A Generative AI-Based Virtual Physician Assistant

Geoffrey W. Rutledge and Alexander Sivura
HealthTap Inc.
Geoffrey.rutledge@healthtap.com

Abstract

We describe "Dr. A.I.", a virtual physician assistant that uses generative AI to conduct a pre-visit patient interview and to create a draft clinical note for the physician. We document the effectiveness of Dr. A.I. by measuring the concordance of the actual diagnosis made by the doctor with the generated differential diagnosis (DDx) list. This application demonstrates the practical healthcare capabilities of a large language model to improve efficiency of doctor visits while also addressing safety concerns for the use of generative AI in the workflow of patient care.

Introduction

Generative AI applications that leverage the largest language models (LLM) have demonstrated remarkable levels of performance in answering challenging health information queries (Erikson et. al. 2023). GPT-4 was shown to outperform other LLMs tailored for medical tasks (Nori et. al. 2023). Others have investigated the ability of LLMs to conduct a patient interview to produce a DDx, testing, treatment and follow up (Tu et. al. 2024), with an approach that duplicates what the doctor does.

The practical use of generative AI in healthcare has so far been limited by the lack of a clearly useful workflow that would improve patient care, by the fact that the foundation models behavior and measurements of performance change over time, and by concerns related to confabulation or "hallucinations." (Lee et. al. 2023)

We describe an innovative application that supports the physician's role in patient care by automating the history taking and by writing a draft of the clinical note for the doctor. We evaluate the quality of the history-taking by measuring how often the LLM predicts the diagnoses that the doctors make after they evaluate their patients.

Methods

We implemented "Dr. A.I." in HealthTap's consumer healthcare mobile application. Users who request (and pay for) a scheduled visit with a board-certified primary care physician are given the option to conduct a pre-visit interview. Dr. A.I. accesses GPT-4 via the Azure API, where it provides as a prompt the patient's age, gender, chief complaint (reason for visit) and summary of previous answers, along with instructions for returning the next question that a doctor would ask. This approach helps provide conversation context in a concise form and avoids the issue of losing context in lengthy conversations.

We do not show patients any feedback after each answer. The interview ends when GPT-4 indicates there are no more questions needed, or after a maximum of 15 questions. GPT-4 writes a patient note suitable for the subjective section of the clinical note, suggests questions to ask the doctor, and records a list of diagnoses for the patient. At the time of the visit, the doctor adds, edits, or ignores the AI-generated clinical note for their documentation. Doctors are not shown Dr. A.I.'s DDx.

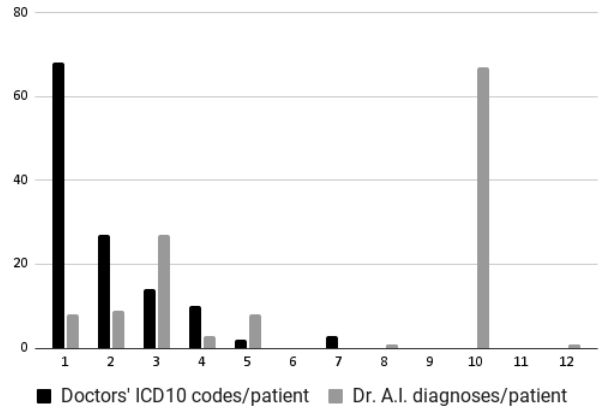

Figure 1. Counts of diagnoses/patient

Results

Dr. A.I. was first implemented for patients who access HealthTap via the HealthTap iOS consumer application in November 2023. We analyzed the first 124 visits for which

the patient completed the Dr. A.I. interview. The doctors documented from 1 to 7 International Classification of Diseases(ICD10) codes for each patient. 96%(119/124) of cases had from 1–4 codes (see Figure 1). Doctors entered the first diagnosis as the main diagnosis for the visit, and the remaining codes for other co-occurring conditions.

Dr. A.I. recorded from 1 to 12 diagnoses, with 35% (44/124) of the cases having 1–3 diagnoses and 54% (67/124) having exactly 10 diagnoses (see Figure 1). We recorded the rank in the Dr. A.I. list of each diagnosis entered by the treating physician. The rates that the doctors' diagnoses appeared in the top 1, top 3, and top 10 list of Dr. A.I. diagnoses are shown in Table 1.

	Top10 Dr. A.I.	Top3 Dr. A.I.	Top1 Dr. A.I.
1st ICD10	87.9% (109/124)	80.6% (100/124)	62.1% (77/124)
2nd ICD10	80.4% (45/56)	60.7% (34/56)	23.2% (13/56)
3rd-7th ICD10	45.5% (25/55)	27.3% (15/55)	3.6% (2/55)

Table 1. Doctors' ICD10 vs. Dr. A.I. diagnoses

Discussion

We take advantage of the interactive dialog and medical reasoning that the LLM is capable of providing without incurring safety concerns, because we only show the doctor a summary of what the patient has said in response to questions. The questions asked are relevant to the presumptive DDx, and so the summary of answers saves the doctor the time otherwise required to ask those questions and document the answers.

If for some reason the questions were not relevant, or were incomplete, we would expect the DDx generated to be inaccurate or incomplete. Because the doctor completes the history and is solely responsible for the diagnosis and treatments, the main concern is the possibility that the AI-generated summary is not helpful.

It is remarkable that in 62.1% of cases, the first diagnosis entered by the doctor was also the first ICD10 diagnosis selected by Dr. A.I., and that the doctors' first diagnoses were present in the top 10 of Dr. A.I.'s list for 87.9% of patients. Similarly, when a second ICD10 code was present, that diagnosis was found 80.4% of the time in the top10 of Dr. A.I.'s diagnoses. On the other hand, Dr. A.I.'s first diagnosis was found in one of the ICD10 codes in 74.2% (92/124) cases. This demonstrates that GPT-4 must be asking appropriate questions—because the answers to those questions usually allow the construction of a DDx list that includes the diagnosis assessed by the doctor.

Because the doctors have access to past medical records for patients, and are able to ask additional questions that are not related to the reason for a given visit, we are not surprised that doctors entered additional diagnosis codes that could not be identified by Dr. A.I.'s medical interview. However, we note that 45% (25/55) of those additional doctors' diagnoses were also present in Dr. A.I.'s DDx list.

A next step will be to evaluate what impact showing Dr. A.I.'s DDx to doctors will have. Prior research demonstrates that prompting doctors to consider multiple alternative diagnoses improves their diagnostic accuracy (Rutledge 2020; McDuff et. al. 2023; Friedman et. al. 1999).

The questions asked and the diagnoses generated by a GPT-4 based patient interview correspond to a high degree to the diagnoses assessed by doctors who evaluated the patients. It is possible to use generative AI based on the largest foundation models to engage patients in a medically relevant dialog that identifies the likely causes of a patient's medical symptoms.

Limitations

A limitation of this study is the lack of a gold standard for the correct final diagnosis for each patient case. We have not measured the reasonableness of the alternate diagnoses suggested by Dr. A.I., or what the impact on a doctors' diagnosis that such a form of decision support would cause.

References

Tu, T. et al 2024. Toward Conversational diagnostic AI. arXiv preprint. arXiv:2401.05654 [cs.AI]. Ithaca, NY

Erikson, A.V., Mollwe, A. and Ryg, G. 2023. Use of GPT-4 to Diagnose Complex Clinical Cases. NEJM AI 1(1) doi.org/10.1056/AIp2300031

Friedman, C.P., Elstein, A.S., and Wolf, F.M. 1999. Enhancement of Clinicians' Diagnostic Reasoning by Computer-Based Consultation. JAMA 282(19):1851-1856 doi:10.1001/jama.282.19.1851

Lee, L., Bubeck, S., and Petro, J. 2023. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. NEJM 388:1233-1239 doi: 10.1056/NEJMsr2214184

McDuff, D. et al. 2023. Towards Accurate Differential Diagnosis with Large Language Models. arXiv preprint. arXiv:2312.00164 [cs.CL]. Ithaca, NY

Nori, H. et al. 2023. Can Generalist Foundation Models Outcompete Special-Purpose Tuning? arXiv preprint. arXiv:2311.16452[cs.CL]. Ithaca, NY

Rutledge, G. 2020. Doctors' diagnostic accuracy with and without diagnostic decision support. Virtual online presentation at the AMIA Clinical Informatics Conference, May 19-21