Title: Large Language Models for Automated AGREE II Quality Appraisal of Sepsis Clinical Practice Guidelines: A Methodological Comparative Study

Tiantian Zhang[†], Xiaoming Zhang[†], Youji Zhu[†], Nuoyi Zhang, Liyin Zheng, Kaifeng Ye, Danping Wang[*]

Department of Intensive Care Unit, Xianju People's Hospital, Zhejiang Southeast Campus of Zhejiang Provincial People's Hospital, Affiliated Xianju's Hospital, Hangzhou Medical College, Xianju, Taizhou, Zhejiang, China.

[†]These authors contributed equally to this work.

Corresponding Author

Danping Wang

Department of Intensive Care Unit, XianJu People's Hospital, Zhejiang Southeast Campus of Zhejiang Provincial People's Hospital, Affiliated Xianju's Hospital, Hangzhou Medical College, Xianju, Taizhou, Zhejiang, China.

East Chengbei Road 53, Taizhou, Zhejiang, China

Email: xjicuwdp@163.com

ORCID: 0009-0001-4750-9715

Supplementary Table S1. Human expert AGREE II domain scores for 11 sepsis clinical practice guidelines

Guideline	Scope and Purpose	Stakeholder Involvement	Rigour of Development	Clarity of Presentation	Applicability	Editorial Independence
G1	88.89	83.33	77.08	94.44	45.83	66.67
G2	94.44	94.44	83.33	94.44	79.17	91.67
G3	83.33	55.56	85.42	94.44	62.50	91.67
G4	94.44	88.89	89.58	94.40	66.67	83.33
G5	94.44	55.56	58.33	94.44	54.17	66.67
G6	77.78	50.00	58.33	88.89	33.33	50.00
G7	94.44	94.44	70.83	94.44	66.67	66.67
G8	77.78	38.89	29.17	72.22	25.00	58.33
G9	83.33	88.89	54.17	77.78	37.50	66.67
G10	77.78	50.00	50.00	88.89	45.83	91.67
G11	94.44	83.33	50.00	88.89	25.00	8.33
(Median, IQR)	88.89 (77.78, 94.44)	83.33 (50.00, 88.89)	58.33 (50.00, 83.33)	94.40 (88.89, 94.44)	45.83 (33.33, 66.67)	66.67 (58.33, 91.67)

Note. Scores are presented as percentages. Data extracted from Li et al[1].

References

1. Li, H.Y., Jiang, S.L., Wang, J., Wang, H.S. & Wang, L.H. Quality assessment of clinical practice guidelines for sepsis and variations in recommendations. *BMC Infect Dis* **26**, 158 (2025).

Supplementary Table S2. Intraclass correlation coefficients (ICC) of six large language models for repeated AGREE II ratings across domains

Domain	Doubao-2.0	DeepSeek-V4	GPT-5.5	Gemini-3.1-Pro	Qwen-3.7	Mimo-v2.5
Scope and Purpose	0.837 (0.733, 0.909)	0.485 (0.278, 0.675)	0.835 (0.73, 0.908)	0.795 (0.672, 0.884)	0.803 (0.683, 0.889)	0.371 (0.158, 0.586)
Stakeholder Involvement	0.916 (0.856, 0.954)	0.801 (0.679, 0.888)	0.964 (0.937, 0.981)	0.941 (0.898, 0.968)	0.918 (0.86, 0.955)	0.742 (0.596, 0.851)
Rigour of Development	0.912 (0.878, 0.939)	0.847 (0.791, 0.892)	0.959 (0.943, 0.972)	0.94 (0.916, 0.959)	0.923 (0.893, 0.947)	0.749 (0.667, 0.819)
Clarity of Presentation	0.691 (0.528, 0.819)	0.291 (0.079, 0.518)	0.744 (0.599, 0.853)	0.567 (0.372, 0.735)	0.536 (0.335, 0.713)	0.235 (0.027, 0.468)
Applicability	0.907 (0.853, 0.945)	0.507 (0.331, 0.669)	0.868 (0.795, 0.921)	0.812 (0.714, 0.885)	0.925 (0.88, 0.955)	0.52 (0.345, 0.678)
Editorial Independence	0.932 (0.869, 0.969)	0.87 (0.758, 0.939)	0.971 (0.943, 0.987)	0.962 (0.926, 0.983)	0.919 (0.846, 0.963)	0.417 (0.158, 0.667)
Overall	0.939 (0.926, 0.951)	0.832 (0.798, 0.862)	0.956 (0.946, 0.964)	0.926 (0.91, 0.94)	0.938 (0.924, 0.95)	0.756 (0.711, 0.798)

Supplementary Table S3. Statistical details of LLM-human bias by AGREE II domain: median differences with raw and FDR-corrected p-values.

Model	Domain	Median Bias (LLM-Human)	Raw P-value	BH-FDR P-value	Significance
DeepSeek-V4	Scope and Purpose	1.85	0.40	0.49	ns
DeepSeek-V4	Stakeholder Involvement	-14.82	0.01	0.03	*
DeepSeek-V4	Rigour of Development	5.56	0.52	0.58	ns
DeepSeek-V4	Clarity of Presentation	5.56	0.01	0.02	*
DeepSeek-V4	Applicability	-18.05	0.01	0.03	*
DeepSeek-V4	Editorial Independence	-5.56	0.21	0.29	ns
Doubao-2.0	Scope and Purpose	3.71	0.20	0.28	ns
Doubao-2.0	Stakeholder Involvement	-12.96	<0.01	0.01	*
Doubao-2.0	Rigour of Development	-6.25	0.15	0.28	ns
Doubao-2.0	Clarity of Presentation	3.75	0.20	0.28	ns
Doubao-2.0	Applicability	-25.00	<0.01	0.01	*
Doubao-2.0	Editorial Independence	-16.67	<0.01	0.01	*
GPT-5.5	Scope and Purpose	0.00	0.82	0.87	ns
GPT-5.5	Stakeholder Involvement	-12.96	0.02	0.07	ns
GPT-5.5	Rigour of Development	1.39	0.93	0.96	ns
GPT-5.5	Clarity of Presentation	1.85	0.04	0.10	ns
GPT-5.5	Applicability	-5.56	0.14	0.28	ns
GPT-5.5	Editorial Independence	-16.67	<0.01	0.01	*
Gemini-3.1-Pro	Scope and Purpose	5.56	0.20	0.28	ns
Gemini-3.1-Pro	Stakeholder Involvement	5.55	0.46	0.54	ns
Gemini-3.1-Pro	Rigour of Development	9.72	0.17	0.28	ns
Gemini-3.1-Pro	Clarity of Presentation	5.56	<0.01	0.02	*
Gemini-3.1-Pro	Applicability	1.39	0.45	0.54	ns
Gemini-3.1-Pro	Editorial Independence	-5.56	0.20	0.28	ns

Mimo-v2.5	Scope and Purpose	5.55	0.27	0.34	ns
Mimo-v2.5	Stakeholder Involvement	-9.26	<0.01	0.02	*
Mimo-v2.5	Rigour of Development	9.03	0.14	0.28	ns
Mimo-v2.5	Clarity of Presentation	0.04	0.96	0.96	ns
Mimo-v2.5	Applicability	-22.22	<0.01	0.01	*
Mimo-v2.5	Editorial Independence	8.33	0.17	0.28	ns
Qwen-3.7	Scope and Purpose	5.56	0.05	0.14	ns
Qwen-3.7	Stakeholder Involvement	-1.85	0.64	0.70	ns
Qwen-3.7	Rigour of Development	7.64	0.08	0.18	ns
Qwen-3.7	Clarity of Presentation	5.56	<0.01	0.02	*
Qwen-3.7	Applicability	-16.67	0.12	0.28	ns
Qwen-3.7	Editorial Independence	-2.77	0.27	0.34	ns

Note. P-values were derived from two-sided Wilcoxon signed-rank tests comparing LLM and human expert scores (n = 11 guidelines; each guideline score represents the mean of 3 independent replicates). Raw p-values were adjusted for multiple comparisons across all model–domain combinations using the Benjamini-Hochberg false discovery rate (BH-FDR) procedure. Significance notation: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$; ns = not significant after BH-FDR correction (adjusted $p \geq 0.05$). Median Bias = median difference (LLM − Human); positive values indicate LLM overestimation relative to human ratings.

Supplementary Table S4. AGREE II quality classification concordance between LLMs and the human expert panel for 11 guidelines.

Guideline	Human	Doubao-2.0	DeepSeek-V4	GPT-5.5	Gemini-3.1-Pro	Qwen-3.7	Mimo-v2.5
G1 (China 2025)	B	B	B	B	A ↑	A ↑	B
G2 (J-SSCG 2024)	A	B ↓	B ↓	B ↓	A	B ↓	B ↓
G3 (KSCCM 2024)	B	B	B	B	B	B	B
G4 (NICE 2024)	A	B ↓	B ↓	B ↓	B ↓	B ↓	B ↓
G5 (SWAB 2022)	B	B	B	B	B	B	B
G6 (China 2022)	B	B	B	B	B	B	B
G7 (SSC 2021)	A	B ↓	B ↓	B ↓	B ↓	B ↓	B ↓
G8 (JAID/JSC 2021)	B	C ↓	B	B	B	B	B
G9 (China 2019)	B	B	B	B	B	B	B
G10 (AGIHO 2018)	B	B	B	B	B	B	B
G11 (China 2018)	B	C ↓	C ↓	C ↓	C ↓	C ↓	B
Concordant / Total	-	6/11	7/11	7/11	7/11	6/11	8/11
Concordance rate	-	54.50%	63.60%	63.60%	63.60%	54.50%	72.70%

Note: Guidelines were classified according to the AGREE II three-tier criteria: Grade A (strongly recommended): all six domain scores ≥ 60%; Grade B (recommended with modification): ≥ 3 domains between 30% and 60%; Grade C (not recommended): ≥ 3 domains < 30%. Human reference grades were derived from Li et al.[1]. Arrows indicate discordant classifications relative to the human expert consensus: ↑ indicates the LLM upgraded guideline quality (B→A); ↓ indicates the LLM downgraded guideline quality (A→B or B→C). Concordance rate represents the proportion of guidelines (out of 11) where the LLM classification matched the human expert consensus.

References

1. Li, H.Y., Jiang, S.L., Wang, J., Wang, H.S. & Wang, L.H. Quality assessment of clinical practice guidelines for sepsis and variations in recommendations. *BMC Infect Dis* **26**, 158 (2025).

Supplementary Table S5. Z-score standardization and weighted composite scoring for overall ranking of six LLMs

Model	Z_{ICC}	$Z_{\|Bias\|}$	Z_{MAE}	Z_{Tokens}	Z_{Time}	Composite Score	Overall Rank
Mimo-v2.5	1.262	0.543	0.607	1.303	0.83	0.912	1
Gemini-3.1-Pro	0.397	1.022	0.662	−1.397	0.478	0.435	2
GPT-5.5	0.613	−0.480	1.042	0.012	0.478	0.352	3
Qwen-3.7	−0.036	0.835	−0.100	0.502	−0.780	0.148	4
DeepSeek-V4	−0.685	−0.292	−0.535	0.502	0.669	−0.303	5
Doubao-2.0	−1.550	−1.627	−1.677	−0.922	−1.673	−1.544	6

Z-scores were computed using sample standard deviations (n = 6). For each metric, Z-scores were oriented so that higher values indicate better performance (i.e., reverse-transformed for MAE, |Mean bias|, tokens, and thinking time). The missing mean thinking time for GPT-5.5 was imputed using the median of the observed values from the other five models (31.2 s). Composite scores were calculated as weighted sums: ICC(A,1) (35%), |Mean bias| (25%), MAE (20%), mean tokens (10%), and mean thinking time (10%). Higher composite scores indicate better overall performance.

Comprehensive Score Formula: $S = 0.35 \times Z_{ICC} + 0.25 \times Z_{|bias|} + 0.20 \times Z_{MAE} + 0.10 \times Z_{tokens} + 0.10 \times Z_{time}$

Supplementary Table S6. Sensitivity analysis of model ranking under alternative weight schemes

Weight scheme	ICC	|Bias|	MAE	Tokens	Time	Ranking (1st → 6th)
Primary (proposed)	**0.35**	**0.25**	**0.20**	**0.10**	**0.10**	**Mimo > Gemini > GPT > Qwen > DeepSeek > Doubao**
Accuracy-only	0.40	0.30	0.30	0.00	0.00	Mimo > Gemini > GPT > Qwen > DeepSeek > Doubao
Efficiency-heavy	0.10	0.10	0.10	0.35	0.35	Mimo > GPT > DeepSeek > Qwen > Gemini > Doubao
Equal weights	0.20	0.20	0.20	0.20	0.20	Mimo > GPT > Gemini > Qwen > DeepSeek > Doubao
ICC-focused	0.50	0.20	0.20	0.05	0.05	Mimo > Gemini > GPT > Qwen > DeepSeek > Doubao
Bias-focused	0.25	0.50	0.25	0.00	0.00	Gemini > Mimo > Qwen > GPT > DeepSeek > Doubao
ICC-only	1.00	0.00	0.00	0.00	0.00	Mimo > GPT > Gemini > Qwen > DeepSeek > Doubao
Bias-only	0.00	1.00	0.00	0.00	0.00	Gemini > Qwen > Mimo > DeepSeek > GPT > Doubao

Note: Weights were applied to the five Z-standardized metrics: ICC(A,1), absolute mean bias (|Bias|), mean absolute error (MAE), mean token consumption (Tokens), and mean thinking time (Time). The "Primary (proposed)" scheme reflects the predetermined 4:1 ratio prioritizing accuracy-related metrics (80%) over computational efficiency (20%). Alternative schemes were designed to test the robustness of the ranking under different prioritizations: Accuracy-only (excluding efficiency metrics), Efficiency-heavy (prioritizing computational cost), Equal weights (uniform 0.20 distribution), ICC-focused and Bias-focused (emphasizing specific accuracy dimensions), and single-metric extremes (ICC-only and Bias-only). Models were ranked from highest to lowest based on the resulting weighted composite score.

Supplementary Text S1: Two-stage standardized AGREE II appraisal prompt templates

Stage 1: role and general output instructions

You are now a top-tier clinical evidence-based medicine expert and guideline methodologist, with over a decade of extensive experience using the AGREE II (Appraisal of Guidelines for Research & Evaluation II) instrument to evaluate clinical practice guidelines. I have already attached the AGREE II tool for your reference.

Next, I will provide you with the full text (or an excerpt) of a clinical practice guideline. Your task is to evaluate it strictly according to the 6 domains and 23 items of the AGREE II instrument.

Scoring Rules:

The score for each item ranges from 1 to 7.

1 (Strongly Disagree): The guideline provides absolutely no relevant information for this item, or the information provided is extremely vague.

7 (Strongly Agree): The guideline comprehensively and excellently meets all requirements of this item, with clear and extremely detailed information.

2 to 6: Intermediate scores based on the completeness and quality of the information provided in the guideline.

Output Format Requirements:

When I subsequently ask you to evaluate each domain, you must strictly follow the format below without omitting anything:

[Item X Analysis]: Briefly describe the evidence found in the text in 2-3 sentences (key original phrases may be quoted) or point out any missing content.

[Item X Score]: X points

If you understand your identity, task, and output format, please reply: "I am ready! Please provide the guideline text and the domain(s) to be evaluated, and I will strictly follow the AGREE II criteria to conduct a professional scoring."

Stage 2: guideline upload and item-level appraisal request

Send evaluation instructions by domain (for a single guideline PDF, ensuring precision)

(After uploading the guideline file, please send the following prompt)

Item 1. The overall objective(s) of the guideline is (are) specifically described.
Item 2. The health question(s) covered by the guideline is (are) specifically described.
Item 3. The population (patients, public, etc.) to whom the guideline is meant to apply is specifically described.
Item 4. The guideline development group includes individuals from all relevant professional groups.
Item 5. The views and preferences of the target population (patients, public, etc.) have been sought.
Item 6. The target users of the guideline are clearly defined.
Item 7. Systematic methods were used to search for evidence.
Item 8. The criteria for selecting the evidence are clearly described.
Item 9. The strengths and limitations of the body of evidence are clearly described.
Item 10. The methods for formulating the recommendations are clearly described.
Item 11. The health benefits, side effects, and risks have been considered in formulating the recommendations.
Item 12. There is an explicit link between the recommendations and the supporting evidence.
Item 13. The guideline has been externally reviewed by experts prior to its publication.
Item 14. A procedure for updating the guideline is provided.
Item 15. The recommendations are specific and unambiguous.
Item 16. The different options for management of the condition or health issue are clearly presented.
Item 17. Key recommendations are easily identifiable.
Item 18. The guideline describes facilitators and barriers to its application.
Item 19. The guideline provides advice and/or tools on how the recommendations can be put into practice.
Item 20. The potential resource implications of applying the recommendations have been considered.
Item 21. The guideline presents monitoring and/or auditing criteria.
Item 22. The views of the funding body have not influenced the content of the guideline.
Item 23. Competing interests of guideline development group members have been recorded and addressed.

Please provide a detailed [Item X Analysis] and the final [Item X Score] (on a scale of 1-7).
"For each AGREE II item, in addition to providing a score (1-7), you must extract 1-2 core original quotes from the guideline text I provide as the basis for your

score. If the guideline does not mention it at all, please explicitly state 'No relevant original text found.'"