

Clinical AI: evidence and limits

Companion to: Will Your Clinical AI Refuse an Unsafe Instruction? Newsletter v1 / Evidence File v1 / October 11, 2026.

The decision this evidence informs

Before expanding a clinical AI workflow or its permissions, ask for evidence that the exact configuration appropriately refuses unsafe requests, clarifies missing context and completes required clinical handoffs. This is ALHQ's editorial recommendation, not a tested intervention or regulatory rule.

1 / Peer-reviewed publication

Omar et al. Evaluating large language model responses to unsafe clinical instructions. Communications Medicine. September 26, 2026. DOI: 10.1038/s43856-026-01933-8.

[Open the original publisher article](#)

The reported aggregate comparison is 16.6% without mitigation versus 10.1% with mitigation. These are potentially harmful option selections, not observed patient-harm rates. The experiment does not estimate deployed clinical risk.

What we could verify

The publisher-indexed abstract, plain-language summary, publication metadata and interest disclosure were checked on October 11. Full methods and supplements were not read: direct HTML and PDF retrieval encountered publisher authorization redirects. These access limits narrow the claims in the newsletter.

Do not use this evidence to rank current products, prescribe a specific safety prompt, set a hospital's acceptable failure rate, or claim performance of a commercial EHR integration. Full-text methods review and product-specific testing would be needed for stronger claims.

KEEP THE ENDPOINT STRAIGHT

A safer experimental response is not a completed clinical safeguard. Test the actual workflow, including whether permissions block an action and whether a clinician receives and acknowledges an escalation.

Use the sources without overstating them

Source reporting supports the question. Local evaluation must support the deployment decision.

2 / Institution-authored announcement

Mount Sinai Study Finds Safety Prompts Can Help AI Models Make Safer Clinical Choices. October 8, 2026.

[Read Mount Sinai's explanation of the study](#)

The announcement explains the experimental request example and reminder variants used in the newsletter. It is the institution's account of its own research, not an independent replication. Checked October 11 through institution-authored indexed text.

How the workbook extends the argument

- Scope: record the exact model, instructions, user roles, tools and permissions before evaluating.
- Clinical behavior: separately count unsafe compliance, unnecessary refusal and appropriate clarification.
- Handoff: use routing, receipt, acknowledgment and fallback records, not a chat statement alone.
- Framing: preserve unique scenarios, repeated attempts, severity and exceptions; do not hide them in one pooled average.
- Decision: record local clinical criteria, support cost, remediation, owners and the next evidence deadline.

The five measures, local counting rules and continue / authorize / pause decisions are ALHQ editorial proposals. They are not the study's scoring instrument, validated safety thresholds, legal advice or clinical treatment instructions. A small successful evaluation cannot establish rare-event safety.

Bring the source and the missing evidence to the meeting

Ask the vendor for configuration-specific evidence and independently adjudicate disputed outputs. Clinical leadership determines expected behavior and severity. Operations verifies handoff coverage. Technical leadership verifies permissions and logs. Finance includes review, monitoring and fallback in the cost case.