# SAFE-Triage explained: evaluating patient-facing LLM triage beyond acuity accuracy

24 November 2026 · 8 min read · blogs.iictp.org/safe-triage-llm-emergency-triage-evaluation-explained

Short answer: SAFE-Triage is an evaluation framework, described in an arXiv preprint submitted on 17 November 2026, for patient-facing AI triage tools. It combines human review, automated metrics, LLM-as-judge scoring and simulation testing across clinical performance, patient interaction and summary quality. In the authors' test, a plain prompted LLM was more accurate and under-triaged less often than a rule-augmented version, which shows why a single accuracy number is not enough. This is an arXiv preprint. It has not been peer reviewed, and its numbers are the authors' own claims.

**TL;DR The short version**

- SAFE-Triage is an evaluation framework for patient-facing AI triage that goes beyond a single accuracy number.
- It combines human review, automated metrics, LLM-as-judge scoring and simulation across clinical performance, patient interaction and summary quality.
- In the authors' test a plain prompted LLM beat a rule-augmented one on accuracy and under-triage, showing that adding rules does not automatically make a tool safer.

Read online, listen to it, or change the text size: https://blogs.iictp.org/safe-triage-llm-emergency-triage-evaluation-explained

## What is SAFE-Triage?

The paper, "Beyond Acuity Accuracy: A Multidimensional Framework for Evaluating Patient-Facing Emergency Triage AI Tools", argues that tools which help people decide when to seek care, or help emergency departments prioritise, need evaluation beyond whether the predicted acuity matches a reference. SAFE-Triage is the framework the authors propose for that, for use before deployment and during monitoring afterwards.

## How it works

- **Four evaluation methods in one framework:** human review, automated metrics, LLM-as-judge evaluation and simulation testing.
- **Three areas assessed:** clinical performance, patient interaction and the quality of the clinician-facing summary the tool produces.
- **Two tool designs compared:** prompted LLM triage (PLT), where a model is prompted to triage, and rule-augmented prompted LLM triage (rPLT), which adds rules to the prompt.

## Results reported by the source

- The authors applied the framework to 40 emergency vignettes derived from the MIMIC-IV dataset and a 10,030-case acuity analysis.
- PLT achieved higher Emergency Severity Index (ESI) accuracy than rPLT, 57.5% versus 42.8%, and lower undertriage, 13.1% versus 29.8%, and received higher clinician ratings in all domains, the abstract says.
- rPLT produced shorter, more stable outputs with higher automated readability scores.
- The framework surfaced trade-offs between acuity accuracy, undertriage risk, communication quality, readability and summary quality.

## Limitations and open questions

- It is a preprint, and the paper is long (82 pages including 34 of supplementary material). Read the methods before relying on the numbers.
- The data are vignettes derived from a public critical-care dataset and a retrospective acuity analysis; real patient behaviour, language and device effects are not in the abstract.
- LLM-as-judge scores inherit the judge's biases, so the human review arm matters. The abstract does not detail inter-rater agreement.
- An accuracy of 57.5% on ESI is the authors' result for one prompt design, not a verdict on all triage tools.

## How to adopt the idea

1. Define the unsafe errors first: under-triage is usually worse than over-triage, so track it separately from overall accuracy.
2. Build a fixed vignette set with clinician-agreed answers, and keep a held-out part for later monitoring.
3. Score the patient-facing text for readability and for whether it gives clear next steps, and have clinicians review the summary they would receive.
4. Re-run the whole set on every change to the prompt, rules or model.

## What it means for clinicians and digital health teams

The practical lesson is that adding rules did not automatically make the tool safer in this test: it made outputs more consistent but less accurate in the authors' setup. Clinical oversight should be built in at the evaluation stage, not only at the point of use. See how to evaluate an LLM application and responsible AI in practice.

## Frequently asked questions

- **What is SAFE-Triage?** An evaluation framework for patient-facing AI triage tools combining human review, automated metrics, LLM-as-judge scoring and simulations.
- **What did it find?** Plain prompted LLM triage had higher ESI accuracy and lower undertriage than the rule-augmented version in the authors' tests.
- **Is it peer reviewed?** No; it is an arXiv preprint.
- **Is it a clinical tool?** No. It is an evaluation method, not a triage product.

## Related reading

- How to evaluate an LLM application
- Gemini for healthcare
- Observability for AI systems
- AI Practitioner

## Sources and currency

Sources were read on 24 November 2026. Products, rules and standards change, so check the linked originals before you rely on any detail. The numbers come from the paper’s abstract.

> Related credential: AI Practitioner. Explore the exam facts and syllabus.

### Sources and further reading


1. Ruio Heerschap and 21 other authors. Beyond Acuity Accuracy: A Multidimensional Framework for Evaluating Patient-Facing Emergency Triage AI Tools. arXiv 2611.22045, 17 November 2026 (preprint). arxiv.org/abs/2611.22045.



Published by IICTP, an independent technical credential body, on 24 November 2026. The current version of this post is at https://blogs.iictp.org/safe-triage-llm-emergency-triage-evaluation-explained. More guides: blogs.iictp.org.