Reliability Note — Epic Sepsis Model Retrospective Evaluation

> **Instrument:** Clinical AI Vendor Evaluation Tool v1.0.0 — Rubric v1.0 **Scoring model:** DeepSeek V3 (`deepseek/deepseek-chat`) via OpenRouter, US-hosted inference (DeepInfra), temperature 0.2 **Consistency protocol:** mandatory two-run double-check on production evaluations; this note additionally reports three five-run consistency studies **Evaluation date:** 13 June 2026 **Evaluator:** Dr Kpakpo Acquaye, MRCEM

This note accompanies the Epic Sepsis Model retrospective evaluation report and documents the measured reliability of the instrument that produced it. It is published in full because the measurement properties of an evaluation instrument are themselves evidence a buyer is entitled to scrutinise. An evaluation tool that does not disclose its own variance is asking to be trusted on faith.

Summary

The published verdict — composite 37/100, Do Not Proceed, safety interlock triggered by Population Validity and Operational Performance — is stable. Across five independent scoring runs of the final evidence pack, every domain score and the recommendation were reproduced without variation that crosses any band or alters the verdict. This stability was not present in the first version of the evidence pack and was achieved through a documented, disclosed method described below. The path to it is reported here in full, including the points at which the instrument produced unreliable or unsafe output and was corrected.

Why this note exists

Large language models used as scorers are not deterministic. The same evidence, scored twice, can yield different numbers — and when those numbers sit near a band boundary, the difference can change a recommendation. Any clinical-AI evaluation pipeline built on an LLM has this property whether or not it is measured. Most are not measured. The risk is not that the model is "wrong" on average; it is that a single run can land on the wrong side of a decision boundary and be published as if it were settled. The double-check protocol and the consistency studies below exist to detect and contain that failure mode.

Three failures caught before publication

The instrument's safeguards fired three times during preparation of this report. Each is recorded because the safeguards working is the evidence that they work.

1. **Verdict instability (caught by the consistency study, below).** The first evidence pack produced a verdict that flipped on re-running: in one run of five, the safety interlock did not arm and the recommendation would have differed. This is the central finding of this note and is detailed in the next section.

2. **Self-contradicting executive summary (caught on review).** When the safety interlock fired, the recommendation was correctly set to Do Not Proceed, but the model-generated summary prose retained composite-band language and read "proceed with caution" beneath the Do Not Proceed banner. The summary generation was made deterministic when the interlock fires: it now states the interlock verdict and names the failing domains, and does not incorporate model prose that could contradict the verdict.
3. **Unsupported red-flag selection (caught on review).** On two separate runs the scorer selected canonical red flags whose factual premises were false for this vendor — first a marketing-versus-regulatory-filing mismatch (no regulatory filing exists), then an assertion that the vendor affirmatively "claimed" a CDS exemption with non-reviewable logic (the public record shows the basis was unarticulated, not affirmatively claimed). Both were corrected by evidence-pack guardrails. The underlying cause — that the validator confirms a flag is canonical but not that the cited evidence supports it — is logged as a v1.1 instrument change (evidence-cited red flags).

The consistency studies

The evidence pack was scored five times, as five independent single passes, at three stages of development. Scores below are the final per-domain values from each of the five runs.

Study 1 — unanchored evidence pack

Domain	Five runs	Spread
Clinical Evidence Quality	40, 40, 40, 40, 50	10
Population Validity	20, 20, 20, 20, 30	10
Operational Performance	20, 30, 20, 20, 40	20
Workflow Integration	60, 70, 60, 80, 70	20
Regulatory Compliance	40, 40, 40, 40, 40	0
Data Governance	40, 50, 40, 60, 60	20
Implementation Maturity	40, 60, 40, 60, 60	20

Composite range 34.0–48.5 (spread 14.5). Safety interlock fired in four runs of five. In the fifth run, Population Validity scored 30 rather than 20, the interlock did not arm, and that run would have returned a "significant concerns / remediate" posture rather than Do Not Proceed. On a single roll of the instrument, the same vendor on the same evidence could have received a materially different procurement verdict. This is the failure mode the instrument exists to expose in other tools, observed in its own output.

Two observations from this study shaped the fix. First, the only domain with zero variance — Regulatory Compliance — was the only domain that carried an explicit evaluator assessment note anchoring its evidence to a band characterisation. Second, the unstable domains were the mixed-signal ones, where evidence combined a defensible element (favourable

architecture, demonstrated integration) with public-record silence, and the scorer had to resolve that tension into a single number without guidance.

Study 2 — Population Validity and Operational Performance anchored

Evaluator assessment notes were added to the two interlock domains, characterising the evidence against the relevant band — for example, that a complete absence of population-validity evidence at deployment scale is an insufficient-evidence posture, not a limited one. The notes anchor to cited facts, not to target scores, and print verbatim in the report's evidence appendix.

Domain	Five runs
Clinical Evidence Quality	40, 40, 40, 40, 40
Population Validity	20, 20, 20, 20, 20
Operational Performance	20, 20, 20, 20, 20
Workflow Integration	60, 60, 61, 60, 61
Regulatory Compliance	40, 40, 41, 40, 41
Data Governance	40, 40, 41, 40, 41
Implementation Maturity	40, 40, 41, 40, 41

Composite range 35.0–38.2 (spread 3.1). Safety interlock fired in all five runs. The verdict flip was eliminated. Both interlock domains were now perfectly stable, and the residual variance was confined to several domains moving by a single point.

Study 3 — remaining mixed-signal domains anchored

Assessment notes were added to the three domains still showing single-point drift (Workflow Integration, Implementation Maturity, and the regulatory red-flag wording), each anchoring the demonstrated-but-unevidenced tension to its band.

Domain	Five runs	Spread
Clinical Evidence Quality	40, 40, 40, 40, 40	0
Population Validity	20, 20, 20, 20, 20	0
Operational Performance	20, 20, 20, 20, 20	0
Workflow Integration	60, 60, 60, 60, 60	0
Regulatory Compliance	40, 40, 40, 40, 40	0
Data Governance	40, 40, 40, 40, 40	0
Implementation Maturity	40, 40, 40, 40, 40	0

Composite range 35.0–36.5 (spread 1.5). Safety interlock fired in all five runs. Six of seven domains showed zero variance; the residual 1.5-point composite movement is sub-band rounding noise that never approaches a boundary or affects the verdict.

The arc

	Composite spread	Interlock fired	Verdict
Unanchored	14.5	4 / 5	flipped once
Two domains anchored	3.1	5 / 5	stable
Mixed-signal domains anchored	1.5	5 / 5	stable

The method: band anchoring

Band anchoring is a disclosed evaluator judgement, embedded in the evidence and printed in the report, that characterises which side of a band boundary the evidence sits on. It does not state a score. It states, for example, that an absence of evidence is an insufficient-evidence posture rather than a limited one, and gives the cited reasons. The scorer then applies the fixed tier definitions to characterised evidence rather than resolving the ambiguity afresh on each run.

The legitimacy of the method rests on three constraints, each of which is testable against this report:

1. **It codifies the majority reading, it does not invent one.** The unanchored instrument already chose the insufficient-evidence band for the interlock domains in four runs of five. The anchoring note pinned the reading the model already reached most often; it did not move the score to a value the evidence did not support.
2. **It anchors to facts, not to numbers.** Every assessment note traces to evidence cited elsewhere in the same pack. No note names or targets a score.
3. **It is disclosed.** Every assessment note prints verbatim in Appendix B of the report. A reader who disagrees with a characterisation can see exactly what it was and contest it. This is the difference between calibration and contamination: calibration is visible and arguable; contamination is hidden.

Known limitations

- **Anchoring is an evidence-side mitigation, not an architectural one.** It stabilises this evaluation but must be applied per evaluation. The structural fix — decomposing mixed-signal domains into per-criterion sub-scores so that no single holistic judgement carries a whole band — is scheduled for Rubric v1.1.
- **Red-flag selection is validated for canonicity, not for evidential support.** The two unsupported flags described above were caught on review and prevented by guardrails; the validator does not yet require that a fired flag cite supporting evidence. This is a v1.1 instrument change.
- **Temperature 0.2 is already low.** The residual sub-band noise is reasoning-granularity, not sampling noise, and will not be removed by lowering temperature further; per-criterion scoring is the route.

- **Single-model scoring.** This evaluation uses one scoring model. Multi-model inter-rater agreement — scoring with two or more models and reporting their concordance — is under consideration for a later version as a stronger reliability measure.
- **Model version pinning and routing.** Inference is restricted to US-hosted providers; the `deepseek/deepseek-chat` slug should be pinned to a dated model version so that any published evaluation is reproducible against fixed weights.

Status

The five datasets behind this note (three consistency studies and the two double-check runs of the production report) are retained in the evaluation folder alongside the report PDF and its audit JSON. The published Epic Sepsis Model evaluation reflects the Study 3 configuration. The instrument's verdict on this vendor is stable, reproducible, and documented; the points at which it was not are recorded above rather than removed.