Shaping Tomorrow's Nephrology: Insight-Driven Kidney Care
KSN 2026 SEOUL, KOREA
The 46th Annual Meeting of the Korean Society of Nephrology
THE KOREAN SOCIETY OF NEPHROLOGY
Jun 11 (Thu) – 14 (Sun), 2026
COEX, Seoul, Korea

Abstract Type : Oral presentation
Abstract Submission No.: A-0720
Abstract Topic : Big Data

Evaluating LLM Agents vs. Human Responses in a Nephrology OSCE Scenario: A Human-Rated Dialogue Analysis

Yongjin Yi[1], Sejoong Kim[2]
[1]Department of Internal Medicine-Nephrology, Dankook University Hospital, Korea, Republic of
[2]Department of Internal Medicine-Nephrology, Seoul National University College of Medicine, Korea, Republic of

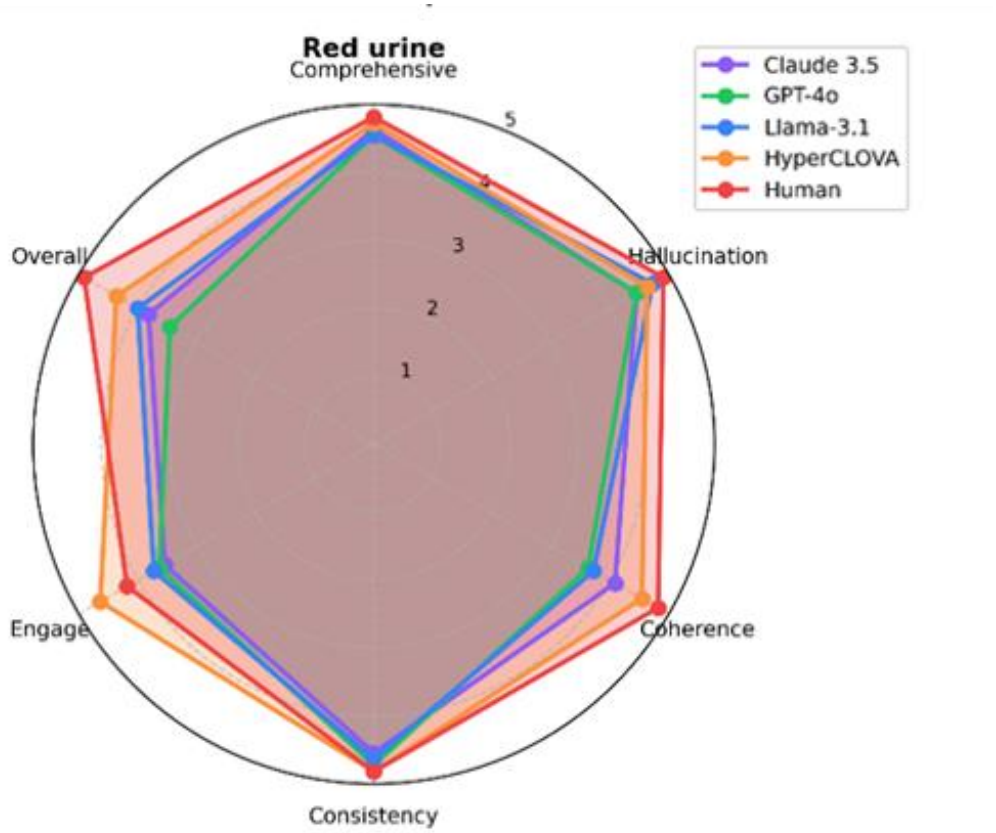
Objectives : Clinical performance examination (CPX) remains the gold standard for assessing clinical reasoning. However, CPX education requires substantial resources, including trained standardized patients, dedicated facilities, and high costs. Recent advances in large language models (LLMs) enable the generation of clinically accurate virtual patients as scalable alternatives to CPX. To evaluate the quality of virtual patient (VP) responses generated by LLMs in CPX scenarios, compared to human-generated responses authored by CPX educators.
Methods : Four state-of-the-art LLMs (OpenAI GPT-4o, Anthropic Claude-3.5 Sonnet, NAVER HyperCLOVA X, and Meta LLaMA-3.1 70B) were used to simulate patient responses across three CPX scenarios (red urine, fever in pregnancy, and chronic cough). Each CPX dialogue contained 23 to 25 question-response sets. An experienced CPX educator provided the reference responses. Eleven senior medical students rated each response on six quality dimensions – comprehensiveness, coherence, accuracy, hallucination avoidance, engagement, and overall satisfaction using a 5-point Likert scale.
Results : Human expert responses achieved the highest mean overall satisfaction score (4.61, IQR [4.40-4.81]), followed by HyperCLOVA X (4.00, IQR [3.73-4.27]) and Claude-3.5 (4.00, IQR [3.67-4.33]). Claude-3.5demonstrated superior hallucination avoidance (4.45, IQR [4.21-4.70]), while HyperCLOVA X excelled in engagement metrics (4.30/5.0). LLaMA-3.1-70B consistently underperformed (2.64, IQR [2.23-3.04]).
Conclusions : LLM-powered virtual patients can produce realistic and clinically relevant responses, demonstrating potential as scalable training tools in medical education. The human-authored responses still outperform LLMs in overall quality. The Korean-optimized LLM showed particular strength in culturally nuanced medical communication, underscoring the importance of domain-specific pretraining for clinical applications.

table1.PNG

Table 1. Comparative evaluation of LLM-generated virtual patient responses. Eleven medical students rated model-generated responses using a 5-point Likert scale. Data are presented as mean [interquartile range].

Model	Comprehensive	Hallucination	Coherence	Consistency	Engagement	Overall satisfaction
Claude 3.5 Sonnet	4.48 [4.27-4.70]	4.45 [4.21-4.70]	4.09 [3.82-4.37]	4.45 [4.18-4.73]	3.88 [3.56-4.20]	4.00 [3.67-4.33]
chatGPT-4o	4.64 [4.45-4.82]	4.39 [4.15-4.63]	4.18 [3.91-4.46]	4.61 [4.40-4.81]	3.88 [3.60-4.16]	3.91 [3.63-4.18]
LLaMA-3.1 70B	3.58 [3.25-3.91]	3.85 [3.50-4.20]	2.79 [2.43-3.15]	3.73 [3.34-4.11]	2.82 [2.48-3.15]	2.64 [2.23-3.04]
HyperCLOVA X	4.48 [4.29-4.68]	4.12 [3.82-4.43]	4.30 [4.09-4.52]	4.45 [4.24-4.67]	4.30 [4.05-4.55]	4.00 [3.73-4.27]
Human answer	4.76 [4.61-4.91]	4.45 [4.21-4.70]	4.76 [4.61-4.91]	4.67 [4.45-4.89]	4.15 [3.88-4.42]	4.61 [4.40-4.81]

table1.PNG


Red urine
Comprehensive
Hallucination
Coherence
Consistency
Engage
Overall
1
2
3
4
5
Claude 3.5
GPT-4o
Llama-3.1
HyperCLOVA
Human