ORIGINAL ARTICLE

DOI: https://doi.org/10.5114/ait/221943

Effect of evaluation prompt strategies on LLM-as-a-judge reliability in critical care

Jia-Yu Yan[1#], Wing-Sum Chan[2#], Ching-Tang Chiu[3], Anne Chao[3], Po-Ni Hsiao[3], Shih-Chi Ku[4], Yu-Chang Yeh[3]

[1]*School of Medicine, College of Medicine, National Taiwan University, Taipei, Taiwan*
[2]*Department of Anesthesiology, Far Eastern Memorial Hospital, New Taipei City, Taiwan*
[3]*Department of Anaesthesiology, National Taiwan University Hospital, Taipei, Taiwan*
[4]*Department of Medicine, National Taiwan University Hospital, Taipei, Taiwan*
[#]These authors contributed equally to this work.

Abstract

Background: Large language model (LLM)-as-a-judge systems offer scalable evaluation of artificial intelligence (AI)-generated clinical outputs, yet their susceptibility to prompt variability raises concerns regarding reproducibility and alignment with expert judgement. This study examined whether evaluation prompt strategies influence scoring patterns and concordance with clinical raters in critical care.

Methods: This post-hoc analysis used 90 structured clinical reports generated in a prior study using an XGBoost ICU mortality prediction model trained on the MIMIC-IV database. GPT-4o (Azure AI, version 2024-11-20) produced structured interpretations from risk estimates and SHAP attributions. These outputs were evaluated using the IMPACT framework under three evaluation prompt strategies: baseline (E1), top-down decremental (E2), and bottom-up incremental (E3). Agreement between clinician ratings and the automated o3-mini evaluator (Azure AI, version 2025-01-31) was assessed using intraclass correlation coefficients (ICC), with strategy comparisons by Fisher's *z*-transformation. Score deviations were examined with repeated-measures ANOVA.

Results: Mean IMPACT scores were 79.9 (SD 9.9) for E1, 83.3 (SD 9.6) for E2, 78.7 (SD 9.1) for E3, and 78.6 (SD 8.9) for clinicians. All strategies demonstrated substantial agreement (ICC > 0.80). E2 showed significantly lower agreement with clinicians (ICC = 0.82) than E1 and E3 (both ICC = 0.94, $p < 0.001$). Score deviations differed significantly across strategies ($p < 0.001$), with E3 showing the smallest mean deviation (0.1) and E2 the largest (4.7).

Conclusions: Prompt design meaningfully affects both IMPACT scoring patterns and the reliability of LLM-based evaluators. Bottom-up incremental scoring showed the closest alignment with human assessment, underscoring the need for standardised prompt architectures in clinical AI evaluation.

Key words: large language model, generative artificial intelligence, prompt engineering, critical care.

Anaesthesiol Intensive Ther 2026; 58: e156–e163

Received: 11.12.2025, **accepted:** 17.05.2026

CORRESPONDING AUTHOR:

Dr. Yu-Chang Yeh, Department of Anaesthesiology, National Taiwan University Hospital, Taipei, 10002, Taiwan, e-mail: tonyyeh@ntuh.gov.tw

The exponential growth of medical data has catalysed the integration of large language models (LLMs) into healthcare, demonstrating promising performance across diverse clinical applications. In high-stakes environments such as critical care, LLM-enabled decision support systems offer potential to enhance evidence-based clinical decision-making [1]. Despite these advances, significant limitations persist. Concerns regarding reproducibility [2, 3], hallucination [4, 5], and algorithmic bias [6] have been well documented in the literature. Given that misinformation in medical contexts can result in life-threatening consequences, human expert evaluation remains the gold standard for validating the accuracy and clinical relevance of LLM outputs [7]. Nevertheless, such expert-based assessments present substantial challenges in terms of resource allocation and scalability.

To address these evaluation challenges, the LLM-as-a-judge paradigm has emerged as a promising automated alternative [8, 9]. Through reinforcement learning from human feedback (RLHF) [10], contemporary LLMs demonstrate substantial alignment with human preferences. Initial

validation studies, including a seminal investigation using GPT-4 as an evaluator, reported over 80% concordance with human annotators in pairwise comparison tasks [11]. However, clinical implementation of LLM-as-a-judge systems remains limited by their susceptibility to prompt variability, wherein modifications to input prompts can produce substantial variations in evaluation outcomes [12-16].

Addressing this methodological gap, we hypothesised that systematic variations in evaluation prompt design would significantly influence both scoring patterns and human-LLM concordance when assessing structured clinical decision-support outputs. By quantifying these prompt-induced effects, this study seeks to explore the methodological implications for LLM evaluators and offer preliminary evidence that may guide their implementation in medical artificial intelligence (AI) evaluation pipelines.

METHODS

Study materials and data source

In this post-hoc analysis, we evaluated the 90 structured clinical reports (hereafter 'LLM-generated responses') produced by GPT-4o in our parent study (Yeh *et al*., 2026, under review), in which we developed an XGBoost mortality prediction model trained on the Medical Information Mart for Intensive Care IV (MIMIC-IV, version 3.1) [17-19]. The use of data was approved by the Institutional Review Boards of MIT (No. 0403000206) and BIDMC (2001-P-001699/14), with waiver of informed consent. Two investigators (J.Y.Y. and Y.C.Y.) obtained CITI certification (Nos. 14536231 and 32697132) for database access. This study was reported in accordance with the TRIPOD-LLM reporting guideline [20]. The completed TRIPOD-LLM Reporting Compliance Checklist is provided in Additional file 1 (Section 1). This post-hoc analysis was not preregistered.

The XGBoost model using 24 clinical features to predict ICU mortality achieved an area under the receiver operating characteristic curve (AUROC) of 0.829 (95% CI: 0.816–0.840) and an area under the precision–recall curve (AUPRC) of 0.327 (95% CI: 0.300–0.356) (Supplementary Figure S1, Additional file 1). In that study, the ICU mortality risk and SHAP-based feature attributions were provided to GPT-4o (Azure AI, version 2024-11-20, knowledge cut-off: October 1, 2023) for clinical interpretation, and three prompting strategies were applied to generate structured clinical support outputs. P1 used baseline zero-shot prompting with defined response structure; P2 applied an expanded instruction format with more explicit task specification, structured output requirements, and controls to improve consistency; and P3 extended this approach into a five-turn conversational sequence to build context progressively. The Python scripts of P1, P2, and P3 are provided in Additional file 1 (Section 2). Thirty cases were selected from the validation dataset using stratified random sampling (Python random seed = 101) based on predicted mortality risk, including 10 low-risk cases (< 10%), 10 moderate-risk cases (10 to 19.99%), and 10 high-risk cases (≥ 20%). For each case, GPT-4o produced five components covering risk factor interpretation, diagnostic recommendations, management planning, follow-up assessment, and an integrated summary, yielding 90 LLM-generated responses (30 cases × 3 prompting strategies). The 90 GPT-4o responses used in this analysis were generated between February and March 2025, using cases drawn from MIMIC-IV v3.1, which contains admissions from 2008 to 2022. In the current post-hoc analysis, these 90 responses served as test materials for evaluating prompt-induced variability in automated assessment systems.

IMPACT evaluation framework and prompt strategies

The IMPACT (Integration, Mastery, Precision, Applicability, Comprehensiveness, and Timeliness) scoring criteria were developed through expert consensus meetings involving the research committee of the Taiwan Society of Emergency and Critical Care Medicine, in collaboration with the Taiwan Society of Critical Care Medicine and the Taiwan Society of Anesthesiologists. The steering team defined six evaluation domains and used the validated DISCERN instrument [21] as a foundational reference, adapting relevant quality assessment concepts while adding new domain-specific items tailored to AI-generated clinical decision-support outputs in critical care. Each domain targets a distinct aspect of output quality: Integration assesses coherence with patient-specific clinical data; Mastery evaluates the depth and accuracy of clinical reasoning; Precision examines the correctness of recommendations; Applicability addresses relevance to bedside decisions; Comprehensiveness captures completeness of assessment; and Timeliness reflects appropriateness for acute care timelines. The detailed framework development and validation methodology are reported separately (Yeh *et al*., AICOJ-D-26-00329, under revision). The complete scoring criteria are provided in Supplementary Table S1 (Additional file 1). In the parent study, we developed a 25-item assessment instrument aligned with the six domains of the IMPACT framework. Each item was scored from 0 to 4, yielding a total possible score of 0–100. All raters received the same standardised scoring prompt with item-level scoring criteria to ensure consistency [22]. The complete scoring criteria are

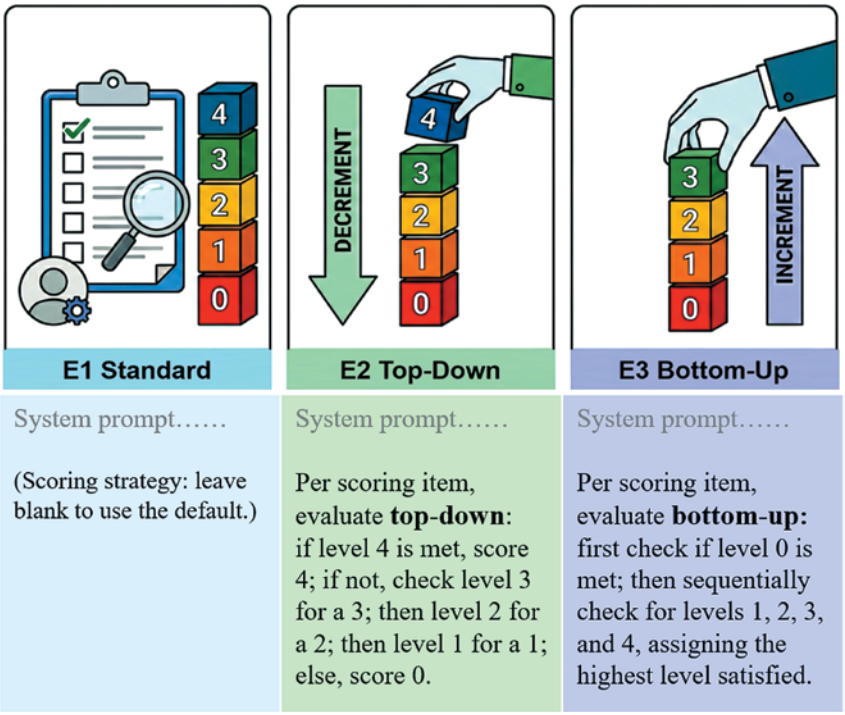

FIGURE 1. Comparison of three evaluation architectures for large language model scoring. This figure compares three scoring architectures used for structured evaluation of model-generated responses. E1 Standard reflects conventional direct scoring based on overall impression. E2 Top-Down applies a decrement approach that begins with the maximum score and subtracts points for missing or incorrect elements. E3 Bottom-Up applies an increment approach that starts at zero and adds points only when required criteria are present. This figure was generated using Gemini 3 with prompts designed by YCY and reviewed for accuracy

provided in Additional file 1. The IMPACT scoring system demonstrated high reliability: human-to-human ICC was 0.836 (95% CI: 0.792–0.876), and the ICC between the human mean and o3-mini mean was 0.975 (95% CI: 0.969–0.982) [22]. Three evaluation prompt strategies were developed to assess the 90 LLM-generated responses. The original evaluation prompt (E1) followed standard scoring procedures. The top-down approach (E2) instructed the model to begin with the maximum score of 4 for each item and decrement based on unmet criteria, while the bottom-up approach (E3) required starting from 0 and incrementing as criteria were satisfied (Figure 1). The full Python evaluation script (E1) and example system prompts for E1, E2, and E3 are provided in Additional file 1 (Sections 5.1 and 5.2). The three prompts share the same IMPACT rubric and differ only in scoring direction. E1 was used as the primary evaluation prompt in the parent study (Yeh *et al*., 2026, under review) and in the related post-hoc study [22]. The current study tested whether alternative scoring directions (E2 and E3) produced different agreement patterns. The overall study flow is illustrated in Supplementary Figure S2 (Additional file 1). LLM responses were assessed by an automated evaluator (o3-mini [Azure AI, version 2025-01-31, knowledge cut-off: October 1, 2023]). As a reasoning model, o3-mini has most sampling parameters fixed by the API (temperature = 1, top-p = 1, frequency penalty = 0, presence penalty = 0; seed not supported). The reasoning_effort parameter was set to 'medium,' and max_completion_tokens was set to 4000. For each case, four sequential API calls were issued to o3-mini, corresponding to Part 1 (Interpretation, 9 items), Part 2 (Examination, 6 items), Part 3 (Management, 6 items), and Parts 4 and 5 combined (Follow-up, 3 items; Summary, 1 item). Each call returned a structured JSON object containing item-level scores (0–4) and justification comments. Subtotals from the four calls were summed to yield the overall IMPACT total score (0–100). Malformed responses were handled through an automated retry mechanism (up to ten retries with exponential backoff) using the Tenacity library. The complete evaluation pipeline code is provided in Additional file 1 (Section 5). o3-mini was selected as the evaluator model because of its reasoning-optimised architecture, its availability through our institutional Azure AI deployment with zero data retention, and its cost-efficiency for triple evaluation of all 90 responses. The compute and inference details of o3-mini evaluations are provided in Supplementary Table S2 (Additional file 1). Human evaluation was performed using the E1 (standard scoring) prompt. Six clinical raters participated: two intensivists (21 and 3 years of ICU experience), one resident (3 years), and three ICU nurses (16, 10, and 2 years). All raters received an introduction

to the scoring criteria and one scored example for calibration before evaluation. Each response was independently assessed by three clinical raters to ensure multiple expert perspectives and minimise individual bias.

Outcomes measurement

The primary outcome was the intraclass correlation coefficient (ICC) between the mean scores of three clinical raters (reference standard) and the o3-mini evaluator across the three evaluation strategies (E1, E2, and E3). The secondary outcome was the mean difference in IMPACT scores between clinical raters and the o3-mini evaluator for each evaluation condition.

Statistical analysis

Continuous variables were summarised as mean (standard deviation) for normally distributed data and as median (interquartile range, IQR) for skewed data. Between-group comparisons were assessed using the independent-samples t-test for continuous variables and Fisher's exact test for categorical variables, as appropriate. Absolute agreement between human raters and o3-mini was quantified using a two-way mixed-effects, absolute-agreement, single-measure ICC (ICC[A,1]) and interpreted according to the benchmarks proposed by Koo and Li [23].

To test for statistical differences between evaluation strategies, ICCs were compared using Fisher's *z*-transformation. Each ICC was converted to its Fisher's *z* value to stabilise variance, followed by pairwise comparisons (E1 vs. E2, E1 vs. E3, and E2 vs. E3). The 30 patients were selected by stratified random sampling across three risk groups, and the three generation prompts (P1, P2, P3) produce substantially different clinical reports in structure, content, and detail, supporting their analysis as independent evaluation targets within each ICC analysis (see example responses of P1, P2, and P3 in Additional file 1, Section 3). Differences in systematic bias of total IMPACT scores across evaluation prompts were assessed using repeated-measures ANOVA, followed by Bonferroni-adjusted paired t-tests for post hoc comparisons. Statistical significance was defined as two-sided $p < 0.05$ for repeated-measures ANOVA and Bonferroni-adjusted $p < 0.017$ for paired comparisons. All analyses were performed using Python 3.12.7.

RESULTS

Baseline characteristics of the 30 selected cases are presented in Table 1. The median age was 72 years (IQR 68–80) for survivors and 86 years (IQR 71–87) for non-survivors. The mean IMPACT scores varied significantly by evaluation strategy. The top-down approach (E2) yielded the highest scores (mean 83.3 [9.6]), followed by the baseline approach (E1; mean, 79.9 [9.9]) and the bottom-up approach (E3; mean 78.7 [9.1]). The reference standard, derived from three independent clinical raters, was 78.6 (8.9). Repeated-measures ANOVA demonstrated a significant effect of evaluation strategy on the difference between o3-mini and clinician scores ($F(2, 178) = 90.46$, $p < 0.001$, $\eta^2G = 0.262$), indicating a large effect. Post hoc analyses revealed significant differences between all strategy pairs. The bottom-up approach (E3) showed the smallest mean deviation from clinician ratings (0.1), followed by the baseline approach (E1, 1.3) and the top-down approach (E2, 4.7). Score deviation distributions are shown in Figure 2.

Agreement analysis revealed substantial variation in alignment between o3-mini and clinician assessments across strategies (Table 2). The bottom-up and baseline approaches demonstrated excellent agreement with clinicians (ICC[A,1] > 0.90), while the top-down approach showed good agreement (ICC[A,1] = 0.82) according to established criteria [23]. Fisher's z-transformation confirmed that the top-down strategy had significantly lower reliability compared with both the baseline ($p < 0.001$) and bottom-up approaches ($p < 0.001$). No significant difference was observed between the baseline and bottom-up approaches ($p = 0.82$). Agreement patterns are illustrated in Figure 3. The bottom-up evaluation strategy (E3) demonstrated optimal performance, achieving the highest ICC (0.94) and minimal deviation from clinician ratings.

DISCUSSION

This study demonstrated that prompt design significantly influences the concordance between automated LLM-based evaluation and clinical expert assessment of structured medical information. The bottom-up evaluation strategy (E3) achieved the highest agreement with clinical raters (ICC = 0.94) and minimal score deviation (mean difference = 0.1), while maintaining excellent reliability. The baseline approach (E1) also demonstrated excellent agreement (ICC > 0.90), though with slightly greater score deviation. In contrast, the top-down strategy (E2) showed significantly lower reliability (ICC = 0.82) and systematically overestimated scores compared with clinical ratings (mean difference = 4.7). These findings suggest that incremental scoring approaches, which progressively build scores from zero based on fulfilled criteria, more accurately replicate clinical evaluative reasoning than decremental approaches that begin with maximum scores. The observed differences in both magnitude and consistency of LLM evaluations across prompt-

TABLE 1. Patient characteristics

Characteristic	Survivors ($n = 24$)	Non-survivors ($n = 6$)	Total ($n = 30$)
Age (years)	72.0 (68.2–80.2)	85.5 (71.2–87.0)	73.0 (67.5–85.5)
Weight (kg)	71.9 (61.5–77.2)	85.2 (66.7–91.7)	73.0 (62.0–86.2)
Heart rate (bpm)	97.5 (84.0–105.5)	110.0 (89.5–125.2)	100.5 (86.0–107.8)
Respiratory rate (bpm)	21.0 (18.8–28.8)	27.5 (26.2–28.8)	23.0 (20.0–28.8)
Temperature (°C)	36.8 (36.6–37.1)	36.5 (36.4–36.9)	36.7 (36.4–37.0)
SpO_2 (%)	95.5 (92.0–100.0)	94.0 (91.5–97.2)	95.0 (92.0–100.0)
SBP (mmHg)	123.0 (110.5–142.5)	109.5 (90.5–116.5)	116.5 (104.5–137.2)
DBP (mmHg)	73.5 (55.8–80.2)	69.5 (61.5–72.2)	72.0 (58.2–80.0)
MAP (mmHg)	89.5 (74.2–100.2)	79.0 (73.5–83.8)	86.0 (72.8–98.2)
Glasgow Coma Scale	15.0 (15.0–15.0)	15.0 (15.0–15.0)	15.0 (15.0–15.0)
Congestive heart failure	6 (25.0%)	5 (83.3%)	11 (36.7%)
Cerebrovascular disease	6 (25.0%)	0 (0.0%)	6 (20.0%)
Chronic pulmonary disease	5 (20.8%)	1 (16.7%)	6 (20.0%)
Mild liver disease	4 (16.7%)	1 (16.7%)	5 (16.7%)
Severe liver disease	2 (8.3%)	0 (0.0%)	2 (6.7%)
Diabetes	7 (29.2%)	3 (50.0%)	10 (33.3%)

Data are presented as median (interquartile range [IQR]) for continuous variables and number (percentage, %) for categorical variables. Other comorbidities, including rheumatic disease, dementia, malignant cancer, and metastatic solid tumour, as well as clinical factors such as intubation at 0 h, emergency admission, and first ICU stay, were recorded but did not show significant associations with survival group and are therefore not displayed.
MAP – mean arterial pressure, SBP – systolic blood pressure, DBP – diastolic blood pressure

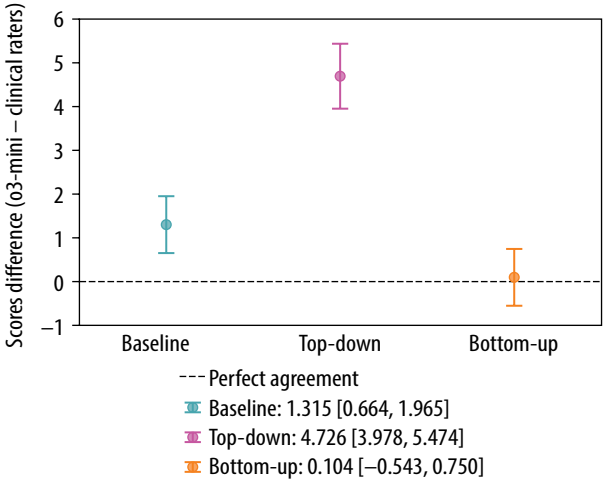

FIGURE 2. Score deviation from clinical raters. Each point stands for the mean difference between o3-mini and clinical rater scores, with error bars denoting 95% confidence intervals ($n = 90$)

ing strategies highlight the critical importance of prompt optimisation for ensuring reliable automated assessment in clinical applications.

Despite lower alignment of the top-down approach with clinical ratings, all evaluation strategies achieved ICCs exceeding 0.80, demonstrating that LLM-based evaluators can maintain substantial concordance with human experts. This supports the feasibility of automated evaluation for clinical decision support systems. However, the significant variation in performance across prompting strategies emphasises the critical need for standardised prompt protocols in clinical applications. The superior performance of baseline (E1) and bottom-up (E3) strategies compared with top-down (E2) prompting can be attributed to two factors. First, the baseline prompt's lack of directional anchoring permitted bidirectional score adjustment, minimising systematic deviation from clinical judgements. Second, the alignment between the bottom-up approach and the human rating form – which positioned zero at the top of the scale – may have naturally primed similar cognitive processing in both human and machine evaluators. This correspondence sug-

TABLE 2. Agreement between o3-mini and clinical raters across evaluation strategies

Rating strategy	ICC[A,1]	Interpretation[a]	Comparison[b]	ΔICC
Baseline (E1)	0.94 (0.90–0.96)	Excellent	E1-E2	0.118#
Top-down (E2)	0.82 (0.74–0.88)	Good	E1-E3	–0.004
Bottom-up (E3)	0.94 (0.91–0.96)	Excellent	E2-E3	–0.123#

Intraclass correlation coefficients (ICC[A,1]) with 95% confidence intervals (CIs) are reported. [a]Interpretations follow Koo and Li's thresholds (excellent: ICC > 0.90; good: 0.75–0.90). [b]Fisher's z transformations were used to test for significant differences between correlated ICCs (*$p < 0.05$; #$p < 0.001$)

gests that evaluation consistency depends not only on LLM prompt design but also on the structure of human rating instruments. These findings indicate that both human and machine evaluators are susceptible to similar cognitive framing effects. Future research should systematically investigate how scoring orientation, midpoint anchoring, and scale directionality affect evaluation outcomes across both human and automated assessors to develop optimal assessment protocols that minimise positional and directional biases.

Building upon these findings, we next investigated why the top-down strategy consistently produced inflated ratings in comparison with the baseline and bottom-up approaches. Based on the existing literature, three mechanisms may explain why the top-down approach consistently produced inflated ratings. First, anchoring bias, documented in both human cognition and LLM reasoning, causes initial reference points to disproportionately influence subsequent judgements [15, 16, 24]. This bias persists despite chain-of-thought prompting [16, 25]. Second, attentional decay occurs due to token-level limitations. While modern LLMs possess extensive context windows, their effective context length remains shorter than theoretical maximums, reducing sensitivity to later information [26, 27]. Third, token-specific priors associated with option identifiers systematically bias output probabilities. Zheng *et al*. demonstrated that removing rather than shuffling option identifiers substantially reduced positional bias [16]. These mechanisms mirror human cognitive biases, suggesting that LLMs exhibit analogous information processing patterns. The parallel between human and machine evaluation biases indicates that psychological frameworks may inform LLM optimisation. Strategies effective for mitigating human cognitive biases, including structured decision protocols and debiasing techniques, may prove equally applicable to LLM-based clinical evaluation systems. These mechanistic explanations are interpretive hypotheses informed by the general LLM behaviour literature and were not directly tested in this study. Future work using model-internal interpretability methods or controlled ablation experiments would be needed to establish causal pathways. Our team has used Claude 3.7 (Anthropic) as an alternative evaluator with the same IMPACT framework [22]. Future studies may examine whether the effects of scoring direction (E1, E2, E3) are consistent across evaluator models from different vendors.

Although the primary aim of this study was to investigate the effect of scoring strategy on LLM-based evaluation performance, the broader LLM-evaluator pipeline could serve as a gatekeeper to ensure the quality of AI-generated clinical support

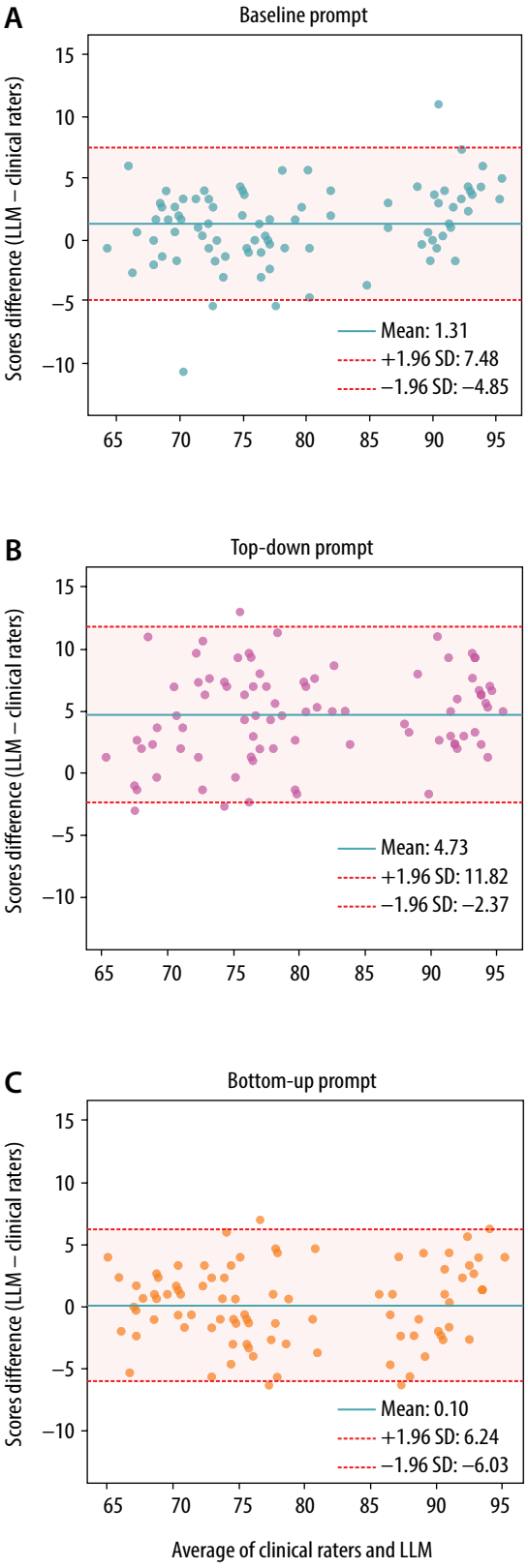

FIGURE 3. Bland-Altman analysis of agreement patterns. Bland-Altman plots for each LLM strategy (Baseline, Top-down, and Bottom-up) illustrate the agreement pattern between LLM-generated and clinical scores. The solid blue line represents the mean difference, and the dashed red lines represent the limits of agreement (LoA, mean difference ± 1.96 SD). Shaded areas show the 95% LoA regions.

CI – confidence interval

information. Such a pipeline would be used by clinical informaticists or AI researchers with expertise in both clinical medicine and LLM behaviour, operating under institutional oversight. Reports flagged as low-quality by the automated evaluator (e.g. total IMPACT scores below a predefined threshold) should be regenerated or revised and then rescored.

To our knowledge, this is one of the first pilot studies examining prompt-induced variability in LLM-based evaluations within critical care contexts. Our findings demonstrate how prompt architecture affects both absolute scoring levels and inter-rater concordance, providing essential groundwork for reproducible, human-aligned evaluation frameworks. These insights support LLMs' potential to augment clinical assessment workflows while maintaining reliability. However, several limitations merit consideration. First, data originated from a single academic centre, necessitating external validation across diverse clinical populations to confirm generalisability. Second, the clinician scoring form design may have influenced alignment patterns. The scale's bottom-up format, with zero positioned at the top, may have unconsciously primed raters toward bottom-anchored evaluation, potentially explaining the superior performance of bottom-up LLM prompts. Third, future studies should explore alternative scale orientations, including midpoint-anchored or symmetrical layouts, to minimise positional bias in both human and automated assessments. Fourth, both models share a knowledge cut-off of October 1, 2023, and the MIMIC-IV records (2008–2022) fall within this period. PhysioNet permits the use of MIMIC data with Azure OpenAI service provided that users opt out of human review of the data, and Microsoft's Azure OpenAI service does not use customer prompts or completions to train models. All data in this study were processed through Azure OpenAI in compliance with these requirements. Although direct patient-level contamination is therefore unlikely, LLM familiarity with published MIMIC-IV-related content (e.g. peer-reviewed studies using MIMIC-IV data) cannot be entirely excluded. Fifth, this study used a single automated evaluator (o3-mini). Both the content generator (GPT-4o) and the evaluator are products of the same vendor (OpenAI), introducing potential same-vendor bias, whereby the evaluator may favour outputs with stylistic characteristics typical of its own model family [28, 29]. Sixth, the 90 responses were nested within 30 patients, each contributing three responses from different prompting strategies. As a post-hoc analysis, the sample size was determined by the parent study, and a formal a priori power calculation was not performed. Future confirmatory studies with larger, independently sampled datasets are warranted. Seventh, the sample size of 30 patients was too small to support reliable subgroup analyses by survival status or other case-level characteristics. Further studies with larger cohorts are warranted.

CONCLUSIONS

This study demonstrated that prompt design substantially affects IMPACT scores and alignment between LLM-based evaluators and clinical raters. Among the tested strategies, the bottom-up configuration achieved the best agreement with minimal bias, underscoring prompt structure as a key determinant of reliability. These findings highlight the need for standardised, reproducible evaluation frameworks to ensure trustworthy LLM-based clinical assessments. Given the rapid evolution of large language models, future studies using newer architectures should re-evaluate potential effects on scoring patterns and concordance with human raters.

ACKNOWLEDGEMENTS

1. Assistance with the article: We thank Miss Hsiao-Lan Shih, Miss Pei-Tsen Ko and Miss Hsiu-Hui Tsai for their assistance in evaluating the LLM responses. In addition to the investigations and application of generative AI described in the Methods section, GPT-5 (OpenAI) was used during manuscript preparation for English-language editing. All AI-generated suggestions were critically reviewed, edited, and approved by the authors, who take full responsibility for the content and integrity of this manuscript.
2. Financial support and sponsorship: This work was funded by National Science and Technology Council (Taiwan) grant NSTC 112-2314-B-002-274 and National Taiwan University Hospital grant NTUH 112-FY0002. The funders had no role in study design, analysis, interpretation, or decision to publish.
3. Conflicts of interest: None.
4. Presentation: None.

REFERENCES

1. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. npj Digit Med 2020; 3: 17. DOI: https://doi.org/10.1038/s41746-020-0221-y.
2. Goh S, Goh RSJ, Chong B, Ng QX, Koh GCH, Ngiam KY, et al. Challenges in implementing artificial intelligence in breast cancer screening programmes: systematic review and framework for safe adoption. J Med Internet Res 2025; 27: e62941. DOI: 10.2196/62941.
3. Hassan M, Kushniruk A, Borycki E. Barriers to and facilitators of artificial intelligence adoption in health care: scoping review. JMIR Hum Factors 2024; 11: e48633. DOI: 10.2196/48633.
4. Bang Y, Cahyawijaya S, Lee N, Dai W, Su D, Wilie B, et al. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. arXiv:2302.04023; 2023. DOI: https://doi.org/10.48550/arXiv.2302.04023.

5. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. ACM Comput Surv 2023; 55: 248. DOI: https://doi.org/10.1145/3571730.
6. Schramowski P, Turan C, Andersen N, Rothkopf CA, Kersting K. Large pre-trained language models contain human-like biases of what is right and wrong to do. Nat Mach Intell 2022; 4: 258-268.
7. Tam TYC, Sivarajkumar S, Kapoor S, Stolyar AV, Polanska K, McCarthy KR, et al. A framework for human evaluation of large language models in healthcare derived from literature review. NPJ Digit Med 2024; 7: 258. DOI: https://doi.org/10.1038/s41746-024-01258-7.
8. Gu J, Jiang X, Shi Z, Tan H, Zhai X, Xu C, et al. A survey on LLM-as-a-judge. Innovation (Camb) 2026; 7: 101253. DOI: 10.1016/j.xinn.2025.101253.
9. Liu Y, Iter D, Xu Y, Wang S, Xu R, Zhu C. G-Eval: NLG evaluation using GPT-4 with better human alignment. arXiv:2303.16634; 2023. DOI: https://doi.org/10.48550/arXiv.2303.16634.
10. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, et al. Training language models to follow instructions with human feedback. Adv Neural Inform Proc Syst 2022; 35: 27730-27744.
11. Zheng L, Chiang WL, Sheng Y, Zhuang S, Wu Z, Zhuang Y, et al. Judging LLM-as-a-judge with mt-bench and chatbot arena. Adv Neural Inform Proc Syst 2023; 36: 46595-46623.
12. Sivarajkumar S, Kelley M, Samolyk-Mazzanti A, Visweswaran S, Wang Y. An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing: algorithm development and validation study. JMIR Med Inform 2024; 12: e55318. DOI: https://doi.org/10.2196/55318.
13. Wei H, He S, Xia T, Liu F, Wong A, Lin J, Han M. Systematic evaluation of LLM-as-a-judge in LLM alignment tasks: Explainable metrics and diverse prompt templates. arXiv:2408.13006; 2025. DOI: https://doi.org/10.48550/arXiv.2408.13006.
14. He J, Rungta M, Koleczek D, Sekhon A, Wang FX, Hasan S. Does prompt formatting have any impact on LLM performance? arXiv:241110541; 2024. DOI: https://doi.org/10.48550/arXiv.2411.10541.
15. Pezeshkpour P, Hruschka E. Large language models sensitivity to the order of options in multiple-choice questions. arXiv:2308.11483; 2023. DOI: https://doi.org/10.48550/arXiv.2308.11483.
16. Zheng C, Zhou H, Meng F, Zhou J, Huang M. Large language models are not robust multiple choice selectors. arXiv:2309.03882; 2024. DOI: https://doi.org/10.48550/arXiv.2309.03882.
17. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet. Components of a new research resource for complex physiologic signals. Circulation 2000; 101: E215-E220.
18. Johnson A, Bulgarelli L, Pollard T, Gow B, Moody B, Horng S, et al. 'MIMIC-IV' (version 3.1), PhysioNet RRID:SCR_007345. DOI: https://doi.org/10.13026/kpb9-mt58.
19. Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data 2023; 10: 1. DOI: 10.1038/s41597-022-01899-x.
20. Gallifant J, Afshar M, Ameen S, Aphinyanaphongs Y, Chen S, Cacciamani G, et al. The TRIPOD-LLM reporting guideline for studies using large language models. Nat Med 2025; 31: 60-69. DOI: 10.1038/s41591-024-03425-5.
21. Charnock D, Shepperd S, Needham G, Gann R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. J Epidemiol Community Health 1999; 53: 105-111.
22. Yeh YC, Yang HY, Chiu CT, Chao A, Chuang YC, Chan WS. Enhancing large language model clinical support information with machine learning risk and explainability: a feasibility study. Intensive Care Med Exp 2026; 14: 51.
23. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016; 15: 155-163.
24. Wang P, Li L, Chen L, Cai Z, Zhu D, Lin B, et al. Large language models are not fair evaluators. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics 2024; 1: 9440-9450. DOI: 10.18653/v1/2024.acl-long.511.
25. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, et al. Chain-of-thought prompting elicits reasoning in large language models. arXiv:2201.11903; 2022. DOI: https://doi.org/10.48550/arXiv.2201.11903
26. An C, Zhang J, Zhong M, Li L, Gong S, Luo Y, et al. Why does the effective context length of LLMs fall short? arXiv:2410.18745; 2024. DOI: https://doi.org/10.48550/arXiv.2410.18745.
27. Hsieh C-P, Sun S, Kriman S, Acharya S, Rekesh D, Jia F, et al. RULER: What's the real context size of your long-context language models? arXiv:2404.06654; 2024. DOI: https://doi.org/10.48550/arXiv.2404.06654.
28. Panickssery A, Bowman SR, Feng S. LLM evaluators recognise and favour their own generations. arXiv:2404.13076; 2024. DOI: https://doi.org/10.48550/arXiv.2404.13076.
29. Wataoka K, Takahashi T, Ri R. Self-preference bias in LLM-as-a-judge. arXiv:2410.21819; 2025. DOI: https://doi.org/10.48550/arXiv.2410.21819.