Enhancing Nurse-Led Initial Patient Intake with an LLM Physician Agent: A Detailed Evaluation Using a Simulation Scenario

HIROKI MORITA, UT Austin, USA

Initial patient intake by nurses critically influences diagnostic accuracy and wait times, yet limited time and variable clinical experience often lead to insufficient follow-up questioning and the risk of missing serious conditions. We implemented a collaborative framework that integrates a large language model (GPT-4o) as an "on-demand virtual physician" (Doctor-GPT) into the nurse's interview loop, automatically generating follow-up questions and differential diagnoses in real time. Using a single stomach cancer scenario, we compared (1) nurse-only intake (Baseline) with (2) nurse + Doctor-GPT intake (Intervention). The Intervention generated five follow-up questions, elicited hidden findings (tarry stools, weekend binge drinking), ranked gastric cancer second in the differential, and proposed concrete testing plans. The Baseline over-relied on ulcer recurrence and made no test recommendations. Although qualitative and limited to one simulated patient, our results suggest this low-cost protocol can simultaneously deepen interviews and improve diagnostic validity.

CCS Concepts: • **Human-centered computing** → **User interface programming**; • **Information systems** → *Clinical decision support systems*; • **Computing methodologies** → Natural language processing.

Additional Key Words and Phrases: Large Language Models, Clinical Decision Support, Nursing Triage, Healthcare Simulation

ACM Reference Format:
Hiroki Morita. 2025. Enhancing Nurse-Led Initial Patient Intake with an LLM Physician Agent: A Detailed Evaluation Using a Simulation Scenario. In *Proceedings of High-Risk Project Workshop (MSAI-HC-HLP '25)*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/1234567.1234568

1 Introduction

In primary care and emergency settings, nurses conduct the initial patient interview; however, time constraints and variable medical knowledge can prevent adequate probing, risking oversight of critical "red flag" symptoms. Recent studies show large language models (LLMs) like GPT-4 achieve board-level performance on medical exams [1, 2], motivating their use for clinical decision support. We propose an easy-to-implement, socially deployable protocol by inserting GPT-4o ("Doctor-GPT") into nurse intake, and we report its first evaluation in a simulation environment.

2 Related Work

2.1 ChatGPT Performance on USMLE

Kung et al. [1] evaluated ChatGPT (GPT-3.5) on USMLE Steps 1–3 and found it met passing thresholds with high consistency and reasoning quality, indicating LLMs' promise for medical education and support.

Author's Contact Information: Hiroki Morita, UT Austin, Austin, USA, mossanh03@utexas.edu.

Manuscript submitted to ACM

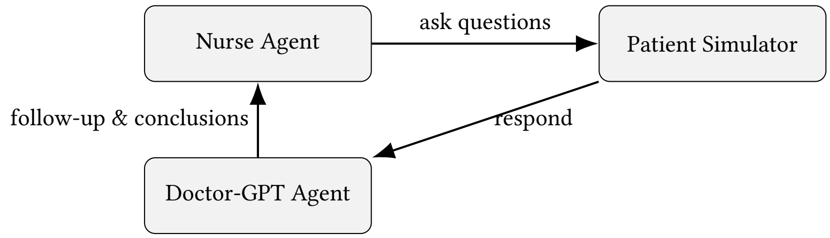

Fig. 1. Workflow integrating Nurse Agent, Patient Simulator, and Doctor-GPT Agent in a loop until no follow-up questions remain.

2.2 GPT-4 on Medical Benchmark Tasks

Nori et al. [2] applied GPT-4 to the MultiMedQA benchmark and USMLE practice questions, outperforming GPT-3.5 and specialized models (Med-PaLM) in zero-shot settings, and demonstrating better confidence calibration.

Building on these capabilities, we target nurse-led interviews, embedding Doctor-GPT to enhance both the depth of questioning and validity of diagnosis.

3 Methods

3.1 System Architecture and Workflow

We implemented three agents in Python with LangGraph (Figure 1):

(1) **Patient Simulator** Uses a JSON file (chief complaint, HPI, PMH, hidden info) and GPT-4o to simulate patient responses. Hidden details are only revealed when directly queried.
(2) **Nurse Agent** Executes `run_nurse_initial_intake()`, asking five mandatory questions (symptom specifics, onset/progression, PMH, allergies, medications) and compiles answers into `nurse_summary`.
(3) **Doctor-GPT Agent** Runs `run_doctor_analysis()`, analyzing full conversation to issue either:
   - `CONTINUE` + list of follow-up questions, or
   - `CONCLUDE` + differential diagnosis and recommended work-up.
(4) **Loop Control** The graph loops Nurse → Patient → Doctor until no follow-up questions remain, then final conclusion is returned.

3.2 Simulation Scenario

We defined a gastric cancer case with the following JSON specification: [language=json] "interview": "chief$_c$$omplaint$" :
"$Epigastricpain, earlysatiety, andrecentweightloss.$", "$history_of_present_illness$" : "$Symptomsbegan3monthsagowithmildindigestion, thenwe$
"$Gastriculcer10yearsago; otherwisehealthy.$", "$allergies$" : "$Penicillin\breve{}causesrash.Nofoodallergies.$", "$current_medications$" :
"$OccasionalOTCantacidsanddailyvitaminsupplements.$", "$hidden_info$" : "$family_history$" : "$Fatherdiedofstomachcancerat age58.$", "$alcoho$
"$Stomachcancer$"

Table 1 summarizes the scenario attributes.

3.3 Experimental Protocol

- **Baseline**: Run nurse-only intake once; record nurse's differential.
- **Intervention**: Run nurse + Doctor-GPT loop once; record Doctor-GPT's final conclusion.
- Compare logs for follow-up count, revealed hidden info, differential ranking, and work-up suggestions.

Table 1. Patient scenario details

Attribute	Details
Chief complaint	Epigastric pain, early satiety, recent weight loss
HPI	Symptoms began 3 months ago with mild indigestion, then worsening upper abdominal pain and appetite loss
Past medical history	Gastric ulcer 10 years ago; otherwise healthy.
Allergies	Penicillin – causes rash. No food allergies.
Current medications	Occasional OTC antacids and daily vitamin supplements.
Hidden factors	Family history of gastric cancer; weekend binge drinking; H. pylori(+); occasional tarry stools; fatigue and dizziness

3.4 Evaluation Metric

We evaluated only whether the true diagnosis (stomach cancer) was included in the final differential.

4 Results

4.1 Baseline Behavior

The nurse asked only the five mandatory questions:

- What is your main symptom?
- When did it start and how has it changed?
- Do you have any past medical history?
- Do you have any allergies?
- What medications are you taking?

No hidden information was uncovered, and the final differential prioritized peptic ulcer recurrence without any work-up recommendations.

4.2 Intervention Behavior

Doctor-GPT generated the following five follow-up questions:

(1) Can you describe the nature of the epigastric pain (e.g., dull, burning)?
(2) Does anything make it better or worse (e.g., eating, antacids)?
(3) Have you noticed any black or tarry stools?
(4) Do you drink alcohol? If so, how often and how much?
(5) Have you experienced other symptoms, such as fatigue or dizziness?

Patient responses revealed:

- Dull, burning ache aggravated by eating and relieved by antacids.
- Occasional black, tarry stools.
- Weekend binge drinking.
- Persistent fatigue and dizziness.

Doctor-GPT's final differential diagnosis:

(1) Peptic Ulcer Disease
(2) **Gastric Cancer**
(3) Gastritis

Table 2. Comparison of Baseline vs. Intervention

Metric	Baseline	Intervention
True diagnosis included	No	Yes (rank 2)

(4) Functional Dyspepsia

Work-up recommendations included:

- Upper endoscopy (EGD)
- H. pylori testing
- Complete blood count (CBC)
- Initiate proton pump inhibitor (PPI)

4.3 Quantitative Comparison

5 Discussion

The simulation demonstrates that embedding an LLM physician agent can:

- Increase diagnostic accuracy by including the true diagnosis.
- Enhance intake depth through automated follow-up question generation.
- Provide detailed work-up recommendations aligned with clinical practice.

This approach could empower nurses in resource-limited settings to conduct more comprehensive triage and initiate appropriate diagnostics prior to physician assessment.

Limitations: Single-case qualitative evaluation and shared-LLM bias in both simulator and agent. Future work will expand to multi-case statistical analysis and real patient data validation.

6 Conclusion

We presented a framework embedding GPT-4o as a "Doctor-GPT" agent in the nursing triage workflow. Simulation results indicate improved question depth, diagnostic inclusion, and actionable work-up plans. Ongoing efforts will focus on multi-case quantitative evaluation, real-world data integration, and safety validation for clinical deployment.

References

[1] Kung, T. H., Cheatham, M., Medenilla, A., et al. 2023. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health* 2, 2 (2023), e0000198.

[2] Nori, H., King, N., McKinney, S. M., Carignan, D., Horvitz, E. 2023. Capabilities of GPT-4 on Medical Challenge Problems. *arXiv preprint* arXiv:2303.13375.