

Large Language Models in Emergency Medicine: A Critical Appraisal of Validity, Reproducibility, and Clinical Utility (2020-2025)

Ahmet Aykut, Cem Yıldırım, Ertuğ Günsoy

University of Health Sciences Türkiye, Van Education and Research Hospital, Clinic of Emergency Medicine, Van, Türkiye

Abstract

Recent studies on large language models (LLMs) in emergency medicine (EM) have expanded rapidly, yet core threats to validity and reproducibility remain under-addressed. We critically synthesized the methods, reporting quality, and clinical relevance of LLM-focused work in emergency care published between January 2020 and April 2025. We conducted a PubMed search and verified journal indexing in the Web of Science (WoS) to restrict screening to EM-relevant studies published in journals indexed under the WoS 'EM' category, excluding editorials that lacked primary or secondary analysis. Two reviewers independently coded protocol availability; prompt transparency; data realism; reference standards; calibration and decision-curve reporting; external validation; and expert benchmarking, resolving discrepancies by consensus. Ninety-one studies met the inclusion criteria; sixty were original investigations. Prompt disclosure was complete in roughly one-third of studies, and real-world clinical data were used less often than synthetic or examination-style vignettes. Calibration, decision-curve analysis, and demonstrations of incremental value over parsimonious clinical baselines were infrequently reported. Expert benchmarking appeared inconsistently across journal strata, and "near-expert" claims often relied on proxy tasks with limited ecological validity. External validation was uncommon, and model/version identifiers were frequently incomplete, undermining reproducibility. Overall, the current LLM literature within this core EM journal corpus is method-lean and report-light: high-level accuracy claims rarely translate into decision-useful evidence. A minimum reporting set—transparent prompts, code, and versioning; calibration; decision-curve analysis; and expert benchmarking on real data—is needed; absent these elements, deployment in time-critical emergency care remains premature.

Keywords: Large language models, emergency informatics, decision support, methodological evaluation, artificial intelligence

Introduction

Large language models (LLMs), such as OpenAI's GPT and Google's Gemini, have rapidly entered clinical discourse not through careful, protocolized integration but largely propelled by enthusiasm and marketing narratives about their capacity to parse unstructured data, aid diagnostics, and streamline documentation across specialties, including emergency medicine (EM) (1,2). Yet the pace of adoption has outstripped the maturation of evaluation science, creating a widening gap between performance claims and decision-useful, reproducible evidence (3).

Unlike traditional clinical tools, LLMs are non-stationary systems. Their behavior evolves due to silent backend updates,

architectural shifts, and access-path changes; the same prompt can yield materially different outputs over time—a form of version drift that undermines reproducibility and weakens causal attribution of observed effects (4). Compounding this instability is prompt sensitivity: minor phrasing changes can produce large, unpredictable swings in outputs, challenging any assumption of determinism or reliability in time-critical EM settings (5).

Despite these well-described failure modes, published LLM evaluations in medicine, including those appearing in core EM journals, frequently omit core methodological details. Model/version identifiers, release dates, full prompt templates, and data provenance are inconsistently reported; synthetic prompts, exam-



Corresponding Author: Ahmet Aykut MD, University of Health Sciences Türkiye, Van Education and Research Hospital, Clinic of Emergency Medicine, Van, Türkiye
E-mail: ahmet.aykut@gmail.com **ORCID ID:** orcid.org/0009-0001-3173-8994

Cite this article as: Aykut A, Yıldırım C, Günsoy E. Large language models in emergency medicine: a critical appraisal of validity, reproducibility, and clinical utility (2020-2025). Eurasian J Emerg Med.2026;25: 117-24.

Received: 23.10.2025
Accepted: 20.12.2025
Published: 26.01.2026



©Copyright 2026 The Author(s). The Emergency Physicians Association of Turkey / Eurasian Journal of Emergency Medicine published by Galenos Publishing House.
Licensed by Creative Commons Attribution-NonCommercial-NoDerivatives (CC BY-NC-ND) 4.0 International License

style questions, or proxy tasks often substitute for authentic clinical inputs and adjudicated reference standards (3,6). Consequently, reported accuracies are rarely accompanied by calibration, decision-curve analysis (net benefit), or incremental value over parsimonious clinical baselines—elements required to judge whether a tool improves triage, resource allocation, or patient-relevant outcomes under real EM constraints (7).

This review systematically evaluates peer-reviewed LLM studies pertinent to EM, as represented by studies published in journals indexed under the Web of Science (WoS) “emergency medicine” category (SCI-E/ESCI), focusing on methodological transparency, empirical rigor, and clinical grounding. Using a structured framework centered on input fidelity, evaluation strategy, and functional use case, we (1) characterize the evidentiary landscape within this core EM journal corpus, (2) identify recurrent threats to validity and reproducibility, and (3) outline priorities for future research that make EM deployment transparent, calibrated, externally validated, and resilient to technological volatility (3,6,7).

Review Scope and Approach

Scope and Conceptual Approach

We conducted a retrospective, descriptive review of the peer-reviewed literature to evaluate how LLMs [including ChatGPT/Generative Pre-trained Transformer (GPT)-4 and Gemini] have been applied, evaluated, and reported methodologically in EM. The review emphasized three core dimensions aligned with EM workflows: (1) transparency in model use (versioning, access modality, prompt disclosure), (2) empirical rigor (real-world data, expert benchmarking, calibration, and decision-curve analysis), and (3) clinical applicability (task-workflow fit). A structured, three-axis framework—input fidelity, evaluation strategy, and functional use case—guided classification and subsequent comparisons.

Literature Identification and Study Selection

Information Sources and Search Strategy

A structured PubMed search was performed in May 2025 directly via the PubMed web search interface using the Boolean string below to capture EM-relevant LLM studies between 01/01/2020 and 04/30/2025: “ChatGPT” or “(GPT)-3” or “GPT-3.5” or “GPT-4” or “GPT-4.5” or “large language model” or “LLM” or “generative AI” or “generative artificial intelligence” or “transformer-based model” or “foundation model” or “instruction-tuned model” or “GPT” and [2020.01.01 (date-publication): “2025.04.30” (date-publication)]. The search returned 12.125 records prior to de-duplication and journal-scope filtering.

Journal Corpus Definition (WoS Emergency Medicine Category)

To define a core corpus of EM journals and to ensure comparability of scope across included venues, we retained only articles published in journals indexed in WoS under SCI-E or ESCI. WoS was used solely to verify the SCI-E/ESCI indexing status of journals; no literature search was conducted within WoS. A curated list of 57 EM-category journal titles was used for this index-verification filter.

Screening and Eligibility Assessment

Two independent reviewers screened titles and abstracts, and subsequently full texts, against pre-specified criteria. Reasons for exclusion included: (1) passing mention of LLMs without evaluative content, (2) general AI discussion lacking LLM-specific evaluation, and (3) abstract-only or unavailable full text. Disagreements were resolved by consensus after discussion. Following screening, 91 studies were retained for full-text analysis. Operational definitions, decision rules, and study-level coding outputs are provided in Supplementary Table S1.

Data Abstraction and Methodological Appraisal

All 91 studies were assessed using a structured codebook covering five methodological indicators: model identification (model name, release/version, and access pathway), prompt disclosure (full, partial, absent, or ambiguous), use of real clinical data (authentic EM inputs vs. synthetic/proxy tasks), human expert benchmarking (comparison with domain experts where clinical judgment is implicated), and methodological consistency (alignment between stated aims and data/evaluation choices). Articles were independently coded by two reviewers; disagreements were adjudicated by consensus. Because our primary aim was appraisal of transparency and reporting rather than effect estimation, we did not perform a formal risk-of-bias assessment. The full codebook (variable definitions and decision rules) and the per-study coding matrix are provided in Supplementary Table S1.

High-Rigor Subset (Predefined Methodological Benchmark)

Among the 60 studies classified as original research, we applied a secondary filter to identify a high-rigor subset that met all of the following criteria: (1) real-world clinical data, (2) expert-based benchmarking, and (3) no reliance on standardized multiple-choice exams as surrogate evaluations (to mitigate data leakage and poor ecological validity). Seventeen studies met these criteria.

Three-Axis Classification Framework

To map methodological diversity across original investigations, we applied a three-axis schema: input fidelity (real-world EM

data vs. simulated/synthetic content), evaluation strategy (e.g., expert benchmarking, external validation, calibration/decision-curve reporting vs. internal metrics alone), and functional use case (clinical decision support, documentation/communication, education/training, or other EM-relevant tasks). Operational definitions and coding rules appear in Supplementary Table S1. For visualization, the evaluation strategy and the use case were collapsed into broader themes; these groupings were visualized in the alluvial (Sankey) plot (Figure 1).

Positioning Relative to Reporting Guidance

We compared our classification outputs with healthcare Minimum reporting items for Clear Evaluation of Accuracy Reports (MI-CLEAR)-LLM reporting frameworks MI-CLEAR-LLM and Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD)-LLM. While these guidelines prioritize documentation of model metadata and evaluation design, our schema extends them by explicitly encoding functional role and data realism, enabling granular, metric-specific appraisal. A side-by-side summary is presented in Supplementary Table S2.

Journal-Tier Summaries

Original research articles were grouped according to indexing in SCI-E or ESCI. For each stratum, we calculated the proportions

reporting full prompt disclosure, use of real clinical data, and human expert benchmarking. Results are summarized in Table 1 with risk differences (SCI-E-ESCI) and 95% confidence intervals (CI) (Newcombe method). Given the small denominators, particularly in ESCI, findings were interpreted cautiously and treated as descriptive. Because several tier-stratified 2x2 tables involved small expected cell counts (<5), large-sample z- or χ^2 -approximations were not used for inference.

Exploratory Cross-tabulations

Exploratory Fisher’s exact tests were prespecified for the interaction between journal tier and reporting items (full prompt, real clinical data, expert benchmarking), and for the interaction between prompt disclosure (full vs. not-full) and high-rigor status. Two-sided Fisher’s exact tests were used, and the corresponding 2x2 tables and p-values are provided in Table 2. Fisher’s exact test was selected because it remains valid when expected cell counts are small, unlike χ^2 - and z-based procedures that rely on large-sample assumptions. Other planned tests (e.g., real data x expert benchmarking, use case x prompt) required joint counts not available from the final abstraction set and were not performed.

Alluvial (Sankey) map of the three-axis classification (n=60 original studies)

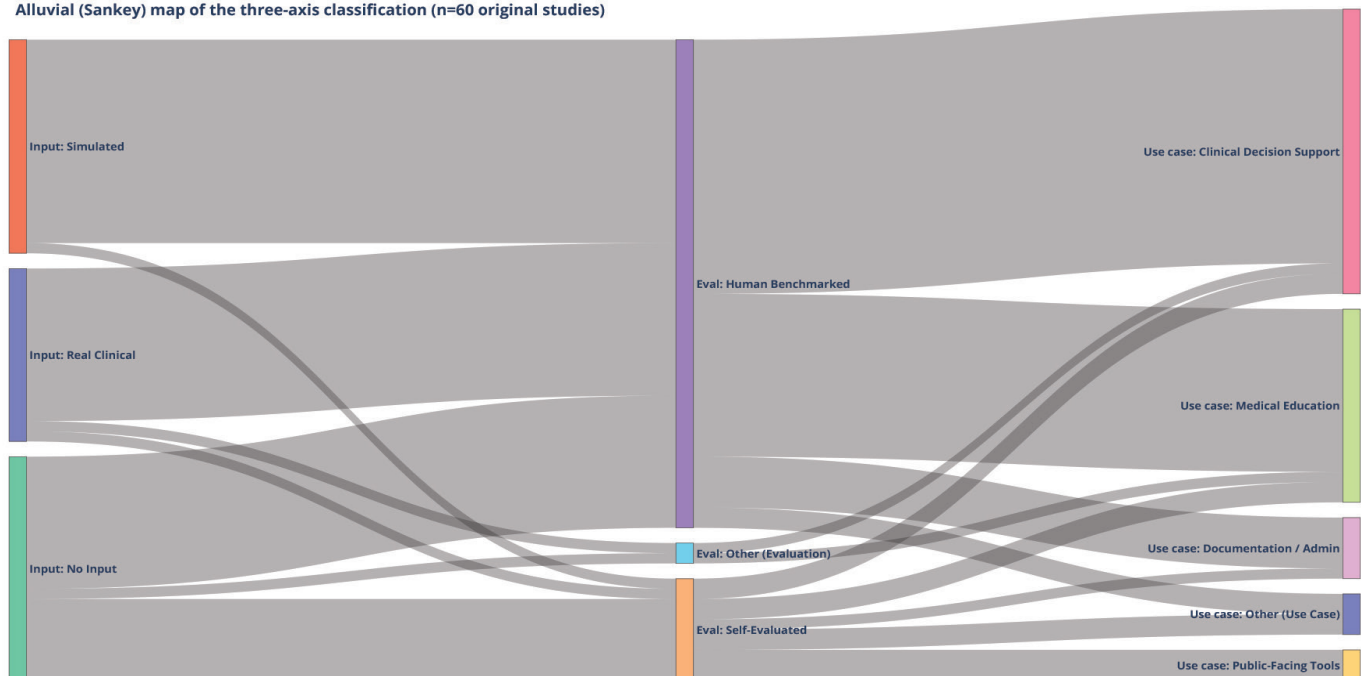


Figure 1. Alluvial (Sankey) map of the three-axis classification of original investigations (n=60)

Nodes represent input fidelity (real clinical / simulated/no input), evaluation strategy (human benchmarked/self-evaluated/other), and functional use cases (clinical decision support/medical education/documentation/admin/public-facing tools/other). Link widths are proportional to the number of studies in each pathway

Evidence Synthesis

Publication Trends and Study Types

Ninety-one peer-reviewed articles met the inclusion criteria, of which sixty were classified as original research. Publications were concentrated between 2023 and 2025, peaking in 2024. Across all included articles, most were published in SCI-E-indexed journals (SCI-E: 80/91; ESCI: 11/91). Among original research articles, 53 appeared in SCI-E-indexed journals and 7 in ESCI-indexed journals. The distribution of publication types by year and journal tier is summarized in Table 3.

Methodological Rigor Subset

Original research articles were evaluated against three a priori criteria: use of real-world clinical data, inclusion of human expert benchmarking, and avoidance of standardized multiple-choice examinations as surrogate tasks. Seventeen studies satisfied all three criteria and were designated as methodologically robust. The remaining forty-three studies did not meet these criteria due to methodological incompleteness or ambiguity (n=32), reliance on simulated data and/or the absence of expert comparison (n=6), or use of examination-style proxies (n=5). Non-robust studies were assigned a primary reason using a mutually exclusive hierarchy, as reflected in the study-level coding (Supplementary Table S1). Counts and definitions are presented in Table 3, with study-level coding in Supplementary Table S1.

Prompt Transparency and Implications for Reproducibility

Prompt reporting was inconsistent across the original research corpus. Only 20 studies (33.3%) disclosed complete prompt text, 36 (60.0%) provided partial or paraphrased examples, 3 (5.0%) did not report prompts, and 1 (1.7%) used ambiguous formatting. Even among the methodologically robust subset, full prompt disclosure was observed in six of the seventeen studies. This limited transparency constrains reproducibility for tasks in which model behavior is highly sensitive to input design, including diagnostic reasoning, triage, and clinical documentation. Tier-stratified prompt disclosure is summarized in Table 1, and the exploratory 2x2 comparison of full prompt disclosure versus high-rigor status is reported in Table 2; study-level prompt coding is provided in Supplementary Table S1.

Journal-tier Summary

Descriptive comparisons between SCI-E and ESCI strata are shown in Table 1. Full prompt disclosure was reported in 18 of 53 SCI-E studies (34.0%) and in 2 of 7 ESCI studies (28.6%), yielding a risk difference of +5.4 percentage points (95% CI: -41.4 to +39.2; Newcombe method). The use of real clinical data was nearly identical across tiers—15/53 (28.3%) in SCI-E and 2/7 (28.6%) in ESCI—corresponding to a risk difference of -0.3 percentage points (95% CI: -46.1 to +33.3 percentage points). Human-expert benchmarking was reported in 17 of 53 SCI-E studies (32.1%) and 3 of 7 ESCI studies (42.9%), yielding a risk difference of -10.8

| Metric | SCIE (x/n) | SCIE (%) | ESCI (x/n) | ESCI (%) | Risk difference (SCIE-ESCI), % | 95% CI (% , Newcombe) |
|---------------------------|------------|----------|------------|----------|--------------------------------|-----------------------|
| Full prompt disclosure | 18/53 | 34 | 2/7 | 28.6 | 5.4 | (-41.4, 39.2) |
| Use of real clinical data | 15/53 | 28.3 | 2/7 | 28.6 | -0.3 | (-46.1, 33.3) |
| Human expert benchmarking | 17/53 | 32.1 | 3/7 | 42.9 | -10.8 | (-53.9, 29.7) |

Values are presented as x/n (%) for SCIE and ESCI original research; the risk difference (SCIE-ESCI) and 95% CI (Newcombe method) are also reported. Positive RD favors SCIE; negative RD favors ESCI. Interpret with caution due to the small ESCI denominator. SCIE: Science Citation Index-Expanded, ESCI: Emerging Sources Citation index, CI: Confidence interval

| Comparison | Group 1 (row 1) | a | b | Group 2 (row 2) | c | d | Odds ratio (Fisher’s exact tests) | p (two-sided) |
|--|-----------------------------------|----|----|----------------------------------|----|----|-----------------------------------|---------------|
| Journal tier × Full prompt disclosure | SCIE (full, not full) | 18 | 35 | ESCI (full, not full) | 2 | 5 | 1.286 | 1.000 |
| Journal tier × Real clinical data | SCIE (real, not real) | 15 | 38 | ESCI (real, not real) | 2 | 5 | 0.987 | 1.000 |
| Journal tier × Human expert benchmarking | SCIE (Yes, No) | 17 | 36 | ESCI (Yes, No) | 3 | 4 | 0.630 | 0.676 |
| Prompt (full vs not-full) × High-rigor (Yes, No) | High-rigor = Yes (full, not full) | 6 | 14 | High-rigor = No (full, not full) | 11 | 29 | 1.130 | 1.000 |

Cells follow the 2x2 convention: [(a, b), (c, d)]. For journal-tier comparisons, row 1 =SCIE and row 2 =ESCI; columns indicate feature presence and absence. For the prompt × high-rigor comparison, row 1 = High-rigor= Yes, row 2 = High-rigor= No; columns are (full, not full). Two-sided Fisher’s exact tests p-values are reported, SCIE: Science Citation Index-Expanded, ESCI: Emerging Sources Citation index

percentage points (95% CI: -53.9 to +29.7 percentage points). Overall, CIs were wide and included the null, limiting inference about tier-level differences and highlighting the imprecision resulting from the small ESCI denominator. Accordingly, tier comparisons are interpreted descriptively, and inferential testing relies on Fisher’s exact test given the small expected cell counts.

Exploratory Cross-tabulations

Pre-specified exploratory 2×2 tests are reported in Table 2. Two-sided Fisher’s exact tests did not identify statistically significant associations between journal tier and full prompt disclosure (odds ratio: 1.286, p=1.000), real clinical data (odds ratio: 0.987, p=1.000), or human expert benchmarking (odds ratio: 0.630, p=0.676). Likewise, the association between full prompt disclosure and inclusion in the high-rigor subset was not significant (odds ratio: 1.130, p=1.000). Planned analyses for real data × expert benchmarking, and use case × prompt transparency could not be conducted because joint counts were not available in the final abstraction; these omissions are noted in Methods 2.8. All contingency tables and p-values are presented in Table 2.

Methodological Structure Across Studies

Using the three-axis framework (input fidelity, evaluation strategy, functional use case), we characterized methodological patterns across the sixty original research articles. Real clinical data were used in 28.3% of studies (17/60), and human expert benchmarking was reported in 33.3% (20/60). Clinical decision support was the most frequent use case, followed by educational and administrative applications (Supplementary Table S1).

Overall, the corpus exhibits rapid growth alongside substantial heterogeneity in reporting and design, with a small but identifiable subset meeting more stringent criteria for empirical and clinical rigor. The distribution of studies across the three axes and their interrelationships are visualized in the alluvial (Sankey) plot (Figure 1).

Discussion

Fragile Foundations Amid Rapid Growth

The present synthesis reveals a corpus expanding faster than its evidentiary substrate. Following the release of contemporary LLMs (e.g., GPT-4), publication volume rose sharply; however, reporting quality and clinical grounding have not kept pace within the core EM journal set examined. Much of the literature remains anchored in capability demonstrations rather than decision-useful evidence, a gap reflected in incomplete description of model provenance, insufficient transparency around inputs, and limited assessment of clinical impact. These findings align with prior critiques of methodological shortfalls and checklist non-adherence in medical LLM research. (1-3,6,7).

Version Instability and Prompt Sensitivity as Threats to Reproducibility

LLMs are non-stationary: silent model updates, undocumented architectural changes, and evolving access pathways introduce version drift, whereby identical prompts can yield divergent outputs across time (4,5). Together with prompt sensitivity, defined as large output variability induced by minor input changes, this dynamic undermines reproducibility and complicates

Table 3. Tier totals

| Journal tier | Commentary (n) | Original research (n) | Review (n) | Total (n) |
|--------------|----------------|-----------------------|------------|-----------|
| ESCI | 1 | 7 | 3 | 11 |
| SCIE | 27 | 53 | 0 | 80 |
| All | 28 | 60 | 3 | 91 |

Counts are numbers of articles by publication year and Web of Science Emergency Medicine journal tier. “Original research” denotes empiric investigations; “commentary” includes editorials, viewpoints, and letters; “review” comprises narrative or systematic reviews. SCIE: Science Citation-Index Expanded, ESCI: Emerging Sources Citation index. Row totals equal the sum of commentary, original research, and review for each year-tier combination; tier totals and the grand total (SCIE: 80, ESCI: 11, overall n=91) are reported beneath the main table. Figures reflect the curated Emergency Medicine journal set used for study selection and may not generalize beyond this corpus

Table 3. Distribution of studies by year, journal tier, and publication type (2023-2025)

| Year | Journal tier | Commentary (n) | Original research (n) | Review (n) | Total (n) |
|------|--------------|----------------|-----------------------|------------|-----------|
| 2023 | ESCI | 1 | 2 | 1 | 4 |
| 2023 | SCIE | 13 | 10 | 0 | 23 |
| 2024 | ESCI | 0 | 4 | 2 | 6 |
| 2024 | SCIE | 10 | 27 | 0 | 37 |
| 2025 | ESCI | 0 | 1 | 0 | 1 |
| 2025 | SCIE | 4 | 16 | 0 | 20 |

SCIE: Science Citation Index-Expanded, ESCI: Emerging Sources Citation index

longitudinal interpretation. Reproducibility challenges are not unique to LLMs; variability across runs and pipelines has also been documented in other medical AI domains such as deep learning-based image segmentation (8). Studies that omit model/version identifiers, release timing, or access modality produce findings that are difficult to replicate or reconcile with subsequent evaluations in this corpus (3-5).

Transparency Deficits and Their Implications

Prompt disclosure was inconsistent in the included EM-journal corpus: complete prompts were reported in only one-third of original studies and even less frequently in the methodologically robust subset. Absence or partial reporting of prompts constrains replication, especially for tasks in which outputs are highly sensitive to input design (diagnostic reasoning, triage, documentation). These observations are consistent with emerging guidance that positions transparent prompts, versioning, and code availability as prerequisites for credible evaluation (3,6,7).

Methodological Weaknesses in Study Design

A recurrent limitation was the reliance on standardized multiple-choice or examination-style instruments as proxy measures. Such tasks are vulnerable to training-data contamination and fail to capture EMs complexity, thereby risking optimistic but clinically uninformative estimates (6). By contrast, the high-rigor subset combined real clinical inputs with expert benchmarking and avoided using exam proxies, illustrating that empirically grounded LLM evaluation in EM is feasible when the study design is aligned with clinical workflow, as demonstrated in studies published in core EM journals.

Journal-tier Comparisons Interpreted with Caution

Tier-stratified summaries suggested a slightly higher prompt disclosure in SCI-E, no meaningful difference in the use of real clinical data, and no consistent advantage in human expert benchmarking (the ESCI proportion was numerically higher); CIs were wide across all metrics. However, these tier-stratified comparisons are constrained by the small ESCI subgroup (n=7) relative to SCI-E (n=53), which limits precision and precludes robust inference. These patterns argue against tier-level generalizations and should be interpreted as descriptive patterns within the WoS “EM” journal category rather than as field-wide differences. Accordingly, tier patterns are not used to support the study’s primary conclusions.

Beyond Checklists: A Structural Lens on Study Quality

Reporting frameworks such as MI-CLEAR-LLM and TRIPOD-LLM have improved transparency, but cannot, by themselves, remedy deeper design deficits (3,7). Our three-axis schema—input fidelity, evaluation strategy, and functional use case—

provides a complementary structural framework, highlighting the persistence of synthetic tasks, the infrequency of calibration and decision-curve analyses, and the variability of expert benchmarking, even among studies presented as rigorous within the examined EM-journal corpus. The framework’s relationship to existing guidance is detailed in Supplementary Table S2.

Structural Barriers to External Validation

Opaque model internals, undisclosed training data, and proprietary constraints impede error analysis and durable external validation (9-11). Absent auditable versioning and accessible artifacts (prompts/code), even well-designed evaluations risk becoming brittle as platforms evolve. A version-aware, documentation-first approach is therefore integral to any clinically credible assessment strategy for EM-facing use cases, including those evaluated in core EM journals.

Toward a More Reliable Evaluation Paradigm

The field should pivot from leaderboard-style claims of accuracy to process-oriented evaluation that prioritizes version tracking, prompt transparency, calibration, decision-curve analysis, external validation, and expert benchmarking using authentic EM data (1,3,6,7,9). Studies exemplifying these elements already exist, demonstrating the feasibility and value of methodologically disciplined LLM research in EM, as reflected by a subset of rigorously designed studies identified in this journal corpus (12,13). Consolidating such practices will be essential to move from exploratory promise to defensible deployment.

Methodological Considerations

This review was restricted to journals indexed in the WoS “EM” category (SCI-E and ESCI); therefore, our findings should be interpreted as describing the LLM evidence base within this core EM journal corpus, rather than the entire EM literature. Relevant EM-facing LLM studies published in general medicine, critical care, resuscitation, trauma, prehospital/EMS, radiology, and medical informatics venues may have been overlooked. This venue restriction may introduce publication-location (venue) bias, potentially affecting the observed distribution of use cases, evaluation designs, and reporting practices. For example, clinically grounded or implementation-focused studies may be more likely to appear outside EM-category journals, whereas capability demonstrations may cluster differently across venues.

Study-level coding was performed systematically, yet some classifications, particularly along the evaluation-strategy and use-case axes, required interpretive judgment due to inconsistent reporting in source articles. To enable visual synthesis, we collapsed several categories in the alluvial mapping, a simplification that may obscure finer methodological distinctions. We did not

perform a sensitivity analysis using an expanded journal set beyond the WoS EM category; future work should assess the robustness of these patterns by broadening sampling frames (e.g., predefined acute-care-relevant journal lists and/or topic-based retrieval with EM-relevance adjudication). Tier-stratified summaries should also be interpreted cautiously because the ESCI subgroup was small ($n=7$) relative to SCI-E ($n=53$), limiting precision and precluding robust inference about tier differences. In addition, some tier-stratified 2×2 tables had expected cell counts below conventional thresholds (e.g., <5), which preclude reliable χ^2 - or z-based inference and z-score calculations. We therefore relied on Fisher's exact tests; this reliance contributes to wide CIs and low power to detect differences between tiers. Consequently, these tier-stratified estimates have limited clinical usability and generalizability, and should not be used to draw tier-level inferences. Finally, we did not undertake a longitudinal analysis of temporal drift in model behavior or prompt practices; given ongoing platform evolution, future meta-research should explicitly incorporate time-stamped versioning and repeated-measures designs.

Recommendations and Conclusion

Report the Model-Precisely and Time-Stamped

Responsible evaluation begins with unambiguous model metadata. Every study should document the model's canonical name, release or build identifier, access pathway (e.g., API versus web interface), and the exact date and time of use. Absent these anchors, results are not reproducible in a non-stationary ecosystem and cannot be meaningfully compared across replications or versions. Contemporary guidance (MI-CLEAR-LLM, TRIPOD-LLM) already positions version transparency as a first-order requirement; clinical LLM research must normalize it as standard practice (3,7).

Treat Prompts as Experimental Conditions

Prompts and associated parameters shape outputs as deterministically as any intervention in a clinical experiment. Studies should provide complete prompt text (including system instructions), retry logic, and sampling parameters. Summaries or partial exemplars are insufficient for replication and hinder meta-research on prompt sensitivity. This expectation is both methodologically necessary and aligned with emerging recommendations on reproducible LLM evaluation (3,5,7).

Replace Exam Proxies with Clinical Reality

Standardized multiple-choice items are convenient but of low ecological validity and vulnerable to training-data contamination. Evaluations should be grounded in authentic clinical artifacts — electronic health records, clinician notes, and imaging reports

— and should mirror real emergency workflows and constraints. Synthetic benchmarks frequently overstate performance and obscure operational failure modes; they should be used only for preliminary exploration and not presented as evidence of clinical readiness (3,6,12).

Benchmark Against Human Experts

For high-stakes tasks (diagnosis, triage, disposition), human expert comparators and adjudicated reference standards are indispensable. Model-to-model comparisons are insufficient for clinical inference and pose a risk of circular validation. The clinical literature and reporting guidance converge on this point: expert benchmarking is a prerequisite for decision-useful claims (3,7,12,13).

Shift the Emphasis from Point Accuracy to Process Understanding

Leaderboard accuracy is an inadequate proxy for safety or usefulness. Future studies should interrogate reasoning pathways and failure modes, incorporating calibration, decision-curve analysis, error taxonomies, and qualitative audits. Robustness, not isolated point estimates, must become the currency of clinical evaluation (9,10).

Anticipate—and Measure—Drift

Because LLMs evolve over time, validation should include repeated assessments across versions and time points, with explicit cross-version comparisons and time-stamped artifacts. Version drift is no longer hypothetical; it is a documented source of variance that threatens longitudinal interpretability if not prospectively measured (4,7,9,10).

Enforce Minimum Reporting at Publication

Editorial policies should require, as a condition of acceptance, complete model metadata, full prompt disclosure, and clinically grounded evaluation design. Manuscripts lacking these elements are methodologically incomplete regardless of novelty or apparent performance. Reproducibility and transparency —not speed to publication— should define publication worthiness (1,3,7,9).

Conclusion

LLMs hold clear promise for EM, particularly in decision support, documentation, and clinical education. Realizing that promise, however, demands a decisive shift in scientific practice: version-aware reporting, full prompt transparency, expert-anchored evaluation on real clinical data, and routine use of calibration and decision-curve analysis. Within the body of LLM studies published in core EM journals (WoS "EM" category), our synthesis

shows that the current evidence base remains fragmented, with critical gaps in documentation and external validity that limit clinical translation. The path forward is not to celebrate higher scores on synthetic tasks but to establish rigorous, reproducible, and clinically coherent methods. If EM embraces this standard—prioritizing process over performance and accountability over expediency—LLM integration can proceed with the trust and durability required for high-stakes care.

Ethics

Footnotes

Authorship Contributions

Concept: A.A., Design: A.A., Data Collection or Processing: A.A., C.Y., E.G., Analysis or Interpretation: A.A., C.Y., E.G., Literature Search: A.A., Writing: A.A., C.Y., E.G.

Conflict of Interest: No conflict of interest was declared by the authors.

Financial Disclosure: The authors declared that this study received no financial support.

References

- Haghani M. Riding the wave of ChatGPT research: an analysis of early-stage scholarly output and the associated authorship anomalies. SSRN. Available from: <https://www.ssrn.com/abstract=4553479>
- Karabacak M, Margetis K. Embracing large language models for medical applications: opportunities and challenges. *Cureus*. 2023;15:e393305.
- Ko JS, Heo H, Suh CH, Yi J, Shim WH. Adherence of studies on large language models for medical applications published in leading medical journals according to the MI-CLEAR-LLM Checklist. *Korean J Radiol*. 2025;26:304-12.
- Chen L, Zaharia M, Zou J. How is ChatGPT's behavior changing over time? arXiv [Preprint]. Available from: <http://arxiv.org/abs/2307.09009>
- Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. *ACM Comput Surv*. 2023;55:35.
- Shool S, Adimi S, Saboori Amlashi R, Bitaraf E, Golpira R, Tara M. A systematic review of large language model (LLM) evaluations in clinical medicine. *BMC Med Inform Decis Mak*. 2025;25:117.
- Gallifant J, Afshar M, Ameen S, Aphinyanaphongs Y, Chen S, Cacciamani G, et al. The TRIPOD-LLM statement: a targeted guideline for reporting large language models use. medRxiv. 2024;25:2024.
- Renard F, Guedria S, Palma N, Vuillerme N. Variability and reproducibility in deep learning for medical image segmentation. *Sci Rep*. 2020;10:13724.
- Semmelrock H, Kopeinik S, Theiler D, Ross-Hellauer T, Kowald D. Reproducibility in machine learning-driven research. arXiv [Preprint]. Available from: <http://arxiv.org/abs/2307.10320>
- Barberis A, Aerts HJWL, Buffa FM. Robustness and reproducibility for AI learning in biomedical sciences: RENOIR. *Sci Rep*. 2024;14:1933.
- Roustan D, Bastardot F. The clinicians' guide to large language models: a general perspective with a focus on hallucinations. *Interact J Med Res*. 2025;14:e59823.
- Lee KL, Kessler DA, Caglic I, Kuo YH, Shaida N, Barrett T. Assessing the performance of ChatGPT and Bard/Gemini against radiologists for Prostate Imaging-Reporting and Data System classification based on prostate multiparametric MRI text reports. *Br J Radiol*. 2025;98:368-74.
- Sonoda Y, Kurokawa R, Nakamura Y, Kanzawa J, Kurokawa M, Ohizumi Y, et al. Diagnostic performances of GPT-4o, Claude 3 Opus, and Gemini 1.5 Pro in "diagnosis please" cases. *Jpn J Radiol*. 2024;42:1231-5.

Supplementary Table S1 - Table S2: <https://d2v96fxpocvxx.cloudfront.net/580eb5e7-1480-44a6-9404-b8b7446acbc/bde585b6-1102-437f-bfa1-2f70f12a1d2b.pdf>