authors; (4) content validity assessment by external experts; (5) finalization of items; and (6) assessment of inter-rater

reliability by independent raters.

First, candidate items were derived from the characteristics extracted from the 91 included studies across the ten domains (44 binary items) of the analytic framework (Supplementary Tables 4 and 12). Characteristics whose reporting status or level of detail varied across studies were compiled as the initial candidate items.

Second, the candidate items were grouped into four domains according to the data extraction framework's structure. The domains were arranged to follow the LaaJ study sequence: defining the role and task, setting the data and evaluation criteria, designing the judge system and its output, and validating its performance.

Third, the coauthors independently reviewed the initial item list drafted by the first author (J.B.L.) and commented on each item regarding (a) whether its link to the scoping review findings was clear, (b) whether it overlapped with other items or was ambiguous, and (c) whether it was worded such that compliance could be checked in an actual study. Items for which opinions differed were reworded after discussion among the authors.

Fourth, content validity was assessed by an external panel of seven experts recruited through the authors' professional networks, including two clinicians, three researchers in AI or health informatics, and two regulatory experts who formerly served at the Ministry of Food and Drug Safety of the Republic of Korea. All experts were affiliated with institutions other than those of the authors, and they all provided consent before participating. Each expert independently rated the 24 items and nine sub-items on a 4-point scale (1 = not at all relevant; 2 = somewhat relevant; 3 = relevant; and 4 = very relevant)[131-133] and provided free-text comments on clarity, wording, and whether items should be deleted or merged.

Fifth, the items were finalized based on expert ratings and comments. The I-CVI was calculated as the number of experts rating an item 3 or 4 divided by the number of experts who rated that item (n = 7 for all items except item 24, which was rated by six experts). At the scale

level, this study calculated S-CVI/Ave (the mean I-CVI across all items) and S-CVI/UA (the proportion of items rated 3 or 4 by all experts). Items with an I-CVI < 0.83 were revised, merged, or deleted based on the experts' comments. This threshold is more conservative than the commonly recommended minimum of 0.78 for panels of six or more experts[131,133] and was chosen because the panel was small. The scale was considered acceptable if the S-CVI/Ave was at least 0.90[133] and the S-CVI/UA was at least 0.80[132].

Sixth, inter-rater reliability was assessed by five raters who had not participated in the content validity assessment or the development of the framework. Before rating, the raters received written instructions on the items and the definitions of "yes," "no," and "not applicable," without practice on the target studies. Five LaaJ studies were randomly selected from the included studies. Each rater independently applied the final items to the five studies without discussion.

Statistical analysis

All analyses were performed using Stata version 16. The 95% CI for the acceptance rate of AI classifications was calculated using the Wilson score method[134], which did not account for the clustering of classifications within the 91 studies. Agreement between the initial AI classification and the final human-corrected classification was assessed using Cohen's kappa[135], and the precision, recall, specificity, and F1 score of the initial AI classification were calculated with the presence of a characteristic ("O") as the positive class.

The inter-rater reliability of the RULE-LLM was assessed using the ratings of the five raters on the 21 items and eight sub-items (29 in total) applied to five LaaJ studies, with each rating classified into three categories (yes, no, and not applicable). Agreement across all item–study combinations was summarized as the percentage of agreement and Fleiss' kappa[136]. Given that kappa can be low when ratings are concentrated in one category (the kappa

paradox)[137], Gwet's AC1[138] was also calculated. The 95% CIs for kappa and AC1 were derived from their standard errors. Agreement statistics were calculated using the kappaetc package[139] in Stata, and kappa values were interpreted according to Landis and Koch's criteria[118].

DATA AVAILABILITY

The coding sheet for the 91 included studies, screening decisions, and LLM prompts and outputs used for screening, data extraction, and reconciliation are provided in the Supplementary. All other data supporting these findings are available in the published articles cited in this review.

CODE AVAILABILITY

The Python code used for LLM-assisted title and abstract screening, data extraction, and reconciliation, including the API calls and the voting procedure, is provided as Supplementary.

ACKNOWLEDGMENTS

This study was supported by grants from the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: RS-2025-02223416), and the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (grant number: RS-2026-25580682). The funders had no role in the study design, data collection and analysis, decision to publish, or manuscript preparation.

During the preparation of this manuscript, the authors used Claude (Anthropic) to improve the clarity and readability of the English language, to assist in generating the data visualizations in Figures 2 and 3, and to cross-check the consistency between the main text and the supplementary data. All underlying data were derived from the study's own dataset; apart from the uses described in the Methods, Claude was not used to generate, alter, or estimate any data values. The authors reviewed and edited all AI-assisted text and figures and take full

responsibility for the content of the manuscript. The manuscript was subsequently reviewed by a professional English-language editing service (Editage). The use of LLMs for study screening, data extraction, and reconciliation is described in the Methods.

AUTHOR CONTRIBUTIONS

Concept and design: J.B.L., H.Y.Y., and Y.B.J.

Design and execution of LLM-assisted screening and data extraction pipeline: H.Y.Y.

Acquisition, analysis, or interpretation of data: J.B.L., H.Y.Y., and Y.B.J.

Drafting of the manuscript: J.B.L. and H.Y.Y.

Critical revision of the manuscript for important intellectual content: J.B.L., H.Y.Y., and Y.B.J.

Administrative, technical, or material support: J.B.L. and Y.B.J.

Supervision: J.B.L. and Y.B.J.

All authors have read and approved the final manuscript and agreed to be accountable for all aspects of the work, ensuring that questions related to the accuracy or integrity of any part of the work were appropriately investigated and resolved.

COMPETING INTERESTS

All authors declare no financial or non-financial competing interests.

References

1. Thirunavukarasu, A. J. et al. Large language models in medicine. *Nat Med* 29, 1930–1940 (2023).
2. Bedi, S. et al. Holistic evaluation of large language models for medical tasks with MedHELM. *Nat Med* 32, 943–951 (2026).
3. Chang, H., Jung, J. W. & Kim, Y. Comparative Analysis of Large Language Models for Answering Cancer-Related Questions in Korean. *Yonsei Med J* 66, 405–411 (2025).
4. Lim, J. Y., Kim, K. H. & Mun, S. K. How Does Medical Artificial Intelligence Revolutionize Physician Productivity? *Yonsei Med J* 67, 1–8 (2026).
5. Lekadir, K. et al. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ* 388, e081554 (2025).
6. Tejani, A. S. et al. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 Update. *Radiol Artif Intell* 6, e240300 (2024).
7. Collins, G. S. et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* 385, e078378 (2024).
8. Liu, Y. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* 2511–2522 (2023).
9. Tam, T. Y. C. et al. A framework for human evaluation of large language models in healthcare derived from literature review. *NPJ Digit Med* 7, 258 (2024).
10. Fraile Navarro, D. et al. Expert evaluation of large language models for clinical dialogue summarization. *Sci Rep* 15, 1195 (2025).
11. Van Veen, D. et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nat Med* 30, 1134–1142 (2024).
12. Bedi, S. et al. Testing and Evaluation of Health Care Applications of Large Language Models: A Systematic Review. *JAMA* 333, 319–328 (2025).
13. Kocaman, V., Kaya, M. A., Feier, A. M. & Talby, D. Clinical Large Language Model Evaluation by Expert Review (CLEVER): Framework Development and Validation. *JMIR AI* 4, e72153 (2025).
14. Zheng, L. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36* 46595–46623 (2023).
15. Mallinar, N. et al. A scalable framework for evaluating health language models. *NPJ Digit Med* 9, 437 (2026).
16. Yang, E., Ko, S. & Woo, H. The Reliability of Human Evaluation of Large Language Models in Health Care Settings: Scoping Review. *J Med Internet Res* 28, e98184 (2026).
17. Chiang, C. H. & Lee, H. y. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 15607–15631 (2023).
18. Gallifant, J. et al. The TRIPOD-LLM reporting guideline for studies using large language

models. *Nat Med* 31, 60–69 (2025).
19. Khoruzhaya, A. N. et al. MEDAI-LLM-SUMM: a reporting checklist for medical text summarization studies using large language models. *Front Digit Health* 8, 1761601 (2026).
20. Lee, J., Shin, J. & Cho, B. Structured taxonomy and framework for developing medical benchmark in large language models derived from scoping review. *NPJ Digit Med* 9, 548 (2026).
21. Li, C. A scoping review of LLM-as-a-judge in healthcare and the MedJUDGE framework. In *arXiv* 2026).
22. Li, L. LLM-as-a-judge in healthcare: a scoping analysis of applications, methods, and human alignment. In *arXiv* 2026).
23. Wang, P. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2024).
24. Li, Z. Split and merge: aligning position biases in LLM-based evaluators. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (2024).
25. Arksey, H. & O'Malley, L. Scoping studies: towards a methodological framework. *Int J Soc Res Methodol* 8, 19–32 (2005).
26. Levac, D., Colquhoun, H. & O'Brien, K. K. Scoping studies: advancing the methodology. *Implement Sci* 5, 69 (2010).
27. Peters, M. D. J. et al. Updated methodological guidance for the conduct of scoping reviews. *JBI Evid Implement* 19, 3–10 (2021).
28. Tricco, A. C. et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Ann Intern Med* 169, 467–473 (2018).
29. Wang, Q. et al. A feasibility study of automating radiotherapy planning with large language model agents. *Phys Med Biol* 70, 075007 (2025).
30. Choi, D. et al. Automated chain-of-thought evaluation framework for large language model-generated emergency department documentation: a simulation-based study. *Clin Exp Emerg Med* 13, 53–64 (2026).
31. Matsui, K. et al. Human-Comparable Sensitivity of Large Language Models in Identifying Eligible Studies Through Title and Abstract Screening: 3-Layer Strategy Using GPT-3.5 and GPT-4 for Systematic Reviews. *J Med Internet Res* 26, e52758 (2024).
32. Curran, C., Neehal, N., Murugesan, K. & Bennett, K. P. Examining Trustworthiness of LLM-as-a-Judge Systems in a Clinical Trial Design Benchmark. In *2024 IEEE International Conference on Big Data (BigData)* 4627–4631 (IEEE, 2024).
33. He, Z. et al. Quality of Answers of Generative Large Language Models Versus Peer Users for Interpreting Laboratory Test Results for Lay Patients: Evaluation Study. *J Med Internet Res* 26, e56655 (2024).
34. Heilmeyer, F. et al. Viability of Open Large Language Models for Clinical Documentation in German Health Care: Real-World Model Evaluation Study. *JMIR Medical Informatics* 12, e59617 (2024).

35. Jin, Y. et al. Better to Ask in English: Cross-Lingual Evaluation of Large Language Models for Healthcare Queries. In *Proceedings of the ACM Web Conference 2024* 2627–2638 (ACM, 2024).
36. Kim, S. H. et al. Performance of Open-Source LLMs in Challenging Radiological Cases – A Benchmark Study on 1,933 Eurorad Case Reports. *medRxiv* https://doi.org/10.1101/2024.09.04.24313026 (2024). (2024).
37. Li, Z. et al. CMCOQA: A Chinese Medical Complex Open-Question Answering Benchmark. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* 3402–3407 (IEEE, 2024).
38. Liu, L. et al. Towards Automatic Evaluation for LLMs' Clinical Capabilities: Metric, Data, and Algorithm. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* 5466–5475 (ACM, 2024).
39. Wang, Z., Zhou, B. & Liu, W. ATSCOT: Chain of Thought for Structuring Anesthesia Medical Records. In *2024 International Conference on New Trends in Computational Intelligence (NTCI)* 145–149 (IEEE, 2024).
40. Zhang, G. et al. Closing the gap between open source and commercial large language models for medical evidence summarization. *NPJ Digit Med* 7, 239 (2024).
41. Zhu, Q. et al. Leveraging Professional Radiologists' Expertise to Enhance LLMs' Evaluation for AI-generated Radiology Reports. In *2024 IEEE 12th International Conference on Healthcare Informatics (ICHI)* 402–411 (IEEE, 2024).
42. Adib, S. E. S. et al. Assessing Large Language Models for Medical QA: Zero-Shot and LLM-as-a-Judge Evaluation. In *2025 28th International Conference on Computer and Information Technology (ICCIT)* 3081–3086 (IEEE, 2025).
43. Banerjee, O. et al. ReXamine-Global: A Framework for Uncovering Inconsistencies in Radiology Report Generation Metrics. In *Biocomputing 2025* 185–198 (WORLD SCIENTIFIC, 2024).
44. Bayram, M. A., Diri, B. & Yildirim, S. Healthcare-Focused Turkish Medical LLM: Training on Real Patient-Doctor Question-Answer Data for Enhanced Medical Insight. *ACM Transactions on Asian and Low-Resource Language Information Processing* 24, 1–13 (2025).
45. Bui, N. et al. Fine-tuning large language models for improved health communication in low-resource languages. *Computer Methods and Programs in Biomedicine* 263, 108655 (2025).
46. Changpun, R. et al. Agentic Stage-Based LLM Framework for Multi-Turn Mental Health Support Conversations in Thai. In *2025 20th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)* 1–6 (IEEE, 2025).
47. Chen, Y., Wen, B. & Zulkernine, F. A Multiagent Summarization and Auto-Evaluation Framework for Medical Text: Development and Evaluation Study. *JMIR AI* 4, e75932 (2025).
48. Croxford, E. et al. Evaluating clinical AI summaries with large language models as judges. *NPJ Digit Med* 8, 640 (2025).
49. Croxford, E. et al. Automating Evaluation of AI Text Generation in Healthcare with a Large

Language Model (LLM)-as-a-Judge. *medRxiv* https://doi.org/10.1101/2025.04.22.25326219 (2025). (2025).

50. Fukui, Y. et al. Evaluation of a retrieval-augmented generation system using a Japanese Institutional Nuclear Medicine Manual and large language model-automated scoring. *Radiological Physics and Technology* 18, 861–876 (2025).

51. Jarchow, H. et al. Benchmarking large language models for personalized, biomarker-based health intervention recommendations. *NPJ Digit Med* 8, 631 (2025).

52. Kaumenova, T. et al. Evaluating Large Language Models for Colonoscopy Preparation Assistance: Correctness and Diversity in Synthetic Dialogues. *medRxiv* https://doi.org/10.1101/2025.11.19.25340596 (2025). (2025).

53. Li, C.-Y. et al. Towards a holistic framework for multimodal LLM in 3D brain CT radiology report generation. *Nature Communications* 16, 2258 (2025).

54. Liu, S. et al. Optimizing Order Sets With a Large Language Model–Powered Multiagent System. *JAMA Network Open* 8, e2533277 (2025).

55. Liu, S. & Kim, J. MedGemma-Critic: Fine-Tuning Medical Language Models for Domain-Specialised Text Evaluation. In *2025 International Conference on Digital Image Computing: Techniques and Applications (DICTA)* 1–5 (IEEE, 2025).

56. Madrid-García, A. et al. 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 (2025).

57. Magnini, M., Aguzzi, G. & Montagna, S. Open-source small language models for personal medical assistant chatbots. *Intelligence-Based Medicine* 11, 100197 (2025).

58. Mahadik, R., Mandal, A., Tripathi, S. & Jaiswal, S. LLM Evaluations for Emotional Expressiveness and Factual Consistency in Medical Dialogue Systems Using LLM-as-a-Judge. In *2025 International Conference on Responsible, Generative and Explainable AI (ResGenXAI)* 1–6 (IEEE, 2025).

59. Ozgun, M. C. et al. Trustworthy AI Psychotherapy: Multi-Agent LLM Workflow for Counseling and Explainable Mental Disorder Diagnosis. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management* 2263–2272 (ACM, 2025).

60. Saito, T., Yamanaka, R., Wakabayashi, Y. & Kitaoka, N. Generation and Automatic Evaluation of Soap Notes from Medical Dialogue Using Large Language Models. In *2025 28th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)* 1–6 (IEEE, 2025).

61. Sarvari, P. & Al-fagih, Z. Rapidly Benchmarking Large Language Models for Diagnosing Comorbid Patients: Comparative Study Leveraging the LLM-as-a-Judge Method. *JMIRx Med* 6, e67661 (2025).

62. Schiezaro, M., Rosa, G., Campos, B. A. G. & Pedrini, H. Guardians of the data: NER and LLMs for effective medical record anonymization in Brazilian Portuguese. *Frontiers in Public Health*

13, 1717303 (2026).

63. Sesen, M. B., Au Yeung, J. & Asgari, E. Development and validation of Retrieval Augmented Generation (RAG) and GraphRAG for complex clinical cases. Preprint at https://doi.org/10.1101/2025.11.25.25341010 (2025).

64. Su, X. et al. MedCritical: Enhancing Medical Reasoning in Small Language Models via Self-Collaborative Correction. Preprint at https://doi.org/10.2139/ssrn.5457009 (2025).

65. Suzić, S. et al. Enhancing Call Center Support for Child Care: Retrieval-Augmented Generation in Serbian. In Proc. 2025 33rd *Telecommunications Forum (TELFOR)* 1–4 (IEEE, 2025).

66. Thomas, J. et al. Preserving Privacy, Increasing Accessibility, and Reducing Cost: An On-Device Artificial Intelligence Model for Medical Transcription and Note Generation. Preprint at https://doi.org/10.1101/2025.07.01.25330679 (2025).

67. Tian, W. et al. A Medical Multimodal Large Language Model for Pediatric Pneumonia. *IEEE J. Biomed. Health Inform.* **29**, 6869–6882 (2025).

68. Tripathi, A. et al. Employing Consensus-Based Reasoning with Locally Deployed LLMs for Enabling Structured Data Extraction from Surgical Pathology Reports. Preprint at https://doi.org/10.1101/2025.04.22.25326217 (2025).

69. Vasilev, Y. et al. Evaluating Medical Text Summaries Using Automatic Evaluation Metrics and LLM-as-a-Judge Approach: A Pilot Study. *Diagnostics* **16**, 3 (2025).

70. Wang, B. et al. From Scores to Steps: Diagnosing and Improving LLM Performance in Evidence-Based Medical Calculations. In Proc. 2025 *Conference on Empirical Methods in Natural Language Processing* 10820–10844 (Association for Computational Linguistics, 2025).

71. Wang, Z. et al. PALLM: Evaluating and Enhancing Palliative Care Conversations with Large Language Models. *ACM Trans. Comput. Healthc.* **7**, 1–24 (2026).

72. Williams, G., Rutunda, S., Nzabakira, F. & Mateen, B. A. Human Evaluators vs. LLM-as-a-Judge: Toward Scalable, Real-Time Evaluation of GenAI in Global Health. Preprint at https://doi.org/10.1101/2025.10.27.25338910 (2025).

73. Wu, D. J. H. et al. Automated Evaluation of Large Language Model Response Concordance with Human Specialist Responses on Physician-to-Physician eConsult Cases. In *Biocomputing 2026* 372–387 (World Scientific, 2025).

74. Wu, J. et al. Dual-stage and Lightweight Patient Chart Summarization for Emergency Physicians. In Proc. 2025 IEEE Annual *Congress on Artificial Intelligence of Things (AIoT)* 43–50 (IEEE, 2025).

75. Yang, J., Shu, L., Duan, H. & Li, H. RDguru: A Conversational Intelligent Agent for Rare Diseases. *IEEE J. Biomed. Health Inform.* **29**, 6366–6378 (2025).

76. Yang, Z. & Rahmani, A. M. Personalized Causal Graph Reasoning for LLMs: An Implementation for Dietary Recommendations. *IEEE J. Biomed. Health Inform.* **29**, 8767–8774 (2025).

77. Yao, Z. et al. ChatCLIDS: Simulating Persuasive AI Dialogues to Promote Closed-Loop Insulin

Adoption in Type 1 Diabetes Care. *Proc. AAAI Conf. Artif. Intell.* **40**, 39539–39547 (2026).
78. Yao, Z. et al. DischargeSim: A Simulation Benchmark for Educational Doctor–Patient Communication at Discharge. In Proc. 2025 *Conference on Empirical Methods in Natural Language Processing* 10783–10809 (Association for Computational Linguistics, 2025).
79. Wang, Y., Meng, K., Deng, J. & Ren, F. MedGuard-KG: Dual-Stream Knowledge Graph Verification for Safe Clinical Reasoning in Medical LLMs. In Proc. 2025 22nd *International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)* 1–7 (IEEE, 2025).
80. Zhang, C. et al. TeenEmpath: Towards Adolescent Psychological Counseling With Multiple Personas and Strategies. *IEEE Trans. Affect. Comput.* https://doi.org/10.1109/taffc.2025.3638958 (2025).
81. Zhou, S. et al. Automating expert-level medical reasoning evaluation of large language models. *npj Digit. Med.* **9**, 34 (2025).
82. Antaki, F. et al. Performance of GPT-5 Frontier Models in Ophthalmology Question Answering. *Ophthalmol. Sci.* **6**, 101034 (2026).
83. Borisov, A. et al. Key aspects of fine-tuning and applying LLM-as-a-judge for clinical data summaries in the radiological workflow. *Front. Artif. Intell.* **9**, 1768005 (2026).
84. Cornelius, J. & Rinaldi, F. Measuring the gap: correlating synthetic-to-real drift with PHI de-identification performance. *Genomics Inform.* **24**, 10 (2026).
85. Croitoru, I., Turcu, C. E. & Turcu, C. O. Privacy-by-Design in AI-Assisted Systems for Caregivers of Children with Autism: A Secure Multi-Agent Architecture. *Appl. Sci.* **16**, 2157 (2026).
86. Fajardo V, D., Proniakin, O., Gruber, V.-E. & Marinescu, R. MedPI: Evaluating AI Systems in Medical Patient-facing Interactions. Preprint at https://doi.org/10.64898/2025.12.24.25342982 (2026).
87. Fouda, A. E., Hassan, A. A., Hanafy, R. J. & Fouda, M. E. PsychiatryBench: a multi-task benchmark for LLMs in psychiatry. *npj Digit. Med.* **9**, 320 (2026).
88. Huang, W. et al. LLM-AutoDP: Automatic Data Processing via LLM Agents for Model Fine-tuning. *Proc. VLDB Endow.* **19**, 794–807 (2026).
89. Li, H. et al. Scaling medical device regulatory science using large language models. *npj Digit. Med.* **9**, 221 (2026).
90. McCoy, L. G. et al. Asking the Right Questions: Benchmarking Large Language Models in the Development of Clinical Consultation Templates. In *Biocomputing 2026* 400–416 (World Scientific, 2025).
91. Onovo, A. A. & Cherima, Y. J. AlignInsight: A Three-Layer Framework for Detecting Deceptive Alignment and Evaluation Awareness in Healthcare AI Systems. Preprint at https://doi.org/10.64898/2026.01.17.26344330 (2026).
92. Prem, P. et al. MedEvalarena: A Self-Generated, Peer-Judged Benchmark for Medical Reasoning. Preprint at https://doi.org/10.64898/2026.01.27.26344905 (2026).
93. Sengupta, D. & Panda, S. Disagreement between human and AI evaluation of treatment

plans. *Sci. Rep.* **16**, 4798 (2026).
94. She, D. et al. MindChat-R0: A Large Language Model for Emotionally Supportive Dialogue through Reinforcement Learning. In *Companion of the 2025 ACM International Joint Conference on Pervasive and Ubiquitous Computing* 1209–1216 (ACM, 2025).
95. Taskin, K. & Yilmaz Ak, H. Artificial intelligence vs. human evaluation of anesthesia education videos: a comparative analysis using validated quality scales. *Front. Med.* **13**, 1752664 (2026).
96. Yan, C. et al. GEN-KnowRD: Reframing AI for Rare Disease Recognition. Preprint at https://doi.org/10.64898/2026.03.02.26347469 (2026).
97. Zhao, M. et al. Automating evaluation of LLM-generated responses to patient questions about rare diseases. *JAMIA Open* **9**, ooag054 (2026).
98. Abrar, M., Sermet, Y. & Demir, I. An Empirical Evaluation of Large Language Models on Consumer Health Questions. *BioMedInformatics* **5**, 12 (2025).
99. Badawi, A. et al. When Can We Trust LLMs in Mental Health? Large-Scale Benchmarks for Reliable LLM Evaluation. In Proc. 19th *Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)* 3873–3896 (Association for Computational Linguistics, 2026).
100. Bolpagni, M. et al. VALISE: A Virtual Agent Laboratory for Instruction-Following Simulation and Evaluation of LLM-Powered Digital Health Interventions. In *Frontiers in Artificial Intelligence and Applications* (IOS Press, 2025).
101. De la Iglesia, I. et al. Ranking Over Scoring: Towards Reliable and Robust Automated Evaluation of LLM-Generated Medical Explanatory Arguments. In Proc. 31st *International Conference on Computational Linguistics* 9456–9471 (Association for Computational Linguistics, 2025).
102. Ellis, Z. et al. WER is Unaware: Assessing How ASR Errors Distort Clinical Understanding in Patient Facing Dialogue. In Proc. 16th *International Workshop on Spoken Dialogue System Technology* 391–417 (Association for Computational Linguistics, 2026).
103. Feng, Y. et al. PASS: Probabilistic Agentic Supernet Sampling for Interpretable and Adaptive Chest X-Ray Reasoning. *Proc. AAAI Conf. Artif. Intell.* **40**, 30717–30725 (2026).
104. Jang, D. et al. MedTutor: A Retrieval-Augmented LLM System for Case-Based Medical Education. In Proc. 2025 *Conference on Empirical Methods in Natural Language Processing: System Demonstrations* 319–353 (Association for Computational Linguistics, 2025).
105. Kim, E., Foty, R., Shrestha, M. & Seyfert-Margolis, V. Conformal Prediction and Verification of Large Language Model Extractions in EHR Data. *Proc. AAAI Symp. Ser.* **7**, 539–546 (2025).
106. Laskar, M. T. R. et al. Improving Automatic Evaluation of Large Language Models (LLMs) in Biomedical Relation Extraction via LLMs-as-the-Judge. In Proc. 63rd *Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 25483–25497 (Association for Computational Linguistics, 2025).
107. Liu, S., Chen, Y. & Wang, J. A Lightweight Large Language Model for Personal Sleep Quality Estimation. In Proc. 2025 IEEE 20th *Conference on Industrial Electronics and Applications*

(ICIEA) 1–6 (IEEE, 2025).

108. Madani, N. & Srihari, R. ESC-Judge: A Framework for Comparing Emotional Support Conversational Agents. In Proc. 2025 *Conference on Empirical Methods in Natural Language Processing* 16059–16076 (Association for Computational Linguistics, 2025).

109. Miranda, M., Bratières, S., Patarnello, S. & Lilli, L. Mamma Mia! Where's My Name? De-Identifying Italian Clinical Notes with Large Language Models. In Proc. Eleventh *Italian Conference on Computational Linguistics (CLiC-it 2025)* 735–746 (CEUR Workshop Proceedings, 2025).

110. Niculae, A., Cosma, A., Dumitrache, C. & Radoi, E. Dr. Copilot: A Multi-Agent Prompt Optimized Assistant for Improving Patient-Doctor Communication in Romanian. In Proc. 2025 *Conference on Empirical Methods in Natural Language Processing: Industry Track* 1780–1792 (Association for Computational Linguistics, 2025).

111. Saggar, A. et al. AI-simulated clinical consultations: Assessing the potential of ChatGPT to support medical training. *Arch. Dis. Child.* **111**, 638–644 (2026).

112. Shah, R. S. et al. TN-Eval: Rubric and Evaluation Protocols for Measuring the Quality of Behavioral Therapy Notes. In Proc. 63rd *Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)* 179–199 (Association for Computational Linguistics, 2025).

113. She, W. J. et al. EmplifAI: a Fine-grained Dataset for Japanese Empathetic Medical Dialogues in 28 Emotion Labels. In Proc. 14th *International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics* 1116–1131 (The Asian Federation of Natural Language Processing and The Association for Computational Linguistics, 2025).

114. Steinigen, D. et al. Fact Finder - Enhancing Domain Expertise of Large Language Models by Incorporating Knowledge Graphs. In Proc. 19th *Conference of the European Chapter of the Association for Computational Linguistics (Volume 3: System Demonstrations)* 101–110 (Association for Computational Linguistics, 2026).

115. Wada, A. et al. Retrieval-augmented generation elevates local LLM quality in radiology contrast media consultation. *npj Digit. Med.* **8**, 395 (2025).

116. Wang, Z. et al. HealthQ: Unveiling questioning capabilities of LLM chains in healthcare conversations. *Smart Health* **36**, 100570 (2025).

117. Yao, Z. et al. MedQA-CS: Objective Structured Clinical Examination (OSCE)-Style Benchmark for Evaluating LLM Clinical Skills. In Proc. 19th *Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)* 6183–6257 (Association for Computational Linguistics, 2026).

118. Landis, J. R. & Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics* **33**, 159–174 (1977).

119. Hopewell, S. et al. CONSORT 2025 statement: updated guideline for reporting randomised trials. *BMJ* **389**, e081123 (2025).

120. Chen, G. H. Humans or LLMs as the judge? A study on judgement bias. In Proc. 2024 *Conference on Empirical Methods in Natural Language Processing* (2024).
121. Koo, R. Benchmarking cognitive biases in large language models as evaluators. In *Findings of the Association for Computational Linguistics: ACL 2024* (2024).
122. Panickssery, A., Bowman, S. R. & Feng, S. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems 37* 68772–68802 (2024).
123. Szymanski, A. Limitations of the LLM-as-a-judge approach for evaluating LLM outputs in expert knowledge tasks. In Proc. 30th *International Conference on Intelligent User Interfaces (IUI '25)* (2025).
124. Ji, Z. et al. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* **55**, 1–38 (2023).
125. Asgari, E. et al. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. *npj Digit. Med.* **8**, 274 (2025).
126. Issaiy, M. et al. Methodological insights into ChatGPT's screening performance in systematic reviews. *BMC Med. Res. Methodol.* **24**, 78 (2024).
127. Khraisha, Q. et al. Can large language models replace humans in systematic reviews? Evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. *Res. Synth. Methods* **15**, 616–626 (2024).
128. Scherbakov, D. et al. The emergence of large language models as tools in literature reviews: a large language model-assisted systematic review. *J. Am. Med. Inform. Assoc.* **32**, 1071–1086 (2025).
129. Goddard, K., Roudsari, A. & Wyatt, J. C. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *J. Am. Med. Inform. Assoc.* **19**, 121–127 (2012).
130. Moher, D., Schulz, K. F., Simera, I. & Altman, D. G. Guidance for developers of health research reporting guidelines. *PLoS Med.* **7**, e1000217 (2010).
131. Lynn, M. R. Determination and quantification of content validity. *Nurs. Res.* **35**, 382–386 (1986).
132. Davis, L. L. Instrument review: getting the most from a panel of experts. *Appl. Nurs. Res.* **5**, 194–197 (1992).
133. Polit, D. F. & Beck, C. T. The content validity index: are you sure you know what's being reported? Critique and recommendations. *Res. Nurs. Health* **29**, 489–497 (2006).
134. Wilson, E. B. Probable Inference, the Law of Succession, and Statistical Inference. *J. Am. Stat. Assoc.* **22**, 209–212 (1927).
135. Cohen, J. A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.* **20**, 37–46 (1960).
136. Fleiss, J. L. Measuring nominal scale agreement among many raters. *Psychol. Bull.* **76**, 378–382 (1971).
137. Feinstein, A. R. & Cicchetti, D. V. High agreement but low kappa: I. The problems of two paradoxes. *J. Clin. Epidemiol.* **43**, 543–549 (1990).

138. Gwet, K. L. Computing inter-rater reliability and its variance in the presence of high agreement. *Br. J. Math. Stat. Psychol.* **61**, 29–48 (2008).
139. Klein, D. Implementing a General Framework for Assessing Interrater Agreement in Stata. *Stata J.* **18**, 871–901 (2018).

Table 1. RULE-LLM reporting framework

No	Item	Answer (Yes, No, N/A)	Page
R – Role & Task Specification: The medical task type and functional purpose of the LLM-as-a-Judge should be clearly defined before application.			
1	**Is the medical task type specified?** e.g.) clinical decision support, clinical note generation, patient communication and education, medical research assistance, administration and workflow		
2	**Are the primary function and purpose of the Judge system clearly defined?** e.g.) response quality assessment, model comparison, output verification, improvement feedback, data annotation, synthetic data filtering and screening		
3	**Is the end-user group of the evaluated outputs described?** e.g.) clinicians, nurses, patients and caregivers, healthcare administrators		
4	**Is the clinical scope of the evaluated outputs specified?** e.g.) medical specialty, disease group, care setting (outpatient, inpatient, emergency)		
U – Underlying Data & Criteria: To ensure reproducibility, the characteristics of the input data and evaluation sample, research ethics procedures, and the source of the evaluation criteria should be transparently reported.			
5	**Are the composition, source, and availability of the data input to the Judge system for evaluation specified?** e.g.) original questions, clinical cases, medical records, reference answers		
6	**Are the evaluation targets and sample described?** e.g.) models used to generate the evaluated outputs (name, version), number of outputs evaluated, sampling method (random, stratified, convenience), rationale for sample size		
7	**Are research ethics approval and data protection procedures described?** e.g.) IRB approval number or reason for exemption (e.g., use of synthetic data), waiver of consent, de-identification/pseudonymization methods, data use approvals and agreements		
	7A†. If data were transmitted to an external API, are the data protection measures described? e.g.) data retention and training-use exclusion terms (zero data retention), de-identification before transmission		
8	**Is the source of the evaluation criteria explicitly specified?** e.g.) reference-based (gold standard, clinical guidelines), self-defined criteria		

	8A†. If reference-based criteria are used, is the gold standard or reference clearly presented?		
	8B†. If self-defined criteria are used, are the evaluation dimensions and items of those criteria completely and clearly described?		
L – LLM-Judge Design & Output: To enable reproducible and consistent evaluation, the Judge system's model and execution settings, architecture, prompting strategy and its development process, and evaluation method and output format should be described in sufficient detail.			
9	**Are the type of Judge model, the specific name and version of the model used, and the access date or evaluation time point reported?** e.g.) type (proprietary, open-source, medically fine-tuned), model name and version (e.g., gpt-4o-2024-08-06, claude-3-5-sonnet-20241022, Llama-3.1-70B-Instruct), access date or evaluation time point (e.g., API calls from March 10 to 24, 2025), and for medically fine-tuned models, the base model and an overview of the fine-tuning data		
10	**Are the access method and execution settings of the Judge model reported?** e.g.) access method (API, local deployment), generation parameters (temperature, top_p), use and level of reasoning mode		
11	**Is the Judge system's architecture described?** e.g.) single-LLM system, multi-LLM system (panel/ensemble)		
12†	**If a multi-LLM or human–AI collaborative architecture is used, are the aggregation/consensus methods or division of roles described?** e.g.) majority voting, score averaging, hierarchical review, criteria for escalation to human review		
13	**Is the prompting strategy specified, and is the full prompt text included in a publicly accessible form?** e.g.) zero-shot, few-shot, chain-of-thought (CoT), rubric-based prompting		
	13A†. If prompts or rubrics were developed or revised, are the process and the separation between development data and performance validation data described? e.g.) number of and criteria for revisions, size of the development and validation sets, exclusion of overlap between few-shot examples and validation data		
14	**Are the Judge system's evaluation method and output format clearly defined?** e.g.) evaluation method (pointwise, pairwise, listwise), judgment type (score, binary judgment, preference/ranking), output encoding (free text, structured output [JSON])		
	14A†. If a rating scale is used, are the scale range, descriptors (anchors) for each score, and method of interpretation described? e.g.) 1–5 Likert scale with explicit descriptors for each score		
	14B†. If pairwise comparison or ranking is used, are the judgment criteria, tie-handling method, and ranking method clearly defined? e.g.) whether ties are allowed, win rate, Elo or Bradley-Terry models		
	14C†. If multiple evaluation dimensions are used, is the method for combining dimension-level results into an overall judgment described? e.g.) simple average, weighted sum (including rationale for the weights), fail rules when a mandatory dimension is not met		

E – Evaluation of Judge Performance: The validity and reliability of the Judge system should be verified through expert review and quantitative meta-evaluation metrics, and known limiting factors should be analyzed.			
15†	**If expert evaluation is performed, are the qualifications of the experts involved in validating the Judge system described?** e.g.) profession (physician, nurse, etc.), specialty, years of experience, number of experts, independence from the research team (authorship)		
16†	**If expert evaluation is performed, are the expert review procedures and methods described?** e.g.) rater training and calibration, independent and blinded review, adjudication procedure and consensus process for disagreements, number of reviewed items and selection method, whether experts had access to the Judge's judgments before their own review		
17†	**If expert evaluation is performed, is the agreement between the Judge system and expert raters reported using metrics appropriate to the scale type, with confidence intervals?** e.g.) nominal (Cohen's κ, Fleiss' κ), ordinal (weighted κ, Krippendorff's α), continuous (ICC, MAE), ranking (Kendall's τ), method for computing confidence intervals (e.g., bootstrap)		
	17A†. Is inter-rater reliability among the expert raters also reported as a benchmark for comparison?		
18†	**If expert evaluation is performed, is an error analysis reported for cases in which the Judge system and expert judgments disagreed?** e.g.) classification of error types, direction of error (overestimation/underestimation), sources of error (differences in prompt/rubric interpretation, model limitations or bias, input data quality, expert judgment error, run-to-run variability), severity of potential clinical harm from judgment errors		
19†	**If repeated evaluations of the same input were performed, is the consistency (intra-rater reliability) of the Judge system's judgments reported?** e.g.) number of repeated runs, time interval between runs (e.g., consecutive runs, day-level intervals), agreement across runs, proportion of cases with changed judgments, method for deriving the final judgment		
20†	**If measures were taken to assess or mitigate known biases of the Judge model, are the methods described?** e.g.) position swap, control of response length, use of a Judge model from a different model family than the evaluated model, blinding of response source (model name)		
21†	**If the Judge model generates rationales, are whether and how the factuality of those rationales (e.g., presence of hallucinations) was verified described?** e.g.) review of rationale faithfulness against input materials, check of consistency between rationale and score, expert review of a sample of rationales		

† Conditional item: Only applies when the stated condition is met; otherwise, answer "N/A". Items without † apply to all studies; if the information is not reported, answer "No" rather than "N/A".

Abbreviations: AI, artificial intelligence; API, application programming interface; CI, confidence interval; CoT, chain-of-thought; ICC, intraclass correlation coefficient; IRB, institutional review board; JSON, JavaScript Object Notation; LLM, large language model; MAE, mean absolute error; N/A, not applicable.

Figure 1. PRISMA-ScR flow diagram of study selection

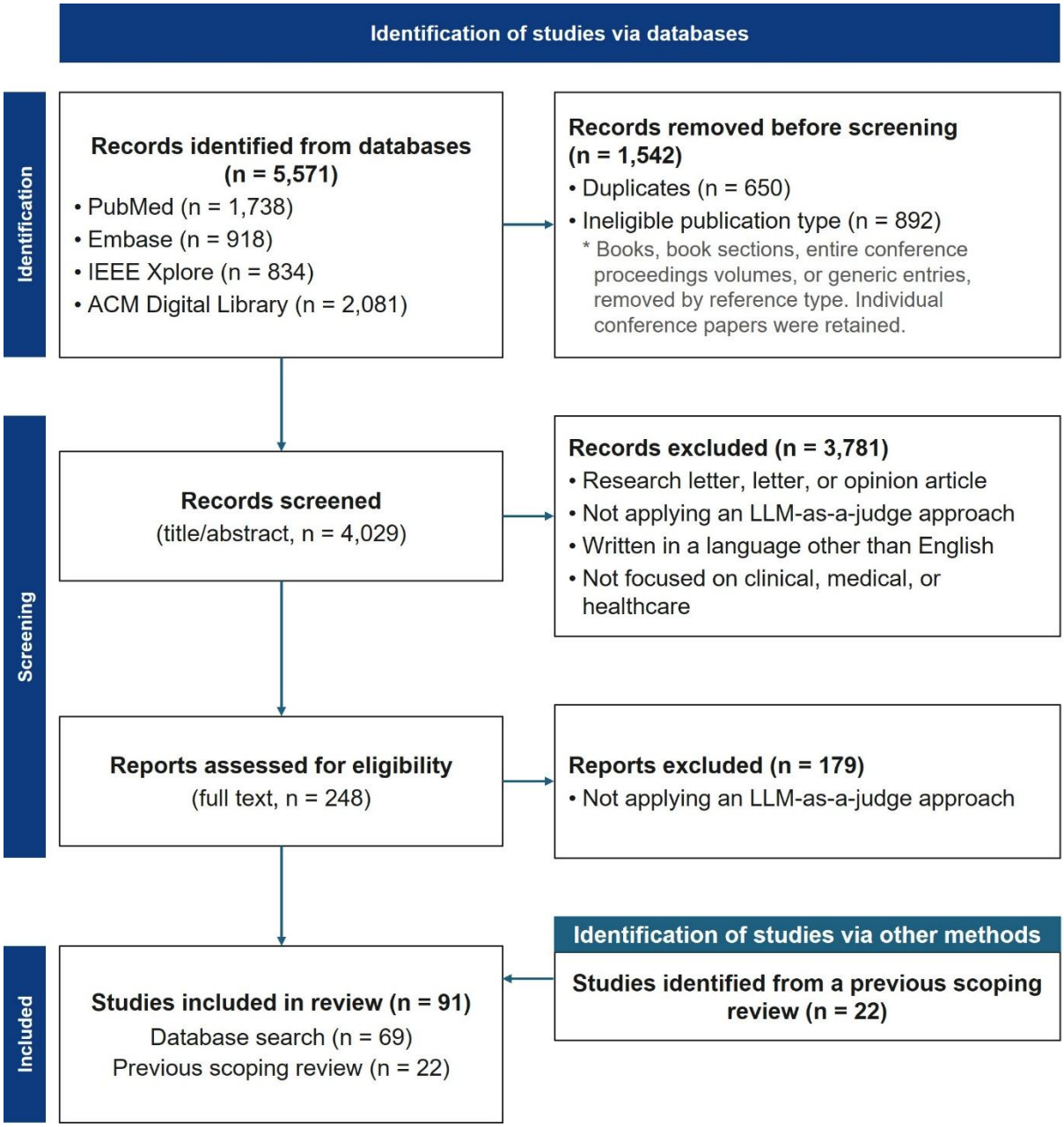
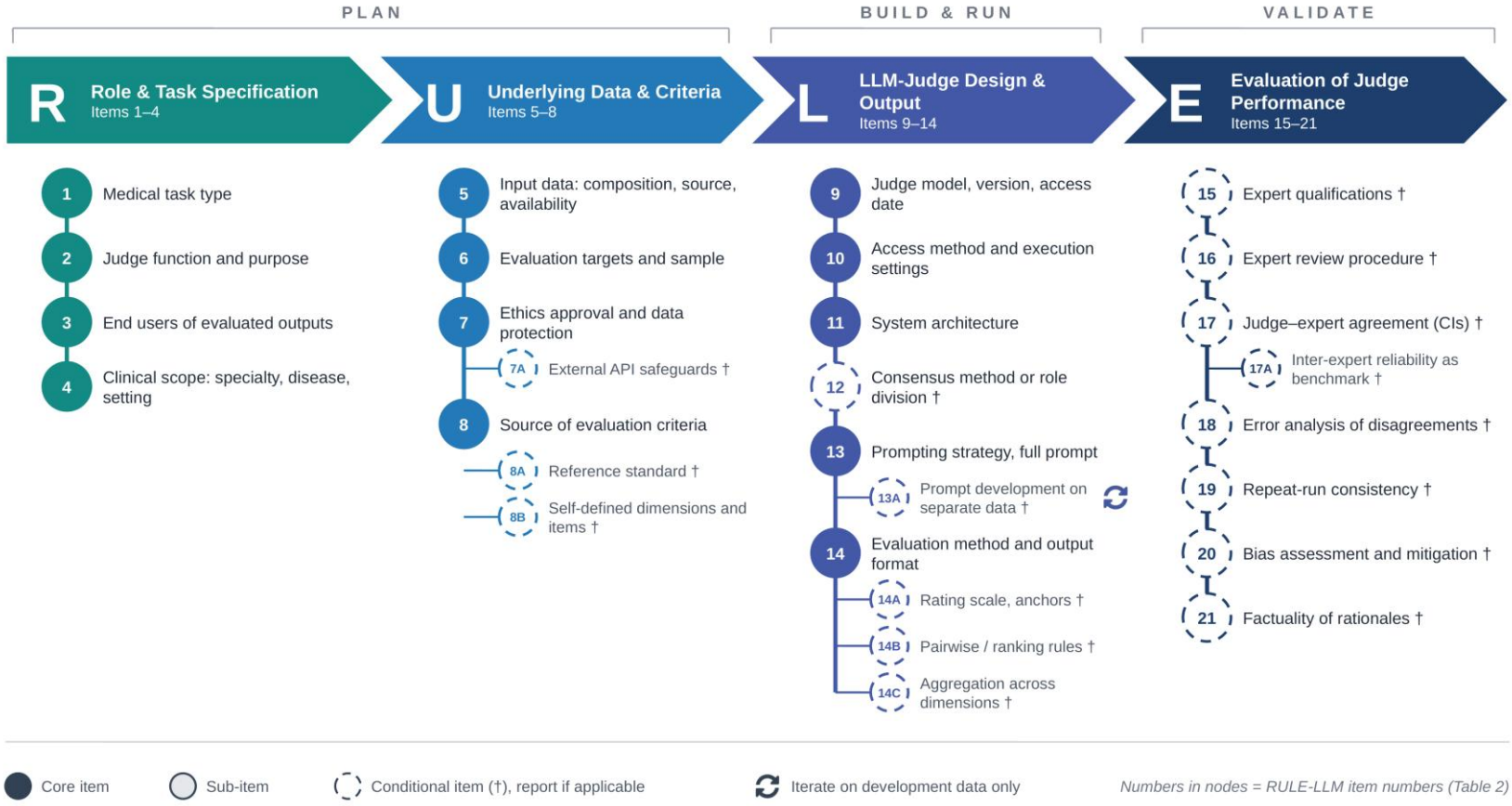


Figure 2. RULE-LLM reporting items along the course of an LLM-as-a-Judge study in medicine



The 21 RULE-LLM items are grouped into four domains and arranged in the order in which they arise when planning, building and running, and validating a Judge. Filled circles indicate core items, light circles indicate sub-items, and dashed circles indicate conditional items (†), which are reported when applicable. The circular arrow at item 13A denotes prompt or rubric refinement, which should be performed on development data kept separate from validation data. Numbers in the circles refer to the item numbers in Table 1 of the main text.

Figure 3. Item-by-item application of RULE-LLM to two LLM-as-a-Judge use cases

	RULE-LLM item	USE CASE 1 Judge for ED discharge summaries	USE CASE 2 Judge for a patient-facing diabetes chatbot
R	1 Medical task type	Clinical note generation: ED discharge summaries	Patient communication and education
	2 Judge function & purpose	Response quality assessment of LLM-generated summaries	Model comparison (chatbot v1 vs v2) and safety gatekeeping
	3 End users	ED physicians and nurses	Patients and caregivers
	4 Clinical scope	Emergency medicine; adult ED visits	Diabetes self-management; outpatient care
U	5 Input data	De-identified ED notes (source documents); not publicly shareable	Anonymized patient-portal questions; question set publicly released
	6 Evaluation targets & sample	3 summarization models (names and versions reported); 300 summaries, stratified random sampling by triage level; size justified for κ precision	2 chatbot versions; 400 random questions; size from power analysis for win-rate difference
	7 Ethics & data protection	IRB approved, consent waived; pseudonymized notes 7A† enterprise API with zero data retention	IRB exemption (no identifiable data) 7A† personal information removed before API transmission
	8 Criteria source	Combined 8A† source note as reference 8B† self-defined dimensions (accuracy, completeness, omission)	Reference-based 8A† national diabetes guideline 8B† *N/A*
L	9 Judge model	Proprietary LLM; exact name, version, and access date reported	Panel of 3 LLMs from different families (2 proprietary, 1 open-source); versions and access dates reported
	10 Access & execution settings	API; temperature 0, top_p 1; reasoning mode not used	API and local deployment; temperature 0; reasoning mode not used
	11 Architecture	Single-LLM	Multi-LLM panel with human–AI collaboration
	12† Consensus / roles	*N/A*	Majority vote for preference; any "unsafe" flag escalated to physician review
	13 Prompting strategy	Zero-shot, rubric-based, CoT; full prompt in supplement 13A† rubric revised on 50 development notes excluded from validation	Few-shot (3 examples), rubric-based; full prompt in supplement 13A† examples from a separate development set, no overlap
	14 Evaluation method & output	Pointwise, per-dimension scores, JSON 14A† 1–5 Likert with anchors 14B† *N/A* 14C† fail on any critical omission, else mean	Pairwise preference + binary safety label, JSON 14A† *N/A* 14B† ties allowed; win rate, Bradley–Terry 14C† unsafe response loses
E	15† Expert qualifications	3 emergency physicians (≥ 5 years of experience), none of whom are study authors	2 endocrinologists and 1 certified diabetes educator nurse, independent of the study team
	16† Expert review procedure	2-h calibration; blinded independent rating of all 300 summaries without access to Judge scores; third physician adjudicates	Blinded independent review of 200 random pairs without access to Judge verdicts; disagreements resolved by consensus meeting
	17† Judge–expert agreement	Weighted κ and ICC with bootstrap 95% CI 17A† inter-physician ICC as benchmark	Cohen's κ (safety) and percent agreement (preference) with bootstrap 95% CI 17A† inter-expert κ as benchmark
	18† Error analysis	Direction of error, error types, sources of error (e.g., rubric misinterpretation), and potential harm severity of misjudgments	Missed unsafe responses classified by type and potential harm severity
	19† Repeat-run consistency	3 consecutive runs per summary; proportion of changed scores; median used as final	5 runs on 100 pairs, 1 day apart; preference flip rate
	20† Bias assessment & mitigation	Verbosity bias: length-controlled analysis; model names blinded	Position bias: position swap; self-preference bias: Judges from families other than the chatbot's
	21† Rationale factuality	50 rationales reviewed for faithfulness to the source note	50 rationales checked against the guideline for hallucinations

Item-by-item application of RULE-LLM to two hypothetical use cases: a single-LLM Judge evaluating emergency department discharge summaries (use case 1) and a multi-LLM Judge panel with physician escalation comparing two versions of a patient-facing diabetes chatbot (use case 2). Entries are hypothetical and illustrate the level of detail expected for each item. Item numbers correspond to those in Table 1. † Conditional item, reported when applicable; N/A, not applicable.

Abbreviations: ED, emergency department; ICC, intraclass correlation coefficient; CI, confidence interval; CoT, chain-of-thought; JSON, JavaScript Object Notation; IRB, institutional review board; API, application programming interface.

Supplementary Files

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

- Supplementary.docx