

Diagnostic Safety Guardrails for Large Language Models in Pediatric Emergency Medicine

Jeremiah Feiner, Elizabeth Nisper MD, Emmy Grinstead MD, Nicholas Castle, Bhawana Arora MD, Jack Brown, Michelle Padley PhD, Tiffany Fleeger, Jeffrey Jones MD

MSU College of Human Medicine, Department of Emergency Medicine; MSU-Corewell Emergency Medicine Residency; Corewell Health Hospitals/Helen DeVos Children's Hospital, Grand Rapids, MI

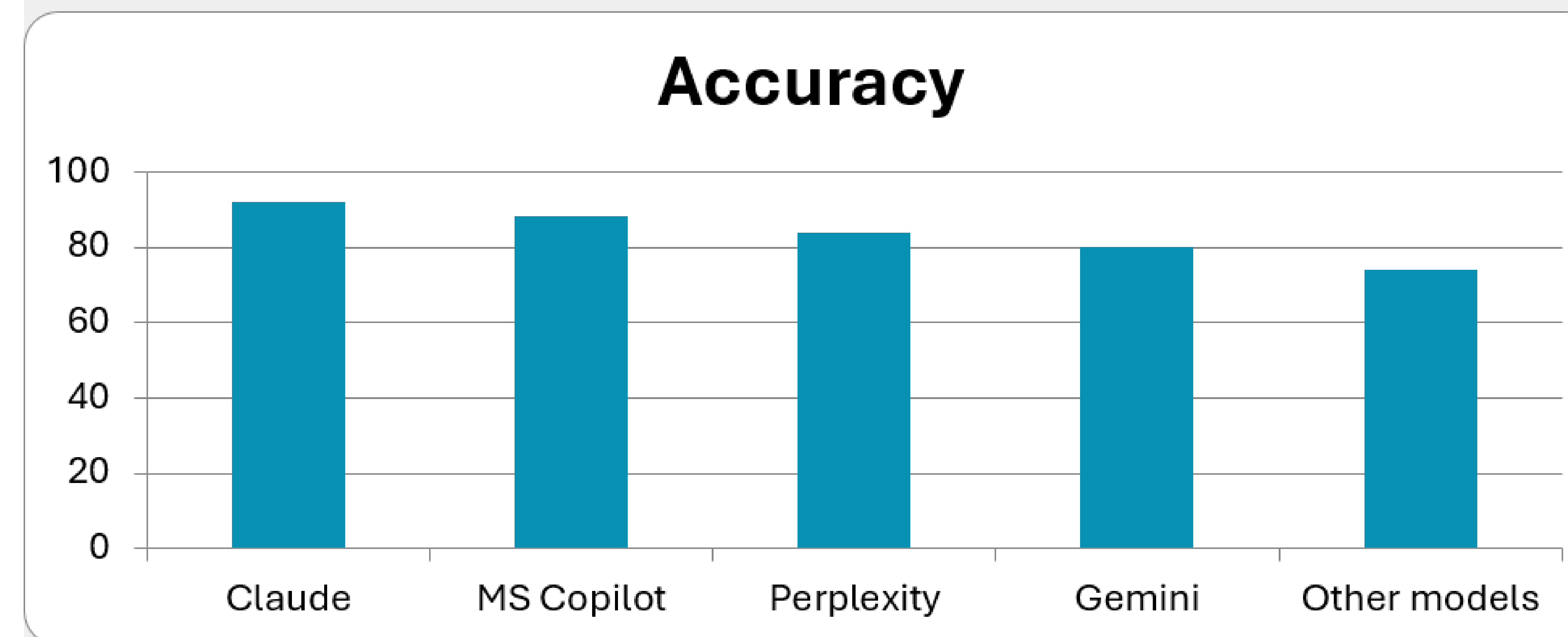
Study Objectives

- Large Language Models (LLMs) are increasingly used for clinical reasoning and education in pediatric emergency medicine (PEM).
- Safe integration remains uncertain because performance, reasoning quality, and evidence transparency vary across systems.
- This study moved beyond simple benchmarking to develop a practical diagnostic safety framework for PEM workflows.

Methods

- Prospective cross-sectional comparison of 7 widely used LLMs.
- Testing performed in October 2025 at a university-affiliated children's hospital.
- 25 complex or rare PEM case scenarios selected by an academic faculty panel.
- 5 expert raters scored 8 domains, including diagnosis, differential, testing, treatment, complications, prognosis, and overall utility.
- Findings were then reviewed to identify model-specific performance patterns, failure modes, and gaps in evidence transparency.

Results



- The LLM with the highest diagnostic accuracy was Claude (92%), followed by Microsoft Copilot (88%), Perplexity (84%), and Gemini (80%), with the remaining three models achieving accuracies of 72–76%.
- Accuracy in recommending initial diagnostic testing ranged from 84–88%, while initial treatment plans were judged appropriate in 88–96% of cases, anticipated complications in 84–96%, and prognostic comments in 72–92%.
- LLMs generated stepwise illness-script reasoning; some were rated highly useful for trainees due to structured formatting.

Observed Weaknesses and Failure Modes

- Atypical symptom recognition was inconsistent across models.
- Laboratory and imaging interpretation remained a meaningful weakness.
- Evidence transparency was variable; some outputs were clinically plausible but insufficiently sourced.
- Two models produced clinically plausible guidance without inline references, effectively functioning as unattributed expert opinion (DeepSeek, Gemini).
- These issues are significant in high-stakes pediatric emergency decisions, where overconfidence can create diagnostic risk.

Seven Diagnostic Safety Guardrails

- Clinician review of all LLM outputs before action.
- Use LLMs for appropriate, lower-risk or supervised tasks.
- Match the model to the specific clinical task.
- Verify recommendations against evidence and cited sources.
- Require explicit acknowledgment of diagnostic uncertainty.
- Train clinicians and trainees in appropriate use.
- Monitor errors, near-misses, and performance changes over time.

Conclusions

- LLMs show meaningful promise as adjuncts for supervised clinical reasoning and trainee education in PEM.
- The key contribution of this study is not just ranking models, but proposing a practical governance framework for safer adoption.
- Implementation should emphasize supervision, evidence verification, and ongoing monitoring rather than automation.