ELSEVIER

Feature

Text summarization with ChatGPT for drug labeling documents

Lan Ying [1], Zhichao Liu [1,2], Hong Fang [1], Rebecca Kusko [3], Leihong Wu [1], Stephen Harris [1], Weida Tong [1,*]

[1] FDA National Center for Toxicological Research, Jefferson, AR 72079, USA
[2] Nonclinical Drug Safety, Boehringer Ingelheim Pharmaceuticals, Inc, Ridgefield, CT 06877, USA
[3] Cellino Biotech, Boston, MA, USA

Text summarization is crucial in scientific research, drug discovery and development, regulatory review, and more. This task demands domain expertise, language proficiency, semantic prowess, and conceptual skill. The recent advent of large language models (LLMs), such as ChatGPT, offers unprecedented opportunities to automate this process. We compared ChatGPT-generated summaries with those produced by human experts using FDA drug labeling documents. The labeling contains summaries of key labeling sections, making them an ideal human benchmark to evaluate ChatGPT's summarization capabilities. Analyzing >14 000 summaries, we observed that ChatGPT-generated summaries closely resembled those generated by human experts. Importantly, ChatGPT exhibited even greater similarity when summarizing drug safety information. These findings highlight ChatGPT's potential to accelerate work in critical areas, including drug safety.

Keywords: ChatGPT; text summarization; document summarization; natural language processing (NLP); artificial intelligence (AI); large language models (LLMs); drug information; drug safety

Introduction

The need for text summarization is not only common across various scientific research endeavors, drug discovery, and drug development but is also essential in the regulatory drug review processs.[(p1),(p2),(p3)] For the latter, text summarization holds particular importance, where precise summarization of drug documents is essential for communication and decision making. Text summarization must correctly distil down essential ideas and findings from a large document or corpus of documents, which requires at a minimum language fluency and expertise in the field. It has long been recognized that natural language processing (NLP) could play a role in assisting text summarization in biomedical applications.[(p4),(p5),(p6),(p7)]

The most significant breakthrough in NLP within artificial intelligence (AI) is the development of large language models (LLMs). These models are trained using extensive corpuses to achieve a level of intelligence akin to human understanding.[(p8),(p9),(p10),(p11),(p12),(p13)] Specifically, an LLM engages in a human-like learning process, predicting not only individual words but also higher-level sentence meanings. Notably, the knowledge acquired in an LLM for one task can be fine-tuned and transferred over to other tasks, mirroring the way the human brain transfers knowledge from one activity to another.[(p14),(p15),(p16),(p17)] This revolution in knowledge transfer has greatly

https://doi.org/10.1016/j.drudis.2024.104018

enhanced NLP for diverse text analytics. Consequently, AI-based LLMs have found active applications in real-world scenarios spanning fields including chatbots, automated translations, customer experience, sentiment-based news aggregation, and language identification. The arrival of ChatGPT marked a sensational milestone, showcasing the ability of AI to perform numerous human-like tasks in the public theater, particularly in the NLP domain.[p18],[p19]

Each FDA-approved drug is required to have a prescription drug labeling document to provide relevant safety and efficacy information.[p20] These documents have been an essential part of the drug review process of the FDA.[p20],[p21] Moreover, drug labeling has been extensively used to address a broad range of drug discovery and development questions such as drug safety,[p22] drug repositioning,[p21],[p23] and precision medicine.[p24] Drug labeling documents are long and complex, with around 20 pages containing 17 main sections with 40–50 subsections, posing a significant challenge in terms of navigating through the documents to find relevant information. To address this issue, the 2006 Physician Labeling Rule (PLR) amended regulations to mandate that major labeling sections have 'Highlights' providing immediate access to a summary of key information.[p25] The summary, which is done by drug-labeling experts and provided in the labeling Highlights, includes sections on most key aspects of drug safety and drug efficacy such as Boxed Warning, Contraindications, Warnings and Precautions, Adverse Reactions, Drug Interactions, Indications and Usage, Dosage and Administration, Dosage Forms and Strength, and Use in Specific Populations. This summary–section pairwise structure of drug labeling documents offers an unprecedented opportunity to assess the utility of ChatGPT in supporting drug discovery development and regulatory science via text summarization of drug information.

ChatGPT generated summaries closely resembling those of human experts

The FDA has close to 10 000 PLR drug labeling documents that summarize key labeling sections in the Highlights hosted in the FDALabel database.[p21],[p26] Since the same drug products usually contain similar labeling information, we deduplicated the documents by the product generic name, resulting in 1730 PLR labeling documents. We then paired the summary texts from labeling Highlights with the summarization texts generated by ChatGPT for nine labeling sections: Boxed Warning (838 pairs), Adverse Reactions (1730 pairs), Warnings and Precautions (1558 pairs), Dosage and Administration (1703 pairs), Indications and Usage (1716 pairs), Dosage Forms and Strength (1730 pairs), Drug Interactions (1636 pairs), Contraindications (1730 pairs), and Use in Specific Populations (1723 pairs). A total of 14,364 pairs of summaries were analyzed. Each of these nine sections contains information on drug safety, efficacy, and other critical data for a drug product. For example, the first three sections (i.e. Boxed Warning, Adverse Reactions, and Warnings and Precautions) list adverse events associated with a drug from the clinical trials and postmarket surveillance, while Indication and Usage provides FDA-approved indications along with other relevant information (e.g. patient conditions) that might be useful for studying drug efficacy, such as drug repositioning and off-label use.

ChatGPT is an LLM-based chat model that takes a set of prompts in dialog form as input and returns model-generated answers as output. We first instructed ChatGPT (i.e. gpt-3.5-turbo-16k API) to play a role using the prompt 'You are a drug labeling reviewer who reads a paragraph from labeling documents and summarizes the most important information about adverse reactions in a few concise sentences'. Thereafter, we input section text and recorded the summary generated by ChatGPT. The analysis was conducted in July 2023.

To assess the similarity between human and ChatGPT summaries, we tokenized both summaries, which were further converted into vectors. The Bidirectional Encoder Representations from Transformers (BERT) model was used to calculate the cosine similarity score of two vectors. As shown in Figure 1a, close to 87.99% of the pairs had a similarity of between 0.8 and 1.0, of which 59.59% of pairs had more than 0.9 similarity between human and ChatGPT summaries.

As shown in Figure 1b, the similarity between humans and ChatGPT was ranked in the order of Boxed Warning >Indications and Usage = Warnings and Precautions >Dosage and Administration >Drug Interactions >Contraindications >Adverse Reactions >Use in Specific Populations = Dosage Forms and Strength. The order was largely in accordance with the length of the section text; labeling sections with shorter text like Boxed Warning had the highest similarity between ChatGPT- and expert-generated summaries. On the contrary, sections with larger text sizes and versatile text, such as 'Use in Specific Populations', clearly posed challenges for ChatGPT for accurate summarizations. Furthermore, if the text was too short, with only one or two sentences like 'Dosage Forms and Strengths', ChatGPT's performance was also significantly reduced. Clearly, there could be an optimal text size range in which ChatGPT is likely to produce better results compared with other sizes. In addition, we observed that the format of the section text also played an important role. Taking 'Drug Interactions' as an example where the drug interaction data are organized in a tabular format, although the section text is lengthy, ChatGPT performed well, achieving a higher similarity score because the information is closely grouped together.

ChatGPT reliably summarizes drug safety information from labeling documents

Boxed Warning, Adverse Reactions, and Warnings and Precautions are three important labeling sections that are widely used in research to assess drug adverse reactions[p27] and organ toxicity.[p22],[p28],[p29] The adverse events highlighted in Boxed Warning are of special significance to drug development and regulatory review since they are scientifically proven with causal assessment. While causal assessment is not required, an adverse event mentioned in Warnings and Precautions is often based on reliable scientific evidence. In contrast, Adverse Reactions usually point out adverse events based on drug class or 'guilt by association' without strong scientific support or minor adverse events. Consequently, the degree of accuracy is of most importance to Boxed Warning, followed by Warnings and Precautions and Adverse Events. As demonstrated in Figure 2, ChatGPT's performance followed the same order of Boxed Warning >Warnings and Precautions >Adverse Events.

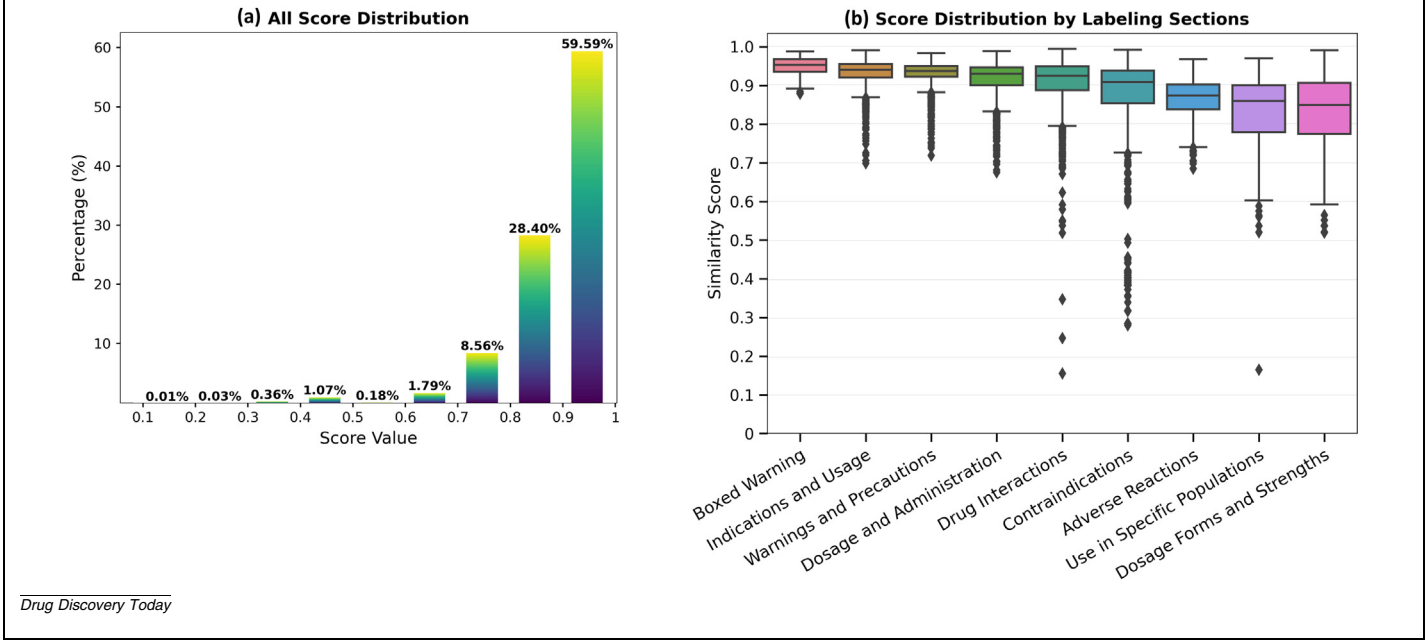

FIGURE 1

Similarity score distributions of ∼14 000 pairs: comparison between human and ChatGPT summaries by **(a)** score value and **(b)** labeling section. **(a)** The histogram represents the frequency distribution of 14,364 pairs of similarity scores. The x-axis shows the similarity scores (range: 0–1) grouped in increments of 0.1 points, while the y-axis indicates the percentage of pairs within the score range. The distribution highlights that over 87.99% of the scores are above 0.8. **(b)** The boxplot displays the similarity scores for nine labeling sections having labeling Highlights produced by humans. Each section name is displayed on the x-axis, and the score values, ranging from 0 to 1.0, are plotted on the y-axis. The results show that the Boxed Warning section has the highest similarity scores among the nine sections.

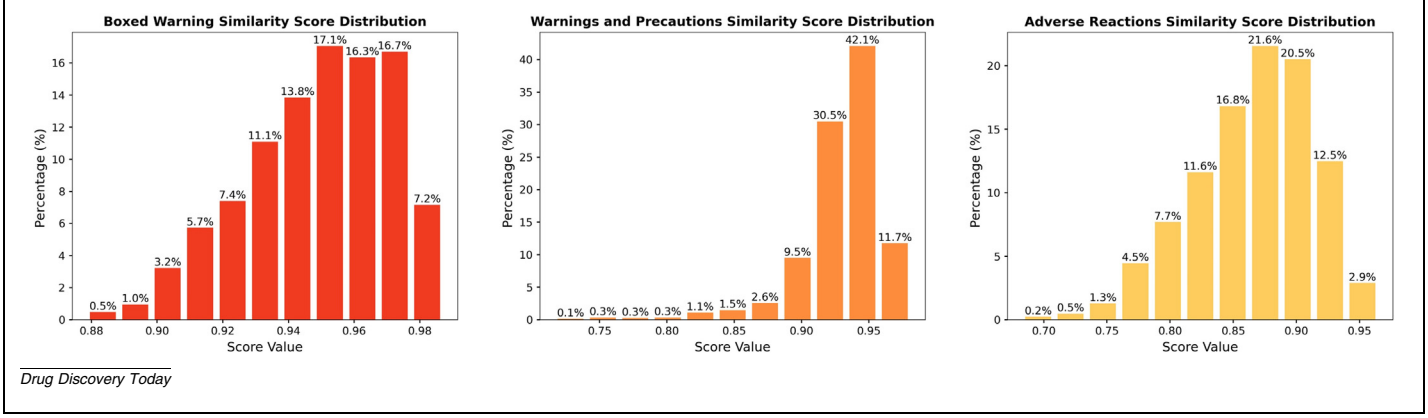

FIGURE 2

Similarity score distributions for Boxed Warning, Warnings and Precautions, and Adverse Reactions. The histogram shows the frequency distribution of similarity scores for Boxed Warning (red bar, 838 pairs), Warnings and Precautions (orange bar, 1558 pairs), and Adverse Reactions (yellow bar, 1730 pairs). The x-axis shows the similarity scores. Scores for each section are evenly grouped into 10 bins. The y-axis indicates the percentage of pairs within the score range. The figure demonstrates that the scores for the Boxed Warning section ranged from 0.88 to 1 >Warnings and Precautions section (0.70 to 1) >Adverse Reactions section (0.65 to 1).

Specifically, for Boxed Warning, >98% pairs of drug labels had a similarity score ≥0.9. The scores for Warnings and Precautions were in the range of 0.85 to 1, of which >90% were ≥0.9. For Adverse Reactions, close to 90% of pairs had similarity scores >0.8. These results demonstrated that ChatGPT could generate accurate summaries for these three labeling sections, which are important to assess drug safety.

Several representative examples with different similarity scores for each of these three sections are provided in Table 1. For Boxed Warning, we noticed that the section text starts with a capitalized warning subtitle, the importance of which was recognized by ChatGPT given its distinct format compared with the rest of text. Moreover, ChatGPT successfully included most information except, in some

TABLE 1

Comparing ChatGPT-generated summaries with original labeling Highlights summarized by human experts. Representative use cases are provided as examples with the highest, intermediate, and lowest scores for three drug safety labeling sections (Boxed Warning, Warnings and Precautions, and Adverse Reaction sections).

Example index	Similarity score	Original labeling Highlights	GPT-generated summary
1.Boxed Warning section			
Case 1: ICLOFENAC CP	0.99 (highest score)	WARNING: RISK OF SERIOUS CARDIOVASCULAR AND GASTROINTESTINAL Events See full prescribing information for complete boxed warning. Nonsteroidal anti-inflammatory drugs (NSAIDs) cause an increased risk of serious cardiovascular thrombotic events, including myocardial infarction and stroke, which can be fatal. This risk may occur early in treatment and may increase with duration of use. (5.1) Diclofenac sodium is contraindicated in the setting of coronary artery bypass graft (CABG) surgery. (4, 5.1) NSAIDs, cause an increased risk of serious gastrointestinal (GI) adverse events including bleeding, ulceration, and perforation of the stomach or intestines, which can be fatal. These events can occur at any time during use and without warning symptoms. Elderly patients and patients with a prior history of peptic ulcer disease and or GI bleeding are at greater risk for serious GI events. (5.2)	BOXED WARNING: RISK OF SERIOUS CARDIOVASCULAR AND GASTROINTESTINAL EVENTS. Diclofenac sodium, a nonsteroidal anti-inflammatory drug (NSAID), carries a heightened risk of severe cardiovascular thrombotic events such as myocardial infarction and stroke, which can be fatal. This risk is present from the early stages of treatment and increases with longer usage. It is contraindicated for individuals who have undergone coronary artery bypass graft (CABG) surgery. Additionally, NSAIDs pose an increased risk of serious gastrointestinal (GI) adverse events, including bleeding, ulceration, and perforation of the stomach or intestines, which can be fatal. These events can occur at any point during treatment and without any warning signs. Elderly patients and those with a history of peptic ulcer disease or GI bleeding are at a higher risk.
Case 2: ELAHERE	0.95	WARNING: OCULAR TOXICITY See full prescribing information for complete boxed warning. ELAHERE can cause severe ocular toxicities, including visual impairment, keratopathy, dry eye, photophobia, eye pain, and uveitis. (5.1, 6.1) Conduct an ophthalmic exam including visual acuity and slit lamp exam prior to initiation of ELAHERE, every other cycle for the first 8 cycles, and as clinically indicated. (2.3) Administer prophylactic artificial tears and ophthalmic topical steroids. (2.3, 5.1) Withhold ELAHERE for ocular toxicities until improvement and resume at the same or reduced dose. (2.4, 5.1) Discontinue ELAHERE for Grade 4 ocular toxicities. (2.4, 5.1)	OCULAR TOXICITY: ELAHERE may cause severe ocular toxicities, including visual impairment, keratopathy, dry eye, photophobia, eye pain, and uveitis. Prior to starting ELAHERE, an ophthalmic exam, including visual acuity and slit lamp exam, should be conducted. Prophylactic artificial tears and ophthalmic topical steroids should be administered. ELAHERE should be withheld in case of ocular toxicities until improvement and can then be resumed at the same or reduced dose. Discontinuation of ELAHERE is recommended for Grade 4 ocular toxicities.
Case 3: DEXTROAMPHETAMINE SACCHARATE, AMPHETAMINE ASPARTATE MONOHYDRATE, DEXTROAMPHETAMINE SULFATE AND AMPHETAMINE SULFATE	0.90	WARNING: ABUSE AND DEPENDENCE See full prescribing information for complete boxed warning CNS stimulants, including dextroamphetamine saccharate, amphetamine aspartate monohydrate, dextroamphetamine sulfate and amphetamine sulfate extended-release capsules, other amphetamine-containing products, and methylphenidate, have a high potential for abuse and dependence (5.1, 9.3) Assess the risk of abuse prior to prescribing and monitor for signs of abuse and dependence while on therapy (9.2, 9.3).	WARNING: ABUSE AND DEPENDENCE: This drug and similar CNS stimulants have a high potential for abuse and dependence. Before prescribing, evaluate the risk of abuse and monitor for signs of abuse and dependence during treatment.
2.Warnings and Precautions section			
Case 4: ATOVAQUONE AND PROGUANIL HCL	0.98 (highest score)	• Atovaquone absorption may be reduced in patients with diarrhea or vomiting. If used in patients who are vomiting, parasitemia	WARNINGS AND PRECAUTIONS: – The absorption of atovaquone may be reduced in patients with diarrhea or

TABLE 1 (*CONTINUED*)

Example index	Similarity score	Original labeling Highlights	GPT-generated summary
		should be closely monitored and the use of an antiemetic considered. In patients with severe or persistent diarrhea or vomiting, alternative antimalarial therapy may be required. (5.1) • In mixed P. falciparum and Plasmodium vivax infection, P. vivax relapse occurred commonly when patients were treated with atovaquone and proguanil hydrochloride alone. (5.2) • In the event of recrudescent P. falciparum infections after treatment or prophylaxis failure, patients should be treated with a different blood schizonticide. (5.2) • Elevated liver laboratory tests and cases of hepatitis and hepatic failure requiring liver transplantation have been reported with prophylactic use. (5.3) • Atovaquone and proguanil hydrochloride has not been evaluated for the treatment of cerebral malaria or other severe manifestations of complicated malaria. Patients with severe malaria are not candidates for oral therapy. (5.4)	vomiting. Close monitoring of parasitemia is recommended in patients who are vomiting, and the use of an antiemetic should be considered. – Vomiting occurred in up to 19% of pediatric patients and 15.3% of adults receiving atovaquone and proguanil hydrochloride. An antiemetic was successfully used in 98.3% of adults. – In cases of severe or persistent diarrhea or vomiting, alternative antimalarial therapy may be necessary. – In mixed infections with P. falciparum and P. vivax, relapse of P. vivax infection is common when treated with atovaquone and proguanil hydrochloride alone. Different blood schizonticides should be used for recrudescent P. falciparum infections or chemoprophylaxis failure. – The prophylactic use of atovaquone and proguanil hydrochloride has been associated with elevated liver laboratory tests, hepatitis, and hepatic failure requiring liver transplantation. – Atovaquone and proguanil hydrochloride has not been evaluated for the treatment of severe malaria complications, and patients with severe malaria are not candidates for oral therapy.
Case 5: PROVOCHOLINE	0.85	Healthcare provider and any other personnel involved in the administration of the methacholine challenge test should take precautions to minimize inhalation of Provocholine powder and nebulized aerosol (5.2)	The drug Provocholine carries several warnings and precautions. Firstly, it can cause severe bronchoconstriction, especially in patients with low baseline lung function. Emergency equipment and medications should be readily available to treat respiratory distress. Provocholine should not be used in patients with asthma or wheezing. If severe bronchoconstriction occurs, it should be reversed immediately with a rapid-acting inhaled β-agonist. Additionally, healthcare providers administering Provocholine should take precautions to avoid inhaling the powder and should use low resistance filters to prevent the release of Provocholine into the air. The drug is not recommended for patients with uncontrolled hypertension, aortic aneurysm, or a history of heart attack or stroke. Patients with certain conditions, such as epilepsy, peptic ulcer disease, and thyroid disease, should undergo methacholine challenge only if the benefits outweigh the potential risks.
Case 6: KLOR-CON	0.72 (lowest score)	Gastrointestinal Irritation: Take with meals. (5.1)	The drug labeling for potassium chloride (Klor-Con) includes warnings and precautions regarding gastrointestinal adverse reactions. Solid oral dosage forms of potassium chloride can cause ulcerative and stenotic lesions in the gastrointestinal tract, especially if the drug remains in contact with the gastrointestinal mucosa for a long time.

(continued on next page)

TABLE 1 (*CONTINUED*)

Example index	Similarity score	Original labeling Highlights	GPT-generated summary
			Patients with dysphagia, swallowing disorders, or severe gastrointestinal motility disorders may benefit from using liquid potassium instead. If severe symptoms such as vomiting, abdominal pain, distention, or gastrointestinal bleeding occur, Klor-Con should be discontinued and the possibility of ulceration, obstruction, or perforation should be considered. It is also advised not to take Klor-Con on an empty stomach due to the potential for gastric irritation.
3.Adverse Reactions section			
Case 7: PACLITAXEL	0.97 (highest score)	The most common adverse reactions (≥20%) in metastatic breast cancer are alopecia, neutropenia, sensory neuropathy, abnormal ECG, fatigue/asthenia, myalgia/arthralgia, AST elevation, alkaline phosphatase elevation, anemia, nausea, infections, and diarrhea. (6.1) The most common adverse reactions (≥20%) in NSCLC are anemia, neutropenia, thrombocytopenia, alopecia, peripheral neuropathy, nausea, and fatigue. (6.1) The most common (≥20%) adverse reactions of protein bound paclitaxel in adenocarcinoma of the pancreas are neutropenia, fatigue, peripheral neuropathy, nausea, alopecia, peripheral edema, diarrhea, pyrexia, vomiting, decreased appetite, rash, and dehydration. (6.1) To report SUSPECTED ADVERSE REACTIONS, contact American Regent, Inc. at 1-888-532-7998 or FDA at 1-800-FDA-1088 or https://www.fda.gov/medwatch.	The most common adverse reactions to protein-bound paclitaxel in metastatic breast cancer include alopecia, neutropenia, sensory neuropathy, abnormal ECG, fatigue/asthenia, myalgia/arthralgia, AST elevation, alkaline phosphatase elevation, anemia, nausea, infections, and diarrhea. In non-small cell lung cancer, the most common adverse reactions include anemia, neutropenia, thrombocytopenia, alopecia, peripheral neuropathy, nausea, and fatigue. In pancreatic adenocarcinoma, the most common adverse reactions include neutropenia, fatigue, peripheral neuropathy, nausea, alopecia, peripheral edema, diarrhea, pyrexia, vomiting, decreased appetite, rash, and dehydration.
Case 8: TYKERB	0.90	The most common (greater than 20%) adverse reactions during treatment with TYKERB plus capecitabine were diarrhea, palmar-plantar erythrodysesthesia, nausea, rash, vomiting, and fatigue. The most common (greater than or equal to 20%) adverse reactions during treatment with TYKERB plus letrozole were diarrhea, rash, nausea, and fatigue. (6.1) To report SUSPECTED ADVERSE REACTIONS, contact Novartis Pharmaceuticals Corporation at 1-888-669-6682 or FDA at 1-800-FDA-1088 or https://www.fda.gov/medwatch.	The most common adverse reactions observed in clinical trials of Tykerb include gastrointestinal symptoms such as diarrhea, nausea, and vomiting, as well as dermatologic symptoms such as palmar-plantar erythrodysesthesia (hand-foot syndrome) and rash. Fatigue and laboratory abnormalities in liver function tests were also common. In addition, decreased left ventricular ejection fraction, hepatotoxicity, and interstitial lung disease/pneumonitis have been reported with Tykerb use. Postmarketing experience has identified hypersensitivity reactions, nail disorders, severe cutaneous adverse reactions, and cardiac disorders as potential adverse reactions.
Case 9: TARGINIQ ER	0.69 (lowest score)	Most common adverse reactions (incidence ≥5%) were nausea and vomiting. (6.1) To report SUSPECTED ADVERSE REACTIONS, contact Purdue Pharma L.P. at 1-888-726-7535 or FDA at 1-800-FDA-1088 or https://www.fda.gov/medwatch.	The most common adverse reactions or events associated with TARGINIQ ER (oxycodone/naloxone) are addiction, abuse, and misuse; life-threatening respiratory depression; neonatal opioid withdrawal syndrome; interactions with benzodiazepine or other CNS depressants; adrenal insufficiency; severe hypotension; gastrointestinal adverse reactions; seizures; withdrawal; and drug withdrawal syndrome. Other commonly observed treatment-emergent adverse reactions include nausea,

TABLE 1 (*CONTINUED*)

Example index	Similarity score	Original labeling Highlights	GPT-generated summary
			vomiting, headache, constipation, abdominal pain, pruritus, anxiety, insomnia, back pain, lacrimation increased, fatigue, arthralgia, dizziness, rash, cold sweat, hot flush, hypertension, diarrhea, inadequate analgesia, malignant neoplasm progression, tremor, somnolence, restlessness, confusional state, dyspnea, hyperhidrosis, serotonin syndrome, and androgen deficiency. Less frequently reported adverse reactions include diarrhea, nausea, vomiting, abdominal pain, constipation, drug withdrawal syndrome, fatigue, pain, malaise, inadequate analgesia, malignant neoplasm progression, dizziness, headache, tremor, somnolence, restlessness, confusional state, anxiety, dyspnea, hyperhidrosis, pruritus, serotonin syndrome, adrenal insufficiency, anaphylaxis, and androgen deficiency.

Abbreviations: AST, aspartate transaminase; CNS, central nervous system; ECG, electrocardiogram; P. falciparum, *Plasmodium falciparum*.

instances, the specific drug names (Case 3 in Table 1). In terms of Warnings and Precautions, the section text is larger than that for Boxed Warning and thus, presumably, more difficult for ChatGPT. However, we found that ChatGPT still included almost all of the important information in the section, plus other details that were not mentioned in the Highlights.

Compared with Boxed Warning and Warnings and Precautions, the section text of Adverse Reactions is larger in size. Moreover, uniquely, the section contains lots of numbers, mainly the percentages of subjects that reported certain adverse reactions. Notably, our results show that human experts are still required to discern the threshold percentage with which an adverse reaction can be ranked by significance. Since this analytical process highly depends on expertise and knowledge, the summarization is more challenging for ChatGPT compared with the other two sections. Consequently, its similarity score was lower compared with the other two sections. In addition, ChatGPT often included significant details in the summarization beyond those summarized by human experts in the Highlights.

Challenges to overcome

While ChatGPT had promising comparison numbers, they may still not meet the threshold to replace the human reviewer. There were challenges when the section text was lengthy, and the technology struggled with drug names. For the examples with low similarity scores, manual examinations suggested that errors could have been derived from both humans and ChatGPT. While modifications are necessary to improve the quality of the dataset according to a benchmark standard, the text summarization pairs curated in this study could serve as a valuable resource for further fine-tuning of ChatGPT or other LLMs to enhance performance in regulatory applications.

In the community, concerns have been raised about the inaccuracies, unreliability, and potential fabrications that may be introduced by LLMs. In our study, we noticed that some ChatGPT-generated summaries included a lot of 'unnecessary' details not highlighted by human experts. In addition, ChatGPT's performance was affected by prompts, indicating that significant input from human experts is still required to construct a prompt with deep understanding of 'rules' or 'knowledge' about the Highlights labeling standards to achieve a result that could be used in regulatory settings.

In this study, we have applied cosine similarity based on BERT output to assess the consistency of generated results in comparison with those of human experts. We acknowledge that there are several other methods (such as BLEU and ROUGE scores) that serve a similar purpose, which may lead to variation in results but should not affect the general conclusions. Nonetheless, this benchmark dataset offers an opportunity for more statistical approaches to be tested and compared to assess the generated text.

Concluding remarks

This report marks the first assessment of ChatGPT's proficiency in summarizing drug information, utilizing the FDA Labeling document as a benchmark. We draw three main conclusions from this study:

- Facilitating drug discovery and development: The study showcases ChatGPT's capability to summarize drug information, suggesting that its future iterations, like GPT-4, may offer similar or enhanced performance in this domain.
- Enhancing drug labeling access: Drug labeling documents predating 2001 (non-PLR documents) do not have 'Highlights', such that significant effort and time must be spent by a human expert to tediously wade through a long document to find relevant information. The FDA/Center for Drug Evaluation and Research (CDER) has encouraged the voluntary submission of PLR conversion information but such conver-

sions have been limited (only 234 reported as of late 2019).[p30] ChatGPT demonstrates potential in generating summaries for non-PLR labeling documents, reducing the time and effort needed to extract relevant data.

- Supporting FDA review processes: The FDA accumulates a vast repository of regulatory documents, which are essential for decision making and provide supporting evidence. ChatGPT presents an opportunity to streamline and facilitate text summarization for these documents, superseding the manual processes currently dependent on human regulatory expertise. This study encourages further exploration of ChatGPT's role in supporting regulatory applications and helping to confirm, with humans, text summarization.

Disclaimer

The views presented in this article do not necessarily reflect those of the US Food and Drug Administration. Any mention of commercial products is for clarification and is not intended as an endorsement.

Declarations of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as potential conflicts of interest.

CRediT authorship contribution statement

Lan Ying: Writing – original draft, Formal analysis, Data curation. **Zhichao Liu:** Formal analysis, Conceptualization. **Hong Fang:** Writing – review & editing, Supervision. **Rebecca Kusko:** Writing – review & editing, Writing – original draft. **Leihong Wu:** Writing – review & editing, Formal analysis, Data curation. **Stephen Harris:** Data curation. **Weida Tong:** Writing – review & editing, Supervision, Conceptualization.

Data availability

Data will be made available on request.

Acknowledgment

We wish to acknowledge Lan Ying's contribution, which was made possible by an appointment through the Research Participation Program of the National Center for Toxicological Research administered by the US Food and Drug Administration through the Oak Ridge Institute for Science and Education.

References

1. Tang L et al. Evaluating large language models on medical evidence summarization. *NPJ Digit Med*. 2023;6:158.
2. Liu Z, Roberts RA, Lal-Nag M, Chen X, Huang R, Tong W. AI-based language models powering drug discovery and development. *Drug Discov Today*. 2021;26:2593–2607.
3. Thakkar S et al. Artificial intelligence and real-world data for drug and food safety – a regulatory science perspective. *Regul Toxicol Pharmacol*. 2023;140, 105388.
4. 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.
5. El-Kassas WS, Salama CR, Rafea AA, Mohamed HK. Automatic text summarization: a comprehensive survey. *Expert Syst Appl*. 2021;165, 113679.
6. Van Veen D et al. Clinical text summarization: adapting large language models can outperform human experts. *Res Sq*. 2023. rs.3.rs-3483777.
7. Adams G, Alsentzer E, Ketenci M, Zucker J, Elhadad N. What's in a summary? Laying the groundwork for advances in hospital-course summarization. *Proc Conf*. 2021;2021:4794–4811.
8. Brown T et al. Language models are few-shot learners. *Adv Neural Inf Process Syst*. 2020;33:1877–1901.
9. Chowdhery A et al. Palm: scaling language modeling with pathways. *J Mach Learn Res*. 2023;24:1–113.
10. Touvron H et al. Llama 2: open foundation and fine-tuned chat models. *arXiv*. 2023. preprint arXiv:2307.09288.
11. Introducing Gemini: Our Largest and Most Capable AI Model. Google.com. https://blog.google/technology/ai/google-gemini-ai/#availability. Published December 06, 2023. Accessed January 12, 2024.
12. Singhal K et al. Towards expert-level medical question answering with large language models. *arXiv*. 2023. preprint arXiv:2305.09617.
13. Anil R et al. Palm 2 technical report. *arXiv*. 2023. preprint arXiv:2305.10403.
14. Raffel C et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J Mach Learn Res*. 2020;21:5485–5551.
15. Lee J et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. 2020;36:1234–1240.
16. Wu L, Gray M, Dang O, Xu J, Fang H, Tong W. RxBERT: enhancing drug labeling text mining and analysis with AI language modeling. *Exp Biol Med*. 2023;248:1937–1943.
17. Ma T, Pan Q, Rong H, Qian Y, Tian Y, Al-Nabhan N. T-bertsum: topic-aware text summarization based on BERT. *IEEE Trans Comput Soc Syst*. 2021;9:879–890.
18. Ouyang L et al. Training language models to follow instructions with human feedback. *Adv Neural Inf Process Syst*. 2022;35:27730–27744.
19. ChatGPT. 2022. OpenAI. https://chat.openai.com/. Accessed October 13, 2023.
20. Fang H et al. FDA drug labeling: rich resources to facilitate precision medicine, drug safety, and regulatory science. *Drug Discov Today*. 2016;21:1566–1570.
21. Fang H et al. FDALabel for drug repurposing studies and beyond. *Nat Biotechnol*. 2020;38:1378–1379.
22. Wu L et al. Study of serious adverse drug reactions using FDA-approved drug labeling and MedDRA. *BMC Bioinformatics*. 2019;20:129–139.
23. Liu Z et al. AI-powered drug repurposing for developing COVID-19 treatments. *Reference Module Biomed Sci*. 2022.
24. Mehta D et al. Study of pharmacogenomic information in FDA-approved drug labeling to facilitate application of precision medicine. *Drug Discov Today*. 2020;25:813–820.
25. Labeling for Human Prescription Drug and Biological Products – Implementing the PLR Content and Format Requirements, U.S. Food & Drug Administration. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/labeling-human-prescription-drug-and-biological-products-implementing-plr-content-and-format. Published February 2013. Accessed October 24, 2023.
26. FDA. *FDALabel: Full-Text Search of Drug Product Labeling*. US Food and Drug Administration; 2023.
27. Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. *Nucleic Acids Res*. 2016;44:D1075–D1079.
28. Qu Y, Li T, Liu Z, Li D, Tong W. DICTrank: the largest reference list of 1318 human drugs ranked by risk of drug-induced cardiotoxicity using FDA labeling. *Drug Discov Today*. 2023;28, 103770.
29. Chen M, Suzuki A, Thakkar S, Yu K, Hu C, Tong W. DILIrank: the largest reference drug list ranked by the risk for developing drug-induced liver injury in humans. *Drug Discov Today*. 2016;21:648–653.
30. Brodsky E. Welcome to the CDER Prescription Drug Labeling Conference 2019. https://www.fda.gov/media/134011/download. Published December 2019. Accessed December 10, 2023.

Lan Ying [1], Zhichao Liu [1,2], Hong Fang [1], Rebecca Kusko [3], Leihong Wu [1], Stephen Harris [1], Weida Tong [1,*]

[1] FDA National Center for Toxicological Research, Jefferson, AR 72079, USA
[2] Nonclinical Drug Safety, Boehringer Ingelheim Pharmaceuticals, Inc, Ridgefield, CT 06877, USA
[3] Cellino Biotech, Boston, MA, USA

* *Corresponding author.*