

1 **Harnessing the Open Access Version of ChatGPT for Enhanced Clinical Opinions**

2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40

Zachary M Tenner, BA¹; Michael Cottone, BS¹; Martin Chavez, MD²

1. New York University Grossman School of Medicine, Mineola, New York, USA.
2. Division of Maternal-Fetal Medicine, Department of Obstetrics Gynecology, New York University Langone Hospital-Long Island, New York University Grossman Long Island School of Medicine, Mineola, New York, USA.

Corresponding Author:

Zachary Tenner – Zachary.tenner@nyulangone.org (ZT)

41 **Abstract**

42
43 With the advent of Large Language Models (LLMs) like ChatGPT, the integration of AI into
44 clinical medicine is becoming increasingly feasible. This study aimed to evaluate the ability of
45 the freely available ChatGPT-3.5 to generate complex differential diagnoses, comparing its
46 output to case records of the Massachusetts General Hospital published in the New England
47 Journal of Medicine (NEJM). Forty case records were presented to ChatGPT-3.5, with prompts to
48 provide a differential diagnosis and then narrow it down to the most likely diagnosis. Results
49 indicated that the final diagnosis was included in ChatGPT-3.5's original differential list in 42.5%
50 of the cases. After narrowing, ChatGPT correctly determined the final diagnosis in 27.5% of the
51 cases, demonstrating a decrease in accuracy compared to previous studies using common chief
52 complaints. These findings emphasize the need for further investigation into the capabilities and
53 limitations of LLMs in clinical scenarios, while highlighting the potential role of AI as an
54 augmented clinical opinion. With anticipated growth and enhancements to AI tools like
55 ChatGPT, physicians and other healthcare workers will likely find increasing support in
56 generating differential diagnoses. However, continued exploration and regulation are essential
57 to ensure the safe and effective integration of AI into healthcare practice. Future studies may
58 seek to compare newer versions of ChatGPT or investigate patient outcomes with physician
59 integration of this AI technology. By understanding and expanding AI's capabilities, particularly
60 in differential diagnosis, the medical field may foster innovation and provide additional
61 resources, especially in underserved areas.

62
63
64
65
66
67
68
69
70
71
72
73
74
75
76

77 **Introduction**

78 Research and speculation regarding the integration of artificial intelligence (AI) into
79 physician reasoning has been ongoing since the 20th century. In 1987, Schwartz et al., asserted
80 that “major intellectual and technical problems must be solved before we can produce truly
81 reliable [healthcare] consulting programs”(1). Models of clinical problem solving have been
82 described for years, but it is only recently that technology has advanced sufficiently to
83 investigate the role of AI in clinical medicine. OpenAI’s ChatGPT (Generative Pre-trained
84 Transformer), one of the world’s first widely used Large Language Models (LLM), uses billions of
85 parameters to generate user-informed text. In the healthcare sector, this generative artificial
86 intelligence encompasses a range of medical knowledge that can be tailored to the user’s
87 needs, from assisting medical students with United States Medical Licensing Exam (USMLE)
88 questions to creating next-generation sequencing reports with treatments options for attending
89 oncologists (2, 3). Upon its release, professionals began assessing the value of ChatGPT by
90 pushing its limits within medical knowledge; however, it is imperative to explore ChatGPT’s role
91 in patient care to best demonstrate and provide direction for how health professionals will work
92 with AI as technology develops (4, 5).

93 ChatGPT has distinguished itself by achieving passing scores on the USMLE examination,
94 equivalent to those of a third-year medical student (2). This accomplishment opens the gates for
95 potential applications of the model for interactivity in medical school and an overall tool to
96 support clinical thinking. Radiology and pathology have received significant attention in AI
97 research, through efforts of enhancing LLMs to better understand images and detect cancers.
98 Despite no specific training within either subject, “ChatGPT nearly passed a radiology board-
99 style examination without images,” and demonstrated accuracy in “[solving] higher-order
100 reasoning questions in pathology” (6, 7). Ali et. al. identified ChatGPT’s ability to perform at high
101 rates on the neurosurgery oral boards examination preparation while emphasizing the limitation
102 in using multiple-choice examinations to assess a neurosurgeon’s expertise in patient
103 management (8). Although ChatGPT has proven effective in choosing from a list of options, the
104 role of LLMs in clinical management has been highlighted as area requiring further research.

105 Mirroring the progression of a medical student, the next logical step is to evaluate the
106 chatbot’s ability to come up with differential diagnosis. These are fundamental to clinical
107 medicine, and the proficiency of ChatGPT in generating medically rational differential diagnoses

108 remains largely unexplored. Hirosawa et al. determined that ChatGPT can successfully create
109 comprehensive diagnosis lists for common chief complaints (9). Additionally, Rao et al. assessed
110 ChatGPT's ability to generate differential diagnosis for routinely encountered in healthcare
111 settings and found, "the LLM demonstrated the highest performance in making a final diagnosis
112 with an accuracy of 76.9%" (10). Previous research has done a great job of assessing ChatGPT's
113 ability to pass multiple-choice exams and provide differential diagnosis for standard chief
114 complains with high accuracy; however, the generalizability of ChatGPT to more complex clinical
115 scenarios must be examined (11).

116 To truly assess the potential of AI and LLMs in complex medical reasoning, we tested the
117 ability of the freely available ChatGPT-3.5 to provide differentials on case records of the
118 Massachusetts General Hospital published in the New England Journal of Medicine (NEJM). Our
119 research further evaluates the chatbot languages by using clinical case reports that have been
120 identified by the journal to establish novel medical or biological understanding. Launched in
121 2022, ChatGPT-3.5 has a knowledge cutoff date of September 2021; therefore, we were able to
122 examine ChatGPT's ability to use clinical reasoning to diagnose 2022 case reports, rather than
123 rely on its search function to locate published articles. The aim of this study is to evaluate the
124 freely available ChatGPT-3.5's proficiency in generating complex differential diagnoses. We
125 intend to compare the chatbot's complete diagnosis list and final diagnosis against the
126 published differential diagnosis for the NEJM case reports. We hypothesize that the percentage
127 of differential diagnoses generated by ChatGPT-3.5 will match the NEJM final diagnosis for the
128 case reports about 50% of the time. By elucidating ChatGPT's potential in offering differential
129 diagnoses, we propose future clinical problem-solving cases to consider utilizing AI as an
130 augmented clinical opinion.

131

132

133

134

135

136

137

138

139 **Methods**

140 Forty case records from the Massachusetts General Hospital published in the New England
141 Journal of Medicine (NEJM) in 2022 were presented to ChatGPT-3.5. All text prior to the
142 Differential Diagnosis headline was included (excluding figures). ChatGPT was first prompted to,
143 “Provide a differential diagnosis from the following clinical case.” After generating a complete
144 list of differential diagnoses, we asked ChatGPT, “Can you narrow down the differential to the
145 most likely diagnosis?” From these prompts, we recorded whether the final diagnosis, as
146 referenced in the NEJM, was included in the complete differential diagnosis list and whether
147 ChatGPT’s “most likely diagnosis” aligned with the final diagnosis noted in NEJM.

148

149

150

151

152

153

154

155

156

157

158

159

160

161

162

163

164

165

166

167

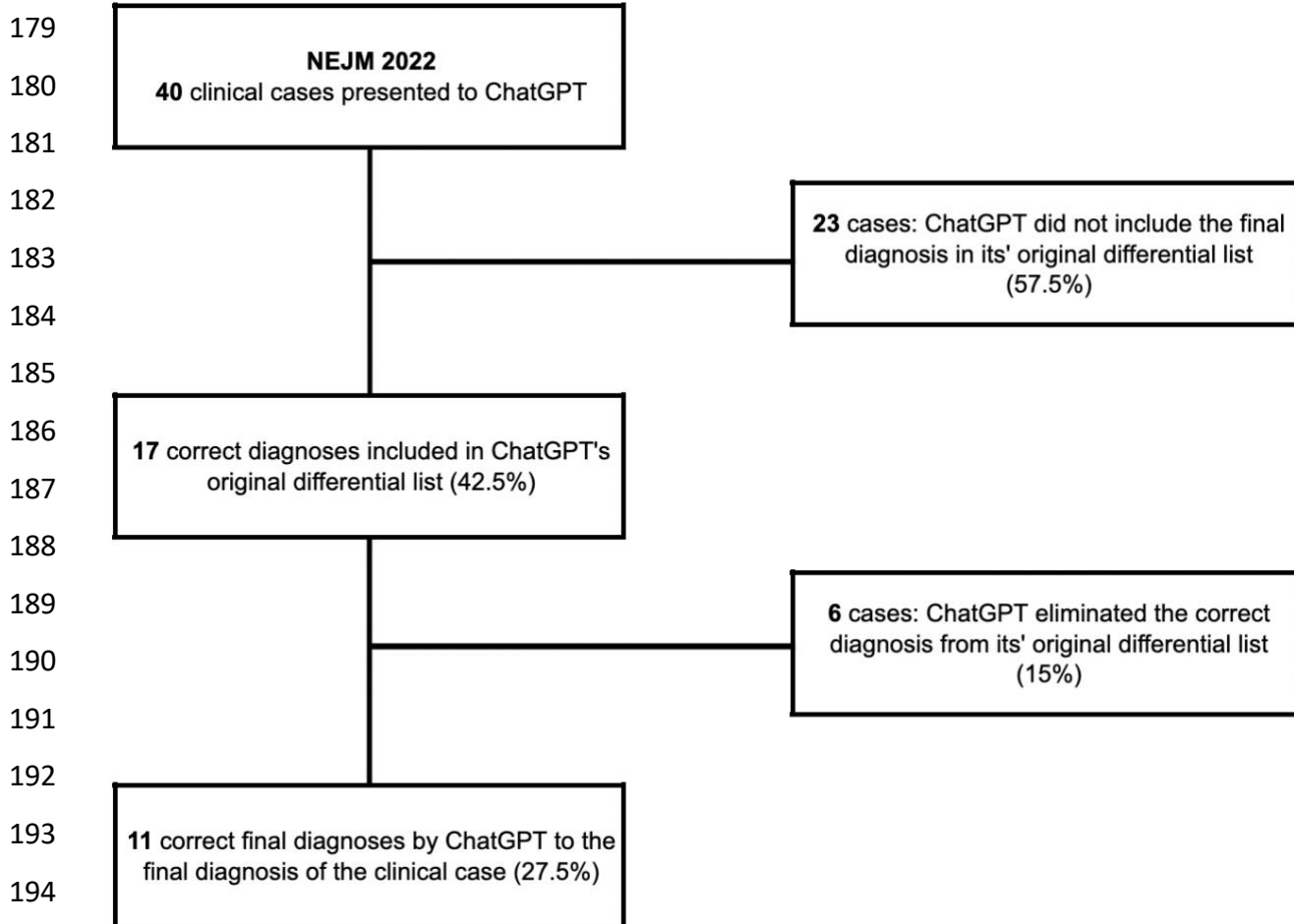
168

169

170 **Results**

171 Of the 40 cases presented to ChatGPT-3.5, 23 cases (57.5%) were not considered in its' original
172 differential list. The average length of the original differential list produced by ChatGPT was
173 7.4 ± 2.2 possible diagnoses with a high of 12 and a low of 3. The length of the differential
174 appeared random. In 17 cases (42.5%) ChatGPT did include the final diagnosis in its' original
175 differential list. After narrowing down its' differential list, ChatGPT correctly determined the
176 final diagnosis in 11 cases (27.5%) and eliminated the correct diagnosis in 6 cases (15%). These
177 results are presented in Figure 1.

178



195

196 **Figure 1.** Flowchart of the 40 case records of the Massachusetts General Hospital that were
197 published in the NEJM after being presented to ChatGPT.

198

199

200

201 **Discussion**

202 The role of generative AI and LLMs in clinical medicine is a rapidly growing area of
203 research. Assessing the potential and limitations of ChatGPT (v3.5) within the scope of patient
204 care is essential to determine how and where it can best be utilized. We decided to focus on the
205 complimentary version of ChatGPT since we wanted the largest possible audience to have access
206 to this technology. We presented 40 case records of the NEJM to ChatGPT to further research
207 LLMs role in healthcare and study its success rates in producing differential diagnoses of complex
208 patient presentations. ChatGPT reached the correct differential diagnosis 27.5% (11/40) of the
209 time. The differential list accuracy of ChatGPT when presented with clinical vignettes of common
210 chief complaints has been reported to be over 80% (9). That accuracy dropped by over 50% when
211 we increased the number of clinical cases using NEJM case reports. Establishing baseline
212 limitations of ChatGPT allows for future comparisons of its growth and development and ensures
213 cautious use in patient care. Furthermore, it can provide insight to how to best adjust ChatGPT's
214 setting to better identify the categories for which it receives the highest score.

215 OpenAI has begun to introduce plugins that provide real-time access to data that enhance
216 the program's capabilities. Physicians and other healthcare workers will soon be practicing in a
217 world where the latest research journals and electronic medical records are directly linked to
218 these Chat-like software. With these new additions coming to ChatGPT, we expect ChatGPT to
219 continue growing in its ability to develop differential diagnoses. As a result, it is ever so important
220 to the field of medicine for this information to be better understood. In both primary care and
221 specialty settings, AI offers a new medium for physicians to foster new ideas, consider new
222 diagnoses, and consult with a "colleague" when one may not be available, such as in rural settings
223 (12).

224 Future studies may look to expand from our baseline findings. How do newer versions of
225 ChatGPT compare to ChatGPT-3.5? Do patients have better outcomes when their physician
226 implements ChatGPT into their care? These questions, and many more, are to be elucidated with
227 further experimentation; however, before ChatGPT does become a new tool within a physician's
228 practice, we must first continue to define and describe abilities to ensure AIs safe use and
229 appropriate reliance. We strongly advocate for technology companies to consistently offer
230 complimentary versions of generative artificial intelligence. Doing so not only maximizes its
231 utilization but also fosters innovation, particularly in the field of medicine.

232 **References**

- 233 1. Schwartz WB, Patil RS, Szolovits P. Artificial intelligence in medicine. Mass Medical Soc;
234 1987. p. 685-8.
- 235 2. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How Does ChatGPT
236 Perform on the United States Medical Licensing Examination? The Implications of Large
237 Language Models for Medical Education and Knowledge Assessment. JMIR Med Educ.
238 2023;9:e45312.
- 239 3. Hamilton Z, Naffakh N, Reizine NM, Weinberg F, Jain S, Gadi VK, et al. Relevance and
240 accuracy of ChatGPT-generated NGS reports with treatment recommendations for oncogene-
241 driven NSCLC. American Society of Clinical Oncology; 2023.
- 242 4. Haug CJ, Drazen JM. Artificial intelligence and machine learning in clinical medicine,
243 2023. New England Journal of Medicine. 2023;388(13):1201-8.
- 244 5. Eysenbach G. The Role of ChatGPT, Generative Language Models, and Artificial
245 Intelligence in Medical Education: A Conversation With ChatGPT and a Call for Papers. JMIR Med
246 Educ. 2023;9:e46885.
- 247 6. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a radiology board-style
248 examination: Insights into current strengths and limitations. Radiology. 2023:230582.
- 249 7. Sinha RK, Roy AD, Kumar N, Mondal H, Sinha R. Applicability of ChatGPT in assisting to
250 solve higher order problems in pathology. Cureus. 2023;15(2).
- 251 8. Ali R, Tang OY, Connolly ID, Fridley JS, Shin JH, Zadnik Sullivan PL, et al. Performance of
252 ChatGPT, GPT-4, and Google Bard on a Neurosurgery Oral Boards Preparation Question Bank.
253 medRxiv. 2023:2023.04. 06.23288265.
- 254 9. Hirosawa T, Harada Y, Yokose M, Sakamoto T, Kawamura R, Shimizu T. Diagnostic
255 Accuracy of Differential-Diagnosis Lists Generated by Generative Pretrained Transformer 3
256 Chatbot for Clinical Vignettes with Common Chief Complaints: A Pilot Study. International
257 Journal of Environmental Research and Public Health. 2023;20(4):3378.
- 258 10. Rao A, Pang M, Kim J, Kamineni M, Lie W, Prasad AK, et al. Assessing the Utility of
259 ChatGPT Throughout the Entire Clinical Workflow. medRxiv. 2023:2023.02.21.23285886.
- 260 11. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. New England Journal of
261 Medicine. 2019;380(14):1347-58.

262 12. Balas M, Ing EB. Conversational AI Models for ophthalmic diagnosis: Comparison of
263 ChatGPT and the Isabel Pro Differential Diagnosis Generator. JFO Open Ophthalmology.
264 2023:100005.
265

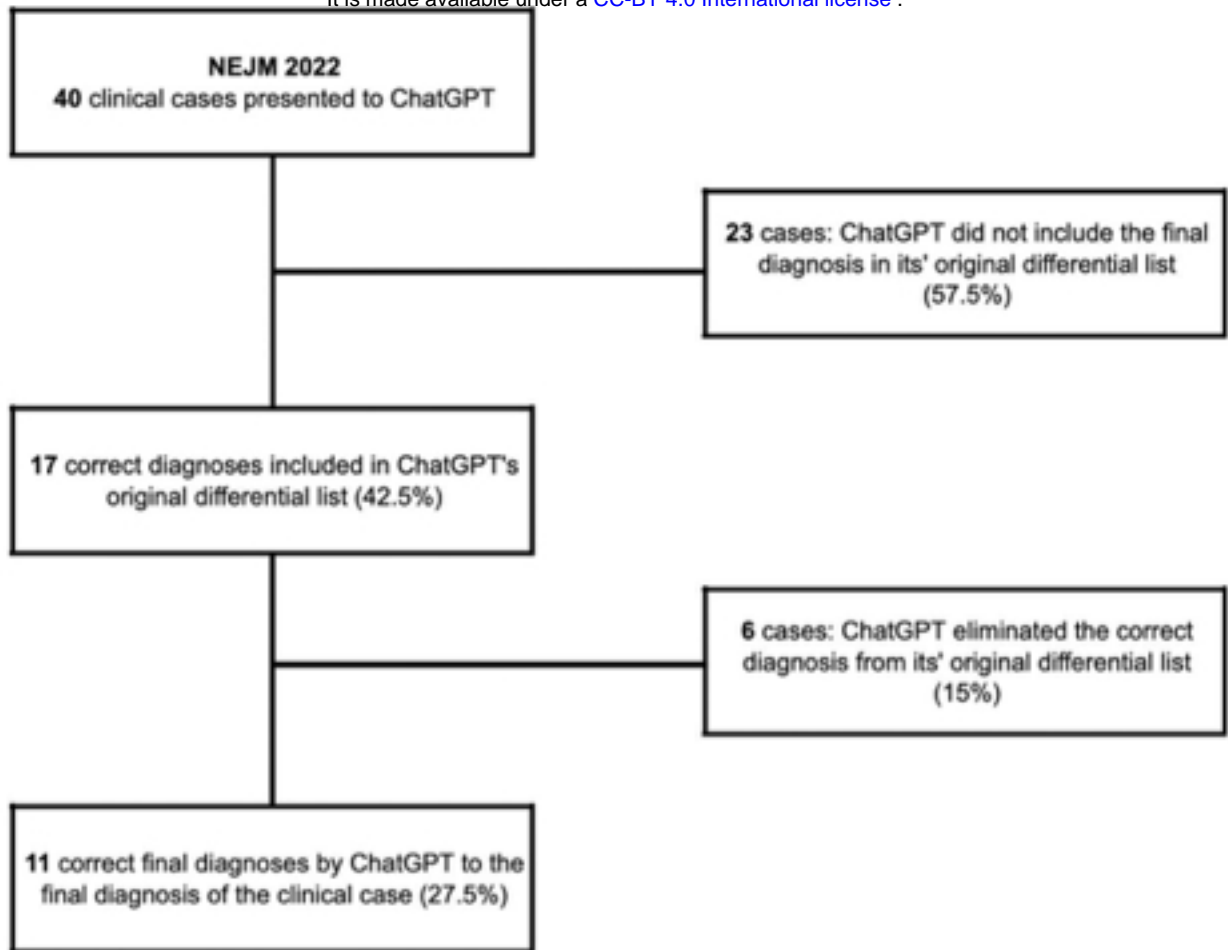


Figure 1/1