Caring is Sharing – Exploiting the Value in Data for Health and Innovation
M. Hägglund et al. (Eds.)

doi:10.3233/SHTI230347

How Good Is ChatGPT for Medication Evidence Synthesis?

Hao LIU[a], Yifan PENG[b] and Chunhua WENG[a,1]
[a]*Dept. of Biomedical Informatics, Columbia University, New York, NY, USA*
[b]*Dept. of Population Health Sciences, Weill Cornell Medicine, New York, NY, USA*

Abstract. With its seeming competence to mimic human responses, ChatGPT, an emerging AI-powered chatbot, has spurred great interest. This study aims to explore the role of ChatGPT in synthesizing medication literature and compare it with a hybrid summarization system. We tested ten medications' effectiveness with reference to their definitions and descriptions extracted from DrugBank. ChatGPT could generate coherent summaries that are not backed by evidence. In contrast, our approach can provide a highly structured and concise synthesis of related evidence, but the resulting summary is not as fluent and convincing as ChatGPT. Therefore, we recommend integrating both techniques to achieve the best performance.

Keywords. Summarization, ChatGPT, Natural Language Processing

1. Introduction

Multi-document summarization is a technique used to create a summary from multiple documents or textual sources. It has long been used for evidence synthesis [1]. Recently, the large language models, exemplified by the recent sensational ChatGPT [2], have taken advantage of pre-training hundreds of billions of parameters on a large corpus of text and have achieved impressive performance in many NLP tasks such as QA or summarization. For example, when asked, "What is the most effective treatment for patients with advanced lung cancer?" ChatGPT can retrieve relevant articles from its training data, summarize the key findings, and provide a summary: "*The most effective treatment for advanced lung cancer can depend on several factors, including the type and stage of cancer, the patient's overall health and medical history, as well as their preferences and goals of care. 1. Chemotherapy … 2. Targeted therapy, … 3. Immunotherapy … 4, Radiation therapy… 5. Surgery…".* The variety of potential text summarization applications and the complexity of model training make ChatGPT particularly appealing for the medical literature synthesis [3]. In this study, we present the first comparative study of ChatGPT and a hybrid multi-documentation summarization method on medication evidence synthesis. We use randomized controlled trials (RCTs) abstracts on PubMed as lengthy documents because they are considered the most reliable source for robust medical evidence for clinical question answering [4] and evidence-based medicine [5]. Our findings suggest the potential and limitations of

[1] Corresponding Author: Department of Biomedical Informatics, Columbia University, 622 West 168th Street, PH-20, Room 407, New York, NY 10032, USA. E-mail: chunhua@columbia.edu.

using ChatGPT for medication evidence synthesis and shed light on the future directions for developing evidence synthesis systems.

2. Method

2.1. Clinical questions on drug effectiveness

We consulted with clinicians and collected ten questions (Table 1) on the effectiveness of drugs for COVID-19, Alzheimer's disease, kidney diseases, and rheumatic diseases.

Table 1. Ten clinical questions on drug effectiveness.

Topic	Questions
COVID-19	Is Hydroxychloroquine (HCQ) effective for treating COVID-19 patients? Is Remdesivir effective for treating COVID-19 patients? Is Tocilizumab effective for treating patients with COVID-19?
Alzheimer's disease	Is Galantamine effective for improving cognitive function for Alzheimer's disease patients? Is Donepezil effective for improving cognitive function for Alzheimer's disease patients?
Rheumatic diseases	Is Tofacitinib effective for treating patients diagnosed with rheumatoid arthritis? Is Belimumab effective for inducing renal remission in patients diagnosed with Systemic Lupus Erythematosus (SLE)? Is Rituximab effective for inducing clinical remission in patients diagnosed with Antineutrophil cytoplasmic antibody (ANCA) vasculitis? Is Cyclophosphamide effective for inducing clinical remission in patients diagnosed with Antineutrophil cytoplasmic antibody (ANCA) vasculitis?
Kidney disease	Is Tolvaptan effective for treating Autosomal dominant polycystic kidney disease (ADPKD)?

2.2. The proposed document summarization system

Our system consists of five modules (Figure 1). **(1) Document collection.** We retrieved PubMed abstracts for 189,648 clinical trial publications by identifying PMIDs labeled "Randomized Controlled Trial" between January 2010 and October 2021. Metadata, such as title, abstract, and metadata, were extracted as JSON files. **(2) Document retrieval.** We employed a two-step approach to retrieve the top-*k* relevant PubMed abstracts for each clinical question, similar to the method used in VERT5ERINI [6]. First, we treated the input question as a "bag of words" and retrieved a set of *n* candidate articles using the BM25 scoring function [7]. Then, we employed an advanced encoder-decoder model (T5 [8]) to estimate the relevance of each candidate article to the question. The top-k articles, ranked by their estimated relevance score, were returned. **(3) Sentence extraction**. Here, we selected the most relevant sentence from each abstract relevant to the question. We used the same T5 model to rank sentences in a study and

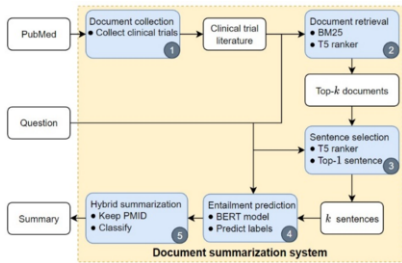

Figure 1. Overview of the document summarization

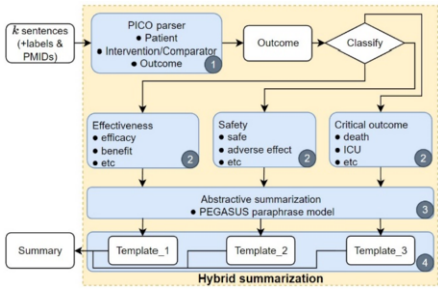

Figure 2. The overview of the hybrid summarization module

chose the top-one sentence as the rationale sentence. **(4) Entailment prediction.** We developed an entailment prediction module that predicts the relationship between the question and the highly ranked sentences. We treated it as a multiclass classification problem, where the output indicates whether a given rationale sentence Supports, (is) Neutral, or Refutes the question. We used a pre-trained model PubMedBERT [9], fine-tuned on entailment datasets. **(5) Hybrid summarization.** This module automatically generates a summary based on all selected sentences from extracted studies via a combination of four components: PICO (Population, Intervention Comparison, and Outcome) parser, study classification, abstractive summarizations, and template formation (Figure 2). The selected sentences were first parsed into PICO entities which are widely-used knowledge representations for clinical questions posed in the natural language [10]. Based on the Outcome entity identified, each sentence is classified into the Effectiveness, Safety, and Critical outcome categories, respectively. Then an abstractive summarization model PEGASUS was used to generate a partial summary of sentences from each category [11]. The three partial summaries were then organized and formatted in a coherent and readable summary following pre-defined templates.

2.3. The comparison study

To get ChatGPT's report on drug effectiveness, each question in Table 1 was queried on the ChatGPT (https://chat.openai.com/chat, version Jan 9, 2023) with responses recorded. In addition, the same questions were also queried against a newly proposed document summarization system (Section 2.2).

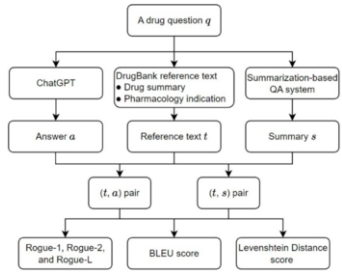

Figure 3. Evaluation process

The generated summaries are compared to the reference texts manually extracted from DrugBank [12] (Figure 3). DrugBank is a widely adopted, free-access, public website that provides information on drugs, including drugs' chemical, pharmacological, and pharmaceutical properties. First, we selected a drug's overall summary and its pharmacological indication as the description of a drug's effectiveness (i.e., reference text). Then, we paired the reference text with the summaries generated by ChatGPT or our proposed method. Finally, we calculated Rouge [13], BLEU [14], and Levenshtein Distance [15] scores. These three metrics are commonly used to evaluate the automatic summarization of texts against a set of reference texts.

3. Results and Discussion

Table 2 shows that ChatGPT achieved consistently higher Rouge-1, Rouge-2, and Rouge-L scores than the summarization-based method. The two methods achieved similar averaged BLUE scores (0.041 vs. 0.038). Our method outperformed ChatGPT with a lower averaged Levenshtein distance (1.42 vs. 1.45). Table 3 shows the description of the drug Galantamine in DrugBank and summaries generated by ChatGPT and our method. ChatGPT only described the function of Galantamine and its target disease and warned users of its usage. In contrast, our method was able to generate reports with references to the findings in each study identified in the clinical publication database. This work employed a multi-document summarization system that summarizes

the drug's effectiveness. The findings in such a summary can be linked to research publications or data from clinical trial publications. We also tested the capability of ChatGPT to generate a summary of the effectiveness of drugs. It is important to note that the report generated by ChatGPT was based on the data provided during its training, without differentiating the validity and reliability of the data source.

Table 2. Comparison between ChatGPT and our method.

		ChatGPT	Our method
Rouge-1	Precision	0.333±0.13	0.242±0.09
	Recall	0.207±0.08	0.160±0.06
	F1	0.236±0.07	0.180±0.04
Rogue-2	Precision	0.135±0.09	0.030±0.02
	Recall	0.070±0.05	0.019±0.02
	F1	0.082±0.04	0.021±0.01
Rogue-L	Precision	0.317±0.13	0.224±0.08
	Recall	0.195±0.08	0.146±0.05
	F1	0.224±0.06	0.165±0.03
BLEU		0.041±0.02	0.038±0.01
Levenshtein Distance		1.450±0.30	1.420±0.20

Table 3. Comparison of drug effectiveness summary for DrugBank, ChatGPT, and our model.

Question	*Is Galantamine effective for improving cognitive function for Alzheimer's disease patients?*
DrugBank	Galantamine is a cholinesterase inhibitor used to manage mild to moderate dementia associated with Alzheimer's Disease. Galantamine is indicated for the treatment of mild to moderate dementia of the Alzheimer's type.
ChatGPT	According to clinical studies, Galantamine appears to have a moderate benefit in improving cognitive function in Alzheimer's disease patients, specifically on measures of attention, memory and thinking, as well as activities of daily living when compared to a placebo. However, it's important to note that results from the studies were mixed. …
Our method	The currently available evidence seems to support the use of Galantamine with a few contradictory evidence. We found five studies investigated the effectiveness of Galantamine: Two studies probably support the efficacy of Galantamine. The randomized control trial (PMID 15209643) found that after 5 months of treatment, Galantamine has a positive effect on ADL performance in patients with AD. The randomized control trial (PMID15525294) found that Galantamine is safe and effective for long-term treatment of mild-to-moderate AD. …

For the evaluation, we manually created drug questions instead of selecting questions and reference texts from the public medical QA dataset to avoid bias because ChatGPT has probably already been pre-trained on them. The results indicate that both our method and ChatGPT have advantages and limitations in summarizing evidence of drug effectiveness. ChatGPT can provide contextually relevant and personalized responses but struggles to extract detail or key information from specific clinical studies. On the other hand, our method can provide highly structured and explainable summaries of clinical studies. In light of these findings, it may be beneficial to use a combination of both summarization and neural language modeling methods to achieve a more comprehensive and accurate summary of the information. This would involve using the document summarization method to extract structured information and the ChatGPT model to generate a more human-like and nuanced summary. In practice, the choice between the two approaches will depend on the specific goals, resources of the research project, and the trade-offs between summary accuracy and comprehensibility. For medical evidence synthesis, combining summarization and neural language modeling methods can be particularly useful. Medical research involves complex terminology and concepts that require domain-specific knowledge to fully understand. Therefore, a summary that is both accurate and easy to understand is essential for medical professionals to make informed decisions.

4. Conclusions

Both ChatGPT and the proposed methods have advantages and limitations, with the former being able to mimic natural and human-like summaries and the latter being highly effective in extracting structured information with links to relevant studies. Nevertheless, combining both methods may lead to a more comprehensive and accurate summary of the drug effectiveness information. One limitation of our work is that we only sampled ten questions to evaluate the systems. A more comprehensive evaluation of large-scale datasets is needed. Next, we only evaluated the models on automatic metrics. Whether they are well-suited to evaluating zero-shot summaries is still being determined. While recent work has shown that classical reference-based scores, such as ROUGE, correlated with human preferences [16], we still need to conduct human evaluations to compare the outputs of models and collect human preferences for quality. We hope our study could encourage future work to address these limitations to further explore the potential of large language model learning on medication evidence synthesis.

Acknowledgments: This work was supported by the National Library of Medicine grant R01LM009886 and 4R00LM013001, National Center for Advancing Clinical and Translational Science award UL1TR001873, NSF CAREER Award No. 2145640, and Amazon Research Award.

References

[1] Afantenos S, Karkaletsis V, Stamatopoulos P. Summarization from medical documents: a survey. J Artificial intelligence in medicine. 2005;33:157-77.
[2] OpenAI. ChatGPT: Optimizing Language Models for Dialogue. Jan 9, 2023 ed. https://chat.openai.com/chat: OpenAI; 2023.
[3] Aydın Ö, Karaarslan E. OpenAI ChatGPT Generated Literature Review: Digital Twin in Healthcare. Emerging Computer Technologies 2. 2022:22-31.
[4] Demner-Fushman D, Lin J. Answering clinical questions with knowledge-based and statistical techniques. Computational Linguistics. 2007;33:63-103.
[5] Bellomo R, Bagshaw SM. Evidence-based medicine: classifying the evidence from clinical trials–the need to consider other dimensions. Critical Care. 2006;10:1-8.
[6] Pradeep R, Ma X, Nogueira R, Lin J. Scientific claim verification with VerT5erini. arXiv preprint arXiv:201011930. 2020.
[7] Robertson S, Zaragoza H. The probabilistic relevance framework: BM25 and beyond. Foundations Trends® in Information Retrieval. 2009;3:333-89.
[8] Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv preprint arXiv:191010683. 2019.
[9] Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH). 2021;3:1-23.
[10] Richardson WS, Wilson MC, Nishikawa J, Hayward RS. The well-built clinical question: a key to evidence-based decisions. 1995.
[11] Zhang J, Zhao Y, Saleh M, Liu P. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. International Conference on Machine Learning: PMLR; 2020. p. 11328-39.
[12] Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic acids research. 2018;46:D1074-D82.
[13] Lin C-Y. Rouge: A package for automatic evaluation of summaries. Text summarization branches out 2004. p. 74-81.
[14] Papineni K, Roukos S, et al. Bleu: a method for automatic evaluation of machine translation. Proceedings of the 40th annual meeting of the Association for Computational Linguistics 2002. p. 311-8.
[15] Yujian L, Bo L. A normalized Levenshtein distance metric. IEEE transactions on pattern analysis machine intelligence. 2007;29:1091-5.
[16] Deutsch D, Dror R, Roth D. Re-Examining System-Level Correlations of Automatic Summarization Evaluation Metrics. Seattle: Association for Computational Linguistics; 2022. p. 6038-52.