

## Examining ChatGPT's Ability to Answer Cirrhosis Related Questions in Arabic Compared to English

Authors: Jamil S. Samaan<sup>1</sup>, Yee Hui Yeo<sup>1</sup>, Wee Han Ng<sup>2</sup>, Peng-Sheng Ting<sup>3</sup>, Hirsh Trivedi<sup>1,4</sup>, Aarshi Vipani<sup>1</sup>, Ju Dong Yang<sup>1,4,5</sup>, Omer Liran<sup>6,7</sup>, Brennan Spiegel<sup>1,7</sup>, Alexander Kuo<sup>1,4</sup>, Walid Ayoub<sup>1,4#</sup>

### Affiliations

<sup>1</sup>Karsh Division of Gastroenterology and Hepatology, Department of Medicine, Cedars-Sinai Medical Center, 8700 Beverly Blvd, Los Angeles, California, 90048, USA.

<sup>2</sup>Bristol Medical School, University of Bristol, 5 Tyndall Ave, Bristol BS8 1UD, United Kingdom.

<sup>3</sup>School of Medicine, Tulane University, 1430 Tulane Ave, New Orleans, Louisiana, 70112, USA.

<sup>4</sup>Comprehensive Transplant Center, Cedars-Sinai Medical Center, 8900 Beverly Blvd, Los Angeles, California, 90048, USA.

<sup>5</sup>Samuel Oschin Comprehensive Cancer Institute, Cedars-Sinai Medical Center, 127 S San Vicente Blvd, Los Angeles, California, 90048, USA.

<sup>6</sup>Department of Psychiatry and Behavioral Sciences, Cedars-Sinai Medical Center, 8700 Beverly Pl, Los Angeles, California, 90048, USA.

<sup>7</sup>Division of Health Services Research, Department of Medicine, Cedars-Sinai Medical Center, 116 N. Robertson Blvd., Los Angeles, California, 90048, USA.

### Corresponding author:

Walid Ayoub, MD

Karsh Division of Gastroenterology and Hepatology

Cedars-Sinai Medical Center,

8700 Beverly Blvd, Los Angeles, California, 90048, USA

(310) 423-3277

Walid.ayoub@cshs.org

Manuscript word count: 1630

Abstract word count: 300/300

Number of tables: 2

### **Author Contributions**

Concept and design: Samaan, Yeo, Ayoub

Acquisition of data: Samaan, Yeo, Ng

Data review: Ayoub

Statistical analysis: Ng

Drafting of the manuscript: Samaan

Critical revision of the manuscript: All authors

All authors approved the final version of the manuscript

**Conflict of Interest Statement:** None declared.

**Funding/Support:** This research did not receive any specific grant from funding agencies in the public, commercial or not-for-profit sectors.

**Ethics Approval:** Since all responses from OpenAI were publicly available, approval from the institutional review board was not sought.

## ABSTRACT

**Background and Study Aims:** Cirrhosis is a chronic progressive disease which requires complex care. Its incidence is rising in the Arab countries making it the 7<sup>th</sup> leading cause of death in the Arab League in 2010. ChatGPT is a large language model with a growing body of literature demonstrating its ability to answer clinical questions. We examined ChatGPT's accuracy in responding to cirrhosis related questions in Arabic and compared its performance to English.

**Materials and Methods:** ChatGPTs responses to 91 questions in Arabic and English were graded by a transplant hepatologist fluent in both languages. Accuracy of responses was assessed using the scale: 1. Comprehensive, 2. Correct but inadequate, 3. Mixed with correct and incorrect/outdated data, and 4. Completely incorrect. Accuracy of Arabic compared to English responses was assessed using the scale: 1. Arabic response is more accurate, 2. Similar accuracy, 3. Arabic response is less accurate.

**Results:** The model provided 22 (24.2%) comprehensive, 44 (48.4%) correct but inadequate, 13 (14.3%) mixed with correct and incorrect/outdated data and 12 (13.2%) completely incorrect Arabic responses. When comparing the accuracy of Arabic and English responses, 9 (9.9%) of the Arabic responses were graded as more accurate, 52 (57.1%) similar in accuracy and 30 (33.0%) as less accurate compared to English.

**Conclusion:** ChatGPT has the potential to serve as an adjunct source of information for Arabic speaking patients with cirrhosis. The model provided correct responses in Arabic to 72.5% of questions, although its performance in Arabic was less accurate than in English. The model produced completely incorrect responses to 13.2% of questions, reinforcing its potential role as an adjunct and not replacement of care by licensed healthcare professionals. Future studies to

refine this technology are needed to help Arabic speaking patients with cirrhosis across the globe understand their disease and improve their outcomes.

Keywords: Arabic, cirrhosis, large language model, ChatGPT, disparity

## INTRODUCTION

Chat Generative Pre-trained Transformer (ChatGPT) is a large language model which was released by the company OpenAI in November 2022 and has exponentially grown in popularity among the general public worldwide<sup>1</sup>. This model was trained on a large dataset from a broad range of topics although the exact data source is unknown. It has the capability to comprehend user questions and respond with easy to understand, conversational and seemingly knowledgeable answers. A recent study examined the knowledge of ChatGPT related to Cirrhosis and Hepatocellular Carcinoma and found the model answered 79% of questions with correct responses, 47% of which were graded as comprehensive<sup>2</sup>. Its ability to answer clinical questions has also been examined in other topics such as bariatric surgery and cancer<sup>3,4</sup>. These studies utilized the English language when prompting ChatGPT, while data regarding its ability to comprehend and respond in other languages is limited.

Cirrhosis is a chronic progressive disease where adherence to medication regimens, lifestyle modifications, regular follow up and routine cancer screening is critical to achieving optimal outcomes. Given the complexity of care required to manage this disease, patient health knowledge is essential, with one study showing an association between poor health literacy and poor liver function<sup>5</sup>. Furthermore, the usefulness of information from online health platforms and large hepatology centers in the United States, Europe and Asia may be limited by lengthy text and complexity of language<sup>6</sup>. Although most of the large hepatology centers publish information in English, a large portion of patients seeking such information don't speak or understand English. Considering these challenges, large language models such as ChatGPT may serve as a valuable easy to understand, quick and accessible source of information for patients with cirrhosis.

Cirrhosis is on the rise in the Arab countries and ranked the 7<sup>th</sup> leading cause of death in the Arab League during the year 2010<sup>7</sup>. Furthermore, displacement and migration of citizens from the middle east sharply rose from 2005 to 2015 with doubling of migrant populations from the region<sup>8</sup>. Large language models may be leveraged to help citizens of Arab countries as well as bridge language barriers for Arabic speaking migrants around the world who have cirrhosis. We examined the ability of ChatGPT to understand liver cirrhosis related questions written in Arabic and assess the accuracy of its responses. To put its performance in context, we compared ChatGPT's performance when responding to English and Arabic questions.

## **MATERIALS AND METHODS**

### ***Question Curation***

A total of 91 questions were included in our study. Patient questions related to liver cirrhosis were collected from professional societies and institutions. Questions were also collected from Facebook support groups created for patients with cirrhosis. Details regarding question curation and selection are described elsewhere<sup>2</sup>. Questions were then translated into Arabic and verified by two authors who are fluent in Arabic and English (JS, WA).

### ***Response Generation and Grading***

ChatGPT is a large language model which was trained on a large dataset containing a broad range of topics. The model was trained using a technique called Reinforcement Learning from Human Feedback or Reinforcement Learning from Human Preference (RLHF/RLHP). This allows the model to fine tune its responses in a way that is coherent and conversational by using human feedback<sup>9</sup>. Each question was entered into the ChatGPT January 30<sup>th</sup> version in both

Arabic and English and responses were recorded. Each question was entered as an individual prompt using the “New Chat” function. The Arabic responses were then graded for accuracy by a board-certified transplant hepatologist reviewer who is fluent in English and Arabic and has more than 15 years of clinical experience in transplant hepatology (WA). The following grading scale was used to grade each response: 1. Comprehensive, 2. Correct but inadequate, 3. Mixed with correct and incorrect/outdated data, and 4. Completely incorrect. The same reviewer then assessed the quality of the Arabic compared to English responses by using the following scale: 1. Arabic response is more accurate, 2. Similar accuracy, 3. Arabic response is less accurate.

### ***Statistical analysis***

For statistical analysis purposes, questions were categorized into multiple subgroups: Basic knowledge, diagnosis, treatment, lifestyle, preventative care and others. Proportion of grades were presented as percentages overall as well as stratified by subgroups. Microsoft Excel (version 16.69.1) was used for all analysis.

## **RESULTS**

A total of 91 questions were included in our study. When examining the accuracy of Arabic responses, the model provided 22 (24.2%) comprehensive, 44 (48.4%) correct but inadequate, 13 (14.3%) mixed with correct and incorrect/outdated data and 12 (13.2%) completely incorrect responses. When stratified by subgroups, the model performed best in the lifestyle subgroup with 36.4% of responses graded as Comprehensive and 45.5% graded as correct but inadequate (Table 1). The diagnosis and treatment subgroups contained the highest proportion of completely incorrect responses with 33.3% and 31.3%, respectively, although the sample size was small in the diagnosis subgroup (n=3).

When comparing the accuracy of Arabic and English responses, 9 (9.9%) of the Arabic responses were graded as more accurate, 52 (57.1%) similar in accuracy and 30 (33.0%) as less accurate compared to English. The model performed best in Arabic compared to English in the preventative care subgroup with 25.0% of responses in Arabic graded as more accurate than English (Table 2). The diagnosis and basic knowledge subgroups contained the highest rate of the grade “Arabic is less accurate than English” with 66.7% and 44.4%, respectively, although the sample size was small in the diagnosis subgroup (n=3). For example the Arabic and English responses provided similar answers on the approach to the management of cirrhosis and vaccination. The Arabic version, however, provided a better explanation for “what is cirrhosis”. Unlike the English version, the Arabic version failed to give an accurate description of varices related to cirrhosis.

## **DISCUSSION**

The large language model ChatGPT has made patient access to artificial intelligence easier than ever and mainstream. Its ability to comprehend clinical questions and provide easy to understand and seemingly knowledgeable answers will likely make it an attractive source for patients seeking information related to their medical care. Our study builds on the current literature showing the potential utility of ChatGPT as an adjunct source of information for patients by describing its ability to answer questions related to cirrhosis in Arabic. The model’s performance in Arabic was impressive, providing correct responses to 72.5% of questions, 24.2% of which were graded as “comprehensive”. When comparing the accuracy of Arabic responses to English, Arabic responses were similar 57.1% the time but less accurate 33.0% of the time, highlighting a disparity in performance. Given the rapid advancement of this technology since its release in



November of 2022, it has the potential to serve as a valuable adjunct source of information for Arabic speaking patients around the world.

Language barriers in medicine have been previously described with studies showing worse outcomes for patients that have language discordance with their healthcare provider<sup>10,11</sup>.

Furthermore, patient's use of online resources for medical information is on the rise with one in three individuals in the United States having searched online to figure out a medical condition and one half of European Union citizens reported using the internet for information related to injury, disease, nutrition, improving health or similar<sup>12,13</sup>. Navigating search engines can be a daunting task for patients, from the time needed to find information, to identifying reliable sources and misinformation. We anticipate the use of large language models such as ChatGPT will significantly increase by patients due to its ease of use and quick access interface as well as its simple and conversational responses. This provides great urgency across the medical field to examine the utility but more importantly the limitations of this technology in order to better counsel patients. Our results are promising given the impressive performance of ChatGPT in Arabic. We anticipate significant improvement in future versions of ChatGPT and with the development of larger and more powerful language models given the rapid growth and advancement in this field.

ChatGPT has important limitations to consider. One important concept is the "hallucinations" or stochastic phenomena where the model produces incorrect statements while sounding confident, leading to increased risk of spreading misinformations<sup>14</sup>. This is demonstrated in our study by the model providing "completely incorrect" answers to 14.3% of questions despite sounding confident and knowledgeable. Due to this important limitation, we emphasize the role of ChatGPT as a potential adjunct and not replacement to care provided by licensed healthcare

professionals. There also appears to be a difference in performance in Arabic compared to English, as evident by the 33.0% of responses in Arabic graded as less accurate than English. The cause of this is unclear although we hypothesize that the source of information utilized by the model may be an important factor and is unknown at the current time. Generally there is more English literature related to diagnosis and treatment compared to lifestyle and basic knowledge which may have contributed to the disparity. It is also unclear if the model is using sources in Arabic to answer questions in Arabic, or translating information from its dataset into Arabic. Future studies examining differences in performance in languages other than English may help shed light on this disparity.

## **CONCLUSION**

ChatGPT has the potential to serve as an adjunct source of information for Arabic speaking patients with cirrhosis. The model provided correct responses in Arabic to 72.5% of questions, 24.2% of which were graded as comprehensive, although its performance in Arabic was less accurate than in English. While its performance was impressive, the model did produce completely incorrect responses to 13.2% of questions, reinforcing its potential role as an adjunct and not replacement to care by licensed healthcare professionals. Future studies to refine this technology are needed to help Arabic speaking patients with cirrhosis across the globe understand their disease and improve their outcomes.

**Acknowledgements:** None.

## REFERENCES

- [1. UBS Editorial Team. Let's chat about ChatGPT. Published online February 23, 2023. Accessed April 5, 2023. <https://www.ubs.com/global/en/wealth-management/our-approach/marketnews/article.1585717.html>
2. Yeo YH, Samaan JS, Ng WH, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. *Clin Mol Hepatol*. Published online March 22, 2023. doi:10.3350/cmh.2023.0089
3. Johnson SB, King AJ, Warner EL, Aneja S, Kann BH, Bylund CL. Using ChatGPT to evaluate cancer myths and misconceptions: artificial intelligence and cancer information. *JNCI Cancer Spectrum*. 2023;7(2):pkad015. doi:10.1093/jncics/pkad015
4. Samaan JS, Yeo YH, Rajeev N, et al. Assessing the Accuracy of Responses by the Language Model ChatGPT to Questions Regarding Bariatric Surgery. *OBES SURG*. Published online April 27, 2023. doi:10.1007/s11695-023-06603-5
5. Kaps L, Hildebrand K, Nagel M, et al. Risk factors for poorer health literacy in patients with liver cirrhosis. Pittet V, ed. *PLoS ONE*. 2021;16(7):e0255349. doi:10.1371/journal.pone.0255349
6. Kaundinya T, Mazumder N, Atiemo K, et al. Health Literacy Gaps in Online Resources for Cirrhotic Patients. *J Curr Surg*. 2020;10(1-2):1-6. doi:10.14740/jcs401
7. Mokdad AH, Jaber S, Aziz MIA, et al. The state of health in the Arab world, 1990–2010: an analysis of the burden of diseases, injuries, and risk factors. *The Lancet*. 2014;383(9914):309-320. doi:10.1016/S0140-6736(13)62189-3
8. Pew Research Center. Middle East's Migrant Population More Than Doubles Since 2005. Published online October 18, 2016.
9. openai. ChatGPT: Optimizing Language Models for Dialogue. Accessed March 15, 2023. <https://openai.com/blog/chatgpt>
10. Al Shamsi H, Almutairi AG, Al Mashrafi S, Al Kalbani T. Implications of Language Barriers for Healthcare: A Systematic Review. *Oman Med J*. 2020;35(2):e122. doi:10.5001/omj.2020.40
11. Divi C, Koss RG, Schmaltz SP, Loeb JM. Language proficiency and adverse events in US hospitals: a pilot study. *Int J Qual Health Care*. 2007;19(2):60-67. doi:10.1093/intqhc/mzl069
12. Eurostat. 53% of EU Citizens Sought Health Information Online. Published online March 27, 2020. Accessed April 4, 2023. <https://ec.europa.eu/eurostat/web/products-eurostat-news/-/ddn-20200327-1>

13. Susannah Fox, Maeve Duggan. Health Online 2013. *Pew Internet and American Life Project*. Published online January 15, 2023. Accessed April 4, 2023. [https://www.pewinternet.org/wp-content/uploads/sites/9/media/Files/Reports/PIP\\_HealthOnline.pdf](https://www.pewinternet.org/wp-content/uploads/sites/9/media/Files/Reports/PIP_HealthOnline.pdf)
14. Ge J, Lai JC. Artificial intelligence-based text generators in hepatology: ChatGPT is just the beginning. *Hepatology Communications*. 2023;7(4). doi:10.1097/HC9.0000000000000097

**Table 1.** Grading of responses generated by ChatGPT to Arabic questions related to cirrhosis categorized by subgroup.

<b>Basic Knowledge (N=36)</b>	
1. Comprehensive	19.4%
2. Correct but inadequate	55.6%
3. Mixed with correct and incorrect/outdated data	8.3%
4. Completely incorrect	16.7%
<b>Diagnosis (N=3)</b>	
1. Comprehensive	0%
2. Correct but inadequate	66.7%
3. Mixed with correct and incorrect/outdated data	0%
4. Completely incorrect	33.3%
<b>Treatment (N=16)</b>	
1. Comprehensive	12.5%
2. Correct but inadequate	31.3%
3. Mixed with correct and incorrect/outdated data	25.0%
4. Completely incorrect	31.3%
<b>Lifestyle (N=22)</b>	
1. Comprehensive	36.4%
2. Correct but inadequate	45.5%
3. Mixed with correct and incorrect/outdated data	18.2%
4. Completely incorrect	0%
<b>Preventative Care (N=4)</b>	
1. Comprehensive	0%
2. Correct but inadequate	50.0%
3. Mixed with correct and incorrect/outdated data	50.0%
4. Completely incorrect	0%
<b>Others (N=10)</b>	
1. Comprehensive	50.0%
2. Correct but inadequate	50.0%
3. Mixed with correct and incorrect/outdated data	0%
4. Completely incorrect	0%

**Table 2.** Grading of responses comparing the accuracy between Arabic and English responses generated by ChatGPT to Cirrhosis related questions categorized by subgroup.

<b>Basic Knowledge (N=36)</b>	
1. Arabic response is more accurate	2.8%
2. Similar accuracy	52.8%
3. Arabic response is less accurate	44.4%
<b>Diagnosis (N=3)</b>	
1. Arabic response is more accurate	0%
2. Similar accuracy	33.3%
3. Arabic response is less accurate	66.7%
<b>Treatment (N=16)</b>	
1. Arabic response is more accurate	12.5%
2. Similar accuracy	43.8%
3. Arabic response is less accurate	43.8%
<b>Lifestyle (N=22)</b>	
1. Arabic response is more accurate	13.6%
2. Similar accuracy	72.7%
3. Arabic response is less accurate	13.6%
<b>Preventative Care (N=4)</b>	
1. Arabic response is more accurate	25.0%
2. Similar accuracy	50.0%
3. Arabic response is less accurate	25.0%
<b>Others (N=10)</b>	
1. Arabic response is more accurate	20.0%
2. Similar accuracy	70.0%
3. Arabic response is less accurate	10.0%

