HealthManagement
Promoting Management and Leadership

LLMs Near Physician Diagnosis but Lag in Triage

Rapid advances in large language models (LLMs) are reshaping conversational decision support in healthcare. An evaluation of eight contemporary systems on concise, single-turn clinical vignettes assessed diagnostic and triage performance with and without structured prompts. The work compared overall accuracy, examined safety-oriented behaviours and applied a difficulty-adjusted capability comparison score (CCS) to distinguish performance beyond raw percentages. Across 48 balanced cases spanning four urgency categories, diagnostic accuracy for the strongest models neared benchmarks previously reported for primary care physicians on the same vignette set, while triage remained more challenging. Structured prompting acted as a training-free lever that lifted accuracy and safety across models, albeit with a trade-off toward more conservative recommendations.

Vignette Design and Evaluation Approach

The vignette set comprised 48 synthetic clinical scenarios, each fewer than 50 words and written at or below a sixth-grade reading level. Cases covered common and severe conditions and were evenly distributed across four triage categories: emergent, within 1 day, within 1 week and self-care. The dataset was drawn from prior validated work, enabling positioning of LLM results against lay and physician baselines established previously. All evaluations of the eight LLMs—ChatGPT-4, ChatGPT-o1, DeepSeek-V3, DeepSeek-R1, Gemini-2.0, Copilot, Grok-2 and Llama-3.1—were conducted between 1 and 15 March 2025.

Must Read: ***Maintaining Prescribing Safety in Pharmacist Absences***

Two testing modes were used. In the non-prompted mode, models received role instruction to act as a senior doctor, then produced a single most likely diagnosis and a triage category. In the prompted mode, models were exposed to 47 solved vignettes with gold-standard answers before independently tackling the remaining target case. This “answer-revealing” setup exceeded typical few-shot formatting guidance by providing a rich, in-context bank of directly relevant exemplars intended to strengthen case-specific reasoning without any model retraining.

Performance was measured by diagnostic and triage accuracy and further characterised using two safety-oriented triage metrics. Over-triage was the share of misclassifications that recommended a more urgent category than the reference standard. Safety of advice was the proportion of recommendations at least as urgent as the standard. CCS complemented accuracy by weighting results according to case difficulty, penalising errors on easy items and rewarding correct answers on hard ones to yield a 0–100 normalised score.

Prompting Improves Accuracy but Increases Conservatism

Structured prompting delivered consistent improvements across tasks, with the clearest gains in triage. Diagnosis, already strong for the leading systems, tightened further and drew closer to physician benchmarks reported on the same vignettes. Triage, historically harder for language models, showed a more pronounced lift under prompting, bringing several systems much nearer to their own diagnostic performance.

Safety shifted alongside these gains. Prompted outputs tended to be more cautious, increasing the likelihood that advice would meet or exceed the reference urgency. The cost of that caution was a higher tendency to recommend more urgent care than necessary. In practice, this trade-off meant fewer potentially unsafe under-calls balanced by more conservative referrals. Among the models, some achieved a more favourable mix of high safety with only moderate over-triage, whereas others leaned heavily toward urgent recommendations to secure very high safety.

Relative performance differed by model and mode. Certain systems, such as ChatGPT-o1 and DeepSeek-R1, demonstrated top-tier diagnostic

results without prompts and remained among the strongest with exemplars. Grok-2 matched the leaders on diagnosis when prompted, illustrating how structured context can elevate peers to the front. On triage, ChatGPT-o1 set the pace under prompting, while Gemini-2.0, Copilot and Grok-2 formed a closely grouped second tier. Even where gains were modest, prompting generally nudged each system in the right direction on both accuracy and safety, reinforcing the value of carefully curated, answer-revealing context for brief clinical problems.

Error Patterns and Difficulty-Adjusted Capability

Misclassifications were dominated by overestimation of urgency. Presentations suitable for near-term outpatient follow-up were often escalated to immediate care, and some self-care scenarios were upgraded to clinic visits. Under-triage appeared less frequently but remained the more clinically concerning failure mode, particularly when acute presentations were assigned to delayed care. Prompting curtailed those under-calls, though the expected price was increased over-triage.

Case-level behaviour underscored the sensitivity of single-turn, text-only reasoning to sparse or ambiguous cues. A dermatology vignette illustrated this variability: only a subset of systems assigned the correct triage under prompting, while another strong performer downgraded the same scenario to self-care. Conversely, a gastrointestinal presentation with a haemato-renal target diagnosis was repeatedly labelled as infectious gastroenteritis, showing how minimal details can steer outputs toward common yet incomplete explanations. These patterns suggest that richer input and iterative clarification are likely to matter as much as model choice in practical use.

The CCS clarified capability beyond raw accuracy. By penalising mistakes on easier items and rewarding correct answers on harder ones, it highlighted systems that maintained performance where reasoning is most demanding. Under both modes, ChatGPT-o1 delivered the strongest overall balance on triage, with DeepSeek-R1, Gemini-2.0, Grok-2 and Copilot closely grouped. Diagnosis showed a similar, though tighter, spread, and rankings were broadly consistent with accuracy-based impressions. The adjusted lens, however, emphasised resilience on challenging cases rather than success on straightforward patterns.

Across concise, balanced vignettes, advanced LLMs achieved high diagnostic performance, approaching previously reported primary care benchmarks on the same material. Triage improved notably with structured prompting yet continued to lag diagnosis, reflecting the added complexity of calibrating urgency in brief, text-only encounters. Safety moved in a favourable direction under prompting, with fewer under-calls and a shift toward conservative recommendations. CCS confirmed these trends by rewarding systems that held their ground on harder cases, with ChatGPT-o1 showing the strongest overall balance and a cluster of peers close behind. The signal is clear: answer-revealing prompts offer a practical, training-free way to enhance reliability today, while triage warrants further validation and refinement on richer, multi-turn and multi-modal inputs.

Source: **Journal of Medical Systems**

Image Credit: **iStock**

Published on : Thu, 23 Oct 2025