Abstract

The advent of accessible and user-friendly large language models (LLMs), such as ChatGPT, Meta AI, or Google Gemini has opened the possibility for use in research, particularly in labor-intensive tasks requiring multiple researchers to manually parse data. One such example of this is in use in systematic reviews and meta-analysis (SR/MA), which both summarize large amounts of relevant studies on a particular topic in order to answer a research question. This project aims to evaluate the use of ChatGPT to parse large amounts of clinical trial papers and extract relevant details such as interventions, sample size, and attrition rate for each intervention. This is done using the data from a recent meta-analysis on medications for generalized anxiety disorder (GAD), using the authors manually parsed data to judge accuracy. Using several packages for text extraction, table extraction, and optical character recognition as well OpenAI's API for batch querying ChatGPT (GPT-4o), prompt engineering was conducted on training data before a final prompt was tested on test data. The output from the test data, organized into a JSON was then compared with the manually parsed data to judge accuracy. Using this workflow, ChatGPT-4.0 was shown the ability to extract data with an overall accuracy of XX%. [short discussion of results].

Introduction

As of 2024, Large Language Models (LLMs) have become increasingly accessible by researchers across various sectors through the use of conversational applications. This front-facing application allows researchers to prompt complex tasks to a LLM as if talking to another person, allowing for feedback, corrections, and memory recollection. Notable examples of these conversational applications include ChatGPT, Google Gemini, and Meta AI. Several studies and reviews have been made for the use of LLMs in helping automate parts healthcare research, emphasizing the need for accuracy and ethics while acknowledging the superiority over traditional natural language processing methods.

One study in 2024 assessed the use of ChatGPT for extracting structured data from clinical notes, in particular pathological classifications from pathology reports. These pathology reports were first transformed from a PDF file to a text format, and then ChatGPT was queried to extract structured data using the API from it's parent company, OpenAI. Data was split into two sets for training and testing, and different kinds of queries were tested on the training data before a final query was used for testing. This editing of queries used is known as prompt engineering and was found to significantly impact accuracy. The final query was found to extract pathological classifications with an overall accuracy of 89%, which outperformed the performance of two traditional NLP methods (WordPiece and Named Entity Recognition (NER)). The results from this study showcase the feasibility of using ChatGPT and other conversational applications to infer and extract complex data from healthcare text.

This study aims to assess the abilities of the newest GPT engine (GPT-4o) to extract structured data features from clinical trial reports. The clinical trials come from a recent meta-analysis on clinical trials on medications for Generalized Anxiety Disorder (GAD). The data was split in two groups in order to develop the final prompt and general training on the first group and test the final prompt on the second group. The accuracy was evaluated by competing the data extracted by ChatGPT to the manually parsed data from the meta-analysis. The ChatGPT output and errors were then analyzed to understand how ChatGPT interprets data from clinical trials.

Results

Data Context

This study aims to assess the use of ChatGPT on another form of healthcare text, clinical trial reports, in order to assess its accuracy in extracting important features such as sample size and attrition rate. For this study I selected the clinical trial data used in a recent meta-analysis on treatments for Generalized Anxiety Disorder (GAD), using the authors own manual data extractions as the basis for accuracy. The data is comprised of 30 studies that each test one or more treatments for GAD in Double-Blind Randomized Controlled Trials, using the [anxiety scale measure]. [add more about the meta analysis here]. [10 of the studies will be used for prompt engineering and general training, and 20 of the studies will be used to test the final prompt and return results].

The specific task given to ChatGPT is to identify main author and year of publication, interventions (and dosing), sample size, main race, percentage of female participants, mean HAMA, attrition rate, sponsor, and diagnosis criteria from a clinical trial report. [feature discussion]

Overall Performance

Using the training data comprised of 8 randomly assigned studies, accuracy testing was done with iterative prompt engineering to create a final prompt. This final prompt was used to evaluate the GPT-4o model were gathered using the testing data comprised of [] randomly assigned studies.

Inference and Interpretation (Case report)

Error Analysis (Case Report then individual feature analysis)

Effect Strength

Sample Size

Drug Identification

Notes

Analyzing irregularities

Reproducibility evaluation

Prompt engineering and training

The 8 studies randomly assigned for training were experimented with to improve the prompt and improve the parsing into Python after generation. The training data was also tested using a semi-final prompt, and received an overall accuracy of 0.89. The two metrics that GPT-4o struggled with the most during training were Female % and Attrition Rate, where it scored an accuracy of 0.71 and 0.58, respectively. The full training data extraction can be found in Train_Data_Extration.ipynb, and the full training accuracy scoring can be found in Train_Data_Analysis.ipynb.

Below are some of the largest improvements from the first prompt and final prompt, and how that impacted the output generated by GPT-4o.

Output format

The initial prompt given to ChatGPT simply provided extracted text a certain study, and asked it to extract a list of certain characteristics from the text "in a structured format". This simple prompt led to ChatGPT returning output in extremely varied formats and interpretations, such as bringing together treatments into one summarized output, or choosing to omit the Placebo results. While these outputs were technically meeting the instructions provided, it made it impossible to compare results across studies and judge it for accuracy by comparing it with the manually extracted characteristic. Note while this sort of varied formats is acceptable for manual review in small quantities, it's difficult to analyze further without carefully providing the structure required. To address this issue, an example structure was specified as well as the data-interchange format (JSON) for parsing into Python for future analysis. An example of this is This can be seen in the final prompt beginning at "Example Format:", with an example provided being "Last Name of Main Author and Year: 'Doe et al., 2021'". This led to fairly standardized outputs from ChatGPT, making it much easier to parse and compare to manually extracted data.

Manually Extracted Data Review

The original manually extracted data set used to compare the generated data set was pulled directly from a figure in the meta-analysis by Kong, 2021, as this is where the selected studies are from. The comparison with the generated output in this study was under the assumption that there were no omissions (as there are none mentioned in the paper), and all the data in the manually extracted data could be found in the PDF for each study. This turned out to be false upon testing, as for example, for Coric et al., 2010, GPT-4o correctly identified Pexacerfont 100 mg/day as one of the interventions in this study, but it was absent in the manually extracted data set. This omission is not noted anywhere in the Kong 2021 study, and had to be added into the manually extracted data set. There were also cases were data for a certain study in from the manually extracted data set could not be found upon manual review of the study, and thus could not be found by GPT-4o, such as in Mezhebovsky 2012, leading to a omission. These changes were documented in the Train_Data_Extration file provided in the GitHub repository. To ensure this would not occur in the testing set, the manually extracted data set and all the studies listed were manually reviewed and edited if necessary. The manually extracted data set lightly edited for training, and the manually extracted data set reviewed completely before testing are available as train_truth_excel and test_truth_excel in the GitHub repository, respectively.

Additional Instructions

Certain characteristics extracted from a study, such as intervention or sample size, benefit from additional instruction to ensure a format or interpretation. For example, certain interventions are divided into three doses a day, and are thus written in the report as "Drug 10 mg/tid" with tid being an abbreviation for "three times a day". Since in the manually extracted data set, all doses were converted to once a day, or "mg/day", this caused the GPT-4o to fail at identifying the correct interventions. To fix this, adding an additional instruction for this characteristic after the example format helped fix this issue. For the case of intervention dosing, the additional instruction was "'Intervention' should be in mg/day, not any other unit of measurement". Additional instructions were included for ensuring the placebo was included and the sample size referred to the sample size for the entire study, and can be seen in the final prompt as the lines after the example format.

Table and Image Extraction

One of characteristics GPT-4o struggled the most with was attrition rate. This is defined as the % of participants in every intervention that dropped out before the study was over. In most cases, GPT-4o instead reported the percentage of participants that dropped out due to adverse effects, as often a section in most clinical trials. While this was partially improved by additional instruction, there was a clear need for additional extracted input, as attrition rate was often put in tables and visuals. For these reasons, the final prompt includes not just extracted text from the clinical trial, but extracted tables and full pdf images of the clinical trials, allowing GPT-4o to consider this in its reasoning. This is seen in the final prompt as the input before the example format, and is discussed more fully in the Inference and Interpretation (Case Study) section.

Discussion

LLMs with conversational applications, such as ChatGPT, Google Gemini, and Meta AI have been tested in the last couple years for use in labor intensive research tasks, such as parsing and organizing large amounts of data. These LLMs may find a new use as a tool in meta-analysis studies that require researchers to manually read through thousands of clinical trials to assess relevance and extract key data. This study aimed to assess the use of ChatGPT in parsing clinical trial reports in order to understand the overall accuracy and effect of prompt engineering on accuracy.

In clinical research, a meta-analysis uses statistical techniques to combine the results of multiple clinical trials to generate an average result that may better estimate the effects of a particular intervention. Meta-analysis is an especially important tool for understanding the efficacy of medications for major depressive disorder and generalized anxiety disorder, as the specific pharmacodynamics behind the first line treatment, selective serotonin reuptake inhibitors (SSRIs), is still not fully understood. One 2006 study, although challenged upon its publication, estimates that 70% of the effect of an anti-depressant medication is due to placebo. This makes it difficult to prove the efficacy of novel medications for psychiatric conditions. The issue of the placebo effect is exacerbated in clinical trials with a small sample size and uncertainty regarding effective medication dosage. Meta-analysis allows researchers to pool the results from multiple clinical studies in order to have a better understanding of the true effect of an anti-depressant, as well as compare the efficacy of a potential SSRI compared to other drugs in its medication class using individual studies.

One of the largest issues stopping meta-analysis is the labor required in manually reading though hundreds or thousands of clinical trials reports in order to extract data. A meta-analysis usually starts by querying large databases, such as PubMed or Google Scholar, with general queries that aim to capture as many relevant studies as possible. A reference system, such as EndNote library, is then used to remove duplicates, and then reports are screened manually by researchers to exclude reports based on criteria. A meta-analysis on meta-analysis studies found that the average time to complete and publish a meta-analysis is approximately 67.3 weeks, with a mean yield rate of studies reviewed to be 2.94%. ChatGPT, the conversational application behind a powerful LLM, may be able to automate these manual processes. However, there is the large issue of accuracy and bias, especially if the meta-analysis is used for advising clinical practice. In a prior study, ChatGPT was found to be able to automate the first manual process, the selection of clinical trials using the titles and abstracts, with a recall greater than 0.9 when using the prompt that performed the best during training, reaching equivalent recall to manual selection. This study aimed to evaluate the use of ChatGPT in the second manual process, the extraction of key features from the selected papers, as well as to further understand the effect of prompt engineering on task accuracy.

Our data supports the use of ChatGPT to accurately process large volumes of clinical trial data and organize it into a structured format without requiring model training or human annotation. [Additionally, the amount of time spent parsing the test data (x amount) took only [time amount] compared to the approximate [time amount] spent by the author and their team to manually parse the data]. These results support the use of LLMs by researchers to parse large amounts of clinical trials reports, lessening the need for manual review and data extraction. Further work is required to understand how accuracy may differ across different types of data and subjects.

There should also be a strong consideration for privacy implications when using ChatGPT and other LLM's. While using these conversational application LLMs allows researchers to easily access extremely powerful pretrained models where the brunt of technology and software requirement are shifted to the providing company, there is an issue of trusting the parent company to protect the data sent. This issue is especially crucial when using medical data, which is governed by strict regulations like HIPPA. While this study is focused on the use of LLMs for meta-analysis, which use public studies as input data, the continued use of LLMs in clinical research may lead to the adaptation of LLM usage in private healthcare settings.

In conclusion, this study and several other recent studies show that ChatGPT and other conversational application LLMs are powerful and accessible tools for researchers to analyze large amounts of healthcare documents. With further studies confirming the accuracy and efficacy of these models, they may become a regular tool of researchers to transform and analyze data.

Methods

Data processing

Each clinical trial report was downloaded as a PDF document and then placed into a folder labeled "testing" or "training" after random assignment. These folders on a local computer, which text, tables, and images were extracted from using Python packages before given to GPT-40 along with a prompt using the OpenAI API.

Python package openai was used to connect to the OpenAI API, select the model (GPT-4o), and prompt ChatGPT to analyze all the PDFs in the "training" folder and extract data in a certain format based on the prompt used. When a final prompt was selected, ChatGPT was prompted to analyze the PDFs in the "testing" folder. The output from the "testing" folder was then used for accuracy metrics.

The final prompt used was selected by iteratively adding and removing subtasks and phrases based on the accuracy on the "training" studies, as well as literature review. This iterative process is known as "prompt engineering," and the initial prompt was heavily inspired by the works of Huang and Xiangming, who each created papers querying ChatGPT for research purposes. The prompt was loosely structured in three parts. The first part is explaining the objective of the task, and all the extracted text, tables, and images from the clinical trial. The second part is a detailed explanation of the output structure wanted with examples. The third part is additional instructions to ensure a standard format and interpretation. The output was specified as in JSON format, and then compared with a excel file manually inputted with the correct data using the Python Pandas package. This study used the latest LLM model from OpenAI, “GPT-4o”, which was queried using the OpenAI API. GPT-4o scored higher on benchmark tests compared to prior models GPT-4 and GPT-3.5. GPT-4 was estimated to have around 1.8 trillion parameters, and GPT-4o is estimated to have a similar or possibly larger amount. All GPT models are pre-trained on various public and authored documents, and corporate clients may further train the model with sensitive documents. The exact code used to select the “GPT-4o” model and query it are available at github.com/brandonorodriguez.

Model evaluation

The performance of ChatGPT and [other models] were evaluated by comparing their output with the manually extracted data provided in the meta-analysis. Classification accuracy, F1, Kappa, recall, and precision were recorded for each attribute of interest, and accuracy and confusion matrices were provided for attributes. The notes attribute was manually discussed in order to assess accuracy and relevance.

Data availability

The 30 clinical trial studies used for this study are [list of referenced studies].

Code availability

The code used in every method tested in this study are available in the walkthrough for this study available at [filename] at github.com/brandonorodriguez. Any additional information or data required is available from the corresponding author upon request.

References

Bagde, H., Dhopte, A., Alam, M. K., & Basri, R. (2023). A systematic review and meta-analysis on ChatGPT and its utilization in medical and dental research. *Heliyon, 9*(12), e23050. https://doi.org/10.1016/j.heliyon.2023.e23050

Borah, R., Brown, A. W., Capers, P. L., & Kaiser, K. A. (2017). Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. *BMJ Open, 7*, e012545. https://doi.org/10.1136/bmjopen-2016-012545

Cai, Y., Deng, Q., & Lv, T. (2024). Impact of GPT on the academic ecosystem. *Science & Education.* https://doi.org/10.1007/s11191-024-00561-9

Cai, X., Geng, Y., Du, Y., Westerman, B., Wang, D., Ma, C., & Vallejo, J. J. G. (2023). Utilizing ChatGPT to select literature for meta-analysis shows workload reduction while maintaining a similar recall level as manual curation. *medRxiv.* https://doi.org/10.1101/2023.09.06.23295072

CNET. (2024). GPT-4o and Gemini 1.5 Pro: How the new AI models compare. *CNET.* https://www.cnet.com/tech/services-and-software/gpt-4o-and-gemini-1-5-pro-how-the-new-ai-models-compare/

Frontiers in Pharmacology. (2019). Sample size calculations for trials: Review and considerations. *Frontiers in Pharmacology.* https://www.frontiersin.org/journals/pharmacology/articles/10.3389/fphar.2019.01701/full

Gupta, M., Akiri, C., Aryal, K., Parker, E., & Praharaj, L. (2023). From ChatGPT to ThreatGPT: Impact of generative AI in cybersecurity and privacy. *IEEE Access, 11*, 80218–80245. https://doi.org/10.1109/ACCESS.2023.3300381

Huang, J., Yang, D. M., Rong, R., et al. (2024). A critical assessment of using ChatGPT for extracting structured data from clinical notes. *npj Digital Medicine, 7*, 106. https://doi.org/10.1038/s41746-024-01079-8

Kong, W., Deng, H., Wan, J., Zhou, Y., Zhou, Y., Song, B., & Wang, X. (2020). Comparative remission rates and tolerability of drugs for generalised anxiety disorder: A systematic review and network meta-analysis of double-blind randomized controlled trials. *Frontiers in Pharmacology, 11*, 580858. https://doi.org/10.3389/fphar.2020.580858

Kwong, J. C. C., Wang, S. C. Y., Nickel, G. C., et al. (2024). The long but necessary road to responsible use of large language models in healthcare research. *npj Digital Medicine, 7*, 177. https://doi.org/10.1038/s41746-024-01180-y

Michelson, M., & Reuter, K. (2019). The significant cost of systematic reviews and meta-analyses: A call for greater involvement of machine learning to assess the promise of clinical trials. *Contemporary Clinical Trials Communications, 16*, 100443. https://doi.org/10.1016/j.conctc.2019.100443

Moncrieff, J., Cooper, R. E., Stockmann, T., et al. (2023). The serotonin theory of depression: A systematic umbrella review of the evidence. *Molecular Psychiatry, 28*, 3243–3256. https://doi.org/10.1038/s41380-022-01661-0

OpenAI. (2023). *GPT-4 Technical Report.* arXiv. https://arxiv.org/pdf/2303.08774.pdf

OpenAI. (n.d.). GPT-4o overview. *OpenAI.* https://openai.com/index/hello-gpt-4o/

PyPI. (n.d.). *OpenAI package.* Retrieved December 7, 2024, from https://pypi.org/project/openai/

PubMed. (2020). Patient-level pooled analysis of enzalutamide efficacy and safety in men with metastatic hormone-sensitive prostate cancer. *PubMed.* https://pubmed.ncbi.nlm.nih.gov/32153391/

Radford, A., et al. (2018). Improving language understanding by generative pre-training. OpenAI. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

Raman, R. (2020). Statistical methods in handling placebo effect. In N. P. Witek, C. G. Goetz, & G. T. Stebbins (Eds.), *International Review of Neurobiology* (Vol. 153, pp. 103–120). Academic Press. https://doi.org/10.1016/bs.irn.2020.04.004

Stahl, S. M. (1998). Mechanism of action of serotonin selective reuptake inhibitors. Serotonin receptors and pathways mediate therapeutic effects and side effects. *Journal of Affective Disorders, 51*(3), 215–235. https://doi.org/10.1016/s0165-0327(98)00221-3

Turner, E. H., & Rosenthal, R. (2008). Efficacy of antidepressants. *BMJ (Clinical Research Edition, 336*(7643), 516–517. https://doi.org/10.1136/bmj.39510.531597.80

Wu, X., Duan, R., & Ni, J. (2023). Unveiling security, privacy, and ethical concerns of ChatGPT. *ArXiv, abs/2307.14192.*

Wu, X., Duan, R., & Ni, J. (2024). Unveiling security, privacy, and ethical concerns of ChatGPT. *Journal of Information and Intelligence, 2*(2), 102–115. https://doi.org/10.1016/j.jiixd.2023.10.007

Zia, A., Aziz, M., Popa, I., Khan, S. A., Hamedani, A. F., & Asif, A. R. (2022). Artificial intelligence-based medical data mining. *Journal of Personalized Medicine, 12*(9), 1359. https://doi.org/10.3390/jpm12091359