Supplemental Online Content

Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med.* Published online April 28, 2023. doi:10.1001/jamainternmed.2023.1838

This supplemental material has been provided by the authors to give readers additional information about their work.

eTable 1

Our original sample included 208 exchanges. Per reviewer feedback we noted that 195 (94%) of these exchanges consisted of a single message and only a single response from a physician. 13 (6%) exchanges consisted of a single message but with 2 separate physician responses. Second responses appeared incidental, e.g., an additional response was given when a post had already been answered.

To this point, we performed sensitivity analyses showing that the results were fundamentally unchanged if we used the original sample or an abridged sample consisting of only single patient/physician exchanges. For readability the revised manuscript uses the abridged data, which we describe as **"**when a physician replied more than once we only considered the first response, although the results were nearly identical regardless of our decision to exclude or include follow-up physician responses" in the manuscript.

Table S1. Primary Results Excluding or Including Additional Responses Authored by Physicians

	Abridged Sample (N = 195)	Original Sample (N = 208)	Difference (Absolute)
Mean Preference (%)	78.6 (ChatGPT)	78.4 (ChatGPT)	0.2 (ChatGPT)
Mean Quality Score	4.132 (ChatGPT), 3.256 (Physicians)	4.128 (ChatGPT), 3.272 (Physicians)	0.004 (ChatGPT) 0.016 (Physicians)
Mean Empathy Score	3.655 (ChatGPT), 2.147 (Physicians)	3.686 (ChatGPT), 2.163 (Physicians)	0.031 (ChatGPT) 0.016 (Physicians)

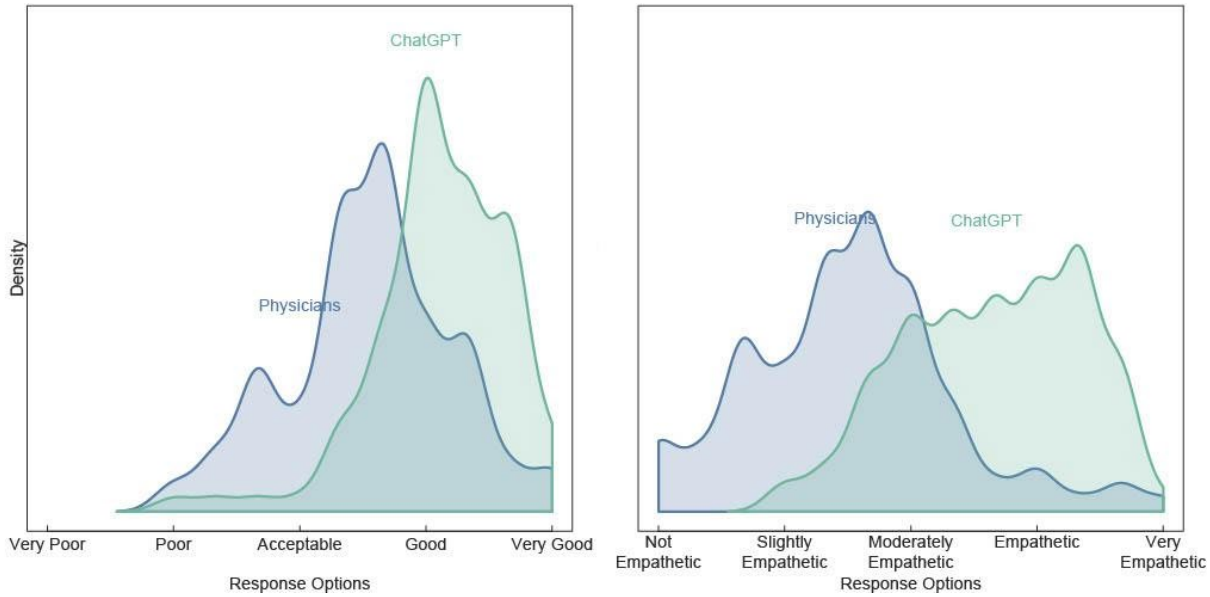
The abridged sample is the subset analyzed in the revised manuscript with a single patient message and single physician response, the original sample is that in the original submission which includes patient messages with two physician responses and treats the additional response as independent.

eFigure 1

Some of the shortest physician responses may not be indicative of an in-clinic response (despite some doctors we know auto replying “schedule an appointment” to all unsolicited patient messages). We performed sensitivity analyses comparing the subset of longer physician responses against ChatGPT, where doctors responding on r/AskDocs potentially put in the most effort, and this may be reflective of better practices in a clinic.

Longer physician responses scored higher for evaluator preference, empathy, and quality but remained statistically significantly below the alternative response authored by ChatGPT. Results for evaluator response preference, ChatGPT vs. physician, is reported in the main text. **Figure S1** shows the results for evaluators mean quality and empathy scores for ChatGPT compared to physicians. Considering the subset of physician responses longer than the median length (≥36 words), ChatGPT had significantly higher quality ($t=6.50$; $p<0.001$) and empathy scores ($t=10.31$; $p<0.001$). Even among the subset of physician responses longer than the 75th percentile of length (≥62 words) ChatGPT had significantly higher quality ($t=2.63$; $p=0.010$) and empathy scores ($t=5.86$;$p<0.001$). As a result, even for the subset of physician responses where more effort was made in responding, ChatGPT responses were preferred by our healthcare professional evaluators and had higher quality and empathy scores.

Physician responses longer than the median (≥36 words)



Physician responses longer than the 75th percentile (≥62 words)

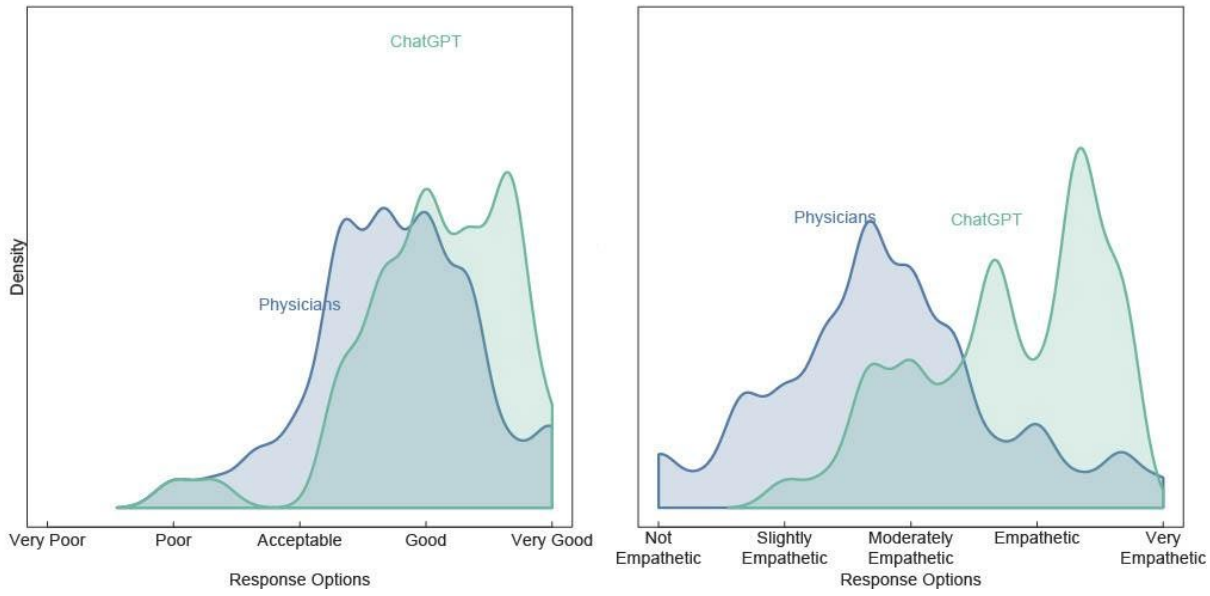


Figure S1. Sensitivity analysis comparing longer physician replies with ChatGPT