



Aligning Large Language Models with Humans: A Comprehensive Survey of ChatGPT's Aptitude in Pharmacology

Yingbo Zhang^{1,2} · Shumin Ren^{1,3} · Jiao Wang^{1,3} · Junyu Lu¹ · Cong Wu¹ · Mengqiao He¹ · Xingyun Liu^{1,3} · Rongrong Wu¹ · Jing Zhao¹ · Chaoying Zhan¹ · Dan Du⁴ · Zhajun Zhan⁵ · Rajeev K. Singla^{1,6} · Bairong Shen¹

Accepted: 11 November 2024 / Published online: 20 December 2024
© The Author(s) 2024

Abstract

Background Due to the lack of a comprehensive pharmacology test set, evaluating the potential and value of large language models (LLMs) in pharmacology is complex and challenging.

Aims This study aims to provide a test set reference for assessing the application potential of both general-purpose and specialized LLMs in pharmacology.

Methods We constructed a pharmacology test set consisting of three tasks: drug information retrieval, lead compound structure optimization, and research trend summarization and analysis. Subsequently, we compared the performance of general-purpose LLMs GPT-3.5 and GPT-4 on this test set.

Results The results indicate that GPT-3.5 and GPT-4 can better understand instructions for information retrieval, scheme optimization, and trend summarization in pharmacology, showing significant potential in basic pharmacology tasks, especially in areas such as drug pharmacological properties, pharmacokinetics, mode of action, and toxicity prediction. These general LLMs also effectively summarize the current challenges and future trends in this field, proving their valuable resource for interdisciplinary pharmacology researchers. However, the limitations of ChatGPT become evident when handling tasks such as drug identification queries, drug interaction information retrieval, and drug structure simulation optimization. It struggles to provide accurate interaction information for individual or specific drugs and cannot optimize specific drugs. This lack of depth in knowledge integration and analysis limits its application in scientific research and clinical exploration.

Conclusion Therefore, exploring retrieval-augmented generation (RAG) or integrating proprietary knowledge bases and knowledge graphs into pharmacology-oriented ChatGPT systems would yield favorable results. This integration will further optimize the potential of LLMs in pharmacology.

1 Introduction

Artificial intelligence (AI) is an interdisciplinary field that trains and develops methods to simulate and extend aid to human intelligence [1–3]. In recent years, with the advancement of machine learning technology and increased computational power, AI has been extensively applied across various disciplines, including pharmacology [1–3]. Machine learning has been used to simulate drug information and parameters in this field. For instance, studies such as those by Mazumdar have explored using neural networks to estimate the drug permeability across the blood-brain barrier, yielding promising results [4]. Similarly, Li et al applied various machine-learning techniques to analyze the toxicity mechanisms of drug combinations. They found that among

the many machine-learning methods, large language models (LLMs) exhibit remarkable capabilities in several areas, demonstrating exceptional parameter learning and extraordinary knowledge-reasoning abilities [5].

In 2022, ChatGPT, a significant AI milestone, was developed as an LLM with over 175 billion parameters. Its training data encompass Large Webtext Corpora, WebText2, books, and Wikipedia content [6–8]. ChatGPT has made a substantial impact on various medical fields, including medicinal chemistry [9, 10], radiology [11], dentistry [12], and otolaryngology [13]. Following its success, prominent technology companies like Google, DeepMind, Meta, and others have entered the LLM space, releasing models such as Llama2, Claude, PaLM, and Gopher [14]. In pharmacology, a series of LLMs, including DrugChat [15], DrugGPT [16], Mol-Instructions [17], and DeepEIK [18], have been developed. These pharmacological models show great potential

Extended author information available on the last page of the article

Key Points

The emergence of general and pharmacology-focused large language models has generated new demands for comprehensive pharmacological test sets (pharmacology-LLM-test-sets).

This study proposes a pharmacology-based large language model test set named 'Pharmacology-LLM-test-set,' consisting of three tasks: basic pharmacology knowledge queries, lead compound structural optimization, and summarization and inference of pharmacological research trends.

In evaluating the 'Pharmacology-LLM-test-set' using the general large language models GPT-3.5 and GPT-4, it was found that these models exhibit significant potential for pharmacology-related queries. However, they also face challenges with knowledge hallucination, limited specialization, and randomness in systematic summarization.

Addressing issues like knowledge hallucination, limited specialization, and randomness in general large language models, exploring pharmacology-specific large language models enhanced with retrieval-augmented generation, integrated knowledge bases, or knowledge graphs represents a potential solution to these problems.

for deciphering drug structure-activity relationships, optimizing lead compound structures, and aiding in drug repurposing, among other pharmacological research areas. However, critics argue that pharmacology is a highly complex domain and caution is advised when applying LLMs in pharmacological settings. The challenge of addressing issues like fact hallucinations, knowledge hallucinations, and answer randomness is critical. Thus, the reasonable application of LLMs in pharmacology has become a pressing concern for AI researchers and pharmacologists.

Since the inception of ChatGPT, researchers have extensively focused on the potential applications of LLMs, including ChatGPT, in pharmacology. Castro et al explored ChatGPT's ability to recognize five types of compound properties: SMILES (simplified molecular input line entry system) identifiers, octanol-water partition coefficients, structural information on coordination compounds, water solubility of polymers, and molecular point groups [9]. The results showed that the accuracy of ChatGPT varied from 25 to 100%, with the variance in accuracy attributed to the knowledge sparsity of the training data. Through interactive questioning, Cloesmeijer et al assessed the potential

of ChatGPT to design population pharmacokinetic (PK) models. They found that ChatGPT could generate R code for predicting PK models, but the code contained several redundancies and errors, which were random [19].

Although several researchers have assessed the potential of LLMs in pharmacology [9, 19], it is essential to note that pharmacology is a complex, scientific, and extensively dynamic field domain [20–22]. Applying general-purpose large language models (LLMs) directly to pharmacological practice can lead to issues such as hallucinations and random errors. Exploring LLMs that integrate specialized datasets, databases, knowledge bases, and knowledge graphs (a retrieval-augmented generation [RAG] technique) can effectively mitigate these issues. Models like DrugChat enhance LLMs by incorporating graph neural networks (GNNs), enabling them to handle molecular graph inputs and facilitating multi-round, interactive Q&A sessions on compound structure-activity relationships and lead compound optimization, thus revolutionizing drug research [15]. DrugGPT and similar models leverage the ZINC20 database, which contains over two billion compounds, as a data augmentation tool to train a drug design-focused LLM based on the GPT-2 model [16]. These specialized models in pharmacology offer potential and value for solving complex tasks. However, the lack of systematic evaluation in pharmacological test tasks makes assessing their accuracy and adaptability difficult. Therefore, constructing a multidisciplinary, multipurpose, and complex pharmacology-LLM-test-set is of significant value and will have clinical application potential. This study aims to build a comprehensive pharmacology-LLM-test-set to thoroughly evaluate the potential of general LLMs, especially GPT-3.5 and GPT-4, in pharmacological research.

2 Methods

2.1 Overall Design

Constructing a comprehensive pharmacology-LLM-test-set should ideally meet and cover the needs of pharmacologists for querying pharmacological knowledge and optimizing plans. Initially, we surveyed numerous pharmacology experts with backgrounds in experimental pharmacology, clinical pharmacology, cheminformatics, pharmacogenomics, AI, or a combination of these to understand their potential needs for the LLMs test set. Based on the survey results and the need for breadth, a core team of pharmacologists (Bairong Shen, Zhajun Zhan, Dan Du, and Rajeev K. Singla) preliminarily constructed the framework of the pharmacology-LLM-test-set that includes 11 subcategories within three types of query tasks. Specifically, the first type of task aims to evaluate the ability of the LLMs to query

basic pharmacological knowledge (Fact Query), the second type assesses their capability in drug structure optimization (Strategy Summarization), and the third type focuses on evaluating the ability of the LLMs to summarize and infer trends and limitations in pharmacological research (Text Generation) (Fig. 1a).

Subsequently, based on Zero-shot (Fig. 1b) and specific text RAG (Fig. 1c), we evaluated the performance of LLMs on the pharmacology-LLM-test-set. We chose OpenAI's GPT-3.5 and GPT-4 as the baseline models for this evaluation. The reason for selecting GPT-3.5 and GPT-4 as representatives of LLMs is that they are among the earliest and most widely used general-purpose LLMs [23–25]. Moreover, evaluations based on several pharmacology and biology datasets have shown that GPT-3.5 and GPT-4 are the most outstanding general-purpose LLMs [17, 26]. These factors were the basis for our selection of GPT-3.5 and GPT-4 as this assessment uses general, foundational LLMs.

2.2 Details of the Pharmacology-LLM-Test-Set

As previously described, the pharmacology-LLM-test-set is designed to evaluate the performance of LLMs in pharmacology, focusing on essential/fundamental, drug structure optimization and systematic summarization and inference capabilities (Fig. 1a). The first task primarily assesses the ability of ChatGPT to handle factual information and attributes related to drugs, such as chemical identifiers, physical and chemical properties, pharmacological properties, drug-drug interaction information, and drug target information. We designed the tasks with the gold standard: the querying of drug identifiers, drug-drug interaction information, and others sourced from the DrugBank database [27]. Based on the distribution of molecular weights (MWs), we classified drugs in DrugBank into three categories: small, medium, and large molecules (based on quartiles). Five drugs were randomly selected for each category to query 29 primary pharmacological attributes. Drugs with MWs in the top 25% were classified as large molecules ($MW > 412.64$), those with $MWs < 255.24$ as small molecules, and the remaining drugs as medium molecules. In this study, apremilast, dequalinium, irbesartan, montelukast, and silodosin were selected as representatives of large molecules; camostat, dimetacrine, naltrexone, pefloxacin, and ropivacaine as representatives of medium molecules; and amobarbital, benzphetamine, butobarbital, chlorzoxazone, and dezocine as representatives of small molecules (Fig. S1).

We divided the primary pharmacological attributes into chemical identifiers, basic physicochemical properties, pharmacological properties, target proteins, and drug-drug interactions. For chemical identifiers, we selected five types as testing standards: IUPAC (International Union of Pure and Applied Chemistry) identifier, InChI identifier (International

chemical identifier), InChIKey identifier (Standard InChI hashes), SMILES identifier (Simplified molecular input line entry system), and molecular formula. For basic physicochemical properties, testing standards included MW, monoisotopic weight, the logarithm of the partition coefficient ($\log P$), bioavailability, and polar surface area (PSA). For pharmacological properties, we chose indications of pharmacological properties, pharmacodynamics, mechanism of action, and toxicity as testing standards. For target proteins, agonists, antagonists, blockers, inhibitors, and modulators were selected as attribute testing standards. For drug-drug interactions, ten interaction risks, including cross-tissue risk, therapeutic efficacy, and nervous system disease, were selected as potential risk indicators.

The second task aimed to determine the ability of ChatGPT to summarize drug strategies, with the primary evaluation method focused on optimizing lead compound structures. Referring to the series of articles on 'Lead compound's structure optimization strategies' published by Professor Hong Liu's team at the Shanghai Institute of Materia Medica, Chinese Academy of Sciences, from 2013 to 2021 [28–31], we investigated ChatGPT's application potential in compound structure optimization schemes. We focused on optimizing compounds with goals such as 'metabolic stability' [30], 'enhanced water solubility' [28], 'reduced cardiac toxicity' [31], and 'minimized adverse effects' [29]. For metabolic stability optimization, compounds like buspirone, paroxetine, and 8-chloro-4-(4-methylpiperazin-1-yl)benzofuro[3,2-d]pyrimidine were selected. To reduce liver toxicity, amodiaquine and ibufenac were chosen. For reducing cardiac toxicity, compounds such as 2-[[[(2R)-4-(4-fluorophenyl)-2-methylpiperazin-1-yl]methyl]-7-methoxy-[1,2,4]triazolo[1,5-c]quinazolin-5-amine and N-(2,3-dihydro-[1,4]dioxino[2,3-c]pyridin-7-ylmethyl)-1-[2-(3-fluoro-6-methoxy-1,5-naphthyridin-4-yl)ethyl]piperidin-4-amine were selected. With regard to 'enhancing water solubility', we focused on compounds like rilpivirine, 8-hydroxyquinoline, and paclitaxel (Taxol). For the selection of these subfields and lead compounds, we took into account both the importance of the subfields and the necessity of selecting these particular compounds. With regard to the subfields, we chose four areas: cardiac toxicity, hepatotoxicity, enhanced water solubility, and metabolic stability. These are urgent issues in lead compound optimization and are also the major factors affecting drug recall [32]. For the drugs, we selected paclitaxel, 10-hydroxycamptothecin, buspirone, and paroxetine as candidate lead compounds, as these are well-known or widely used antitumor drugs, anxiolytic drugs, and antidepressants. Additionally, during the selection process, we considered the performance of lead compound optimization before and after modifications. For example, paclitaxel

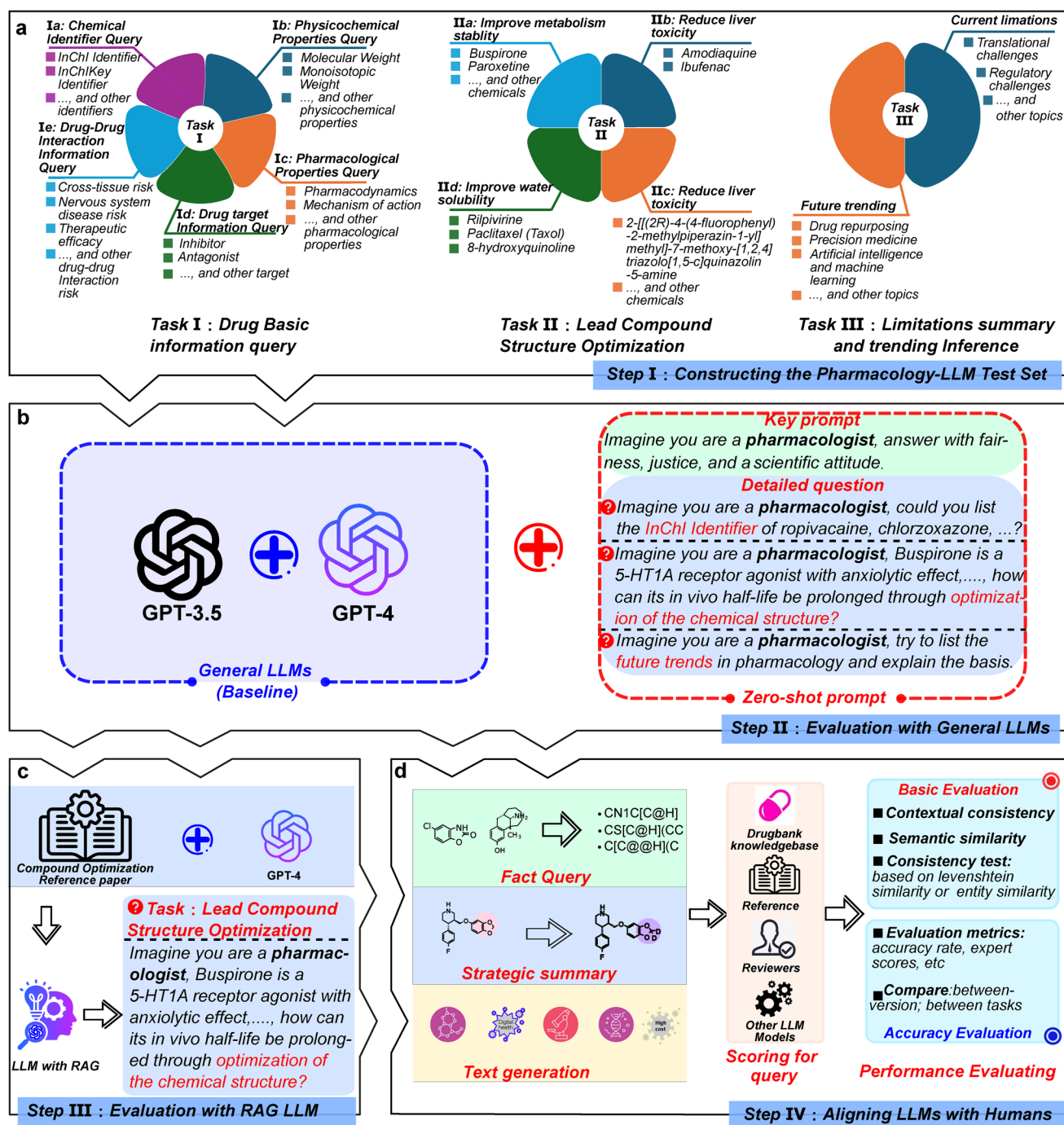


Fig. 1 The flowchart of constructing and comprehensively surveying the pharmacology-based large language model test set 'Pharmacology-LLM-test-set.' **a**: Construction and comprehensive survey of the pharmacology-based large language model test set 'Pharmacology-LLM-test-set'; **b**: Evaluation of general LLMs (including GPT-3.5

and GPT-4 in ChatGPT) based on the pharmacology LLM test set, with details including prompts for each task; **c**: Evaluation of general LLMs based on specific text Retrieval-Augmented Generation (RAG); **d**: Assessment and comparison of the performance of general LLMs in pharmacological tasks

and 10-hydroxycamptothecin, after prodrug optimization through PEGylation, glycosylation, esterification, and other modifications, showed significant improvements in their water solubility and pharmacological effects [28–31].

For the third type of task, we aimed to evaluate the text summarization capabilities of LLMs in a Zero-shot setting. We formulated evaluation tasks focused on exploring and summarizing the 'problems and limitations of current

pharmacology research’ and the ‘future directions and trends of pharmacology research.’

A substantial body of research indicates that LLMs based on the RAG approach can effectively mitigate hallucination issues when answering complex questions [33]. In this study, we attempt to construct a temporary LLM, named *PharmacologyGPT*, for lead compound optimization, using Liu et al.’s series of scientific papers on lead compound optimization (in Chinese) as the RAG data source, and GPT-4 as the base model (Fig. 1c). This model aims to explore the performance of LLMs in lead compound structure optimization under the RAG framework. Specifically, we use the online GPT-4 model as the base and follow OpenAI’s publicly available methodology for building RAG-based LLMs [34]. The purpose and scope of *PharmacologyGPT* are described as ‘primarily for pharmacological LLMs in lead compound structure optimization’. The text embedding model employs the default text-embedding-3-small model of GPT-4, and other parameters such as the data storage vector library, document segmentation parameters, data vectorization model, and associated settings that all follow GPT-4’s default configurations [34].

In addition, to assess the potential of general LLMs in pharmacology, all ‘question-answer’ tasks were introduced with the prompt ‘Imagine you are a pharmacologist; answer with fairness, justice, and a scientific attitude’ to ensure fairness, justice, and a scientific approach in this evaluation. To maintain result consistency, each task was presented in triplicate to GPT-3.5 and GPT-4 (Fig. 1b).

2.3 Evaluation of the Pharmacology-LLM-test-set Based on General LLMs (GPT-3.5 and GPT-4)

The evaluation of ChatGPT in the pharmacology-LLM-test-set was completed in two phases: basic assessment and accuracy assessment (Fig. 1d). The basic assessment aimed to explore the ability of general LLMs to understand pharmacological instructions, with evaluation content including contextual consistency, semantic similarity, and consistency tests. The accuracy assessment aimed to explore the accuracy rate of general LLMs in answering pharmacological tasks, with different evaluation benchmarks used according to the task type. For the first and second types of tasks, the evaluation benchmarks were based on actual data from the DrugBank database or on results recorded by Professor Liu. For the third type of task, which involves summarizing ‘current research limitations and future trends,’ and where there is no standard answer, the evaluation method considers both the recommendation frequency of LLMs and expert scores. Using the percentage scoring method, all three types of tasks

are evaluated by three independent evaluators (Zhaju Zhan, Dan Du, and Rajeev K. Singla) to explore the benchmark scores of general LLMs in pharmacological tasks.

We incorporated contextual consistency, semantic similarity, and consistency tests in the basic assessment phase. Specifically, we used the Reference-Free quality evaluation method reported by Xu [35] and Zhou [36] for contextual consistency and semantic similarity. It used prompt engineering and multimodal LLMs, GPT-4o and Gemini, to evaluate the contextual consistency and semantic similarity of the responses of GPT-3.5 and GPT-4. The prompt engineering followed the approach reported by Xu and Zhou, with the instruction: ‘Imagine you are a pharmacologist, using a 0- to 5-point scale to score the contextual consistency and semantic similarity of the following described answer, where a score of 0 indicates that the answer has no coherence or contextual relevance. A score of 1 to 2 indicates that the answer has some degree of coherence and contextual relevance. A score of 3 indicates a moderate level of coherence and contextual relevance. A score of 4–5 indicates good coherence and contextual relevance.’ For the consistency test, we used a reference-with-quality evaluation method combining Levenshtein Similarity and entity similarity. Specifically, we used the data from the DrugBank database or results recorded by Professor Liu as the gold standard (Reference). We first calculated the Levenshtein similarity and entity similarity between each GPT response and the reference. Then, we used Cronbach’s alpha consistency to assess the consistency of GPT’s results across different repetitions.

For accuracy assessment, we combined the accuracy rate with the percentage scoring method to evaluate different tasks. For instance, when assessing basic drug property inquiry tasks, we used the accuracy rate as the accuracy metric, where the method of calculating the accuracy rate is

$$\text{Accuracy rate} = \frac{N_{\text{correctly_predicted_task}}}{N_{\text{predicted_task}}} \times 100\%$$

Where, $N_{\text{predicted_task}}$ and $N_{\text{correctly_predicted_task}}$, respectively, represent the total number of tasks that needed to be predicted and the number of tasks correctly predicted by GPT-3.5 or GPT-4.

We use the percentage scoring method for accuracy assessment for tasks that include multiple options, such as lead compound optimization (Task II) and summarizing current research limitations and future trends (Task III). It involves accuracy evaluations performed by three independent evaluators (Zhaju Zhan, Dan Du, and Rajeev K. Singla) and uses the mean \pm standard deviation (SD) scoring method. Additionally, to compare the accuracy of GPT-3.5 and GPT-4, we used paired t-tests.

2.4 Data Statistics and Visualization Optimization

Data statistics and visualization were conducted in the R environment [37]. The distribution of drug MWs in Drug-Bank and the selection of drugs for each category were analyzed using the summary and sample functions from the dplyr package [38] in R. Additionally, other data statistics, such as sum, mean, standard error calculations, and paired t-tests, were performed using the dplyr package [38].

We employed quality evaluation methods based on Levenshtein Similarity and entity similarity for consistency tests. For Levenshtein Similarity, we used the Levenshtein distance similarity calculation method from the stringdist package [39] to obtain the Levenshtein Similarity between ChatGPT responses and the Reference. For entity similarity, we used the entity similarity calculation method in the text2vec package [40], where we first convert sentences into entity vectors (using GloVe Word Embeddings) [41] and then calculate the vector similarity between two sentences. Subsequently, we used the ltm package to calculate Cronbach's alpha index based on Levenshtein Similarity and entity similarity to assess the consistency across different repetitions of ChatGPT [42].

The visualization and plotting of the results were primarily accomplished using the ggplot2 package [43]. Furthermore, the figures were enhanced using Inkscape software [44].

3 Results

3.1 Construction of the Pharmacology-LLM-Test-Set

A comprehensive and meticulously designed set of evaluation tasks is crucial for assessing, testing, and enhancing the potential and value of LLMs in specific domains. Constructing an integrative test set that covers a wide range of tasks in pharmacology not only tests the ability of LLMs to process complex pharmacological problems but also stimulates new research and application ideas, advancing the application of AI in drug discovery and development.

To better apply LLMs in pharmacological practice, we propose the 'Pharmacology-LLM-test-set,' a test set designed to evaluate the performance of general or specialized LLMs in pharmacology. This test set consists of three tasks: fact query, strategy summarization, and text generation (Table 1). Specifically,

1. Task I: fact query assesses the LLM's performance in querying basic pharmacological information. It includes 15 compounds across biomacromolecules, mid-sized molecules, and small molecules, covering five subtasks and 18 attributes such as chemical identifiers, MW, iso-

topic mass, bioavailability, surface area, pharmacokinetics, and drug-drug interactions (Table 1).

2. Task II: strategy summarization task aimed at evaluating the potential of LLMs in chemical structure optimization. We selected ten compounds, including buspirone, paroxetine, rilpivirine, 8-hydroxyquinoline, and paclitaxel (Taxol), for optimization. The focus was on three subtasks and four strategies related to metabolic stability, reducing liver toxicity, and others, as outlined in Table 1.
3. Task III: text generation task aimed at trying to assess the ability of LLMs to extract and summarize information in pharmacological texts, focusing on summarizing limitations and trends as two subtasks (Table 1).

Moreover, to further promote the widespread use and continuous improvement of the pharmacology-LLM-test-set, we have uploaded it to both Hugging Face (<https://huggingface.co/datasets/zhangyingbo1984/Pharmacology-LLM-test-set>) and GitHub (<https://github.com/zyb1984/Pharmacology-LLM-test-set>) platforms for easy access by other users. Additionally, based on this test set, the baseline 'question-answer' scenarios and scoring outcomes for GPT-3.5 and GPT-4 can be found in the document's appendix.

3.2 Evaluation of Pharmacologica Test Set Based on General LLMs

3.2.1 The Accessibility of ChatGPT in Pharmacologica Test Set

Essential attribute evaluations, such as contextual consistency, semantic similarity, and consistency tests, are fundamental for assessing the capabilities of general LLMs like GPT-3.5, GPT-4, Llama2, Claude, PaLM, and specialized LLMs like DrugChat, DrugGPT, and Mol-Instructions in handling question-answering tasks [45]. Since LLMs do not require specialized knowledge or terminology for everyday conversations or text generation, general LLMs typically exhibit good contextual consistency and semantic similarity. However, in specialized fields, where executing question-answering tasks or generating text demands extensive professional knowledge or terminology, conducting basic attribute evaluations is the first step towards aligning human expectations with LLMs. In this study, we assessed the primary attributes of LLMs in the field of pharmacology using three fundamental attribute metrics: contextual consistency, semantic similarity, and consistency tests (including Cronbach's alpha consistency based on Levenshtein Similarity and entity similarity).

The evaluation results for contextual consistency, semantic similarity, and consistency tests indicate that

Table 1 Details of constructing the pharmacology large language model test set ‘Pharmacology-LLM-test-set’

Tasks category	Tasks subcategory	The attribute details of the query	The drug details of the query
Task I: Fact query	Task Ia: Drug chemical identifier information query	InChI identifier (IUPAC international chemical identifier) InChIKey identifier (Standard InChI hashes) IUPAC (International Union of Pure and Applied Chemistry) SMILES identifier (simplified molecular input line entry system), and molecular formula	In this test set, we selected apremilast, dequalinium, irbesartan, montelukast, and sildenafil as representatives of large-molecule drugs, camostat, dimetacrine, naltrexone, pefloxacin, and ropivacaine as representatives of medium molecule drugs, and amobarbital, benzphetamine, butobarbital, chlorzoxazone, and dezocine as representatives of small molecule drugs. The accurate chemical identifier (based on DrugBank records) information for these drugs can be found on HuggingFace and Github. For detailed information on these compounds' MW and chemical structure, please refer to Fig. S1. The performance of GPT-3.5 and GPT-4 in the drug chemical identifier tasks can be found in Table 2
	Task Ib: Drug basic properties query	Molecular weight (MW), monoisotopic weight, logS (logarithm of the solubility), logP (logarithm of the partition coefficient), bioavailability, polar surface area (PSA)	We selected the same compounds as in Tasks Ia, and the accurate property information for these drugs (based on DrugBank records) can be found on HuggingFace and Github. The performance of the LLMs GPT-3.5 and GPT-4 in drug basic properties information tasks can be seen in Fig. 3
	Task Ic: Drug pharmacological properties query	Pharmacological properties indication, pharmacodynamics, mechanism of action, and toxicity	We selected the same compounds as in Tasks Ia, and accurate information on the pharmacological properties of these drugs (based on DrugBank records) can be found on HuggingFace and Github. The performance of GPT-3.5 and GPT-4 in drug pharmacological properties tasks can be seen in Table 3
	Task Id: Drug target property query	Five types of drug action targets, including agonist, antagonist, blocker, inhibitor, and modulator	We selected the same compounds as in Tasks Ia, and the accurate target for these drugs (based on DrugBank records) can be found on HuggingFace and Github. The performance of GPT-3.5 and GPT-4 in drug target property tasks can be seen in Fig. 4
	Task Ie: Drug-drug interaction information query	Cross-tissue risk, therapeutic efficacy, nervous system disease, and ten other drug-drug adverse effects risks	We selected the same compounds as in Tasks Ia, and the accurate information on drug-drug interaction for these drugs (based on DrugBank records) can be found on HuggingFace and Github. The performance of GPT-3.5 and GPT-4 in drug-drug interaction information tasks can be seen in Fig. 5

Table 1 (continued)

Tasks category	Tasks subcategory	The attribute details of the query	The drug details of the query
Task II: Strategy summarization	Task IIa: Metabolic stability	— ^a	We selected buspirone, paroxetine, and 8-chloro-4-(4-methylpiperazin-1-yl)benzofuro[3,2-d]pyrimidine as lead compounds for optimization. More detailed information about these drugs can be found in reference 31 [30]. The performance of GPT-3.5 and GPT-4 in the metabolic stability Tasks is shown in Fig. 6
	Task IIb: Reduced toxicity	Reduced liver toxicity	We selected amodiaquine and ibufenac as lead compounds for optimization More detailed information about these drugs can be found in reference 30 [29] The performance of GPT-3.5 and GPT-4 in the reduced liver toxicity tasks is shown in Fig. 6
		Reduced cardiac toxicity	We selected 2-[[[(2R)-4-(4-fluorophenyl)-2-methylpiperazin-1-yl]methyl]-7-methoxy-[1,2,4]triazolol[1,5-c]quinazolin-5-amine and N-(2,3-dihydro-[1,4]dioxino[2,3-c]pyridin-7-ylmethyl)-1-[2-(3-fluoro-6-methoxy-1,5-naphthyridin-4-yl)ethyl]piperidin-4-amine as lead compounds for optimization. More detailed information about these drugs can be found in reference 32 [31]. The performance of GPT-3.5 and GPT-4 in the reduced liver toxicity tasks is shown in Fig. 6
Task III: Text generation	Task IIc: Enhanced water solubility	— ^a	We selected rilpivirine, 8-hydroxyquinoline, and paclitaxel (Taxol) as lead compounds for optimization. More detailed information about these drugs can be found in reference 29 [28]. The performance of GPT-3.5 and GPT-4 in the enhanced water solubility tasks is shown in Fig. 6
	Task IIIa: The current limitations of pharmacological research	— ^a	— ^a
	Task IIIb: The directions and trends of future pharmacological research	— ^a	— ^a

^aMeans 'not involved' or 'not applicable'

ChatGPT demonstrates good human alignment capabilities. Specifically, the contextual consistency score is 4.25 ± 0.63 , the semantic similarity score is 4.15 ± 0.79 , and the Cronbach's alpha consistency based on Levenshtein Similarity or entity similarity is 0.990 (0.980–0.996) and 0.987 (0.983–0.991), respectively (Fig. 2, Table S1). Further comparisons of GPT-3.5 and GPT-4 across the three tasks reveal that GPT-4 outperforms GPT-3.5 in most tasks (Fig. 2, Table S1). However, for the text summarization task (Task III), the performance difference between GPT-3.5 and GPT-4 in contextual consistency is minimal, indicating that even GPT-3.5, as a well-trained LLM, can effectively understand pharmacological instructions issued by humans.

3.2.2 The Accuracy of ChatGPT in the Drug Basic Information Query Tasks

3.2.2.1 The Accuracy of ChatGPT in the Drug Chemical Identifiers Information-Based Query Tasks A chemical identifier is a unique symbol that identifies compounds in computer systems. It plays a vital role in compound retrieval and chemoinformatics [46]. Standard chemical identifiers

include chemical formulas, IUPAC identifiers, CAS identifiers, InChI identifiers, InChIKey identifiers, SMILES identifiers, and more. The DrugBank database systematically records the chemical formula, IUPAC identifier, InChI identifier, InChIKey identifier, and SMILES identifiers of drugs are systematically recorded to standardize the basic information of collected drugs [27].

To assess the adaptability of ChatGPT in associating compound chemical identifiers, we conducted a ‘question and answer’ style query for the chemical identifiers of 15 drugs. The results indicated that, except for the chemical formula, ChatGPT (including both GPT-3.5 and GPT-4) could not provide adequate and accurate answers to the queried InChI, InChIKey, IUPAC name, and SMILES of the 15 drugs (Table 2, Fig. S2).

Among the drug identifiers that ChatGPT (including GPT-3.5 and GPT-4) could effectively answer, the average accuracy rate was $83.33 \pm 37.90\%$, with GPT-3.5 achieving an accuracy rate of $86.67 \pm 35.19\%$ and GPT-4 achieving an accuracy rate of $80.00 \pm 41.40\%$. Therefore, compared to GPT-3.5, GPT-4 did not exhibit a significant improvement in accuracy rate but demonstrated a downward trend. Upon examining the distribution of incorrect answers, they

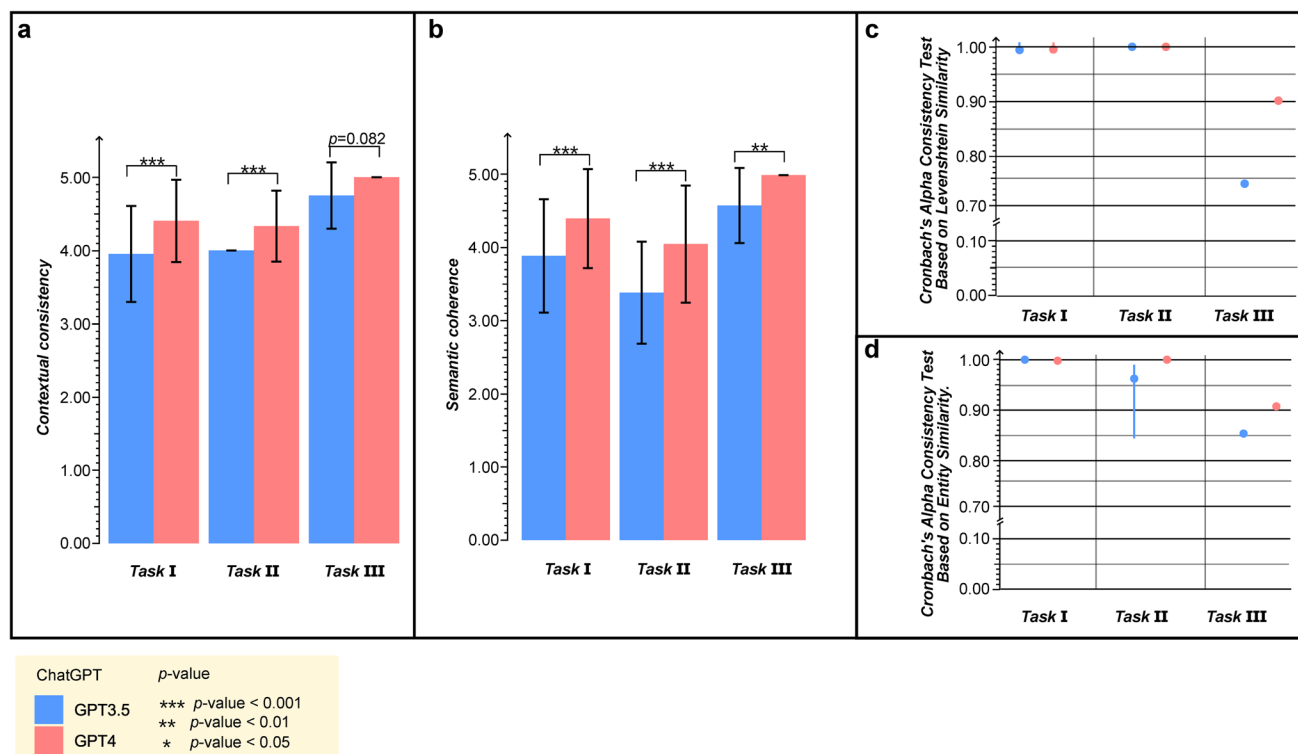


Fig. 2 The accessibility of ChatGPT in the pharmacological test set based on contextual consistency, semantic similarity, and consistency tests. **a:** The accessibility of ChatGPT in the pharmacological test set based on contextual consistency; **b:** The accessibility of ChatGPT in the pharmacological test set based on semantic similarity; **c:**

The accessibility of ChatGPT in the pharmacological test set based on Levenshtein similarity consistency tests; **d:** The accessibility of ChatGPT in the pharmacological test set based on entity similarity consistency tests

were found to be scattered. It is speculated that the reason for these incorrect answers may be associated with the frequency of drug molecular formulas in the training tasks of GPT-3.5 rather than the difficulty level of the queries (Table 2, Fig. S2).

Another critical issue that needs attention is the ‘knowledge hallucination’ with ChatGPT; i.e., when asked about the InChI identifier, InChIKey identifier, IUPAC identifier, and SMILES identifiers of the 15 drugs, GPT-3.5 and GPT-4 explicitly stated that they could not effectively answer this professional information, but instead gave seemingly reasonable but wrong answers (details in Supplementary Data 1).

3.2.2.2 The Accuracy of ChatGPT in the Drug's Physicochemical Properties Query Task The physicochemical properties of drugs significantly impact their absorption, distribution, metabolism, and excretion processes in the body. Hence, these factors also affect drug efficacy and pharmacodynamic characteristics. The physicochemical properties of common drugs in pharmacology and chemoinformatics include MW,

monoisotopic weight, logP, logD, bioavailability, and PSA [20, 21]. The DrugBank database has detailed records of the MW of collected drug, monoisotopic weight, logP, and other physicochemical properties. To test the adaptability of ChatGPT in the physicochemical properties of drugs, we conducted a ‘question and answer’ style query on the physicochemical properties of 15 query drugs. The results have shown that, except for logP and PSA, for which ChatGPT (including GPT-3.5 and GPT-4) explicitly stated its inability to answer, ChatGPT (including GPT-3.5 and GPT-4) effectively answered the other three types of physicochemical attributes (Fig. 3a).

The accuracy of predictions by ChatGPT for MW, monoisotopic weight, and bioavailability were 60.00%, 50.00%, and 53.33, respectively. For GPT-3.5, the accuracy of predictions for MW, monoisotopic weight, and bioavailability were 66.67%, 60.00%, and 60.00%, respectively. while for GPT-4, they were 53.33%, 40.00%, and 40.00%, respectively. Compared to GPT-3.5, GPT-4 exhibited no significant improvement in drug MW, monoisotopic weight, or

Table 2 The consistency performance of ChatGPT across various types of tasks in the drug chemical identifiers information-based query task^a

Query drug	Class	Score _{IUPAC} ^b	Score _{InChI} ^c	Score _{InChIkey} ^d	Score _{SMILES} ^e	Score _{Molecular Formula} ^f (%)	
						GPT-3.5	GPT-4
Amobarbital	A	0	0	0	0	100.00 ± 0.00	100.00 ± 0.00
Benzphetamine	A	0	0	0	0	0.00 ± 0.00	100.00 ± 0.00
Butobarbital	A	0	0	0	0	100.00 ± 0.00	100.00 ± 0.00
Chlorzoxazone	A	0	0	0	0	100.00 ± 0.00	100.00 ± 0.00
Dezocine	A	0	0	0	0	100.00 ± 0.00	100.00 ± 0.00
Camostat	B	0	0	0	0	100.00 ± 0.00	100.00 ± 0.00
Dimetacrine	B	0	0	0	0	100.00 ± 0.00	0.00 ± 0.00
Naltrexone	B	0	0	0	0	100.00 ± 0.00	100.00 ± 0.00
Pefloxacin	B	0	0	0	0	100.00 ± 0.00	100.00 ± 0.00
Ropivacaine	B	0	0	0	0	100.00 ± 0.00	100.00 ± 0.00
Apremilast	C	0	0	0	0	100.00 ± 0.00	100.00 ± 0.00
Dequalinium	C	0	0	0	0	100.00 ± 0.00	100.00 ± 0.00
Irbesartan	C	0	0	0	0	100.00 ± 0.00	100.00 ± 0.00
Montelukast	C	0	0	0	0	100.00 ± 0.00	0.00 ± 0.00
Silodosin	C	0	0	0	0	0.00 ± 0.00	0.00 ± 0.00
Mean		0	0	0	0	86.67 ± 35.19	80.00 ± 41.40

^aThe predictive performance is assessed using a correct/incorrect (100%/0) scoring method and reported as mean ± SD (standard deviation)

^bScore_{IUPAC} represents the prediction score of ChatGPT (GPT-3.5 or GPT-4) in the IUPAC (International Union of Pure and Applied Chemistry, IUPAC) identifiers, where a score of 100% indicates correct prediction and 0 indicates incorrect prediction

^cScore_{InChI} represents the prediction score of ChatGPT (GPT-3.5 or GPT-4) in the InChI identifiers (The IUPAC international chemical identifier, InChI), where a score of 100% indicates correct prediction, and 0 indicates incorrect prediction

^dScore_{InChIkey} represents the prediction score of ChatGPT (GPT-3.5 or GPT-4) in the InChIkey identifiers, and InChIkey is a new format directly derived from InChI, where a score of 100% indicates correct prediction and 0 indicates incorrect prediction

^eScore_{SMILES} represents the prediction score of ChatGPT (GPT-3.5 or GPT-4) in the SMILES (simplified molecular input line entry system) identifiers, where a score of 100% indicates correct prediction, and 0 indicates incorrect prediction

^fScore_{Molecular Formula} represents the prediction score of ChatGPT (GPT-3.5 or GPT-4) in the molecular formula identifiers, where a score of 100% indicates correct prediction and 0 indicates incorrect prediction

bioavailability but displayed a downward trend. Analysis of the distribution of incorrect predictions revealed a scattered distribution pattern without a high concentration in small- or large-molecule drugs (Fig. 3a). Analysis of error values for MW and monoisotopic weight showed an uneven distribution pattern, indicating that the errors were unrelated to the size of the molecule (Fig. 3b, c, d, and e).

3.2.2.3 The Accuracy of ChatGPT in Pharmacological Properties of Drugs

Pharmacological properties of drugs, such as the mechanism of action, pharmacodynamics, and toxicity, play a crucial role in elucidating and determining drug absorption, utilization, distribution, and metabolic patterns of drugs within the body [20, 21]. Understanding these properties is essential for identifying contraindications, determining dosage and administration frequency, and greatly influencing the medical application of drugs. Through a ‘question and answer’ task focusing on the fundamental pharmacological properties, pharmacodynamics, mechanism of action, and toxicity of these 15 queried drugs, the accuracy rates were found to be $93.00 \pm 20.54\%$, $85.00 \pm 3.27\%$, $88.67 \pm 6.94\%$, and $95.00 \pm 0.00\%$, respectively, showcasing higher prediction accuracy compared to other drug properties. The prediction accuracy rates for GPT-3.5 were $95.33 \pm 13.56\%$, $85.00 \pm 3.27\%$, $86.67 \pm 8.59\%$, and $95.00 \pm 0.00\%$, while for GPT-4, the rates were $90.67 \pm 26.04\%$, $85.00 \pm 3.27\%$, $90.67 \pm 4.17\%$, and $95.00 \pm 0.00\%$ (Table 3, Fig. S3).

In predicting basic pharmacological properties, the overall prediction accuracy of GPT-3.5 ($95.33 \pm 13.56\%$) was higher than that of GPT-4 ($90.67 \pm 26.04\%$). Upon detailed comparison of prediction outcomes for each compound, it has been discovered that the performance difference in predicting the basic pharmacological properties of the compound dequalinium is the primary reason GPT-3.5 has demonstrated superior predictive performance over GPT-4. According to the DrugBank database, dequalinium is used in various over-the-counter products to treat mouth infections and inflammation, such as tonsillitis, pharyngitis, and gingivitis. It is also indicated for treating bacterial vaginosis in adult women aged < 55 years in the form of vaginal tablets. It was GPT-3.5 that explicitly provided the information that dequalinium can be used as an antimicrobial and anti-inflammatory agent for treating different infections. However, GPT-4 only mentioned that dequalinium can be used as an antimicrobial agent in lozenges or mouthwashes.

In the task of action mechanism properties, the overall prediction accuracy of GPT-4 ($90.67 \pm 4.17\%$) was higher than that of GPT-3.5 ($86.67 \pm 8.59\%$). Comparing the prediction performance for each compound, GPT-4 exhibited better performance than GPT-3.5 in predicting the

pharmacodynamic properties of benzphetamine, dezocine, camostat, ropivacaine, montelukast, and silodosin (Table 3, Fig. S3).

According to the DrugBank database, benzphetamine is described as follows: ‘The mechanism of action of these drugs is not fully understood; however, it may be similar to that of amphetamines. Amphetamines stimulate norepinephrine and dopamine release in nerve endings in the lateral hypothalamic feeding center, decreasing appetite.’ This release is mediated by the binding of benzphetamine to centrally located adrenergic receptors. GPT-4 not only responded that benzphetamine could increase the release of norepinephrine in the brain (which can be used for short-term treatment of obesity), but also mentioned its similarity to amphetamines as a sympathomimetic amine. However, in GPT-3.5, although it acknowledged that benzphetamine reduces appetite and increases feelings of fullness by enhancing the release of norepinephrine, it did not respond regarding the similarity to other drugs and its use in short-term obesity treatment (refer to Table 3, Fig. S3). Similarly, in predicting the pharmacodynamic properties of dezocine, camostat, ropivacaine, montelukast, and silodosin, GPT-4 demonstrated better performance in specific details than GPT-3.5.

With regard to other pharmacological properties of drugs, such as pharmacodynamics and toxicity, GPT-3.5 and GPT-4 demonstrated similar and excellent performance, achieving accuracy rates of $85.00 \pm 3.27\%$ and $95.00 \pm 0.00\%$, respectively. No significant differences were observed between GPT-3.5 and GPT-4 in these aspects (Table 3 and Fig. S3).

Based on the data analysis of ChatGPT on fundamental pharmacological properties, drug action mechanisms, pharmacokinetics, and toxicity, it has been demonstrated that ChatGPT has a distinct advantage in text-processing tasks than those related to text-numerical association and text-text association.

3.2.2.4 The Accuracy of ChatGPT in Drug-Target Attribute Query Task

The drug's targets, such as antagonists, agonists, blockers, inhibitors, and modulators, are the proteins that drugs directly act upon and are critical to the drug's mechanism of action [20, 21]. In the ‘question-answer’ tasks for the target properties of 15 drugs, GPT-3.5 and GPT-4 exhibited varying performances across different drugs. For drugs with a single target, such as dimetacrine, pefloxacin, ropivacaine, and apremilast, both GPT-3.5 and GPT-4 demonstrated good predictive performance, achieving 100% prediction accuracy (Fig. 4 and Supplementary data 4). However, for drugs with two or more targets, except for dezocine, both GPT-3.5 and GPT-4 could not accurately predict all the targets of the drugs.

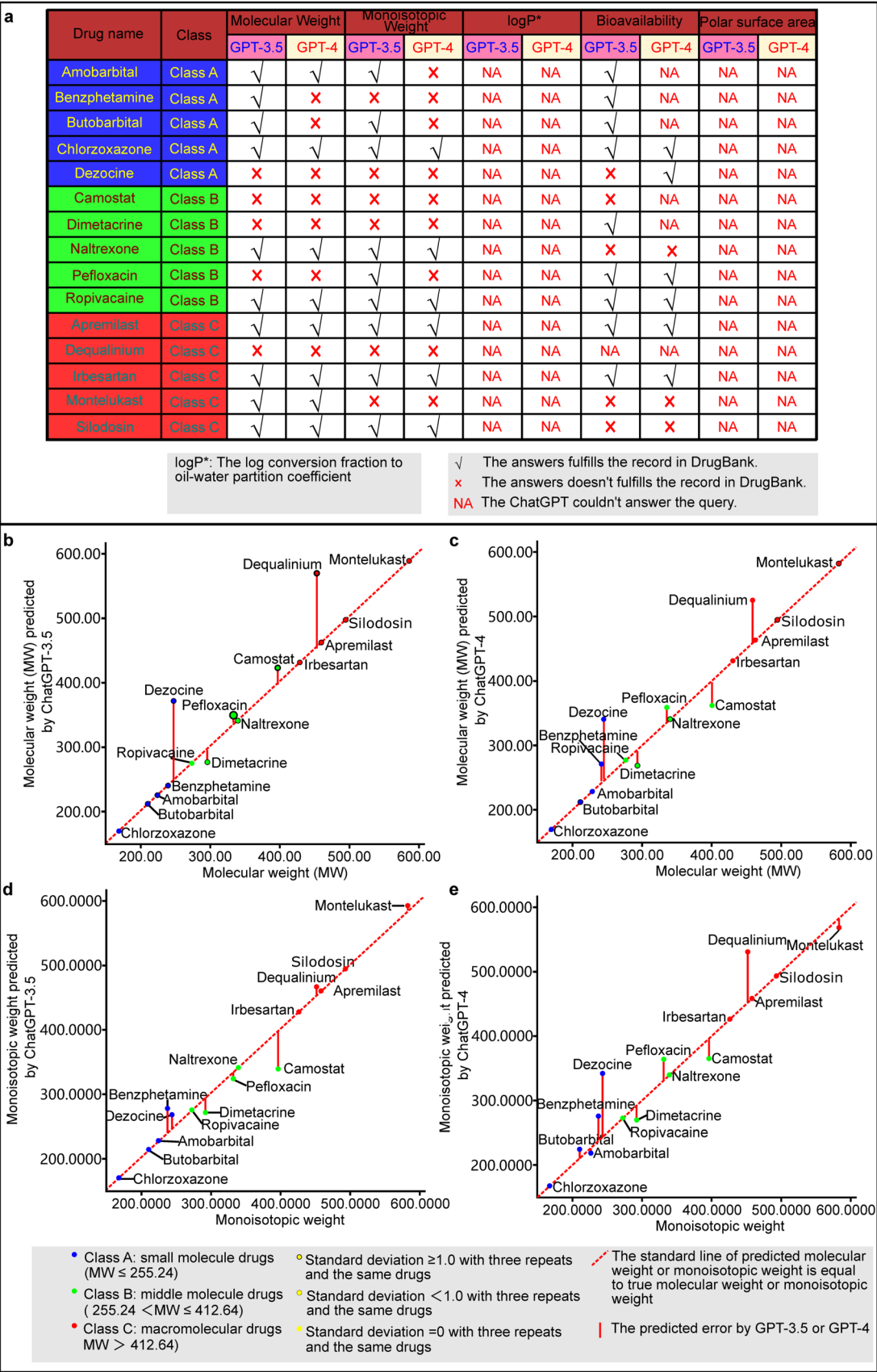


Fig. 3 Investigating the potential of ChatGPT in the ‘question and answer’ task for drug physicochemical properties. **a** Overview of the capability and accuracy of ChatGPT in answering the physicochemical properties of query drugs; **b** The consistency of the predicted molecular weight (MW) by GPT-3.5 with the molecular weight records in the DrugBank database; **c** The consistency of the predicted molecular weight (MW) by GPT-4 with the molecular weight records in the DrugBank database; **d** The consistency of the predicted monoisotopic weight by GPT-3.5 with the monoisotopic weight records in the DrugBank database; **e** The consistency of the predicted monoisotopic weight by GPT-4 with the monoisotopic weight records in the DrugBank database

For example, irbesartan, widely used to treat hypertensive patients with type 2 diabetes, relieves hypertension and reduces blood sugar levels. It has direct targets, including AGTR1 (Angiotensin II receptor type 1) and JUN (Jun proto-oncogene, AP-1 transcription factor subunit). However, GPT-3.5 and GPT-4 only recorded AGTR1 as the target for irbesartan, omitting JUN (refer to Fig. 4). Similar issues were observed in the query tasks for silodosin, dequalinium, camostat, and other drugs (Fig. 4). Comparatively, GPT-4 may exhibit higher accuracy in the task of drug target prediction compared to GPT-3.5. For example, naltrexone, a medication used to manage alcohol or opioid dependence, blocks the effects of opioids in the brain to reduce cravings. It acts as an antagonist for OPRK1 (opioid receptor kappa 1) and as an agonist for OPRM1 (opioid receptor mu 1) and SIGMAR1 (sigma non-opioid intracellular receptor 1). GPT-3.5 only recorded the OPRM1 target, disregarding the other protein targets. In contrast, GPT-4 recorded the OPRM1 target correctly identified the primary target, OPRK1 (Fig. 4).

Another phenomenon observed in the target prediction task is ‘illusory knowledge construction’ and ‘knowledge hallucination.’ When predicting the action target of montelukast, both GPT-3.5 and GPT-4 not only failed to make accurate predictions for its inhibitory target ALOX5 (arachidonate 5-lipoxygenase) but also erroneously predicted new targets LTB4R (leukotriene B4 receptor) and LTC4S (leukotriene C4 synthase) (refer to Fig. 4). Upon searching the Genecards database, it was discovered that LTB4R and LTC4S belong to cysteinyl leukotriene receptors. However, while LTC4S is a valid target of montelukast, LTB4R is a false target. The Genecards database lists five drugs that can interact with the LTB4R receptor, including three confirmed drugs such as gamolenic acid, zafirlukast, and leukotriene B4, and two drugs that have only been demonstrated in experiments, such as cinalukast and morniflumate [29].

3.2.2.5 The Accuracy of ChatGPT in the Querying Tasks of Drug-Drug Interactions Drug-drug interactions (DDIs) occur when two or more drugs are used in combination, and it elicits various risks, including those associated with liver damage, elevated blood pressure, and lowered blood pressure. It is an essential factor influencing drug efficacy and

safety and is also one of the critical issues affecting rational clinical drug use and post-marketing surveillance. The DDIs have become a significant area of interest in pharmacology [47–49]. The DrugBank database provides detailed records of DDI risks. Regarding amobarbital, nine types of DDIs have been documented, including risks of adverse effects, methemoglobinemia, hypotension, central nervous system depression, sedation, constipation, decreased therapeutic efficacy of amobarbital, decreased therapeutic efficacy of other drugs, and decreased metabolism rate of amobarbital (Fig. 5, Table S3).

The analysis of the potential of ChatGPT in predicting DDIs reveals an overall prediction accuracy of $64.50 \pm 21.27\%$, with GPT-3.5 achieving a prediction accuracy of $64.64 \pm 0.00\%$ and GPT-4 achieving a prediction accuracy of $64.33 \pm 0.00\%$ (Table S3). A comparison of ChatGPT's prediction results for large, medium, and small molecules of varying sizes shows that the performance is significantly better for medium- and small-molecule compounds than for large-molecule drugs. For instance, dequalinium, a large molecule compound with a MW of 456.67, exhibits DDIs mainly related to risks of adverse effects, bleeding, viral infections, methemoglobinemia, hypotension, and decreased therapeutic efficacy of other drugs (Fig. 5, Table S3). Both GPT-3.5 and GPT-4 failed to produce precise predictions regarding dequalinium. However, they did emphasize the significance of disclosing all medications to healthcare professionals, especially in cases where dequalinium is predominantly administered topically.

Furthermore, the overall predictive performance of GPT-4 was compared to that of GPT-3.5 in drug predictions. It was observed that GPT-4 outperformed GPT-3.5 in predicting 15 types of drugs. Specifically, GPT-4 demonstrated significantly better predictive performance for six drugs, namely amobarbital, butobarbital, chlorzoxazone, dezocine, irbesartan, and dimetacrine. However, in the case of montelukast and apremilast, GPT-3.5 exhibited better predictive performance than GPT-4 (Table S3).

For instance, let us consider chlorzoxazone as an example. In the DrugBank database, there are four types of interactions between chlorzoxazone and other drugs: the risk of side effects with 103 drugs, the risk of CNS depressant effects with 22 drugs, the risk of sedative effects with one drug, and the risk of changing the rate of metabolism with 214 drugs. GPT-3.5 provided answers regarding the CNS depressant risk and adverse effects risk of drug-drug interactions for chlorzoxazone. However, GPT-4 not only provided answers regarding the CNS depressant risk and adverse effects risk of drug-drug interactions for chlorzoxazone, but it also addressed the risk of affecting the metabolism rate, stating that ‘as chlorzoxazone is primarily metabolized by the liver, drugs that can affect liver enzymes may affect the metabolism of chlorzoxazone. This could alter the drug's

Table 3 The performance of ChatGPT in predicting drug pharmacological properties^a

Query drug	Class	Score _{indication} ^b (%)		Score _{pharmacodynamics} ^c (%)		Score _{mechanism} ^d (%)		Score _{toxicity} ^e (%)	
		GPT-3.5	GPT-4	GPT-3.5	GPT-4	GPT-3.5	GPT-4	GPT-3.5	GPT-4
Amobarbital	A	100.00 ± 0.00	100.00 ± 0.00	85.00 ± 0.00	85.00 ± 0.00	90.00 ± 0.00	90.00 ± 0.00	95.00 ± 0.00	95.00 ± 0.00
Benzphetamine	A	100.00 ± 0.00	100.00 ± 0.00	85.00 ± 0.00	85.00 ± 0.00	90.00 ± 0.00	95.00 ± 0.00	95.00 ± 0.00	95.00 ± 0.00
Butobarbital	A	100.00 ± 0.00	100.00 ± 0.00	85.00 ± 0.00	85.00 ± 0.00	90.00 ± 0.00	90.00 ± 0.00	95.00 ± 0.00	95.00 ± 0.00
Chlorzoxazone	A	100.00 ± 0.00	80.00 ± 0.00	80.00 ± 0.00	80.00 ± 0.00	85.00 ± 0.00	85.00 ± 0.00	95.00 ± 0.00	95.00 ± 0.00
Dezocine	A	100.00 ± 0.00	100.00 ± 0.00	85.00 ± 0.00	85.00 ± 0.00	90.00 ± 0.00	95.00 ± 0.00	95.00 ± 0.00	95.00 ± 0.00
Camostat	B	100.00 ± 0.00	100.00 ± 0.00	90.00 ± 0.00	90.00 ± 0.00	85.00 ± 0.00	95.00 ± 0.00	95.0 ± 0.00	95.00 ± 0.00
Dimetacrine	B	100.00 ± 0.00	100.00 ± 0.00	85.00 ± 0.00	85.00 ± 0.00	90.00 ± 0.00	90.00 ± 0.00	95.00 ± 0.00	95.00 ± 0.00
Naltrexone	B	100.00 ± 0.00	100.00 ± 0.00	85.00 ± 0.00	85.00 ± 0.00	80.00 ± 0.00	85.00 ± 0.00	95.00 ± 0.00	95.00 ± 0.00
Pefloxacin	B	100.00 ± 0.00	100.00 ± 0.00	90.00 ± 0.00	90.00 ± 0.00	95.00 ± 0.00	95.00 ± 0.00	95.00 ± 0.00	95.00 ± 0.00
Ropivacaine	B	100.00 ± 0.00	100.00 ± 0.00	80.00 ± 0.00	80.00 ± 0.00	60.00 ± 0.00	85.00 ± 0.00	95.00 ± 0.00	95.00 ± 0.00
Apremilast	C	80.00 ± 0.00	80.00 ± 0.00	90.00 ± 0.00	90.00 ± 0.00	95.00 ± 0.00	95.00 ± 0.00	95.00 ± 0.00	95.00 ± 0.00
Dequalinium	C	100.00 ± 0.00	100.00 ± 0.00	85.00 ± 0.00	85.00 ± 0.00	90.00 ± 0.00	90.00 ± 0.00	95.00 ± 0.00	95.00 ± 0.00
Irbesartan	C	50.00 ± 0.00	0.00 ± 0.00	85.00 ± 0.00	85.00 ± 0.00	90.00 ± 0.00	90.00 ± 0.00	95.00 ± 0.00	95.00 ± 0.00
Montelukast	C	100.00 ± 0.00	100.00 ± 0.00	80.00 ± 0.00	80.00 ± 0.00	80.00 ± 0.00	85.00 ± 0.00	95.00 ± 0.00	95.00 ± 0.00
Silodosin	C	100.00 ± 0.00	100.00 ± 0.00	85.00 ± 0.00	85.00 ± 0.00	90.00 ± 0.00	95.00 ± 0.00	95.00 ± 0.00	95.00 ± 0.00
Mean		95.33 ± 13.56	90.67 ± 26.04	85.00 ± 3.27	85.00 ± 3.27	86.67 ± 8.59	90.67 ± 4.17	95.00 ± 0.00	95.00 ± 0.00

^aThe predictive performance is assessed using a percentage scoring method and reported as mean ± SD (standard deviation)

^bScore_{indication} represents the prediction score of ChatGPT (GPT-3.5 or GPT-4) in the basic pharmacological properties of drugs. A score closer to 100% indicates higher accuracy rates

^cScore_{pharmacodynamics} represents the prediction score of ChatGPT (GPT-3.5 or GPT-4) in the pharmacodynamics properties of drugs. A score closer to 100% indicates a higher accuracy rate

^dScore_{mechanism} represents the prediction score of ChatGPT (GPT-3.5 or GPT-4) in the action mechanism properties of drugs. A score closer to 100% indicates higher accuracy rates

^eScore_{toxicity} represents the prediction score of ChatGPT (GPT-3.5 or GPT-4) in the toxicity properties of drugs. A score closer to 100% indicates higher accuracy rates

effectiveness or increase the risk of side effects.’ Similar phenomena were also observed for the other five drugs (Fig. 5, Table S3).

3.2.3 Assessing the Potential of ChatGPT in Drug Structure Optimization Tasks

Compound structure optimization is crucial in enhancing the bioavailability of lead compounds or drug candidates, mitigating toxicity, improving metabolic stability, and optimizing pharmacodynamics [28–31]. In order to assess the potential of ChatGPT in the field of compound structure optimization, we established ‘Improving metabolic activity,’ ‘Reducing hepatotoxicity,’ ‘Reducing cardiotoxicity,’ and ‘Increasing solubility’ as the primary optimization objectives. The findings indicate that ChatGPT (GPT-3.5 and GPT-4) solely demonstrates its ability to have general ideas in drug structure optimization tasks. In other words, it can delineate common strategies employed in structure optimization. However, it cannot devise comprehensive optimization plans for specific drugs (Fig. 6).

Metabolic activity optimization encompasses optimization strategies aimed at enhancing the metabolic stability of compounds, prolonging drug action duration in the body, increasing exposure within the body, reducing compound clearance rates, and improving bioavailability. In drug structure optimization tasks targeting ‘improving metabolic activity,’ we selected buspirone, paroxetine, and 8-chloro-4-(4-methylpiperazin-1-yl)benzofuro[3,2-d]pyrimidine as the compounds to be optimized. Similar issues were observed in the structure optimization of paroxetine and 8-chloro-4-(4-methylpiperazin-1-yl)benzofuro[3,2-d]pyrimidine. GPT-3.5 suggests modification of susceptible functional groups, blocking metabolic sites, employing the prodrug approach, and utilizing metabolic stability prediction and modeling. However, it does not provide detailed operational procedures and optimization plans (Fig. 6).

For optimization tasks, including reducing hepatotoxicity, reducing cardiotoxicity, and increasing solubility, both GPT-3.5 and GPT-4 provide generalized answer schemes. For instance, when addressing the solubility improvement of Taxol, GPT-3.5 and GPT-4 propose optimization strategies such as the ‘prodrug approach,’ ‘formulation techniques,’

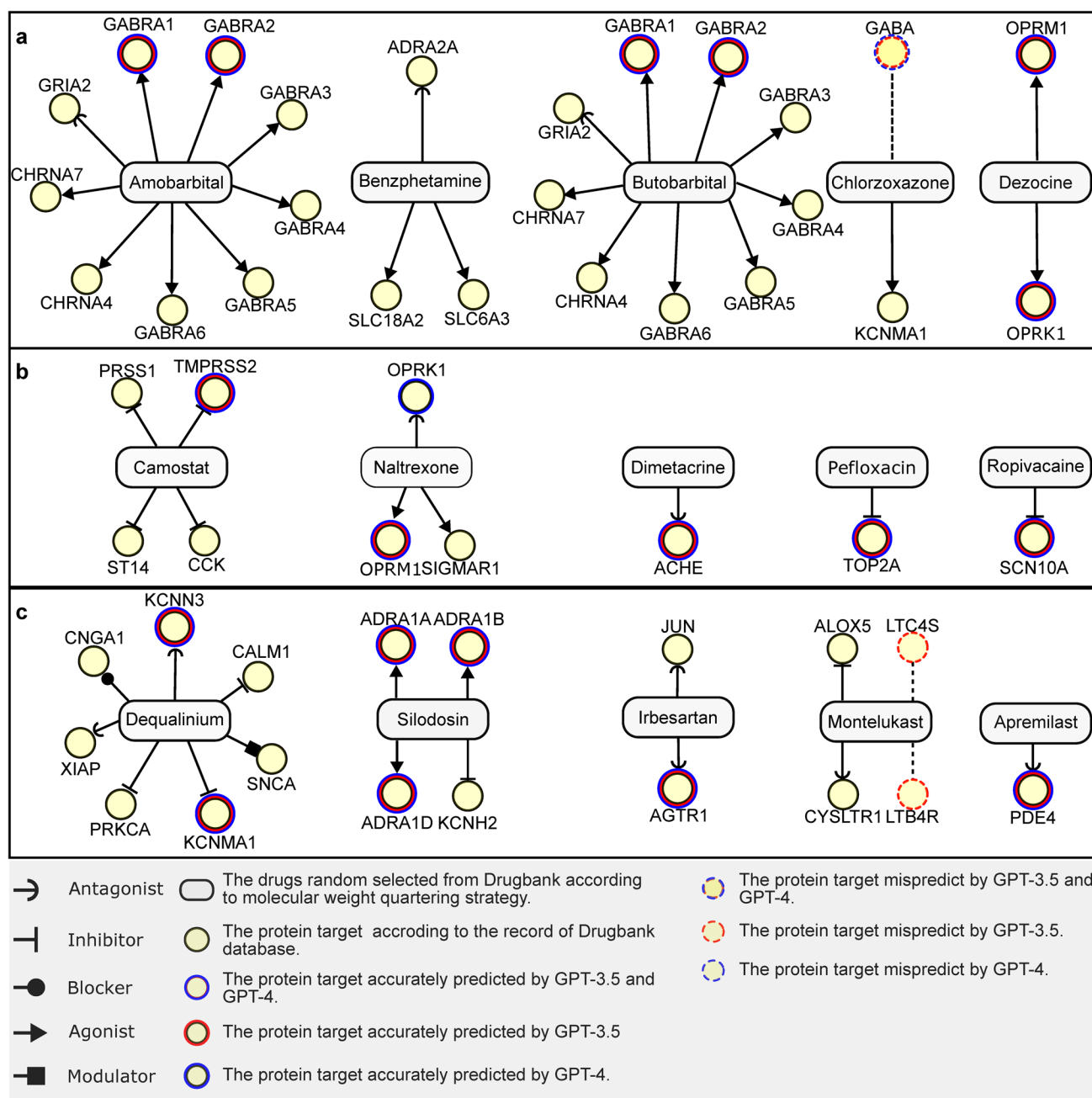
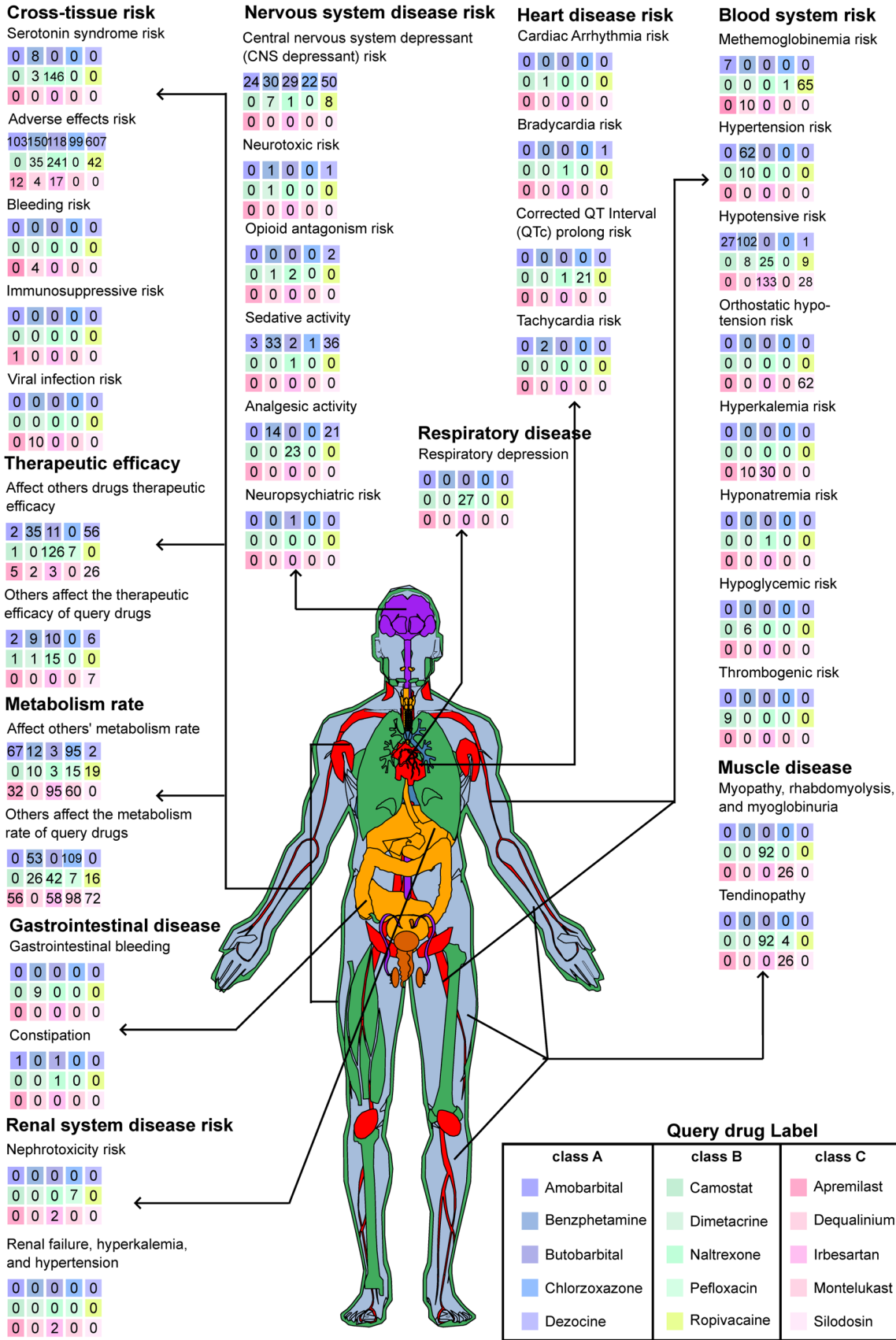


Fig. 4 Exploring the potential of ChatGPT in ‘question-answer’-based tasks for drug-target attributes. **a** Overview of the predicted and actual targets of class A drugs (molecular weight < 255.24); **b**

Overview of the predicted and actual targets of class A drugs (255.24 < molecular weight > 412.64); **c** Overview of the predicted and actual targets of class C drugs (molecular weight > 412.64)

‘structural modifications,’ and ‘combination with solubilizing agents.’ They also suggest methods such as complexation with cyclodextrins, nanoemulsion formulation, or encapsulation in liposomes or nanoparticles to enhance solubility by increasing drug dispersibility and effective surface area in water. However, they do not describe the execution difficulty, specific implementation methods, or successful case studies (Fig. 6).

In summary, ChatGPT demonstrates its generalizing ability in compound structure optimization tasks. It provides structural optimization strategies for improving compound activity but cannot offer effective plans and specific examples. Additionally, it fails to provide adequate literature and data support.



◀**Fig. 5** Investigating the potential of ChatGPT in drug-drug interaction ‘question-answer’-based tasks. The query drugs are represented by colored cells, and the numbers within the colored cells indicate the count of drug-drug interactions with specific items. For example, a value of eight signifies that there are 8 drug-drug interactions resulting in the mentioned side effect when combined with the respective query drug. Further details regarding the drug-drug interactions can be found in Supplementary Data 5

3.2.4 The Accuracy of ChatGPT in Systematically Summarizing and Inferring the Current Limitations and Emerging Trends in Pharmacology

The efficacy of retrieval, comprehension, summarization, and reasoning abilities is vital in evaluating the capabilities of LLM models [6, 14, 50]. To clarify and determine ChatGPT's capability in text summarization, we evaluated its performance in ‘current limitations in pharmacological research’ and ‘future trends in pharmacological research.’

3.2.4.1 The Accuracy of ChatGPT in Systematically Summarizing the Current Limitations in Pharmacology In three repeated inquiries into GPT-3.5 and GPT-4, 16 topics are identified as limitations in current pharmacological research. These topics include ‘limited predictability of preclinical models,’ ‘regulatory challenges,’ ‘limited availability of drug targets,’ ‘translational challenges,’ ‘limited access to human tissue samples,’ ‘limited understanding of disease mechanisms,’ and others. These topics receive an importance score of 80 or above (Fig. 7a and Table S4).

Among all the topics, GPT-3.5 identifies ‘lack of diversity in clinical trials,’ ‘high cost of drug development,’ ‘limited understanding of disease mechanisms,’ and ‘ethical concerns’ (5 times) as the most significant limitations in current pharmacology. For instance, GPT-3.5 highlights that current clinical trials are based on a minority of populations and do not represent a broader population, resulting in potential drug efficacy and safety variations across different patient populations. To address this limitation, conducting clinical trials in the broader population or ethnic group is suggested as a practical approach to improving efficacy and safety in current pharmacological research. The three reviewers concur that this topic is a limitation in current pharmacological research. However, they do not consider it the most significant limitation, assigning it an importance score of 86.67 ± 2.89 (Fig. 7a and Table S4).

The results are inconsistent when comparing the importance scores provided by the three reviewers with the number of recommendations made by ChatGPT. The three reviewers considered limited understanding of disease mechanisms and limited access to human tissue samples as the most significant limitations in current pharmacological research, with an average score of 91.67 ± 2.89 . However, GPT-3.5 only recommended these two topics five times and one time

(Fig. 7a and Table S4), indicating a significant imbalance. The topic ‘limited understanding of disease mechanisms’ is regarded as the most crucial limitation in pharmacological research, possibly due to its frequent mention in the literature. On the other hand, ‘limited access to human tissue samples’ has only been recommended once, which may be related to the relatively low frequency of reports in the literature. However, three reviewers gave it a very high importance score. Most pharmacological studies speculate that it is associated with the urgent need for human tissue samples, including live ones. Unfortunately, such samples are severely scarce, and related research is often restricted.

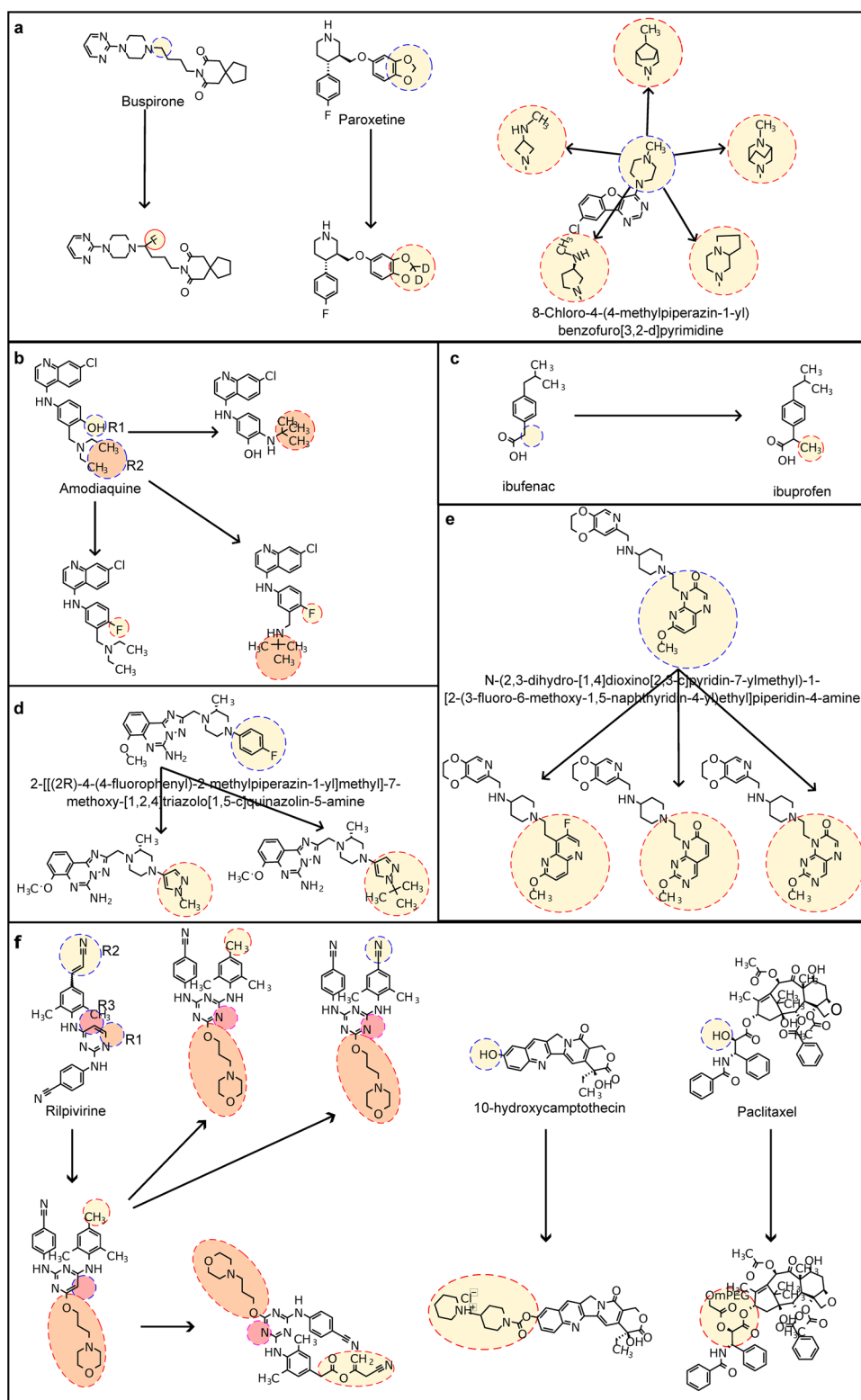
3.2.4.2 The Accessibility and Accuracy of ChatGPT in Systematically Summarizing and Inferring the Emerging Trends in Pharmacology In the three repeated inquiries to GPT-3.5 and GPT-4, 18 topics are identified as trends in future pharmacological research. These topics include ‘Artificial intelligence and machine learning,’ ‘drug repurposing,’ ‘precision medicine,’ ‘nanomedicine,’ ‘gene therapy and gene editing,’ ‘digital health,’ and others. Among them, the ‘Artificial Intelligence and Machine Learning’ topic and the ‘Drug Repurposing’ topic are considered to be hotspots for future pharmacology research, with each being recommended by ChatGPT six times (three times each by GPT-3.5 and GPT-4). However, based on the perspectives of the three reviewers, ‘digital health’ and ‘immunotherapy’ are considered the most important subjects for future research. Each topic receives a high importance score of 91.67 ± 2.89 , making them the highest-scoring topics (Fig. 7b and Table S5).

‘Nanomedicine’ is the most controversial topic among all the covered topics. One reviewer argues that this topic remains a prominent issue in pharmacology, assigning it a high score of 90. However, the other two reviewers assigned importance scores below 80. Except for the ‘nanomedicine’ topic, all other topics received an importance score of more than 85 (Fig. 7b and Table S5).

3.3 Evaluation of Lead Compound Structure Optimization Tasks for LLMs Based on Specific Text RAG Mode

We constructed a transient LLM named PharmacologyGPT, using GPT-4 as the base LLM and Liu et al's literature records as the source for specific text RAG. For three optimization tasks on 10 compounds, such as metabolic stability, reduced toxicity, and enhanced water solubility, the results showed that PharmacologyGPT improved the predictive effectiveness compared to GPT-3.5 and GPT-4 without altering the prompt method. PharmacologyGPT provided answers for the lead compound optimization

Fig. 6 Exploring ChatGPT's potential in drug structure optimization 'Question-answer'-based tasks. **a** The potential of ChatGPT in drug structure optimization is being explored using buspirone, paroxetine, and 8-chloro-4-(4-methylpiperazin-1-yl)benzofuro[3,2-d]pyrimidine as compounds to be optimized; **b, c**: Amodiaquine and ibufenac are being utilized as compounds to be optimized in order to explore their potential in reducing hepatotoxicity; **d, e**: The potential to lower cardiac toxicity is being investigated using 2-[[[(2R)-4-(4-fluorophenyl)-2-methylpiperazin-1-yl]methyl]-7-methoxy-[1,2,4]triazolo[1,5-c]quinazolin-5-amine and N-(2,3-dihydro-[1,4]dioxino[2,3-c]pyridin-7-ylmethyl)-1-[2-(3-fluoro-6-methoxy-1,5-naphthyridin-4-yl)ethyl]piperidin-4-amine as compounds to be optimized; **f**: rilpivirine, 8-hydroxy camptothecin, and taxol are being utilized to investigate the potential of ChatGPT in enhancing water solubility optimization schemes



strategies reported in the literature and explained specific actionable plans (Fig. 8d). Additionally, basic attribute evaluations, such as context consistency (Fig. 8a) and semantic relevance (Fig. 8b), indicated that PharmacologyGPT did

not significantly affect the context consistency and semantic relevance of LLM. Therefore, exploring LLMs based on specific information, such as RAG or fine-tuning, will

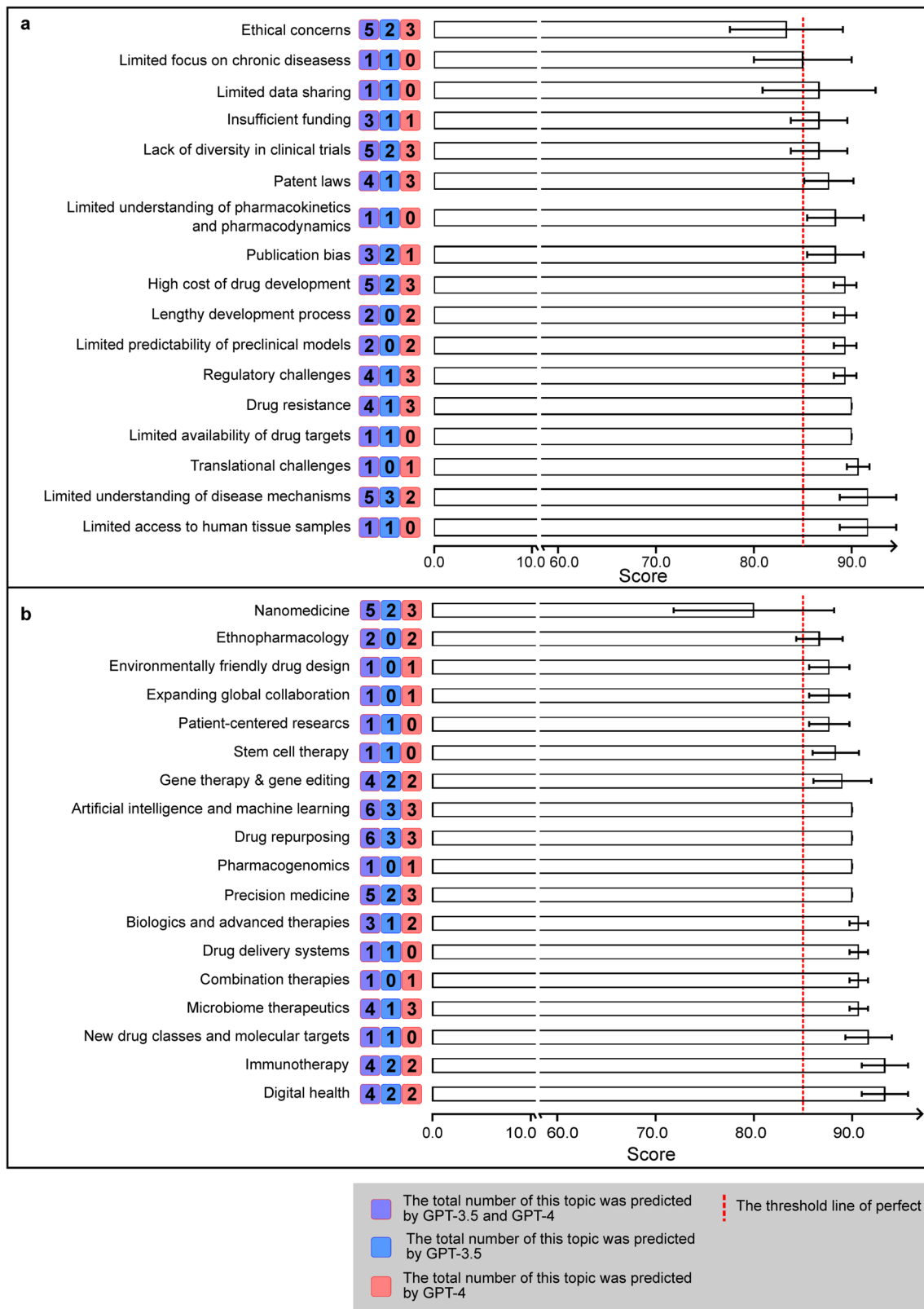


Fig. 7 ChatGPT's ability to systematically summarize and infer the current limitations and emerging trends in pharmacology. **a** Exploring the ability of ChatGPT to summarize the current limitations in phar-

macology systematically. **b** Exploring the ability of ChatGPT to summarize and infer the emerging trends in pharmacology systematically

significantly improve the hallucination issues in general LLMs when handling pharmacology tasks.

Furthermore, for the water solubility improvement optimization task of paclitaxel, the specific text RAG-powered PharmacologyGPT demonstrated exceptional capability. PharmacologyGPT addressed the glycosylation prodrug and poly(ethylene glycol) (PEG) prodrug strategies recorded in the literature but also compared the effectiveness of both optimization strategies and identified the PEG prodrug strategy as the relatively more efficient optimization approach.

Our research results indicate that, under the RAG framework, even GPT-4 based on non-English text can significantly improve the accuracy of answering complex pharmacological questions (specifically, in this study, focusing on lead compound optimization tasks). A tracking analysis of the specific text RAG data flow revealed that when we use a specific data source, such as the RAG resource, the LLMs first segment the data and convert it into vector representations, which are then stored in a vector library. Subsequently, during question-answering tasks, the LLMs retrieve relevant information through information retrieval and enhance the response accuracy with the added context.

4 Discussion

The emergence of LLMs, including ChatGPT, has opened up new avenues for exploring AI-driven drug discovery in the era of AI [51]. These models provide unprecedented opportunities, particularly regarding information retrieval and strategy discovery through human-machine interaction. However, pharmacology, as an exceedingly complex application field, presents challenges not only in the separation, purification, and identification of chemical components but also encompasses a wide range of complex research areas,

including the optimization of lead compound structures, investigation into potential drug toxic mechanisms, analysis of drug targets and their mechanisms of action, and drug-drug interactions [19, 22, 52–54]. Leveraging machine learning strategies, including LLMs, can significantly enhance information retrieval efficiency [55, 56], deepen and broaden data analysis [57], and improve the accuracy of theoretical predictions, thus offering numerous advantages in the drug development process. A comprehensive and meticulously designed artificial pharmacology evaluation task demonstrates the potential of existing LLMs in pharmacology. It establishes a strong foundation for evaluating and testing the performance of these models in pharmacological research. Through such evaluation tasks, we can assess the ability of LLMs to address complex pharmacological challenges and further inspire new research and application ideas, advancing the deep integration of AI in drug discovery and development.

Regarding the application potential of LLMs in pharmacology, although several researchers have previously conducted evaluations [9, 19], the breadth and complexity of pharmacology mean that evaluations based solely on these attributes or characteristics are not sufficient to fully demonstrate the potential of LLMs in handling complex pharmacological tasks [19, 22, 52–54]. Additionally, most of these evaluations have been published only in article form, which limits the further application of these test data in evaluating large pharmacology models. Therefore, to comprehensively evaluate the value and potential of LLMs in pharmacology, we have designed a comprehensive test set that includes comprehensive drug property query tasks and lead compound optimization tasks and tasks for summarizing research trends and limitations. Subsequently, we evaluated the general LLM ChatGPT based on this test set. ChatGPT effectively understood our pharmacological

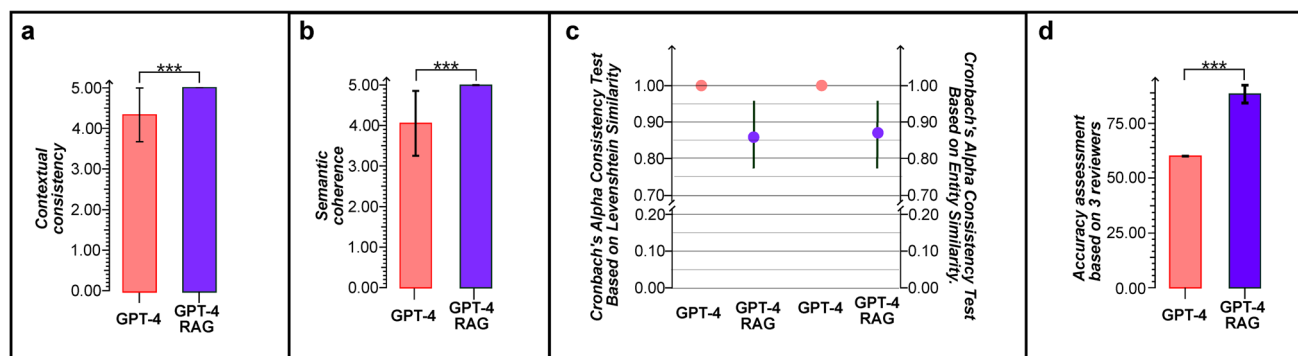


Fig. 8 Evaluating the capability of ChatGPT in lead compound optimization based on the RAG (Retrieval-Augmented Generation) model. **a** The accessibility of ChatGPT RAG in lead compound optimization based on contextual consistency; **b** The accessibility of ChatGPT RAG in lead compound optimization based on semantic

similarity; **c** The accessibility of ChatGPT RAG in lead compound optimization based on Levenshtein similarity consistency tests and entity similarity consistency tests; **d** The accuracy of ChatGPT RAG in lead compound optimization based on expert scores

instructions and provided good responses in drug pharmacological properties, pharmacokinetics, mode of action, and toxicity prediction. However, ChatGPT also exhibits limitations when confronted with drug identifier queries, drug interaction information queries, and drug structure simulation optimization-based queries. It struggles to retrieve interaction information for a single or specific drug effectively and cannot the ability to optimize specific drugs. The most severe issue stems from the ‘knowledge hallucination’ problem inherent in ChatGPT, where the answers may appear logical but contain entirely incorrect information. Research on the mechanisms behind the ‘knowledge hallucination’ phenomenon in LLMs suggests that it may be related to issues with data source noise, model flaws, and unclear user tasks. As summarized by Huang et al in ‘*A Survey on Hallucination in LLMs: Principles, Taxonomy, Challenges, and Open Questions*,’ when general-purpose LLMs like ChatGPT, LaMDA, PaLM, and Gopher are directly applied to fundamental pharmacological questions such as drug side effects-based information retrieval, drug target queries, drug structure optimization suggestions, and summarizing pharmacological research trends, they exhibit knowledge hallucination and randomness in their responses [58]. To address this issue, several researchers have proposed solutions, including the integration of external knowledge bases [59], knowledge graphs [60], multi-agent interaction [61], human-in-the-loop [62], and the Tree of Thoughts framework [63]. Among these, external knowledge bases offer a convenient and effective method to mitigate knowledge hallucination. For instance, GeneGPT [64] exemplifies overcoming ChatGPT’s ‘knowledge hallucination’ in gene or genomic data by accessing the National Center for Biotechnology Information (NCBI) web application programming interface (API), demonstrating improved performance in gene naming, genomic location, gene function analysis, and sequence alignment tasks over existing machine learning models. Additional studies addressing the limitations of LLMs or knowledge hallucinations of LLMs like ChatGPT in handling specialized topics through external knowledge bases or knowledge graphs, include Med-PaLM2, BioMedGPT [65], CancerGPT, and scGPT [66].

Beyond addressing data quality to prevent or solve the problem of insufficient capability or knowledge hallucination in handling specialized topics, some scholars have explored alleviating hallucinations through retrieval enhancement, prompt engineering, and other methods. For instance, Meta AI researchers proposed a model fine-tuning approach called RAG [26], combining an information retrieval component with a text-generation model. The information retrieval component receives inputs and retrieves relevant/supporting documents, indicating their sources. These are then combined with the original prompt as context for the LLM text generator to produce the final output. Therefore, exploring

the integration of external knowledge bases/knowledge graphs and using RAG as a model optimization strategy for large pharmacology models will be a future direction for developing large pharmacology models. This approach can effectively improve issues such as hallucinations and random errors in general LLMs when answering pharmacological questions. This study also explored the potential of LLMs in lead compound structure optimization using the specific text RAG method. The results showed that GPT-4 significantly improved the predictive effectiveness of lead compound structure optimization strategies when specific text RAG was introduced. The LLMs prioritized the compound optimization strategies reported in the literature and summarized and compared multiple strategies and methods from different studies.

5 Conclusion

Large language models, whether they are general-purpose models, like ChatGPT, Llama2, and Claude, or specialized models like DrugChat, DrugGPT, and Mol-Instructions, are fundamentally transforming the knowledge query methods and approaches in drug discovery for pharmacologists, drug researchers, clinical research scientists, and AI researchers. They enable multi-round consultations and queries of pharmacological questions in a ‘question-answer’ format. However, pharmacology is an exceptionally complex field of application. To better apply LLMs in pharmacological practice, we propose a pharmacology test set comprising three pharmacological tasks: drug property queries, lead compound structure optimization, and summaries of trends and limitations. This test set covers a variety of pharmacological application scenarios. The general LLM ChatGPT evaluation also shows that general LLMs can understand pharmacological task instructions such as drug information queries, lead compound optimization, and systematic summaries. However, they encounter issues like ‘knowledge hallucination,’ ‘randomness,’ and ‘generality’ when dealing with compound structure optimization and complex pharmacological property-based queries. Therefore, exploring pharmacology-specific LLMs based on external knowledge bases/knowledge graphs and RAG will offer practical solutions to alleviate issues like knowledge hallucinations and insufficient specialization faced by general LLMs in the pharmacology domain. We anticipate that this study will provide new frameworks and insights for drug researchers, pharmacologists, clinical research scientists, and AI researchers in developing and evaluating pharmacology LLMs.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s40265-024-02124-2>.

Declarations

Funding This study was supported by the National Natural Science Foundation of China (32270690 and 32070671, Natural Science Foundation of Hainan Province of China (820MS102), and a special fund for agro-scientific research in the public interest (1630032020031).

Conflict of interest YZ, SR, JW, JL, CW, MH, XL, RW, JZ, CZ, DD, ZZ, RKS and BS have no conflicts of interest to declare.

Availability of data and materials The dataset and its instructions have been uploaded to Huggingface (<https://huggingface.co/datasets/zhangyingbo1984/Pharmacology-LLM-test-set>) and Github (<https://github.com/zyb1984/Pharmacology-LLM-test-set>), and all users can download and reuse them from Huggingface and Github. The results and scores based on GPT-3.5 and GPT-4 can be downloaded from the attachments of this manuscript.

Ethics approval Not applicable.

Author contributions BS conceived the project. BS and YZ initially designed the entire workflow, and SR, JW, and RKS, among others, extensively reviewed and discussed the rationality of the entire paper framework, providing modification suggestions. BS, RKS, DD, and ZZ designed the pharmacology-LLM-test-set, while YZ and XL constructed the large-scale model test set. SR, JW, and XL collected GPT-3.5 and GPT-4 evaluation results. BS, YZ, RKS, DD, and ZZ reviewed and scored the results. YZ, SR, JL, MH, RW, and CZ analyzed the data results. YZ drafted the initial manuscript, and SR, JZ, and CW edited it. All authors participated in the discussion of the results.


Open Access This article is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License, which permits any non-commercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc/4.0/>.

References

- Sarkar C, Das B, Rawat VS, Wahlang JB, Nongpiur A, Tiewsoh I, et al. Artificial intelligence and machine learning technology driven modern drug discovery and development. *Artificial intelligence and machine learning technology driven modern drug discovery and development*. Int J Mol Sci. 2023;24(3):2026.
- Srivathsa AV, Sadashivappa NM, Hegde AK, Radha S, Mahesh AR, Ammunje DN, et al. A review on artificial intelligence approaches and rational approaches in drug discovery. *Curr Pharm Des*. 2023;29(15):1180–92.
- van der Lee M, Swen JJ. Artificial intelligence in pharmacology research and practice. *Clin Transl Sci*. 2023;16(1):31–6.
- Mazumdar B, Deva Sarma PK, Mahanta HJ, Sastry GN. Machine learning based dynamic consensus model for predicting blood-brain barrier permeability. *Comput Biol Med*. 2023;160: 106984.
- Li T, Shetty S, Kamath A, Jaiswal A, Jiang X, Ding Y, et al. CancerGPT: few-shot drug pair synergy prediction using large pre-trained language models. *ArXiv*. 2024;7:40.
- Bommasani R, Liang P, Lee T. Holistic evaluation of language models. *Ann N Y Acad Sci*. 2023;1525(1):140–6.
- Stokel-Walker C. AI bot ChatGPT writes smart essays - should professors worry? *Nature*. 2022 Dec 9. <https://doi.org/10.1038/d41586-022-04397-7>. Epub ahead of print. PMID: 36494443
- van Dis EAM, Bollen J, Zuidema W, van Rooij R, Bockting CL. ChatGPT: five priorities for research. *Nature*. 2023;614(7947):224–6.
- Castro Nascimento CM, Pimentel AS. Do large language models understand chemistry? A conversation with ChatGPT. *J Chem Inf Model*. 2023;63(6):1649–55.
- Guo T, Guo K, Liang Z, Guo Z, Chawla NV, Wiest O, et al. What indeed can GPT models do in chemistry? A comprehensive benchmark on eight tasks. 2023. *arXiv:2305.18365*.
- Ferres JML, Weeks WB, Chu LC, Rowe SP, Fishman EK. Beyond chatting: the opportunities and challenges of ChatGPT in medicine and radiology. *Diagn Interv Imaging*. 2023;104(6):263–4.
- Eggmann F, Weiger R, Zitzmann NU, Blatz MB. Implications of large language models such as ChatGPT for dental medicine. *J Esthet Restor Dent*. 2023; 35(7):1098–1102.
- Park I, Joshi AS, Javan R. Potential role of ChatGPT in clinical otolaryngology explained by ChatGPT. *Am J Otolaryngol*. 2023;44(4): 103873.
- Zhao WX, Zhou K, Li J, Tang T, Wang X, Hou Y, et al. A survey of large language models. 2023. *arXiv:2402.06196*.
- Liang Y, Zhang R, Zhang L, Xie P. DrugChat: towards enabling ChatGPT-like capabilities on drug molecule graphs. 2023. *arXiv:2309.03907*.
- Li Y, Gao C, Song X, Wang X. DrugGPT: a GPT-based strategy for designing potential ligands targeting specific proteins. 2023. *BioRxiv*. 2023.06.29.543848.
- Fang Y, Liang X, Zhang N, Liu K, Huang R, Chen Z, et al. Mol-instructions: a large-scale biomolecular instruction dataset for large language models. 2023. *arXiv:2306.08018*.
- Luo Y, Liu XY, Yang K, Huang K, Hong M, Zhang J, et al. Towards unified AI drug discovery with multiple knowledge modalities. *Health Data Sci*. 2024;4:0113.
- Cloesmeijer ME, Janssen A, Koopman SF, Cnossen MH, Mathôt RAA. ChatGPT in pharmacometrics? Potential opportunities and limitations. *Br J Clin Pharmacol*. 2024;90(1):360–5.
- Müller M. The discipline of clinical pharmacology. 1st ed. Cham: Springer International Publishing; 2016.
- Zhao L, Peck CC. Impact of clinical pharmacology on the modernization of drug development and regulation. Cham: Springer International Publishing; 2023.
- Liu Q, Ahadpour M, Rocca M, Huang S-M. Clinical pharmacology regulatory sciences in drug development and precision medicine: current status and emerging trends. *AAPS J*. 2021;23:1–10.
- Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med*. 2023;29(8):1930–40.
- Omiye JA, Gui H, Rezaei SJ, Zou J, Daneshjou R. Large language models in medicine: the potentials and pitfalls: a narrative review. *Ann Intern Med*. 2024;177(2):210–20.
- Zhou H, Liu F, Gu B, Zou X, Huang J, Wu J, et al. A survey of large language models in medicine: progress, application, and challenge. 2024. *arXiv:2311.05112v4*.
- Zakka C, Shad R, Chaurasia A, Dalal AR, Kim JL, Moor M, et al. Almannac—retrieval-augmented language models for clinical medicine. *Nejm ai*. 2024;1(2).

27. 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 Res.* 2018;46(D1):D1074–d82.
28. Li Z, Wang J, Zhou Y, Liu H. Lead compound optimization strategy (3)—structure modification strategies for improving water solubility. *Acta Pharm Sin.* 2014;49(9):1238–47.
29. Liu HL, Wang J, Lin DZ, Liu H. Lead compound optimization strategy (2)—structure optimization strategy for reducing toxicity risks in drug design. *Acta Pharm Sin.* 2014;49(1):1–15.
30. Wang J, Liu H. Lead compound optimization strategy (1)—changing metabolic pathways and optimizing metabolism stability. *Acta Pharm Sin.* 2013;48(10):1521–31.
31. Zhou SB, Wang J, Liu H. Lead compound optimization strategy(5)—reducing the hERG cardiac toxicity in drug development. *Acta Pharm Sin.* 2016;51(10):1530–9.
32. Hall K, Stewart T, Chang J, Freeman MK. Characteristics of FDA drug recalls: a 30-month analysis. *Am J Health-Syst Pharm.* 2016;73(4):235–40.
33. Kim D, Kim B, Han D, Eibich M. AutoRAG: automated framework for optimization of retrieval augmented generation pipeline. 2024. arXiv:2410.20878.
34. OpenAI. Optimizing LLM Accuracy. OpenAI Cookbook 2024 [cited 2024 October 25th]. <https://platform.openai.com/docs/guides/optimizing-llm-accuracy#retrieval-augmented-generation-rag>.
35. Chen Y, Wang R, Jiang H, Shi S, Xu R. Exploring the use of large language models for reference-free text quality evaluation: an empirical study. 2023. arXiv:2304.00723.
36. Qi B, Zhang K, Tian K, Li H, Chen Z-R, Zeng S, et al. Large language models as biomedical hypothesis generators: a comprehensive evaluation. 2023. arXiv:2407.08940.
37. Team RC. R: A language and environment for statistical computing. 4.3.1 ed; 2023.
38. Wickham H, François R, Henry L, Müller K, Vaughan D. dplyr: a grammar of data manipulation. 2023.
39. Loo MPJ. The stringdist package for approximate string matching. *R J.* 2014;6(1):111–22.
40. Selivanov D, Bickel M, Wang Q. text2vec: modern text mining framework for R. 2023.
41. Pennington J, Socher R, Manning C. GloVe: global vectors for word representation. In: Moschitti A, Pang B, Daelemans W, editors. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). Doha: Association for Computational Linguistics; 2014. p. 1532–43.
42. Rizopoulos D. ltm: an R package for latent variable modelling and item response theory analyses. *J Stat Softw.* 2006;17(5):1–25.
43. Wickham H. ggplot2: elegant graphics for data analysis. New York: Springer; 2016.
44. Bah T. Inkscape: guide to a vector drawing program. Prentice Hall Press; 2011.
45. Liu Y, Iter D, Xu Y, Wang S, Xu R, Zhu C. G-Eval: NLG evaluation using GPT-4 with better human alignment. 2023. arXiv:2303.16634.
46. White AD. The future of chemistry is language. *Nat Rev Chem.* 2023;7(7):457–8.
47. Hauben M. Artificial intelligence and data mining for the pharmacovigilance of drug-drug interactions. *Clin Ther.* 2023;45(2):117–33.
48. Lin X, Dai L, Zhou Y, Yu ZG, Zhang W, Shi JY, et al. Comprehensive evaluation of deep and graph learning on drug-drug interactions prediction. *Brief Bioinform.* 2023;24(4):bbad235.
49. Zhang Y, Deng Z, Xu X, Feng Y, Junliang S. Application of artificial intelligence in drug-drug interactions prediction: a review. *J Chem Inf Model.* 2024;64(7):2158–2173.
50. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. *Nature.* 2023;620(7973):E19.
51. Chakraborty C, Bhattacharya M, Lee SS. Artificial intelligence enabled ChatGPT and large language models in drug target discovery, drug discovery, and development. *Mol Ther Nucleic Acids.* 2023;12(33):866–8.
52. Sabry Abdel-Messih M, Kamel Boulos MN. ChatGPT in clinical toxicology. *JMIR Med Educ.* 2023;9: e46876.
53. Sharma G, Thakur A. ChatGPT in drug discovery: a case study on anticocaine addiction drug development with chatbots. *ChemRxiv.* 2023.
54. Kothari AN. ChatGPT, large language models, and generative AI as future augments of surgical cancer care. *Ann Surg Oncol.* 2023;30:3174–6.
55. Chen Q, Sun H, Liu H, Jiang Y, Ran T, Jin X, et al. An extensive benchmark study on biomedical text generation and mining with ChatGPT. *Bioinformatics.* 2023;39(9):btad557.
56. Kim HW, Shin DH, Kim J, Lee GH, Cho JW. Assessing the performance of ChatGPT's responses to questions related to epilepsy: a cross-sectional study on natural language processing and medical information retrieval. *Seizure.* 2024;114:1–8.
57. Shin E, Ramanathan M. Evaluation of prompt engineering strategies for pharmacokinetic data analysis with the ChatGPT large language model. *J Pharmacokinet Pharmacodyn.* 2024; 51(2):101–108.
58. Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. 2023. arXiv:2311.05232.
59. Caufield JH, Hegde H, Emonet V, Harris NL, Joachimiak MP, Matentzoglou N, et al. Structured prompt interrogation and recursive extraction of semantics (SPIRES): a method for populating knowledge bases using zero-shot learning. *Bioinformatics.* 2024;40(3):btad104.
60. Remy F, Demuynck K, Demeester T. BioLORD-2023: semantic textual representations fusing large language models and clinical knowledge graph insights. *J Am Med Inform Assoc.* 2024;31(9):1844–1855.
61. Qingyun Wu, Bansal G, Zhang J, Wu Y, Li B, Zhu E, et al. AutoGen: enabling next-gen LLM applications via multi-agent conversation. 2023. arXiv:2308.08155.
62. Yang X, Zhan R, Wong DF, Wu J, Chao LS. Human-in-the-loop machine translation with large language model. 2023. arXiv:2310.08908.
63. Liévin V, Hother CE, Motzfeldt AG, Winther O. Can large language models reason about medical questions? *Patterns (New York, NY).* 2024;5(3): 100943.
64. Jin Q, Yang Y, Chen Q, Lu Z. Genegpt: augmenting large language models with domain tools for improved access to biomedical information. *Bioinformatics.* 2024;40(2):btad075.
65. Zhang K, Yu J, Yan Z, Liu Y, Adhikarla E, Fu S, et al. BiomedGPT: a unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. 2023. arXiv:2305.17100.
66. Cui H, Wang C, Maan H, Pang K, Luo F, Wang B. scGPT: towards building a foundation model for single-cell multi-omics using generative AI. *Nat Methods.* 2024;21:1470–80.

Authors and Affiliations

Yingbo Zhang^{1,2} · Shumin Ren^{1,3} · Jiao Wang^{1,3} · Junyu Lu¹ · Cong Wu¹ · Mengqiao He¹ · Xingyun Liu^{1,3} · Rongrong Wu¹ · Jing Zhao¹ · Chaoying Zhan¹ · Dan Du⁴ · Zhajun Zhan⁵ · Rajeev K. Singla^{1,6} · Bairong Shen¹ 

✉ Bairong Shen
bairong.shen@scu.edu.cn

¹ Department of Pharmacy and Institutes for Systems Genetics, Frontiers Science Center for Disease-related Molecular Network, West China Hospital, Sichuan University, Chengdu 610212, China

² Tropical Crops Genetic Resources Institute, Chinese Academy of Tropical Agricultural Sciences (CATAS), Haikou 571101, China

³ Department of Computer Science and Information Technology, University of A Coruña, 15071 A Coruña, Spain

⁴ Advanced Mass Spectrometry Center, Research Core Facility, Frontiers Science Center for Disease-Related Molecular Network, West China Hospital/West China Medical School, Sichuan University, Chengdu 610041, China

⁵ College of Pharmaceutical Science, Zhejiang University of Technology, Hangzhou 310014, China

⁶ School of Pharmaceutical Sciences, Lovely Professional University, Phagwara Punjab-144411, India