Current Oncology

MDPI

Article

When the Judge Is an LLM: An Oncologist and a Language Model Evaluator Disagree on Responses to Inpatient Oncology Consultations-A Retrospective Blinded Study

Suleyman Sami Guzel *, Kubilay Tay, Nurhan Seda Jeral Evinc, Ebru Cicek, Yalcin Gul, Sukran Erdem Nurcan, Nebi Serkan Demirci and Ozkan Alan

Department of Medical Oncology, Cerrahpaşa Faculty of Medicine, Istanbul University-Cerrahpaşa, 34098 Istanbul, Türkiye; kubilay.tay@iuc.edu.tr (K.T.); nurhan.jeral@iuc.edu.tr (N.S.J.E.); ebru.cicek@iuc.edu.tr (E.C.); yalcin.gul@iuc.edu.tr (Y.G.); sukran.erdem@iuc.edu.tr (S.E.N.); nebiserkan.demirci@iuc.edu.tr (N.S.D.); ozkan.alan@iuc.edu.tr (O.A.)

* Correspondence: suleymanguzel34@hotmail.com

Simple Summary

Artificial intelligence chatbots are being studied as tools that could help oncologists answer consultation requests from other hospital departments. Because expert review is slow, many studies now use another chatbot to grade these answers. We tested whether this shortcut gives the same results as an oncologist. For 108 real consultation requests concerning hospitalised patients with cancer, we compared the original oncologist's note with answers written by two chatbots. An experienced oncologist and a chatbot grader scored all answers without being told their source. The oncologist found only small differences between the three sources, whereas the chatbot grader rated the oncologists' notes far lower than the chatbot answers, gave higher scores to longer answers, and flagged different answers, at different rates, as potentially unsafe. The choice of grader therefore changed the conclusions of the study. Oncologists conducting or reading such studies should check whether the grader, human or chatbot, has been compared against independent expert judgement.

Abstract

Large language models (LLMs) are increasingly tested as aids for oncology consultation and, increasingly, as graders of such outputs ("LLM-as-a-judge"). We assessed whether an LLM judge and an oncologist reach comparable conclusions on real oncology documentation. For 108 consecutive inpatient oncology consultations at a tertiary centre, the oncologist's original note and responses generated by Gemini and ChatGPT were scored by a medical oncologist and an LLM judge (Claude Opus 5), both blinded to source labels, using a nine-domain rubric (0–45). The oncologist separated the three sources by ≤1.6 points (Kendall's W = 0.082), whereas the LLM judge scored the two models 9.8 and 13.4 points above the oncologists' notes (W = 0.772). Agreement was poor overall (ICC(2,1) = 0.167) and within each source (0.003–0.235). Between-evaluator differences were 10.5 points for oncologist notes versus 1.1 and −1.2 points for model responses (distribution-free interaction $p < 0.001$). LLM-judge scores, but not oncologist scores, increased with response length, including among oncologists' notes ($\varrho = 0.56$). Safety flags were assigned at different rates and overlapped no more than expected by chance. In oncology consultation research, evaluator choice can determine study conclusions; no single evaluator,

Received: date
Revised: date
Accepted: date
Published: date

Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license.

human or LLM, should serve as sole assessor without calibration against independent expert judgement.

Keywords: large language models; LLM-as-a-judge; medical oncology; inpatient oncology consultation; clinical decision support; artificial intelligence; inter-rater reliability; evaluation methodology

1. Introduction

Large language models (LLMs) have advanced rapidly in clinical applications, achieving expert-level performance on medical licensing examinations and structured clinical vignettes, and in some studies producing answers to patient questions that were rated as more complete and more empathetic than physician replies [1,2]. These results have driven considerable interest in integrating such models into clinical workflows.

The evidence base, however, rests largely on material that differs from real clinical practice in two respects. First, most evaluations use examination items, synthetic vignettes, or questions curated from online forums—texts that are by construction complete, internally consistent, and written to have a determinate answer. A real inpatient consultation is answered under time pressure, from incomplete data, by a specialist with access to the entire medical record [3,4]. Second, and less often examined, is the nature of the comparator itself. Model performance is measured against the clinician's written output, but a consultation note is a document addressed to a colleague who already holds the chart, and is therefore abbreviated, selective, and dependent on shared context.

This second point connects to a growing methodological problem. Expert evaluation of model outputs is slow and expensive, and to relieve this bottleneck, LLMs are increasingly deployed as evaluators of other models—the "LLM-as-a-judge" paradigm [5]. In general-purpose text evaluation, such judges have been reported to agree reasonably well with human raters [5]. Yet systematic biases have also been documented: a preference for longer responses, favouritism toward outputs from related model families, and greater sensitivity to form than to substance [6,7]. Whether these biases operate when the material being judged is clinical content, and whether they are large enough to alter conclusions, remains largely untested.

Clinical text makes this concern structurally sharper. In general-purpose writing, form and substance are only loosely coupled. In clinical documentation, the coupling runs the other way: an experienced clinician's note is terse and abbreviated because it presupposes what the reader already knows, whereas a model's answer is long, structured, and self-contained because it can presuppose nothing. A judge who sees only the text, without the surrounding clinical context, may read this stylistic difference as a difference in quality. The risk exists for human judges as well; but when the judge is itself a language model, a tendency to reward text resembling its own output may introduce an additional source of bias.

Inpatient oncology consultation offers a demanding setting in which to examine this. Such consultations are high-stakes, multi-component, and frequently concern questions for which guidelines provide no determinate answer—whether to withhold or delay systemic therapy, how to adjust dosing in organ dysfunction, how to manage oncological emergencies, when to redefine goals of care. Most of these decisions hinge on performance status, comorbidity, and treatment intent, so that the correctness of an answer depends on contextual information that cannot be recovered from the text alone [8,9].

We therefore studied 108 consecutive inpatient oncology consultations at a single centre, with two prespecified aims. The first, the primary outcome, was to compare, using

a nine-domain rubric, three responses to each consultation question: the note written by the responding oncologist, and the responses generated by two large language models, Gemini (Gemini 3.6 Flash; Google LLC, Mountain View, CA, USA) and ChatGPT (GPT-5.5; OpenAI, San Francisco, CA, USA). The second, addressed through prespecified secondary outcomes, was to quantify agreement between two independent evaluators blinded to response source—a practising oncologist and an LLM judge. Because the evaluator analyses proved to determine how the comparison between response sources should be interpreted, they form the main emphasis of this report; the primary outcome is reported in full in Section 3.3. The contribution of the study therefore concerns evaluation methodology rather than the clinical safety or efficacy of the models.

2. Materials and Methods

2.1. Study Design and Setting

We conducted a retrospective, comparative evaluation study of responses to inpatient oncology consultations at Istanbul University-Cerrahpaşa, Cerrahpaşa Medical Faculty, a tertiary academic hospital. Consecutive consultation requests received between 1 June 2025 and 1 October 2025 during routine clinical care were included. Permission to use the anonymised data was granted by the hospital administration on 9 April 2026. After ethics approval, model responses were generated, and all evaluations were performed in October 2026; no step of response generation or evaluation preceded the approval.

The study was approved by the Istanbul University-Cerrahpaşa Medical Research Ethics Committee at its meeting on 6 October 2026 (application reference code Jgn7IjEK; approval No. E-24687260-604.01-1800288). The requirement for informed consent was waived because of the retrospective, anonymised study design.

2.2. Case Selection

We included consecutive inpatient oncology consultation requests received between 1 June and 1 October 2025. A total of 108 consultation requests, each corresponding to a distinct patient encounter, were included in the analysis. All consecutive consultations received during the study period were included; no formal sample size calculation was performed, and the precision of the agreement estimates is therefore reported with confidence intervals and considered in the interpretation.

All consultation records were de-identified before analysis. Patient identifiers were replaced with sequential study codes, and the linking key was held separately by the principal investigator and was not shared with the evaluators or included in any analysis dataset.

2.3. Case Characterisation

For each consultation, we recorded patient age, primary tumour diagnosis, the requesting clinical unit, and the stated reason for consultation. Two additional variables were derived for prespecified subgroup analyses. The consultation request was classified as decision-oriented (treatment planning, perioperative assessment, or level-of-care decisions) or acute-problem management (organ dysfunction, oncological emergencies, infection, cytopenia, respiratory, gastrointestinal, or neurological complications, and symptom management). The requesting unit was grouped as the emergency department versus all other units. Tumours were classified into five categories for analysis: gastrointestinal, lung, breast and gynaecological, urological, and other. Classification was performed using the consultation request only, before any response was reviewed.

2.4. Response Generation

Three responses were compared for each consultation. Source A was the consultation note written by the responding medical oncology team and entered into the patient record at the time of the consultation; in accordance with departmental practice, no inpatient consultation is closed without review and approval by an attending medical oncologist, so each note reflects specialist-level judgement. Source B was a response generated by Gemini (Gemini 3.6 Flash; Google LLC, Mountain View, CA, USA) and source C was a response generated by ChatGPT (GPT-5.5; OpenAI, San Francisco, CA, USA).

Both models are proprietary, closed-source models and were used without fine-tuning. They were accessed through their public web interfaces (gemini.google.com and chatgpt.com) in October 2026 using default settings; these interfaces do not expose sampling parameters such as temperature, and no API version identifier is available for web access. An identical prompt was used for every case and with both models (Box 1; original Turkish text in Supplementary File S1). The prompt contained the requesting unit, the consultation request and the case summary written by the requesting team, with patient identifiers removed. Each case was submitted as a single query, with no follow-up turns or regeneration of responses.

The information available to the three sources therefore differed. The responding oncologist had access to the complete electronic health record, including laboratory results, imaging and previous treatment, and could examine the patient, whereas the models received only the text of the consultation request and the case summary. In 22 consultations, the case summary referred to laboratory results that had been reviewed in the hospital information system without reproducing the values; these consultations were identified during evaluation and examined in a sensitivity analysis (Section 2.8).

Box 1. Standardised response-generation prompt (English translation; bracketed fields were completed for each case).

ROLE: You are a medical oncology specialist responding to an oncology consultation request from an inpatient ward.

REQUESTING SPECIALTY: [requesting unit]. CONSULTATION REQUEST NOTE: [verbatim request]. CASE SUMMARY: [verbatim case summary written by the requesting team, including history, current medications and orders, vital signs, physical examination, laboratory values and imaging reports where provided].

TASK: In light of the information above, prepare clinical recommendations in the style an oncology consultant would write in the patient's record. Give your answer in the following format: 1. Assessment: a brief clinical assessment based on the available findings (1–3 sentences). 2. Recommendations: bulleted, concrete and actionable recommendations (required investigations, treatment, follow-up, referral or admission decision, etc.). 3. Conclusion: a clear statement of whether there is an indication for admission or urgent intervention from an oncological standpoint. Base your assessment only on the information provided. If information is missing, state which additional information or investigations are needed. Your answer should be suitable for direct use as a consultation note in clinical practice.

2.5. Evaluation

Each of the 324 responses (108 consultations × 3 sources) was independently scored by two evaluators: evaluator 1, an LLM judge (Claude Opus 5; Anthropic PBC, San Francisco, CA, USA), accessed through its web interface (claude.ai) with default settings in October 2026; and evaluator 2, a board-certified medical oncologist with extensive clinical

experience who was independent of the index consultations and had not participated in the clinical care or authored any of the consultation notes included in the study.

Responses were presented to both evaluators in a standardised format without identifiers indicating their source. Within each consultation, the three responses were presented in a non-fixed order; they were relabelled A, B and C by source only after scoring, for analysis. Neither evaluator was told the source of any individual response; both were informed of the general aim of the study, namely that responses written by an oncologist and generated by LLMs were being compared. The LLM judge was given this description of the study, the full rubric with its anchor descriptions verbatim and, for each consultation, the consultation request, the case summary and the three responses; it returned the nine domain scores and a brief free-text rationale for each consultation. Generation parameters could not be set, and each consultation was scored once. During scoring, the judge noted the 22 consultations in which laboratory values had not been reproduced in the text provided to the models, and it was instructed to score each response on the information available to its author. The human evaluator received the same material and rubric in a spreadsheet. Citation markers generated by the Gemini interface were removed before evaluation. Blinding was therefore to source labels; because responses differed visibly in length and structure, the possibility that evaluators inferred the source was examined post hoc (Section 2.8). For clarity, response sources are identified by name in this manuscript.

Responses were scored using a nine-domain rubric, with each domain rated from 0 to 5, yielding a total score ranging from 0 to 45: clinical accuracy; currency and evidence base; patient-specific appropriateness; safety and potential for harm; medicolegal safety; clarity of recommendation; responsiveness to the question asked; language and comprehensibility; and practical applicability. The full rubric with anchor descriptions is provided in Supplementary File S2. Equal weights were assigned to the nine domains because no empirical basis existed for differential weighting; the effect of this choice was examined in sensitivity analyses (Section 2.8).

Evaluators did not rank responses directly. Rankings reported in this study were derived from total rubric scores, with ties assigned mean rank.

2.6. Outcomes

The primary outcome was the difference in total rubric scores among the three response sources. Prespecified secondary outcomes included individual scores for each of the nine rubric domains; derived response rankings and the proportion of responses ranked first; agreement between the two evaluators; the interaction between evaluator and response source; the proportion of responses scored ≤2 on the safety and clinical accuracy domains; and prespecified exploratory subgroup analyses by requesting unit, question type, tumour group, and patient age group. Analyses added in response to peer review are described in Section 2.8 and are identified as post hoc.

2.7. Statistical Analysis

Primary analyses were prespecified and performed separately for each evaluator, because the study was designed to compare the scoring behaviour of a human expert and an LLM judge as two independent evaluation systems applied to the same responses. Pooling or averaging the two sets of scores would have produced a single consensus score and concealed precisely the between-evaluator differences that the study was designed to examine; agreement between the evaluators and the evaluator × source interaction were therefore analysed using the combined dataset as prespecified secondary outcomes.

Differences between the three sources in total score were tested using the Friedman test with case as the blocking factor, with Kendall's W as the effect size, followed by

pairwise Wilcoxon signed-rank tests with Holm correction. Location shifts were reported as Hodges–Lehmann estimates with distribution-free 95% confidence intervals derived from Walsh averages, computed using all paired differences, including zeros. Rank-biserial correlation was reported as the paired effect size. In parallel, linear mixed models of the form total score ~ source + (1 | case) were fitted for each evaluator. Because total scores were bounded, showed ceiling effects, and were non-normally distributed (Shapiro–Wilk $p < 0.001$), nonparametric results were treated as primary and parametric models as supportive analyses [10,11].

Evaluator–source interaction was tested by comparing a mixed model containing the source × evaluator with the corresponding additive model by likelihood ratio test, with case as a random intercept. Between-evaluator differences within each source are reported as estimated marginal contrasts with 95% confidence intervals.

Inter-rater agreement was assessed at the level of the individual response ($n = 324$) using a two-way random-effects, absolute-agreement, single-measures intraclass correlation coefficient, ICC(2,1) [12,13], interpreted according to established thresholds [12], together with Pearson and Spearman correlations and Bland–Altman analysis including assessment of proportional bias [14]. Domain-level agreement was quantified using quadratic weighted kappa [15] with Benjamini–Hochberg correction [16] across the nine domains. Categorical agreement on the top-ranked source was summarised as raw concordance and Cohen's kappa [17], and full-ranking concordance as the proportion of identical rankings and mean Kendall's tau-b. Because one evaluator produced a high proportion of tied scores, all ranking-based analyses were repeated after excluding consultations with tied rankings.

Proportions of responses scored ≤2 and ≤3 on the safety and clinical accuracy domains were compared across sources using Cochran's Q tests with post hoc McNemar tests and Holm correction; proportions were reported with Wilson 95% confidence intervals [18].

Subgroup analyses were conducted by fitting source × subgroup interaction terms rather than by testing subgroups separately, with Benjamini–Hochberg correction [16] applied across the four interaction tests. These analyses were prespecified but exploratory, and strata containing fewer than 20 cases were not interpreted.

Psychometric behaviour of the rubric was examined through ceiling rates, distributional diagnostics, Cronbach's alpha [19] with item-deleted values, and inter-domain Spearman correlation matrices computed separately for each evaluator.

Two-sided p-values < 0.05 were considered significant. All analyses were performed in Python 3.12 (Python Software Foundation, Wilmington, DE, USA) using pandas 3.0, NumPy 2.4, SciPy 1.17, and statsmodels 0.14; post hoc analyses (Section 2.8) were performed in Python 3.13 using pandas 3.0.5, NumPy 2.5.3, SciPy 1.18.1, statsmodels 0.15.0 and scikit-learn 1.9.1. All reported statistics were independently recomputed from the raw scoring data using an independently written script, and any discrepancies were resolved by inspection of the underlying computation.

Reporting was conducted in accordance with the CHART (Chatbot Assessment Reporting Tool) statement [20]; the completed checklist is provided in Supplementary File S3. Items from the TRIPOD-LLM statement [21] relevant to evaluation studies were also consulted. No study protocol was preregistered.

2.8. Post Hoc Analyses Performed in Response to Peer Review

Because the three sources are structurally different documents and each consultation contributed three responses, inter-rater agreement was additionally estimated within each source, after centring each evaluator's scores on its own source means, which removes between-source differences in level and retains only the ordering of responses within

source. ICC(2,1) and the consistency form ICC(3,1) are reported with 95% confidence intervals from a case-level cluster bootstrap (2000 resamples of consultations, keeping the three responses of each consultation together). Agreement was also examined after excluding responses at the human evaluator's ceiling. Confidence intervals for Kendall's W were obtained by the same bootstrap.

As a distribution-free test of the evaluator × source interaction, the within-consultation between-evaluator difference (human minus LLM) was computed for each source and compared across the three sources with the Friedman test, followed by Wilcoxon signed-rank tests.

Agreement on the top-ranked source was interpreted against the concordance expected by chance from the two evaluators' marginal distributions and against the maximum kappa attainable given those marginals [22], and Gwet's AC1 was reported as a prevalence-robust alternative [23]. For low safety scores (≤2 and ≤3), we report the number of jointly flagged responses expected if the evaluators had flagged independently at their observed rates, the conditional probability that a response flagged by one evaluator was also flagged by the other, positive specific agreement, Cohen's kappa, the prevalence- and bias-adjusted kappa (PABAK) [24] and Fisher's exact test, overall and within source.

Subgroup interactions, which in the original analysis had been estimated in a model pooling both evaluators, were refitted separately for each evaluator, with Benjamini–Hochberg correction across the four tests within each evaluator and Kruskal–Wallis comparison of within-consultation C − A differences between strata as a distribution-free check.

Response length was measured as the number of words, excluding interface-generated citation markers, and structure as the number of list items and the presence of the three-part Assessment/Recommendations/Conclusion format requested in the prompt. Associations between length and total score were examined with Spearman correlation within each source and evaluator, and with linear models of total score on source and log-transformed word count with cluster-robust standard errors by consultation. A length-restricted comparison was performed in consultations whose oncologist note was at least as long as the shortest model response (≥45 words). The LLM judge's free-text rationales were reviewed for statements indicating inference of response origin.

Main analyses were repeated (i) excluding the 22 consultations in which laboratory values had not been provided to the models; (ii) using a clinical-content subscore (clinical accuracy, currency and evidence base, patient-specific appropriateness, and safety; range 0–20); and (iii) excluding the three presentation-related domains (medicolegal safety, clarity of recommendation, and language and comprehensibility; range 0–30), which also tests the sensitivity of the findings to equal domain weighting. The proportion of inter-domain variance explained by the first principal component was computed for each evaluator. Three consultations were selected to illustrate concordant scoring, a safety concern flagged only by the LLM judge, and a safety concern flagged only by the human evaluator; excerpts were translated from Turkish by the authors.

3. Results

3.1. Case Characteristics

A total of 108 inpatient oncology consultation requests, each representing a distinct patient encounter, were analysed. The median patient age was 66 years (range 23–91); 57 patients (52.8%) were aged ≥65 years. Most consultations originated from the emergency department (n = 62, 57.4%), followed by surgical specialities (n = 23, 21.3%), medical specialities (n = 17, 15.7%), and the intensive care unit (n = 6, 5.6%). Gastrointestinal (n = 29) and lung (n = 26) malignancies were the most common primary tumour types, whereas

no haematological malignancies were included. Fifty-seven consultations (52.8%) concerned acute problem management, whereas 51 (47.2%) were decision-oriented. Baseline case characteristics are summarised in Table 1.

Table 1. Case characteristics (*n* = 108).

Characteristic	*n* (%)
Age, median (range)	66 (23–91)
≥65 years	57 (52.8)
<65 years	51 (47.2)
Requesting unit	
Emergency department	62 (57.4)
Surgical specialties	23 (21.3)
Medical specialties	17 (15.7)
Intensive care unit	6 (5.6)
Tumour group	
Gastrointestinal	29 (26.9)
Lung	26 (24.1)
Breast and gynaecological	21 (19.4)
Other	21 (19.4)
Urological	11 (10.2)
Consultation question	
Acute problem management	57 (52.8)
Decision-oriented	51 (47.2)

Each consultation derived from a distinct patient encounter. No haematological malignancies were represented.

Each consultation generated three responses (source A, the oncologist's original note; source B, Gemini; source C, ChatGPT), each scored independently by both evaluators, yielding 648 individual evaluations. No scores were missing, no values fell outside the 0–5 domain range, and total scores matched the sum of the nine domains in every record.

3.2. Total Rubric Scores

Total scores differed between evaluators, and the magnitude and direction of these differences varied across sources (Table 2, Figure 1).

The human evaluator scored the three sources within a narrow range (A 39.4 ± 6.5; B 39.8 ± 5.4; C 41.0 ± 5.5), awarding the maximum score of 45 to responses from each source. The LLM evaluator scored source A lower (28.9 ± 7.0) than sources B (38.7 ± 3.8) and C (42.2 ± 1.8), and its scores for source C were confined to a range of 37–44, with an interquartile range of a single point.

Table 2. Total rubric scores by evaluator and response source (scale 0–45).

Evaluator	Source	Mean ± SD	Median (IQR)	Range	Maximum Score *n* (%)
Human	A—Oncologist	39.36 ± 6.46	41.0 (36–45)	13–45	37 (34.3)
Human	B—Gemini	39.78 ± 5.38	41.5 (36–45)	26–45	32 (29.6)
Human	C—ChatGPT	41.00 ± 5.54	45.0 (38–45)	25–45	56 (51.9)
LLM	A—Oncologist	28.85 ± 6.97	29.5 (24–34)	16–43	0 (0)
LLM	B—Gemini	38.65 ± 3.76	40.5 (37–41)	19–43	0 (0)
LLM	C—ChatGPT	42.21 ± 1.84	43.0 (42–43)	37–44	0 (0)

n = 108 per cell. IQR, interquartile range.

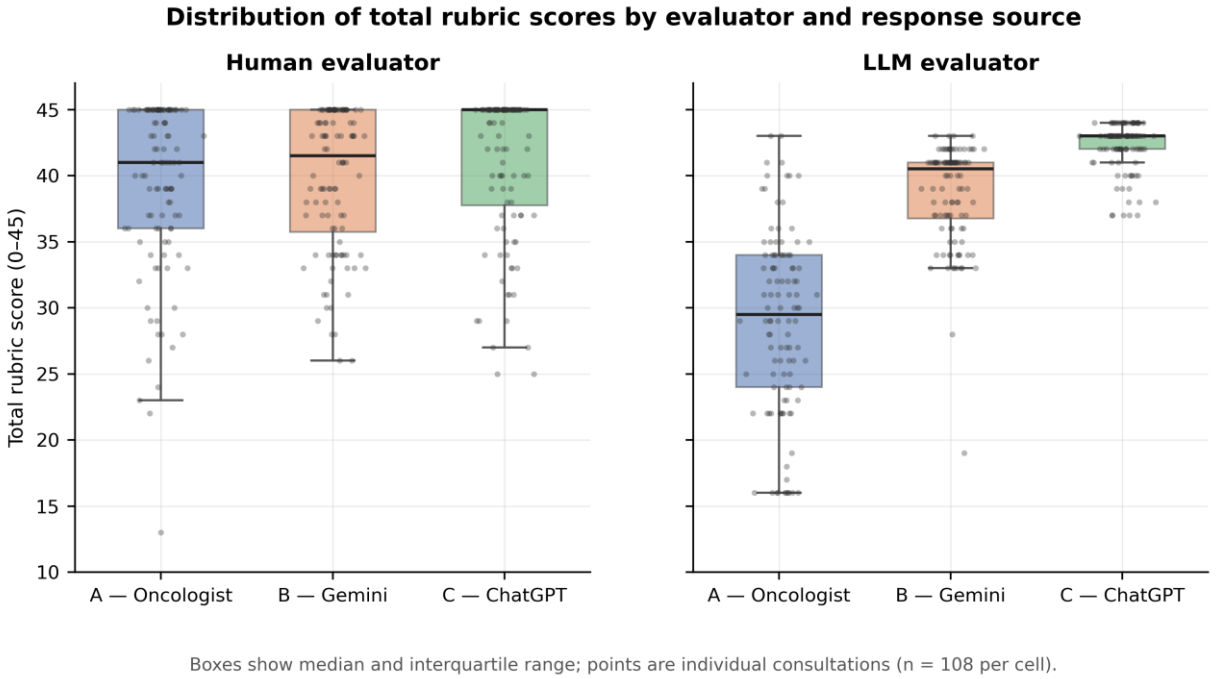

Figure 1. Distribution of total rubric scores by evaluator and response source. Boxes show median and interquartile range; points are individual consultations (n = 108 per cell).

3.3. Primary Outcome: Differences Between Response Sources

Analyses were performed separately for each evaluator.

Human evaluator. Total scores differed significantly across response sources (Friedman $\chi^2(2) = 17.72$, $p < 0.001$), although the overall effect size was small (Kendall's W = 0.082; bootstrap 95% CI 0.029–0.180). In post hoc comparisons with Holm correction, source C scored higher than source A (Hodges–Lehmann difference 1.5 points, 95% CI 0.0–2.5; $p = 0.010$; rank-biserial r = 0.376) and higher than source B (0.5 points, 95% CI 0.0–1.0; $p = 0.009$; r = 0.416). Sources A and B did not differ significantly (0.5 points, 95% CI 0.0–1.5; $p = 0.302$). The mixed model estimated a C-A difference of 1.64 points (95% CI 0.57–2.71; $p = 0.003$) on a 45-point scale, and a B-A difference of 0.42 points (95% CI −0.66 to 1.49; $p = 0.446$).

LLM evaluator. Total scores differed substantially across response sources (Friedman $\chi^2(2) = 166.77$, $p < 0.001$), with a large overall effect size (Kendall's W = 0.772; 95% CI 0.672–0.863). All three pairwise comparisons were significant after Holm correction: source B exceeded source A by 10.0 points (95% CI 8.0–11.5; $p < 0.001$; r = 0.870), source C exceeded source A by 13.0 points (95% CI 11.5–15.0; $p < 0.001$; r = 0.993), and source C exceeded source B by 3.5 points (95% CI 2.5–4.0; $p < 0.001$; r = 0.869). Corresponding mixed model estimates for B-A and C-A were 9.80 points (95% CI 8.54–11.05) and 13.36 points (95% CI 12.11–14.61), respectively, both $p < 0.001$.

Pairwise differences with confidence intervals are shown in Supplementary Figure S1. Full post hoc results are given in Table 3.

The two mixed models differed in their variance structure. For the human evaluator, the case-level intraclass correlation was 0.522. For the LLM evaluator, the random intercept variance was estimated at the boundary of the parameter space, and the case-level intraclass correlation was approximately zero.

Model residuals departed from normality for both evaluators (Shapiro–Wilk $p < 0.001$; skewness −0.81 and −0.54; excess kurtosis ≈ 2.07; Supplementary Figure S2); therefore, the nonparametric analyses were considered primary.

Table 3. Post hoc pairwise comparisons of total score, by evaluator.

Evaluator	Comparison	HL Difference	95% CI	*p* (Holm)	r
Human	B − A	0.5	0.0–1.5	0.302	0.134
Human	C − A	1.5	0.0–2.5	0.010	0.376
Human	C − B	0.5	0.0–1.0	0.009	0.416
LLM	B − A	10.0	8.0–11.5	<0.001	0.870
LLM	C − A	13.0	11.5–15.0	<0.001	0.993
LLM	C − B	3.5	2.5–4.0	<0.001	0.869

Friedman test: human evaluator $\chi^2(2) = 17.72$, $p < 0.001$, Kendall's W = 0.082; LLM evaluator $\chi^2(2) = 166.77$, $p < 0.001$, W = 0.772. HL, Hodges–Lehmann estimate computed on all paired differences including zeros; r, rank-biserial correlation. Positive values favour the first-named source. Bootstrap 95% CIs for W: human 0.029–0.180; LLM 0.672–0.863.

3.4. Evaluator × Source Interaction

The interaction between evaluator and response source was significant (likelihood ratio $\chi^2(2) = 148.6$, $p < 0.001$), indicating that between-evaluator differences varied according to response source (Table 4).

A significant between-evaluator difference was observed only for source A: the human evaluator assigned scores that were, on average, 10.51 points higher than those assigned by the LLM evaluator (95% CI, 9.20–11.82; z = 15.67; $p < 0.001$). In contrast, the between-evaluator difference was 1.13 points for source B (95% CI, −0.18 to 2.44; $p = 0.092$) and −1.21 points for source C (95% CI, −2.53 to 0.10; $p = 0.071$), neither of which reached statistical significance.

The interaction was confirmed by a distribution-free test. The within-consultation between-evaluator difference had a median of 12 points for source A, 2 points for source B and 1 point for source C (Friedman $\chi^2(2) = 116.6$, $p < 0.001$; Kendall's W = 0.54). The difference was larger for source A than for source B or C (both Wilcoxon $p < 0.001$), and was also larger for source B than for source C ($p < 0.001$): the evaluators differed not only in how they scored the oncologists' notes but also in how far they separated the two model sources (Hodges–Lehmann C − B difference 0.5 versus 3.5 points).

Table 4. Between-evaluator difference in total score, by response source.

Source	Difference (Human − LLM)	SE	95% CI	z	*p*
A—Oncologist	10.51	0.671	9.20–11.82	15.67	<0.001
B—Gemini	1.13	0.671	−0.18–2.44	1.68	0.092
C—ChatGPT	−1.21	0.671	−2.53–0.10	−1.81	0.071

Linear mixed model with case as random intercept. Interaction likelihood ratio $\chi^2(2) = 148.6$, $p < 0.001$.

3.5. Agreement Between Evaluators

Agreement was assessed across all 324 responses (Table 5) and, post hoc, within each source (Table 5d).

The intraclass correlation coefficient for total score was 0.167 (95% CI 0.06–0.27; $F(323,323) = 1.467$, $p < 0.001$; case-level bootstrap 95% CI 0.08–0.24), corresponding to poor reliability by conventional thresholds [12], with the upper confidence bound also within the poor range. Pearson correlation was 0.194 (95% CI 0.09–0.30) and Spearman correlation was 0.170 (95% CI 0.07–0.28).

Low agreement was not an artefact of pooling structurally different sources (Table 5d). Within source, ICC(2,1) was 0.097 for oncologist notes, 0.235 for Gemini responses and 0.003 for ChatGPT responses, with Spearman correlations of 0.267, 0.164 and −0.088,

respectively. After each evaluator's scores were centred on its own source means, removing all between-source differences in level, ICC(3,1) was 0.178 (95% CI 0.073–0.278). Source accounted for 59.4% of the variance in LLM-judge scores but for 1.4% of the variance in human scores. Restricting the analysis to responses below the human evaluator's ceiling (n = 199) did not increase agreement (ICC(2,1) = 0.089).

Bland–Altman analysis (Supplementary Figure S3) showed a mean difference of 3.48 points, with limits of agreement of −13.08 to 20.03. The between-evaluator difference decreased as the mean score increased, indicating proportional bias (slope −0.379, $p < 0.001$).

Domain-level agreement, quantified using quadratic weighted kappa, ranged from 0.056 to 0.249 across the nine domains; five domains remained statistically significant after Benjamini–Hochberg correction (Table 5).

Categorical agreement on the top-ranked source must be interpreted in the light of strongly imbalanced marginal distributions—source C was ranked first by the human evaluator in 72.2% and by the LLM evaluator in 90.7% of consultations—under which kappa is bounded well below 1 [22]. Across all consultations, the two evaluators selected the same top-ranked source in 33 of 108 (30.6%; Cohen's κ = 0.016) and produced identical full rankings in 20 (18.5%); mean Kendall's tau-b was 0.343.

Tied total scores were substantially more frequent for the human evaluator (56 of 108 consultations, 51.9%) than for the LLM evaluator (8 of 108, 7.4%). When the 57 consultations containing ties for either evaluator were excluded (n = 51 analysed), raw concordance on the top-ranked source was 50.98%, almost exactly the 50.9% expected by chance from the marginal distributions (κ = 0.002; maximum attainable κ given the marginals 0.32; Gwet's AC1 0.37), and identical full rankings were produced in 39.22% (mean tau-b 0.333). The higher raw concordance in the tie-free set therefore reflects the concentration of both evaluators' first places on source C rather than agreement beyond chance.

Table 5. (a). Agreement between evaluators—continuous measures (n = 324 responses). (b). Agreement between evaluators—domain-level quadratic weighted kappa. (c). Agreement between evaluators—categorical measures. (d). Agreement between evaluators within response source (post hoc).

(a)

Measure	Value	95% CI
ICC(2,1), total score	0.167	0.06–0.27
Pearson r	0.194	0.09–0.30
Spearman ϱ	0.170	0.07–0.28
Bland–Altman bias (points)	3.48	limits of agreement −13.08 to 20.03
Proportional bias slope	−0.379	$p < 0.001$

(b)

Domain	κw	SE	p (FDR)
Responsiveness to the question	0.249	0.058	<0.001
Patient-specific appropriateness	0.214	0.046	<0.001
Safety and potential for harm	0.163	0.051	0.008
Language and comprehensibility	0.120	0.061	0.023
Clarity of recommendation	0.096	0.055	0.118
Currency and evidence base	0.078	0.038	0.050
Medicolegal safety	0.071	0.045	0.133
Clinical accuracy	0.064	0.038	0.133
Practical applicability	0.056	0.041	0.141

(c)

Measure	All consultations (n = 108)	Tie-free (n = 51)
Same top-ranked source	30.6%	50.98%

Cohen's κ, top-ranked source	0.016	0.002
Identical full ranking	18.5%	39.22%
Mean Kendall's tau-b	0.343	0.333

(d)

Responses	ICC(2,1) (95% CI)	ICC(3,1) (95% CI)	Spearman ϱ (95% CI)
All (*n* = 324)	0.167 (0.081–0.241)	0.189 (0.096–0.274)	0.170 (0.078–0.262)
A—Oncologist (*n* = 108)	0.097 (0.022–0.178)	0.215 (0.049–0.383)	0.267 (0.100–0.427)
B—Gemini (*n* = 108)	0.235 (0.050–0.394)	0.241 (0.047–0.403)	0.164 (−0.020 to 0.341)
C—ChatGPT (*n* = 108)	0.003 (−0.125 to 0.137)	0.003 (−0.128 to 0.133)	−0.088 (−0.280 to 0.115)
Source-centred (*n* = 324)	0.179 (0.077–0.275)	0.178 (0.073–0.278)	0.174 (0.078–0.276)
Below human ceiling (*n* = 199)	0.089	0.090	0.080

ICC(2,1): two-way random effects, absolute agreement, single measures. $F(323,323) = 1.467$, $p < 0.001$; FDR, Benjamini–Hochberg correction across the nine domains; Tied total scores occurred in 51.9% of consultations for the human evaluator and 7.4% for the LLM evaluator; 95% CIs from a case-level cluster bootstrap (2000 resamples). Source-centred: each evaluator's scores centred on its own source means, removing between-source differences in level. Below human ceiling: responses to which the human evaluator did not award the maximum total score of 45.

3.6. Domain-Level Comparisons

Mean domain scores by source and evaluator are shown in Figure 2 and Table 6.

For the human evaluator, six of nine domains showed significant differences between sources after correction (clinical accuracy, patient-specific appropriateness, safety, clarity of recommendation, responsiveness to the question, and language and comprehensibility); currency and evidence base, medicolegal safety, and practical applicability did not. Effect sizes were small across domains, with Kendall's W ranging from 0.011 to 0.091.

For the LLM evaluator, eight of nine domains showed significant differences; source A scored lowest in every domain, and source C scored highest in all except clarity of recommendation, in which sources B and C scored similarly (mean 4.96 and 4.90). Effect sizes were largest for language and comprehensibility (W = 0.926), clarity of recommendation (W = 0.871), and medicolegal safety (W = 0.869). Practical applicability was the only domain in which no difference was detected (W = 0.012; corrected $p = 0.270$).

Table 6. Domain-level comparison between response sources, by evaluator.

Domain	Human W	Human *p* (FDR)	LLM W	LLM *p* (FDR)
Clinical accuracy	0.041	0.023	0.481	<0.001
Currency and evidence base	0.011	0.303	0.604	<0.001
Patient-specific appropriateness	0.091	<0.001	0.439	<0.001
Safety and potential for harm	0.033	0.041	0.487	<0.001
Medicolegal safety	0.017	0.196	0.869	<0.001
Clarity of recommendation	0.039	0.026	0.871	<0.001
Responsiveness to the question	0.079	<0.001	0.686	<0.001
Language and comprehensibility	0.052	0.009	0.926	<0.001
Practical applicability	0.021	0.132	0.012	0.270

W, Kendall's W from the Friedman test across the three sources. FDR, Benjamini–Hochberg correction across the nine domains within each evaluator.

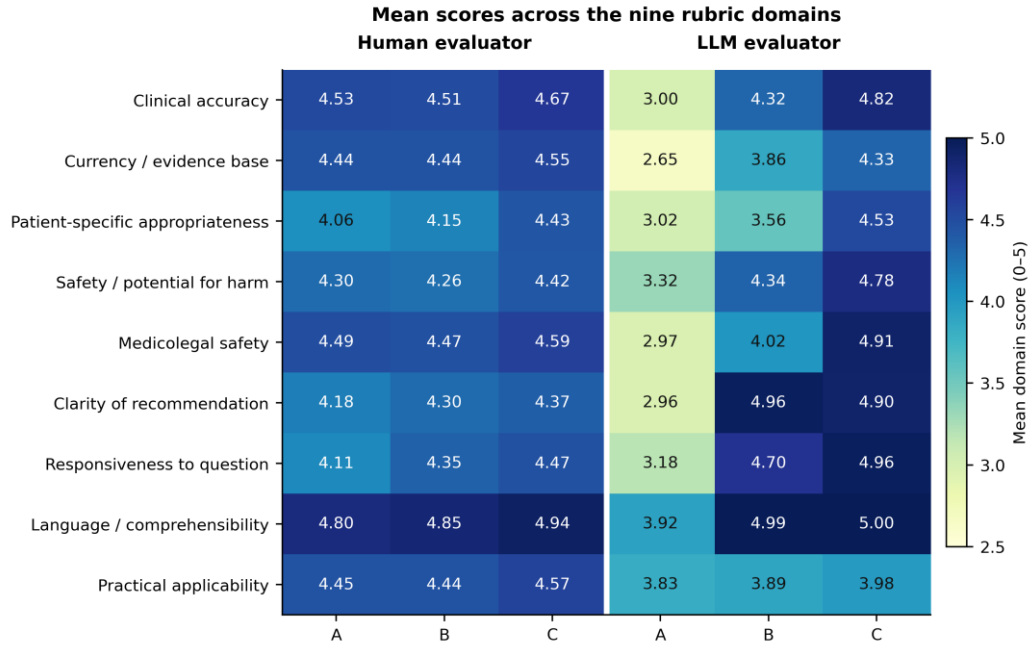

Figure 2. Mean scores across the nine rubric domains, by response source and evaluator.

3.7. Safety and Clinical Accuracy

Proportions of responses receiving low scores in the safety and clinical accuracy domains are given in Table 7 and Figure 3.

The human evaluator found no significant differences between sources for either domain or at either threshold. Fewer than 4% of responses from any source were scored ≤2 for safety and about one in five fell at or below 3; for clinical accuracy, only one response—one source A note—was scored ≤2 (all $p \geq 0.368$).

The LLM evaluator found significant differences between sources at both thresholds for both domains, consistently with source A receiving the lowest scores and source C highest (all $p < 0.001$). Among source A responses, 20.4% were ≤2 for safety and 35.2% ≤2 for clinical accuracy, compared with fewer than 2% of source B and none of the source C responses. At the more inclusive ≤3 threshold, the separation widened: roughly half of source A responses fell below the threshold for safety and roughly two-thirds for clinical accuracy, while no source C response fell below it on either domain. Proportions with confidence intervals and test statistics for all comparisons are given in Table 7.

The two sets of low safety scores largely concerned different responses, but this mainly reflects the evaluators' different rates of flagging. Thirty-four responses were scored ≤2 for safety by at least one evaluator: 24 by the LLM evaluator and 11 by the human evaluator, one of them by both; three-quarters of the 34 (n = 25) were source A responses. Had the two evaluators flagged responses independently at their observed rates, 0.8 jointly flagged responses would have been expected; the observed overlap therefore did not differ from chance (Fisher p = 0.58), and neither evaluator's flag predicted the other's (probability that a response flagged by the human evaluator was also flagged by the LLM evaluator, 9%; in the reverse direction, 4%; positive specific agreement 0.06; κ = 0.011; PABAK 0.80). At the ≤3 threshold, overlap exceeded chance for oncologist notes: 15 of the 20 notes scored ≤3 by the human evaluator were also scored ≤3 by the LLM evaluator (10.4 expected; κ = 0.168; Fisher p = 0.026). For ChatGPT responses, the human evaluator scored 20 responses ≤3 for safety and the LLM evaluator none. Because no reference standard was available, these data describe how often and where each evaluator flagged responses, not which flags were correct.

Table 7. Proportion of responses receiving low domain scores, with Wilson 95% confidence intervals.

Domain/threshold	Evaluator	A—Oncologist	B—Gemini	C—ChatGPT	Q	*p*
Safety ≤ 2	Human	3.7 (1.4–9.1)	3.7 (1.4–9.1)	2.8 (0.9–7.9)	0.22	0.895
Safety ≤ 2	LLM	20.4 (13.9–28.9)	1.9 (0.5–6.5)	0.0 (0.0–3.4)	37.00	<0.001
Safety ≤ 3	Human	18.5 (12.3–26.9)	21.3 (14.6–29.9)	18.5 (12.3–26.9)	0.51	0.773
Safety ≤ 3	LLM	51.9 (42.5–61.0)	15.7 (10.1–23.8)	0.0 (0.0–3.4)	73.82	<0.001
Clinical accuracy ≤ 2	Human	0.9 (0.2–5.1)	0.0 (0.0–3.4)	0.0 (0.0–3.4)	2.00	0.368
Clinical accuracy ≤ 2	LLM	35.2 (26.8–44.6)	1.9 (0.5–6.5)	0.0 (0.0–3.4)	68.60	<0.001
Clinical accuracy ≤ 3	Human	4.6 (2.0–10.4)	7.4 (3.8–13.9)	4.6 (2.0–10.4)	1.29	0.526
Clinical accuracy ≤ 3	LLM	64.8 (55.4–73.2)	19.4 (13.1–27.9)	0.0 (0.0–3.4)	94.42	<0.001

Values are % (95% CI). Q, Cochran's Q with 2 degrees of freedom.

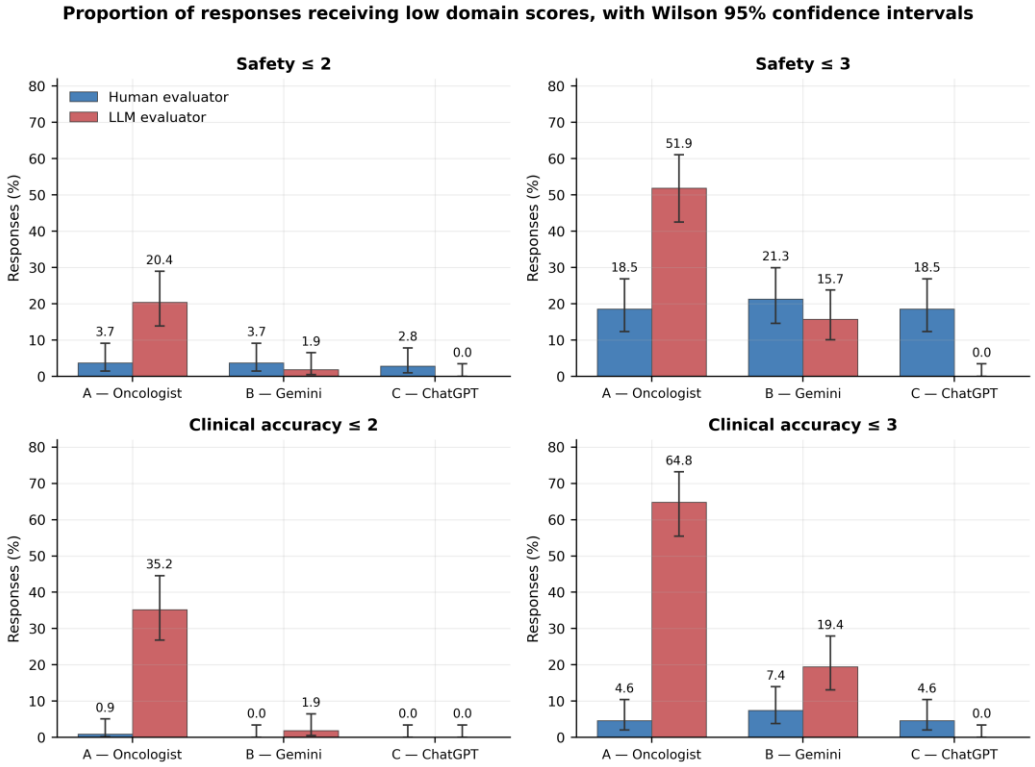

Figure 3. Proportion of responses receiving low safety and clinical accuracy scores, with Wilson 95% confidence intervals.

3.8. Derived Rankings

Mean derived ranks and first-place proportions are given in Table 8 and shown in Supplementary Figures S4 and S5. Because tied scores were permitted, more than one source could be ranked first within the same consultation.

For the human evaluator, mean ranks were 2.19 (A), 2.08 (B), and 1.73 (C); source C was ranked first in 72.2% of consultations, compared with 49.1% for A and 44.4% for B (Cochran's Q = 18.68, $p < 0.001$; C versus A $p = 0.004$; C versus B $p < 0.001$; A versus B $p = 0.568$).

For the LLM evaluator, mean ranks were 2.86 (A), 2.02 (B), and 1.12 (C); source C was ranked first in 98 of 108 consultations (90.7%), compared with 3.7% for A and 11.1% for B (Q = 150.89, $p < 0.001$; C versus A and C versus B both $p < 0.001$; A versus B $p = 0.077$).

Excluding, for each evaluator separately, consultations containing tied scores (human evaluator, $n = 52$ analysed; LLM evaluator, $n = 100$ analysed) did not change the order of sources within either evaluator: for the human evaluator, mean ranks were 2.33 (A), 2.04 (B) and 1.63 (C), with source C ranked first in 55.8% of consultations; for the LLM evaluator, mean ranks were 2.89, 2.01 and 1.10, with source C ranked first in 90.0%. Between-evaluator concordance in the tie-free set is reported in Section 3.5.

Table 8. Derived rankings by evaluator.

Evaluator	Source	Mean Rank	Ranked First, *n* (%)
Human	A—Oncologist	2.19	53 (49.1)
Human	B—Gemini	2.08	48 (44.4)
Human	C—ChatGPT	1.73	78 (72.2)
LLM	A—Oncologist	2.86	4 (3.7)
LLM	B—Gemini	2.02	12 (11.1)
LLM	C—ChatGPT	1.12	98 (90.7)

Rankings derived from total rubric scores; 1 = best. Cochran's Q: human evaluator 18.68, $p < 0.001$; LLM evaluator 150.89, $p < 0.001$. Percentages sum to more than 100% within an evaluator because tied scores permit more than one source to occupy first place.

3.9. Subgroup Analyses

In the original analysis, subgroup interactions were estimated in a model pooling both evaluators, so that the reported C − A differences (approximately 7–8 points) were averages of the human and LLM estimates. Refitted separately for each evaluator (Table 9; Supplementary Figure S6), no source × subgroup interaction remained significant after Benjamini–Hochberg correction within either evaluator (all corrected $p \geq 0.095$).

Two interactions were nominally significant before correction. For the human evaluator, the C − A difference was larger for emergency-department consultations than for other units (2.74 versus 0.15 points; likelihood ratio $\chi^2(2) = 6.53$, $p = 0.038$; Kruskal–Wallis $p = 0.040$); for the LLM evaluator, it was larger in patients younger than 65 years than in older patients (14.67 versus 12.19 points; $\chi^2(2) = 7.49$, $p = 0.024$; Kruskal–Wallis $p = 0.093$). The LLM evaluator's C − A difference ranged from 12.1 to 14.8 points across all strata, and the human evaluator's from −0.1 to 2.7 points outside the urological tumour stratum, which ($n = 11$) was not interpreted because of its small sample size.

Table 9. Prespecified exploratory subgroup interactions, estimated separately for each evaluator.

Subgroup (Strata; n)	Evaluator	C − A Difference by Stratum (Points)	LR χ^2 (df)	*p*	*p* (FDR)
Requesting unit (ED/other; 62/46)	Human	2.74/0.15	6.53 (2)	0.038	0.152
	LLM	12.73/14.22	1.53 (2)	0.465	0.620
Question type (acute/decision; 57/51)	Human	2.14/1.08	1.84 (2)	0.398	0.530
	LLM	12.11/14.76	4.97 (2)	0.083	0.167
Tumour group (GI/lung/breast–gyn/urological/other; 29/26/21/11/21)	Human	1.14/1.77/−0.14/5.45/1.95	12.36 (8)	0.136	0.272
	LLM	12.97/14.00/13.05/13.91/13.14	5.14 (8)	0.743	0.743
Age (<65/≥65 years; 51/57)	Human	2.24/1.11	1.12 (2)	0.570	0.570
	LLM	14.67/12.19	7.49 (2)	0.024	0.095

Likelihood ratio (LR) tests comparing linear mixed models with and without the source × subgroup interaction, fitted separately for each evaluator with case as random intercept. FDR, Benjamini–Hochberg correction across the four tests within each evaluator. ED, emergency department; GI, gastrointestinal. The urological tumour stratum ($n = 11$) was not interpreted.

3.10. Psychometric Behaviour of the Rubric

Score distributions near the upper end of the scale differed between evaluators (Table 10). The human evaluator awarded the maximum total score of 45 to 34.3% of source A responses, 29.6% of source B responses, and 51.9% of source C responses. The LLM evaluator never awarded the maximum total score to any response, but confined source C scores to a narrow range of 37–44.

At the domain level, the human evaluator awarded the maximum domain score most often for language and comprehensibility (83.3%, 87.0%, and 93.5% for sources A, B, and C, respectively). The LLM evaluator awarded the maximum domain score for language and comprehensibility to 99.1% of source B and 100% of source C responses, and for clarity of recommendation to 97.2% and 89.8%, respectively. It awarded the maximum score for medicolegal safety to 90.7% of source C responses, whereas no source A response received the maximum score in this domain. For practical applicability, the maximum domain score was awarded to 24.1% of source A, 0.9% of source B, and none of source C responses.

Internal consistency was high and similar for the two evaluators (Cronbach's α = 0.947 and 0.949, respectively). Item-deleted analyses showed different patterns: for the human evaluator, alpha increased only when language and comprehensibility were excluded (0.947 → 0.955); for the LLM evaluator, alpha increased when practical applicability was excluded (→0.955) or when language and comprehensibility (→0.952) were removed. The first principal component explained 70.8% of inter-domain variance for the human evaluator and 71.9% for the LLM evaluator, indicating that for both evaluators the rubric behaved largely as a single dimension.

Inter-domain Spearman correlations differed substantially between evaluators (Figure 4). For the human evaluator, no domain pair exceeded ϱ = 0.85; the language and comprehensibility domain correlated relatively weakly with all other domains (ϱ = 0.29–0.44), while the remaining eight domains intercorrelated between 0.63 and 0.85. For the LLM evaluator, three domain pairs exceeded ϱ = 0.85, all involving clinical accuracy: with currency and evidence base (ϱ = 0.855), with safety (ϱ = 0.874), and with responsiveness to the question (ϱ = 0.869). In this evaluator, language and comprehensibility correlated strongly with clarity of recommendation (ϱ = 0.84), responsiveness (ϱ = 0.75), and medicolegal safety (ϱ = 0.73), while practical applicability showed weaker correlations with all other domains (ϱ = 0.03–0.54).

Table 10. Psychometric behaviour of the rubric, by evaluator.

Measure	Human Evaluator	LLM Evaluator
Cronbach's α	0.947	0.949
Domain whose removal raises α	Language and comprehensibility (→0.955)	Practical applicability (→0.955); language and comprehensibility (→0.952)
Domain pairs with $\varrho > 0.85$	none	3 (all involving clinical accuracy)
Maximum total score awarded (A/B/C)	34.3%/29.6%/51.9%	0% for all sources
Shapiro–Wilk, total score	$p < 0.001$ (all sources)	A $p = 0.010$; B and C $p < 0.001$

α computed across the nine domains over 324 responses per evaluator.

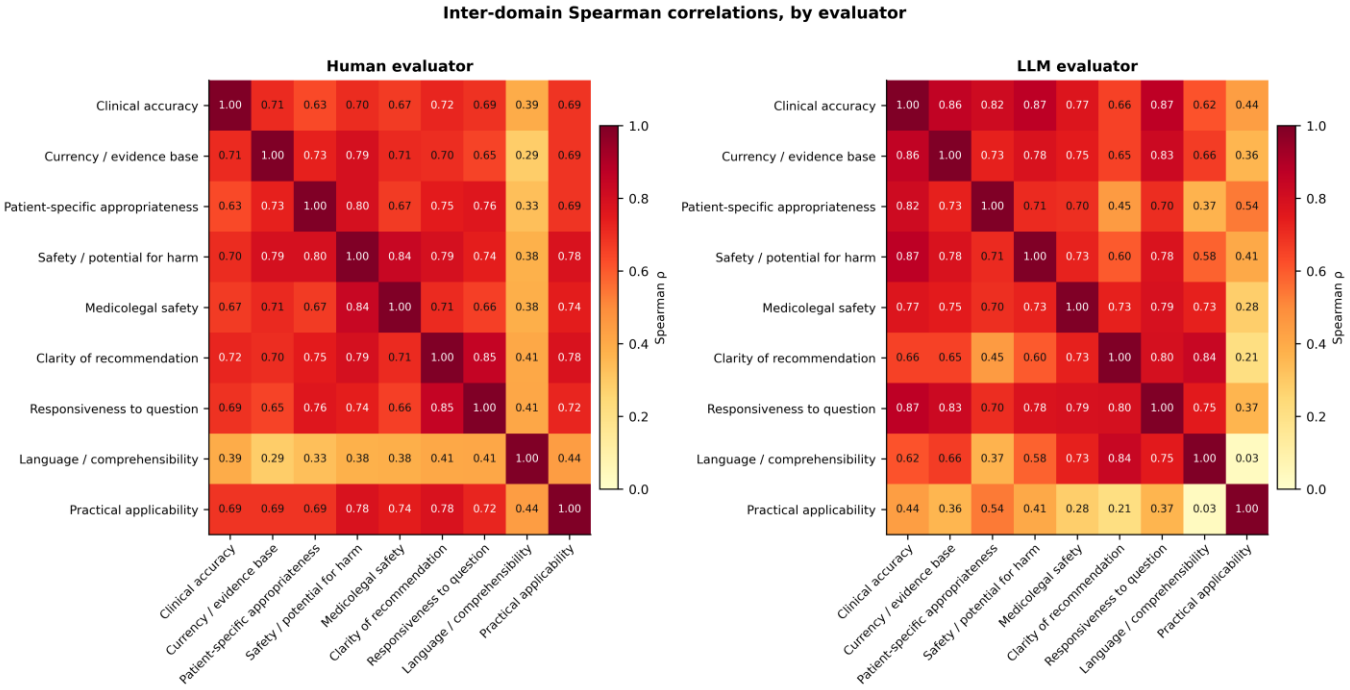

Figure 4. Inter-domain Spearman correlation matrices, by evaluator. For the human evaluator, no domain pair exceeds ϱ = 0.85; for the LLM evaluator, three pairs do, all involving clinical accuracy.

3.11. Response Length, Structure and Source Inference (Post Hoc)

Oncologist notes were shorter and more variable in length (median 39.5 words, range 6–171) than Gemini (median 65) and ChatGPT responses (median 95.5) (Table 11). No oncologist note followed the three-part Assessment/Recommendations/Conclusion structure requested of the models, whereas every model response did; structure was therefore completely confounded with source and could not be analysed separately.

Human-evaluator scores were unrelated to length within every source (Spearman ϱ 0.004, 0.075 and −0.084 for sources A, B and C). LLM-judge scores increased with length within source A (ϱ = 0.560, $p < 0.001$) and source B (ϱ = 0.301, $p = 0.002$), and weakly within source C (ϱ = 0.166, $p = 0.085$) (Figure 5). Among oncologist notes, the LLM judge's mean total score rose from 23.8 in the shortest tertile (6–31 words) to 29.7 and 33.2 in the middle and longest tertiles, whereas the human evaluator's scores were 39.4, 38.8 and 40.0. Adjusting for log word count reduced the LLM judge's estimated advantage of model responses over oncologist notes from 9.8 to 7.0 points (B − A) and from 13.4 to 8.2 points (C − A), corresponding to about 4.1 points per doubling of length; for the human evaluator, adjustment left the estimates unchanged. In the 44 consultations in which the oncologist note was at least as long as the shortest model response, the LLM judge still scored the notes lower (mean 32.8) than Gemini (37.8) and ChatGPT (41.8) responses, whereas the human evaluator did not (39.8, 38.6 and 40.2). Length therefore accounted for part, but not all, of the LLM judge's separation of sources.

Although source labels had been withheld, the LLM judge stated in its free-text report that response A was probably the physician's note and that responses B and C were probably generated by the two models, indicating that the stylistic differences between sources were sufficient for source inference.

Table 11. Response length and structure, and their association with total score (post hoc).

Measure	A—Oncologist	B—Gemini	C—ChatGPT
Words, median (IQR)	39.5 (28–64)	65 (60–72)	95.5 (88–102)

Words, range	6–171	45–97	71–183
List items, mean (SD)	5.2 (3.6)	5.6 (0.5)	8.9 (0.9)
Three-part structure, *n* (%)	0 (0)	108 (100)	108 (100)
Spearman ρ, length vs. total score—human	0.004	0.075	−0.084
Spearman ρ, length vs. total score—LLM	0.560 ***	0.301 **	0.166
Difference vs. A, human: unadjusted/length-adjusted	—	0.42/0.43	1.64/1.67
Difference vs. A, LLM: unadjusted/length-adjusted	—	9.80/6.97	13.36/8.22

Three-part structure: Assessment/Recommendations/Conclusion format requested in the prompt. Length-adjusted differences from linear models of total score on source and log word count with cluster-robust standard errors by consultation. ** $p < 0.01$; *** $p < 0.001$.

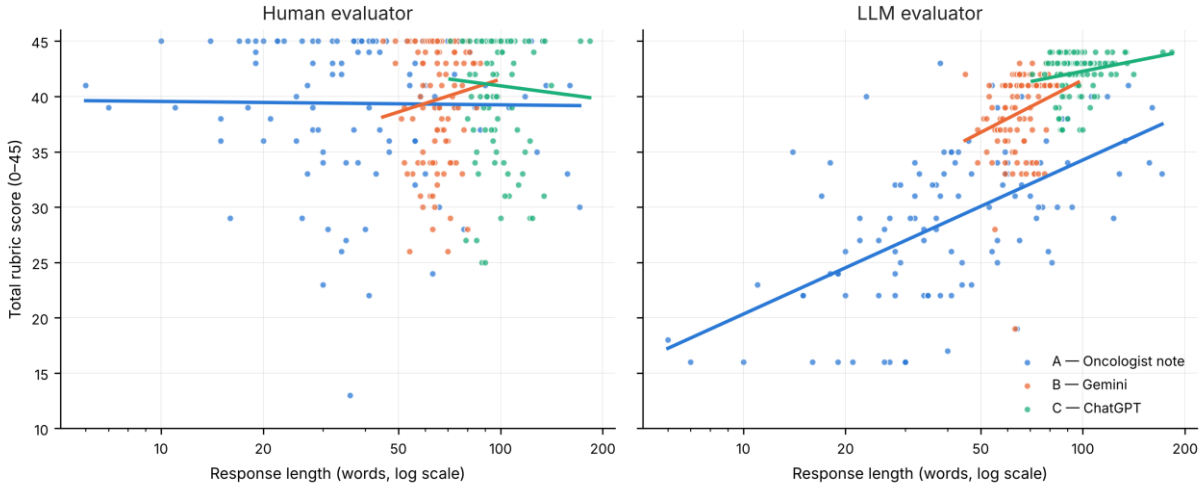

Figure 5. Total rubric score versus response length (log scale), by evaluator and response source. Lines are least-squares fits on log word count within each source; points are individual responses (*n* = 108 per source).

3.12. Sensitivity Analyses (Post Hoc)

Excluding the 22 consultations in which laboratory values had not been provided to the models did not alter the results (Table 12). The LLM judge's separation of sources was largest for the presentation-related domains (W = 0.907) but persisted in the clinical-content subscore (W = 0.566 versus 0.066 for the human evaluator), and the between-evaluator difference for oncologist notes remained in every variant. The discordance was therefore not confined to presentation-related domains, and it was not an artefact of equal domain weighting.

Table 12. Sensitivity analyses (post hoc).

Analysis	Human W	LLM W	ICC(2,1)	Human − LLM Difference, A/B/C
Main analysis (total score 0–45; *n* = 108)	0.082	0.772	0.167	10.51/1.13/−1.21
Excluding 22 consultations without laboratory values for the models (*n* = 86)	0.073	0.822	0.173	10.22/1.16/−1.36
Clinical-content subscore (0–20)	0.066	0.566	0.153	5.34/1.28/−0.41
Excluding presentation-related domains (0–30)	0.078	0.550	0.175	6.90/1.48/−0.31
Presentation-related domains only (0–15)	0.032	0.907	0.111	3.61/−0.35/−0.91

W, Kendall's W across the three sources (Friedman test). Clinical-content subscore: clinical accuracy, currency and evidence base, patient-specific appropriateness, safety. Presentation-related domains: medicolegal safety, clarity of recommendation, language and comprehensibility.

3.13. Illustrative Cases

Concordant scoring (consultation 3, source A). A man in his seventies with hepatocellular carcinoma on a background of cirrhosis presented with massive ascites, oliguria, hyperkalaemia, grade 2 hydronephrosis and ten days of constipation. The oncologist's note recommended close monitoring of fluid balance with a urinary catheter, discontinuation of spironolactone and renal dose adjustment, urinalysis, a fluid balance target of +500 mL, repeat biochemistry for hyperkalaemia, urological assessment of the hydronephrosis with re-consultation for diuretics once post-renal kidney injury was excluded, diagnostic and therapeutic paracentesis with attention to hypotension, enemas and lactulose with monitoring for hepatic encephalopathy, and infectious diseases consultation, and stated that admission was indicated. Both evaluators scored the note 41/45 (safety 5/5 each). The LLM judge noted that the note addressed post-renal kidney injury and the risk of hepatic encephalopathy, and that its language was heavily abbreviated. A concrete, comprehensive note was recognised as such by both evaluators despite its terse form.

Safety concern flagged only by the LLM judge (consultation 95, source A). A patient with high-grade neuroendocrine (small-cell) carcinoma was neutropenic (neutrophils 70/μL) with raised acute-phase reactants. The note recommended monitoring, infectious diseases and pulmonology opinions, caution regarding febrile neutropenia and a single dose of filgrastim, stated that no bed was available on the oncology ward and asked for admission through other units, and ended: "Owing to a system error, 'No indication for admission' has been ticked." The human evaluator scored the note 40/45 (safety 4/5); the LLM judge scored it 19/45 (safety 1/5), describing the final sentence as a discharge-oriented conclusion inconsistent with the clinical picture and a serious safety concern. In local practice, the attending oncologist determines whether admission is indicated; when no oncology bed is available, the 'No indication for admission' option has to be selected in the hospital information system although the indication exists. Read in context, the note recommends admission and active management of neutropenia, and the judge's safety score reflects a misreading of an institutional documentation convention rather than unsafe advice.

Safety concern flagged only by the human evaluator (consultation 88, source C). A woman in her sixties receiving concurrent cisplatin-based chemoradiotherapy for squamous cell carcinoma of the tongue presented with presyncope, loss of consciousness, fever and diarrhoea; leukocytes were 0.29×10^9/L, neutrophils 0.07×10^9/L, platelets 11×10^9/L and C-reactive protein 108 mg/L. The ChatGPT response identified high-risk febrile neutropenia and recommended urgent admission, blood and urine cultures, empirical antipseudomonal antibiotics, granulocyte colony-stimulating factor, platelet transfusion, stool investigations, haemodynamic monitoring and deferral of chemotherapy. The LLM judge scored it 44/45 (safety 5/5); the human evaluator scored it 31/45 (safety 2/5). The human evaluator did not record a rationale. On review by the authors, the response did not address the episode of loss of consciousness, for which intracranial haemorrhage needed to be considered in a patient with a platelet count of 11×10^9/L, and it deferred chemotherapy and recommended granulocyte colony-stimulating factor without addressing the ongoing radiotherapy. The oncologist's note, by contrast, requested a neurological assessment. The LLM judge, which noted that the response named febrile neutropenia explicitly and proposed concrete treatment, did not identify these omissions.

4. Discussion

Two evaluators, blinded to response source, applied the same nine-domain rubric to the same 324 responses and reached markedly different conclusions about the relative performance of the three response sources. The human evaluator separated the three sources by no more than 1.6 points on a 45-point scale (Kendall's W = 0.082), whereas the LLM evaluator produced a consistent hierarchy of source C > source B > source A, scoring the two models 13.4 and 9.8 points above the oncologist's notes (Kendall's W = 0.772). The divergence was concentrated in, but not confined to, the oncologists' notes: the between-evaluator difference was 10.5 points for these notes, and the evaluators also differed in how far they separated the two model sources. Agreement was poor overall (ICC(2,1) = 0.167) and remained poor within each source and after removal of between-source differences in level, so it cannot be attributed to the pooling of structurally different documents. The confidence interval of the overall ICC (0.06–0.27) indicates modest precision, but even its upper bound lies in the poor range. These findings are descriptive: they show that the conclusions of this comparison depend on the evaluator, not that either evaluator was correct.

Evaluator-dependent scoring is increasingly documented outside medicine. In JUDGE-BENCH, which compared 11 models against human annotations across 20 datasets, judge–human agreement varied with the property being assessed, with the expertise of the human annotators, and with whether the text under evaluation was human- or model-generated, leading the authors to conclude that LLMs are not yet ready to replace human judges systematically [25]. In expert knowledge domains, LLM judges agreed with subject-matter experts on only 64–68% of pairwise comparisons and tended to equate clarity with detail and comprehensiveness, whereas the experts rewarded concise, well-ordered explanations [26]. Judgement biases favouring longer and more elaborately formatted answers have been characterised directly [6,27], and evaluators have been shown to recognise and favour text from their own or related model families [7]. Each of these mechanisms is compatible with our data: the two model sources were long, self-contained and stylistically close to the register an LLM judge itself produces, whereas the oncologist's notes were terse and presupposed the medical record. Alignment is not uniformly poor in clinical settings, however; using a validated instrument and a reasoning model, an LLM judge reproduced expert scores of electronic health record summaries with an intraclass correlation of 0.818 [28]. Evaluator behaviour therefore may depend substantially on the task, the rubric, and the model rather than representing a fixed property of LLM judges. Because our design did not include an independent gold standard, we cannot determine which evaluator was closer to clinical reality; our findings demonstrate systematic evaluator-dependent differences, rather than establishing that either evaluator was biased or more accurate.

How low is this agreement relative to agreement between clinicians? Without a second human evaluator, we cannot answer this directly, and the literature advises caution. In a blinded study in which three clinicians rated LLM- and clinician-authored diabetes counselling responses, single-rater ICC(2,1) between clinicians was 0.27 for overall quality and 0.09 for clinical accuracy [29]; in HealthBench, physician–physician agreement was of the same order as model–physician agreement [30]; and clinician–clinician ranking agreement of τ = 0.38–0.43 has been reported for case-specific clinical rubrics [31], compared with τ = 0.34 between our evaluators. Disagreement among physicians evaluating model outputs is largely case-specific and is amplified by missing context [32]. Our overall ICC is therefore below, but not far below, published human–human values, and we do not claim that the LLM judge disagreed with the oncologist more than a second oncologist would have. What distinguishes the present data is the structure of the disagreement rather than its level: a large, systematic, source-dependent offset affecting one type of

document, and a dependence of LLM-judge scores on response length that was absent from the human evaluator's scores.

Part of the disagreement reflects evaluator calibration and scale use rather than judgement of individual responses. The human evaluator awarded the maximum total score to 30–52% of responses from each source and produced tied totals in half of the consultations, whereas the LLM judge never awarded the maximum score and confined ChatGPT responses to a range of 37–44 points. Leniency, severity and range restriction are classical rater effects in performance assessment [33], and they depress absolute-agreement indices independently of how raters order responses. Consistency-type and source-centred analyses, which remove differences in level, raised agreement only to 0.18–0.19, however, and agreement did not improve when responses at the human ceiling were excluded. Calibration therefore explains the size of the between-evaluator offset but not the weak ordering agreement within sources. LLM judges have been shown to be harsh on some clinical tasks and lenient on others, with calibration varying by task and by model [34]; a judge's calibration on one clinical task cannot be assumed to hold on another.

The most plausible mechanism for the LLM judge's pattern is a preference for length and form. LLM judges favour longer and more elaborately formatted answers [6,27] and can recognise model-generated text [7]. Our post hoc analyses are consistent with this: among oncologist notes alone, where source is held constant, LLM-judge scores rose by almost ten points from the shortest to the longest tertile while human scores did not change, and adjustment for length removed about 40% of the LLM judge's advantage of ChatGPT responses over the oncologists' notes. Length did not explain the whole difference, and structure could not be examined separately, because every model response and no oncologist note followed the three-part format requested in the prompt. The judge's free-text report also shows that withholding source labels did not prevent source inference. Blinding in studies of this kind should therefore be regarded as blinding to labels rather than to source, and should be tested, for example, by asking evaluators to state the presumed origin of each response. Position bias [35] is unlikely to have contributed, because responses were presented in non-fixed order; self-preference in the narrow sense could not arise, because the judge did not score outputs of its own model family, although a broader preference for model-like text is compatible with our data. Human raters can also prefer chatbot-written to clinician-written responses [2,29]; in the present study, the human evaluator did not.

The comparison was also asymmetric in information. The oncologist answered with access to the full record and the patient, whereas the models and both evaluators saw only the text of the request and the case summary. A consultation note written for a colleague who holds the chart omits what the reader already knows, and an evaluator who sees only the text may read omission as error. Consultation 95 (Section 3.13) illustrates this: the LLM judge read a sentence explaining a field in the hospital information system as an unsafe discharge decision. Similarly, in consultation 62, the judge classified the oncologist's request for a gynaecological opinion for genital bleeding as a fabrication, because the finding was absent from the case summary, whereas it had been observed by the examining physician at the bedside. Excluding the 22 consultations in which laboratory values were not supplied to the models did not change the results, but subtler asymmetries cannot be excluded. Missing context is the strongest identifiable driver of physician disagreement in evaluations of model output [32], and identical standardised clinical summaries supplied to all sources and evaluators would be needed to remove this asymmetry in future work.

The safety analysis requires particular restraint. The two evaluators flagged responses at very different rates (24 versus 11 responses scored ≤2), and their flags overlapped no more than would be expected had they flagged independently. These data

show that the evaluators applied different thresholds and attended to different features; they do not show that either evaluator missed unsafe responses identified by the other, and without a reference standard neither set of flags can be treated as correct. Resolving this question requires an adjudication design: a panel of at least two independent oncologists, blinded to source and to the original flags and with access to the full medical record, reviewing every flagged response together with a random sample of unflagged responses against a prespecified harm taxonomy, with consensus or a third adjudicator for disagreements. Such a design would allow the sensitivity and positive predictive value of each evaluator to be estimated, and should precede any use of an LLM judge for safety screening.

The rubric also behaved differently in the two evaluators' hands. Internal consistency was high and almost identical (Cronbach's α = 0.947 and 0.949), and for both evaluators the first principal component explained about 71% of inter-domain variance, so neither evaluator produced nine fully separable judgements. For the LLM judge, however, clinical accuracy correlated above 0.85 with currency and evidence base, safety and responsiveness, and language and comprehensibility correlated strongly with clarity of recommendation and medicolegal safety, whereas for the human evaluator language and comprehensibility remained largely independent of the clinical domains (ρ = 0.29–0.44). We interpret this as evidence that an instrument developed for human raters did not transfer intact to the LLM judge, whose domain scores appear to share a general impression in which presentation and clinical content are not kept apart [36,37]. High internal consistency is not evidence of validity, and rubrics intended for LLM judges should be validated in that use.

The contribution of this study concerns evaluation methodology; it provides no direct evidence on the safety or efficacy of the models, or of the oncologists' notes. LLM judges are fast, inexpensive, and scalable, making large-scale repeated evaluation more feasible than reliance on expert review alone [28,38]. Our findings do not argue against their use, but against assuming that any single evaluator, human or LLM, yields conclusions that are independent of the evaluator. Where an LLM judge is used, its agreement with independent experts should be measured on the same material and reported alongside the primary result, in line with current reporting frameworks for chatbot and LLM studies [20,21]; where a single human expert is used, the same caution applies. This is directly relevant to oncology, where LLM applications are under active investigation across clinical care, education and research [39], and where model-generated recommendations, for example for patients reporting treatment-related symptoms, are being assessed against expert ratings [40]. In our study, reliance on either evaluator alone would have produced substantially different estimates of the relative performance of the three response sources.

This study has important limitations. First, we used a single human evaluator, so no human–human reliability estimate is available, and the extent to which this evaluator's scoring pattern is representative of other oncologists cannot be determined; single-rater agreement between clinicians in comparable designs is itself modest [29,30]. Second, we used a single LLM judge in a single configuration—Claude Opus 5, accessed through its web interface with default settings, informed of the study aim and run once—and our findings are specific to that configuration rather than to LLM judges as a class; different judges disagree with physicians and with one another in task-dependent ways [34]. Third, no independent gold standard or blinded adjudication was available to arbitrate the disagreements we report, particularly in the safety domain; the present study therefore demonstrates evaluator discordance but cannot establish which evaluator provided the more clinically valid assessment. Fourth, the generating models and the judge were accessed through web interfaces whose underlying versions and parameters are not

disclosed and may change without notice, and each case was generated and evaluated once, so run-to-run variability, including whether the LLM judge's lower scores for oncologist notes persist across independent runs, was not quantified; the results are a snapshot of these systems in 2026. Fifth, blinding was limited to source labels: length and structure differed systematically between sources, structure was fully confounded with source, and the LLM judge explicitly inferred the source of the responses. Sixth, the sources differed in the information available to them. Seventh, the cohort was a single-centre sample of solid-tumour consultations written in Turkish, without a formal sample size calculation. Finally, several analyses were performed post hoc in response to peer review and should be regarded as exploratory. Future studies should evaluate multiple human and LLM judges, incorporate repeated runs through versioned application programming interfaces, test source inference directly, control for length and format, and include independent expert adjudication of clinically consequential disagreements.

5. Conclusions

In this evaluation of 324 responses to real inpatient oncology consultations, a human expert and an LLM judge applying the same nine-domain rubric to identical clinical text produced markedly different rankings, effect sizes and safety assessments, with poor agreement overall and within each response source. The LLM judge's scores depended on response length and were lowest for the terse, context-dependent notes of practising oncologists; the human evaluator's scores did not depend on length. Evaluator choice can therefore materially influence the conclusions drawn from clinical LLM evaluations. Neither a single LLM judge nor a single human rater should be assumed to provide a reference standard; evaluators should be calibrated against independent expert judgement, and their agreement reported, before either is used as the sole assessor of clinical text.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/doi/s1, Figure S1: Paired differences in total score between response sources; Figure S2: Normal Q–Q plots of linear mixed model residuals; Figure S3: Bland–Altman plot of agreement between the two evaluators; Figure S4: Mean derived rank and proportion ranked first, by evaluator; Figure S5: Distribution of derived ranks by response source and evaluator; Figure S6: Exploratory subgroup analyses, by evaluator; File S1: Standardised response-generation prompt (original Turkish text); File S2: Nine-domain evaluation rubric; File S3: CHART reporting checklist.

Author Contributions: Conceptualization, S.S.G. and O.A.; methodology, S.S.G. and O.A.; formal analysis, S.S.G.; investigation, S.S.G., K.T., N.S.J.E., Y.G. and S.E.N.; resources, K.T. and N.S.D.; data curation, K.T., N.S.J.E., E.C. and S.E.N.; validation, E.C. and Y.G.; writing—original draft preparation, S.S.G.; writing—review and editing, K.T., N.S.J.E., E.C., Y.G., S.E.N., N.S.D. and O.A.; supervision, N.S.D. and O.A.; project administration, S.S.G., O.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Istanbul University-Cerrahpaşa Medical Research Ethics Committee at its meeting held on 6 October 2026 (application reference code Jgn7IjEK; approval No. E-24687260-604.01-1800288).

Informed Consent Statement: Patient consent was waived due to the retrospective nature of the study and the use of anonymised data.

Data Availability Statement: The patient-level data generated and analysed during this study are not publicly available due to patient confidentiality and ethical restrictions. The analysis code is available from the corresponding author upon reasonable request.

Acknowledgments: During the preparation of this manuscript, the authors used Claude Opus 5.5 (Anthropic PBC, San Francisco, CA, USA) for the purposes of correcting typographical errors. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

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

Abbreviations

The following abbreviations are used in this manuscript:

CHART	Chatbot Assessment Reporting Tool
CI	Confidence interval
FDR	False discovery rate
HL	Hodges–Lehmann
ICC	Intraclass correlation coefficient
IQR	Interquartile range
LLM	Large language model
LR	Likelihood ratio
SD	Standard deviation
SE	Standard error
TRIPOD-LLM	Transparent Reporting of a multivariable model for Individual Prognosis Or Diagnosis—Large Language Models

References

1. Kung, T.H.; Cheatham, M.; Medenilla, A.; Sillos, C.; De Leon, L.; Elepaño, C.; Madriaga, M.; Aggabao, R.; Diaz-Candido, G.; Maningo, J.; et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS Digit. Health* **2023**, *2*, e0000198. https://doi.org/10.1371/journal.pdig.0000198.
2. Ayers, J.W.; Poliak, A.; Dredze, M.; Leas, E.C.; Zhu, Z.; Kelley, J.B.; Faix, D.J.; Goodman, A.M.; Longhurst, C.A.; Hogarth, M.; et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern. Med.* **2023**, *183*, 589–596. https://doi.org/10.1001/jamainternmed.2023.1838.
3. Bedi, S.; Liu, Y.; Orr-Ewing, L.; Dash, D.; Koyejo, S.; Callahan, A.; Fries, J.A.; Wornow, M.; Swaminathan, A.; Lehmann, L.S.; et al. Testing and evaluation of health care applications of large language models: A systematic review. *JAMA* **2025**, *333*, 319–328. https://doi.org/10.1001/jama.2024.21700.
4. Huo, B.; Boyle, A.; Marfo, N.; Tangamornsuksan, W.; Steen, J.P.; McKechnie, T.; Lee, Y.; Mayol, J.; Antoniou, S.A.; Thirunavukarasu, A.J.; et al. Large language models for chatbot health advice studies: A systematic review. *JAMA Netw. Open* **2025**, *8*, e2457879. https://doi.org/10.1001/jamanetworkopen.2024.57879.
5. Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.P.; et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 46595–46623.
6. Saito, K.; Wachi, A.; Wataoka, K.; Akimoto, Y. Verbosity bias in preference labeling by large language models. *arXiv* **2023**, arXiv:2310.10076.
7. Panickssery, A.; Bowman, S.R.; Feng, S. LLM evaluators recognize and favor their own generations. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 68772–68802.
8. Chen, D.; Parsa, R.; Hope, A.; Hannon, B.; Mak, E.; Eng, L.; Liu, F.-F.; Fallah-Rad, N.; Heesters, A.M.; Raman, S. Physician and artificial intelligence chatbot responses to cancer questions from social media. *JAMA Oncol.* **2024**, *10*, 956–960. https://doi.org/10.1001/jamaoncol.2024.0836.
9. Chen, D.; Huang, R.S.; Jomy, J.; Wong, P.; Yan, M.; Croke, J.; Tong, D.; Hope, A.; Eng, L.; Raman, S. Performance of multimodal artificial intelligence chatbots evaluated on clinical oncology cases. *JAMA Netw. Open* **2024**, *7*, e2437711. https://doi.org/10.1001/jamanetworkopen.2024.37711.
10. Hodges, J.L.; Lehmann, E.L. Estimates of location based on rank tests. *Ann. Math. Stat.* **1963**, *34*, 598–611. https://doi.org/10.1214/aoms/1177704172.
11. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **1979**, *6*, 65–70.
12. Koo, T.K.; Li, M.Y. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J. Chiropr. Med.* **2016**, *15*, 155–163. https://doi.org/10.1016/j.jcm.2016.02.012.

13. Shrout, P.E.; Fleiss, J.L. Intraclass correlations: Uses in assessing rater reliability. *Psychol. Bull.* **1979**, *86*, 420–428. https://doi.org/10.1037/0033-2909.86.2.420.
14. Bland, J.M.; Altman, D.G. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* **1986**, *1*, 307–310. https://doi.org/10.1016/S0140-6736(86)90837-8.
15. Cohen, J. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol. Bull.* **1968**, *70*, 213–220. https://doi.org/10.1037/h0026256.
16. Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* **1995**, *57*, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.
17. Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46. https://doi.org/10.1177/001316446002000104.
18. Newcombe, R.G. Two-sided confidence intervals for the single proportion: Comparison of seven methods. *Stat. Med.* **1998**, *17*, 857–872. https://doi.org/10.1002/(SICI)1097-0258(19980430)17:8<857::AID-SIM777>3.0.CO;2-E.
19. Cronbach, L.J. Coefficient alpha and the internal structure of tests. *Psychometrika* **1951**, *16*, 297–334. https://doi.org/10.1007/BF02310555.
20. The CHART Collaborative. Reporting guideline for chatbot health advice studies: The Chatbot Assessment Reporting Tool (CHART) statement. *BMC Med.* **2025**, *23*, 436. https://doi.org/10.1186/s12916-025-04274-w.
21. Gallifant, J.; Afshar, M.; Ameen, S.; Aphinyanaphongs, Y.; Chen, S.; Cacciamani, G.; Demner-Fushman, D.; Dligach, D.; Daneshjou, R.; Fernandes, C.; et al. The TRIPOD-LLM reporting guideline for studies using large language models. *Nat. Med.* **2025**, *31*, 60–69. https://doi.org/10.1038/s41591-024-03425-5.
22. Feinstein, A.R.; Cicchetti, D.V. High agreement but low kappa: I. The problems of two paradoxes. *J. Clin. Epidemiol.* **1990**, *43*, 543–549. https://doi.org/10.1016/0895-4356(90)90158-L.
23. Gwet, K.L. Computing inter-rater reliability and its variance in the presence of high agreement. *Br. J. Math. Stat. Psychol.* **2008**, *61*, 29–48. https://doi.org/10.1348/000711006X126600.
24. Byrt, T.; Bishop, J.; Carlin, J.B. Bias, prevalence and kappa. *J. Clin. Epidemiol.* **1993**, *46*, 423–429. https://doi.org/10.1016/0895-4356(93)90018-V.
25. Bavaresco, A.; Bernardi, R.; Bertolazzi, L.; Elliott, D.; Fernández, R.; Gatt, A.; Ghaleb, E.; Giulianelli, M.; Hanna, M.; Koller, A.; et al. LLMs instead of human judges? A large-scale empirical study across 20 NLP evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025)*; Volume 2: Short Papers; Association for Computational Linguistics: Vienna, Austria, 2025; pp. 238–255. https://doi.org/10.18653/v1/2025.acl-short.20.
26. Szymanski, A.; Ziems, N.; Eicher-Miller, H.A.; Li, T.J.-J.; Jiang, M.; Metoyer, R.A. Limitations of the LLM-as-a-judge approach for evaluating LLM outputs in expert knowledge tasks. In *Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI '25)*; ACM: New York, NY, USA, 2025. https://doi.org/10.1145/3708359.3712091.
27. Chen, G.H.; Chen, S.; Liu, Z.; Jiang, F.; Wang, B. Humans or LLMs as the judge? A study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024)*; Association for Computational Linguistics: Miami, FL, USA, 2024; pp. 8301–8327. https://doi.org/10.18653/v1/2024.emnlp-main.474.
28. Croxford, E.; Gao, Y.; First, E.; Pellegrino, N.; Schnier, M.; Caskey, J.; Oguss, M.; Wills, G.; Chen, G.; Dligach, D.; et al. Evaluating clinical AI summaries with large language models as judges. *npj Digit. Med.* **2025**, *8*, 640. https://doi.org/10.1038/s41746-025-02005-2.
29. Guo, Z.; Lai, A.; Korakas, E.; Vagenas, A.; Ahamed, I.; Albor, C.; Zhang, H.; Healy, J.; Li, K. Blinded multi-rater comparative evaluation of a large language model and clinician-authored responses in CGM-informed diabetes counseling. *arXiv* **2026**, arXiv:2604.15124.
30. Arora, R.K.; Wei, J.; Soskin Hicks, R.; Bowman, P.; Quiñonero-Candela, J.; Tsimpourlas, F.; Sharman, M.; Shah, M.; Vallone, A.; Beutel, A.; et al. HealthBench: Evaluating large language models towards improved human health. *arXiv* **2025**, arXiv:2505.08775.
31. Shah, A.; Hines, A.; Downs, A.; Bajet, D.; Mui, P.; Araujo, F.; Offutt, L.; Rutledge, A.; Jimenez, E. Case-specific rubrics for clinical AI evaluation: Methodology, validation, and LLM-clinician agreement across 823 encounters. *arXiv* **2026**, arXiv:2604.24710.
32. Borgohain, S.; Mariathas, R. Decomposing physician disagreement in HealthBench. *arXiv* **2026**, arXiv:2602.22758.
33. Myford, C.M.; Wolfe, E.W. Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *J. Appl. Meas.* **2003**, *4*, 386–422.
34. Buckley, T.A.; Kanjee, Z.; Brodeur, P.G.; Crowe, B.; Pettinato, A.M.; Shah, A.P.; Haimovich, A.D.; McCoy, L.G.; Restrepo, D.; Goh, E.; et al. Scaling clinical judgment to evaluate medical AI. *arXiv* **2026**, arXiv:2609.12822.

35. Wang, P.; Li, L.; Chen, L.; Cai, Z.; Zhu, D.; Lin, B.; Cao, Y.; Kong, L.; Liu, Q.; Liu, T.; et al. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*; Volume 1: Long Papers; Association for Computational Linguistics: Bangkok, Thailand, 2024; pp. 9440–9450. https://doi.org/10.18653/v1/2024.acl-long.511.
36. Thorndike, E.L. A constant error in psychological ratings. *J. Appl. Psychol.* **1920**, *4*, 25–29. https://doi.org/10.1037/h0071663.
37. Cooper, W.H. Ubiquitous halo. *Psychol. Bull.* **1981**, *90*, 218–244. https://doi.org/10.1037/0033-2909.90.2.218.
38. Tam, T.Y.C.; Sivarajkumar, S.; Kapoor, S.; Stolyar, A.V.; Polanska, K.; McCarthy, K.R.; Osterhoudt, H.; Wu, X.; Visweswaran, S.; Fu, S.; et al. A framework for human evaluation of large language models in healthcare derived from literature review. *npj Digit. Med.* **2024**, *7*, 258. https://doi.org/10.1038/s41746-024-01258-7.
39. Caglayan, A.; Slusarczyk, W.; Rabbani, R.D.; Ghose, A.; Papadopoulos, V.; Boussios, S. Large language models in oncology: Revolution or cause for concern? *Curr. Oncol.* **2024**, *31*, 1817–1830. https://doi.org/10.3390/curroncol31040137.
40. Liao, C.; Chu, C.; Lien, M.; Wu, Y.; Wang, T. AI-enhanced healthcare: Integrating ChatGPT-4 in ePROs for improved oncology care and decision-making: A pilot evaluation. *Curr. Oncol.* **2025**, *32*, 7. https://doi.org/10.3390/curroncol32010007.

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