A Network Analysis Approach to Understanding Medical Claims by ChatGPT: Where is the Information Coming From?

Oscar Yuan-Jie Shen, Jayanth Sairam Pratap, Xiang Li, Neal C Chen[1], Abhiram Bhashyam
[1]Massachusetts General Hospital

INTRODUCTION:

ChatGPT has multiple potential applications in medicine, and represents a potential paradigm shift in how information is accessed. Currently, the most popular method of obtaining information is through search engines like Google Search. Unlike Google, ChatGPT is non-deterministic, i.e., the information provided is different each time, and it is a text generator rather than a text retriever. The lay public can be expected to increasingly use ChatGPT as a source of medical information in addition to or as a replacement to traditional search engines. In this study, we aim to investigate: 1) the sources of information that ChatGPT tends to use in response to medical inquiries, and 2) the reliability of ChatGPT responses depending on variations in the level of academic consensus for a particular topic. To study this, we performed a text network analysis comparing the similarity of ChatGPT responses to Google Search results for 3 topics in orthopaedic surgery, determined by the authors to have high, medium, and low consensus in the existing medical literature, respectively.

METHODS:

We used repeated identical queries of ChatGPT, along with the same queries with Google Search, to explore the relationship of ChatGPT's answers to the top 20 Google Search results using a text network analysis.

The following three questions were investigated:

1. What is the cause of carpal tunnel syndrome (CTS)? (High consensus)
2. What is the cause of tennis elbow (TE)? (Medium consensus)
3. Platelet rich plasma (PRP) for thumb arthritis? (Low consensus)

Each question was entered verbatim into ChatGPT-3·5 and 4 each 20 times.

The questions were also entered verbatim into Google without being logged in, and the top 20 search results were identified. Websites were assigned the classification of manuscript for academic papers, academic for academic institutions or organizations, government for websites sponsored/run by the government, or private for any other type of website.

After text processing, the text was converted into a network graph. Each ChatGPT response and Google source was defined as a node with edges based on overlap in words between the nodes. A pairwise similarity metric was generated for each pair of nodes in the network using term-frequency inverse-document frequency (TF-IDF). Clustering was performed using Leiden community detection.

The mean TF-IDF similarity was computed for each of the Google Search results. Averaged across all ChatGPT responses, the aggregate weight vector was sorted to obtain the ranked similarity of each of the search results with the aggregated ChatGPT node.

RESULTS:

The number of academic, government, private, or manuscript Google Search results varied greatly between the three questions. The websites with the highest TF-IDF similarities to aggregate GPT responses remained almost the same between versions, with at least 8 of the top 10 websites being the same.

Using a paired t-test to compare the TF-IDF similarities between GPT-4 and 3·5 for each website for CTS, TE, and PRP, there was a significantly higher TF-IDF similarity for GPT-4 compared to 3·5 with a p-value (95% Confidence Interval) of <0·001 (0·42-1·18), 0·001 (0·41-1·40), and <0·001 (0·44-1·22), respectively.

When stratified by type of website, the network demonstrates which category most closely resembles ChatGPT's responses. For CTS, the highest TF-IDF similarity was Academic websites For TE, the TF-IDF similarity is highest for Private websites. For PRP, the TF-IDF similarity is highest for Private websites. These findings were consistent across GPT versions.

DISCUSSION AND CONCLUSION:

Our study illustrates that the information ChatGPT provides for queries on a specific topic is closely correlated to the top websites for the same query on Google. For conditions like CTS, where there is a widely accepted medical consensus and detailed, patient-oriented resources from a range of academic and government institutions, ChatGPT responses cover topics similar to these more reliable resources and provide a reasonable average of websites on Google. When fewer reliable resources are available in the case of TE and PRP, ChatGPT becomes increasingly similar to non-academic sources and less representative of the information generally available through Google Search. Furthermore, when comparing GPT-4 to 3·5, we can see improvements in TF-IDF similarity to websites regardless of the source, suggesting that the newest version is more representative of the information available through Google Search.

ChatGPT and other pre-trained language models (PLM) are in rapid development, with some developers creating versions specifically for medical information. Ensuring accurate, precise information while preventing misinformation is essential for building trust in these PLMs for both clinicians and patients. Some strategies to mitigate the risks of PLMs include 1) organizing across medical institutions to create more curated websites across a broader range, 2) medical societies should be involved in the development of PLMS and offer guidance to patients on how to interpret results from search engines and PLMs, 3) further understanding and characterizing limitations of PLMs when synthesizing information/documents within medicine, and 4) study the feasibility of GPTs trained on a more constrained and knowledge-guided dataset or algorithms that weight certain datasets more heavily when considering medical information.

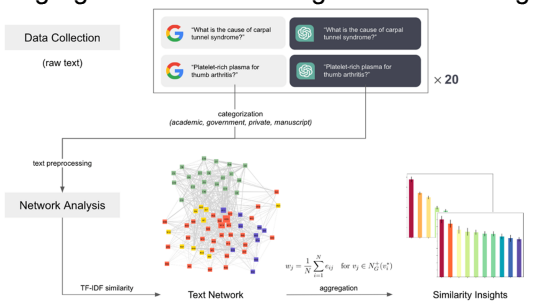

Figure 1. Summary diagram of the study, from the data collection to the analysis stage.

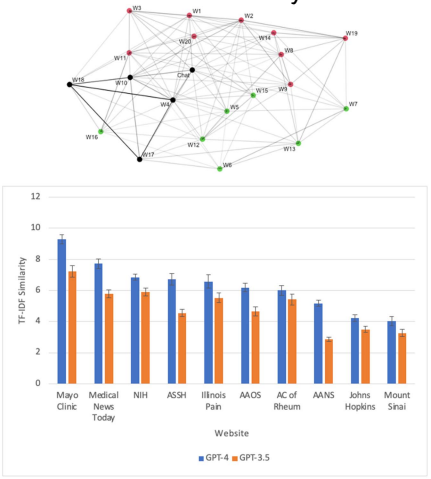

Figure 2. Aggregated text network *(top)* and top 10 information sources, ranked by mean TF-IDF similarity *(bottom)*, for "What is the cause of carpal tunnel syndrome?"