Benchmarking Biomedical Relation Knowledge in Large Language Models

Fenghui Zhang, Kuo Yang(✉), Chenqian Zhao, Haixu Li, Xin Dong, Haoyu Tian, and Xuezhong Zhou(✉)

Institute of Medical Intelligence, Beijing Key Lab of Traffic Data Analysis and Mining, School of Computer Science and Technology, Beijing Jiaotong University, Beijing 100044, China
{kuoyang,xzzhou}@bjtu.edu.cn

Abstract. As a special knowledge base (KB), a large language model (LLM) stores a great deal of knowledge in the form of the parametric deep neural network, and evaluating the accuracy of the knowledge within this KB has emerged as a key area of interest in LLM research. Although lots of evaluation studies of LLM knowledge have been carried out, due to the complexity and scarcity of biomedical knowledge, there are still few evaluation studies on this kind of knowledge. To address this, we designed five specific identification and evaluation tasks for the biomedical knowledge in LLMs, including the identification of genes for diseases, targets for drugs/compounds, drugs for diseases, and effectiveness for herbs. We selected four well-known LLMs, including GPT-3.5-turbo, GPT-4, ChatGLM-std, and LLaMA2-13B, to quantify the quality of biomedical knowledge in LLMs. Comprehensive experiments that include overall evaluation of accuracy and completeness, ablation analysis, few-shot prompt optimization and case study fully benchmarked the performance of LLMs in the identification of biomedical knowledge and assessed the quality of biomedical knowledge implicit in LLMs. Experimental results showed some interesting observations, e.g., the incompleteness and bias of knowledge of different LLMs, which will give us some insight into LLMs for biomedical discovery and application.

Keywords: biomedical knowledge evaluation · large language model · biomedical relationship identification · benchmarking

1 Introduction

In recent years, advancements in deep learning have driven the development of large language models (LLMs) like GPTs [5,24], marking significant progress in Natural Language Processing (NLP). Built on Transformer architecture [38], these models excel in language understanding and generation, showing near-human proficiency. They serve as implicit knowledge bases (KBs), supporting various NLP tasks [14,25,32]. The emergence of models such as LLaMA [37], Vicuna [6], and ChatGLM [8] signifies advancements in LLM technology.

W. Peng et al. (Eds.): ISBRA 2024, LNBI 14955, pp. 1–14, 2024.
https://doi.org/10.1007/978-981-97-5131-0_41

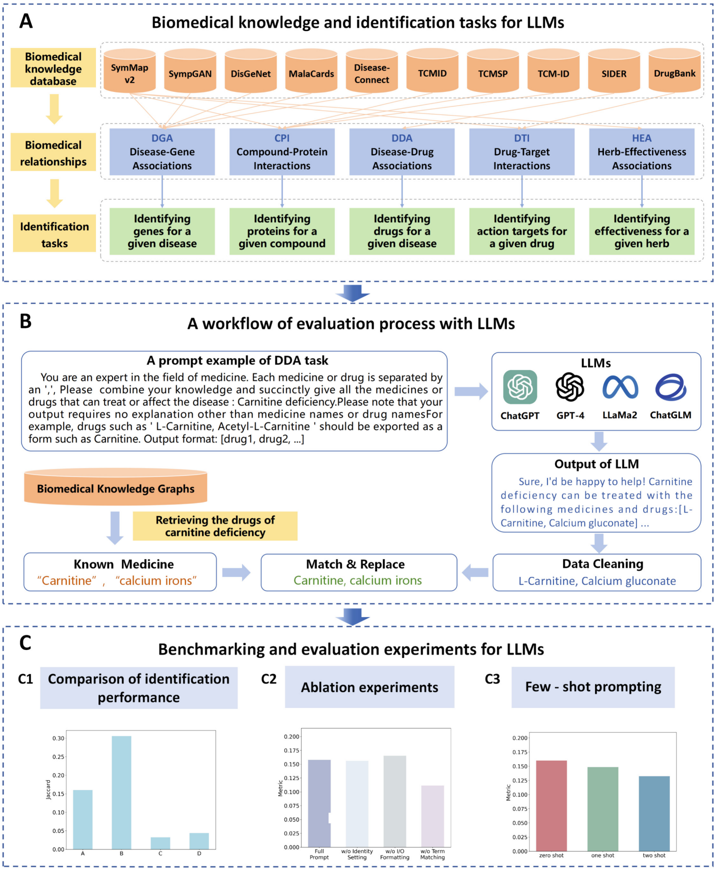
However, alongside their progress, LLMs have been noted to exhibit issues such as the tendency to produce hallucinations or overly confident assertions without sufficient evidence [12,13,21]. This phenomenon presents significant challenges across sectors, especially in fields like medicine and law. Consequently, evaluating the accuracy and reliability of knowledge represented by LLMs across different domains has become a crucial area of research. Researchers such as Kassner et al. [16] and Luo et al. [22] have systematically investigated LLMs' knowledge coverage and precision through comprehensive assessment tasks.

Despite the variety of evaluation tasks for LLMs, two significant issues and challenges remain in evaluating LLMs within the medical domain. First, previous research has focused mainly on the applications of LLMs in medicine. It emphasizes their roles in supporting medical decision-making, diagnosing diseases, and generating summaries of medical literature [30,34,36]. These studies have rarely focused on evaluating the accuracy of the biomedical knowledge contained within LLMs. Second, current evaluation tasks often rely on multiple-choice formats to assess LLMs' knowledge scope and reasoning skills. However, these formats are inadequate in specialized fields like medicine, which require direct and precise answers from LLMs [39]. This is because the complexity and nuance of medical questions require answers with a higher degree of specificity and accuracy than multiple-choice questions [30]. Therefore, constructing an evaluation framework that utilizes biomedical relational knowledge from medical knowledge graphs to evaluate the precision of LLMs' biomedical knowledge is crucial.

To address the above challenges, we drew significant inspiration from established medical evaluation benchmarks such as MedEval [11]. Recognizing the extensive and accurate biomedical relational knowledge in existing medical knowledge graphs (KGs), we collected test data from datasets derived from these KGs. This helped us develop a comprehensive biomedical evaluation framework. Our evaluation framework is grounded in designing five specific identification and evaluation tasks for the biomedical knowledge in LLMs, including the identification of genes of diseases, targets of drugs/compounds, drugs of disease, and effectiveness of herbs (Fig. 1). We selected several well-known LLMs to quantify the quality of biomedical knowledge they contain. We conducted comprehensive experiments that included overall evaluation of accuracy and completeness, ablation analysis, few-shot prompt optimization and case study. These experiments thoroughly benchmarked LLMs' performance in identifying biomedical knowledge and evaluating the quality of biomedical information implicit in LLMs. In summary, the main contributions of our work are three-fold:

1. We constructed a framework for the biomedical evaluation of large language models (LLMs) that contains five specific identification and evaluation tasks.
2. We break new ground by directly evaluating the quality of biomedical knowledge contained in LLMs through data from medical knowledge graphs, moving away from conventional assessment methods that rely on multiple-choice questionnaires.

3. We conducted comprehensive experiments on LLMs that include overall evaluation of accuracy and completeness, ablation analysis, few-shot prompt optimization and case study.


A
Biomedical knowledge and identification tasks for LLMs
Biomedical knowledge database
SymMap v2
SympGAN
DisGeNet
MalaCards
Disease-Connect
TCMID
TCMSP
TCM-ID
SIDER
DrugBank
Biomedical relationships
DGA
Disease-Gene Associations
CPI
Compound-Protein Interactions
DDA
Disease-Drug Associations
DTI
Drug-Target Interactions
HEA
Herb-Effectiveness Associations
Identification tasks
Identifying genes for a given disease
Identifying proteins for a given compound
Identifying drugs for a given disease
Identifying action targets for a given drug
Identifying effectiveness for a given herb
B
A workflow of evaluation process with LLMs
A prompt example of DDA task
LLMs
ChatGPT
GPT-4
LLaMa2
ChatGLM
Output of LLM
Biomedical Knowledge Graphs
Retrieving the drugs of carnitine deficiency
Known Medicine
“Carnitine” , “calcium irons”
Match & Replace
Carnitine, calcium irons
Data Cleaning
L-Carnitine, Calcium gluconate
C
Benchmarking and evaluation experiments for LLMs
C1
Comparison of identification performance
C2
Ablation experiments
C3
Few - shot prompting
Jaccard
0.00
0.05
0.10
0.15
0.20
0.25
0.30
A
B
C
D
Metric
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
Full Prompt
w/o Identity Setting
w/o I/O Formatting
w/o Term Matching
Metric
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
zero shot
one shot
two shot

A prompt example of DDA task

You are an expert in the field of medicine. Each medicine or drug is separated by an ',', Please combine your knowledge and succinctly give all the medicines or drugs that can treat or affect the disease : Carnitine deficiency.Please note that your output requires no explanation other than medicine names or drug namesFor example, drugs such as ' L-Carnitine, Acetyl-L-Carnitine ' should be exported as a form such as Carnitine. Output format: [drug1, drug2, ...]

Output of LLM

Sure, I'd be happy to help! Carnitine deficiency can be treated with the following medicines and drugs:[L-Carnitine, Calcium gluconate] ...

Fig. 1. Overview of the Biomedical Evaluation Framework for LLMs. **(A)** Schematic representation of the biomedical knowledge and identification tasks for LLMs. **(B)** A workflow diagram illustrating the evaluation process of LLMs. **(C)** Benchmarking and evaluation experiments conducted for LLMs, presenting a comparative analysis across multiple dimensions. **(C1)**Identification performance, **(C2)**Ablation experiments, **(C3)** Few-shot prompting coherence.

2 Related Work

2.1 Large Language Models and Hallucinations

The advancement of deep learning has led to LLMs such as the GPT series and LLaMA, achieving breakthroughs in NLP via the Transformer architecture

[7,45,46]. Trained on vast datasets, these models are proficient in generating text that closely mimics human language. However, they sometimes face challenges with factual accuracy due to the ’hallucination problem,’ where their outputs might be incorrect [42]. This issue is particularly concerning in fields that require high precision, such as medicine [28]. To address this, strategies like few-shot learning [5] and Chain of Thought (CoT) reasoning [40] aim to enhance reliability and factual accuracy by encouraging deeper reasoning and utilizing structured knowledge from KGs [2,10].

2.2 Evaluation of LLMs

Despite LLMs’ impressive natural language abilities, it’s crucial to explore their strengths, limitations, and risks fully. Various benchmarks, including BIG-Bench [31], HELM [18], and Promptbench [47], provide thorough evaluations across multiple NLP datasets, facilitating a comprehensive assessment of the performance and biases of LLMs. Besides these general benchmarks, targeted research assesses LLMs in areas such as question answering [3,23], mental health [15], and law [9]. Our research, however, zeroes in on evaluating the completeness and accuracy of LLMs’ biomedical knowledge, addressing a specific gap in LLM assessment.

3 Materials and Methods

This section introduces a framework for evaluating LLMs in biomedical knowledge, covering data acquisition, prompt engineering for guidance, and response processing methods. These are crucial for a comprehensive evaluation of biomedical knowledge retrieval and generation.

3.1 Biomedical Evaluation Framework

This study introduces a comprehensive evaluation framework for LLMs, as illustrated in Fig. 1. Our framework is built around three main components. The first component introduces a set of evaluation tasks designed to probe the LLMs’ capabilities in accurately handling specific biomedical relational knowledge areas. These areas include Disease-Gene Associations (DGA), Compound-Protein Interactions (CPI), Disease-Drug Associations (DDA), Drug-Target Interactions (DTI), and Herb-Effectiveness Associations (HEA). These areas, as detailed in Sect. 3.2, are critical for understanding the scope and depth of biomedical information processed by LLMs. The second component outlines the workflow for evaluating LLMs, including prompt design and processing LLM responses. The methodologies for these are detailed in Sects. 3.3 and 3.4. We designed specific prompts to guide the LLMs in accurately identifying biomedical relationships. The third component details the benchmarking and comprehensive evaluation experiments conducted for LLMs. We compare the results produced by the LLMs against established KBs and utilize various evaluation metrics, as detailed in Sect. 3.5. This allows us to quantify the proficiency of the models in understanding and generating medically relevant information. This framework ensures a rigorous and structured assessment of LLMs' biomedical knowledge retrieval and generation capabilities.

3.2 Knowledge of Biomedical Relationships

To evaluate the performance of LLMs in identifying biomedical relationships, we collected five types of biomedical relationship knowledge from various authoritative biomedical databases, each sourced from one or more KBs. For instance, DGA was compiled from four KBs, namely SymMap v2 [41], DisGeNet [26], MalaCards [27], and DiseaseConnect [20]. Conversely, DDA was obtained exclusively from SIDER [17].

3.3 Prompt Design for Biomedical Tasks

Prompt engineering (PE) plays a crucial role in leveraging the capabilities of LLMs for specialized applications. Through the strategic design of prompts, LLMs can be effectively guided to address specific challenges, which is paramount in scenarios where extensive training datasets are unavailable [5]. In this study, we meticulously designed prompts for each biomedical evaluation task to enhance the performance of LLMs in the medical domain and to rigorously evaluate their proficiency in accurately and comprehensively capturing biomedical knowledge.

3.4 Processing and Matching of LLM Responses

In NLP evaluation tasks, a precise and efficient method for processing LLM responses is crucial to minimize assessment errors. Our study's processing phase includes data cleansing of LLM responses and term matching with KBs. Data cleansing involves extracting target terms from LLM responses, while term matching entails aligning these extracted terms with the entity terms in KBs. This process is essential for evaluating the accuracy of the identification results.

3.5 Experiments and Settings

LLMs for Comparison and Evaluation Metrics. In this study, we selected four typical LLMs, i.e., GPT-3.5-turbo, GPT-4, ChatGLM-std and LLaMA2-13B to conduct the experiments of identifying biomedical relation knowledge. To measure the performance of LLMs in biomedical identification tasks, we adopted evaluation metrics commonly used in this domain [43,44], including precision, recall, F1-score, and the Jaccard Index (JI) to comprehensively measure the accuracy and completeness of the identification results.

4 Results

We evaluated five designed biomedical relation discovery tasks from multiple aspects, including comparing identification performance, conducting component-specific ablation analysis, optimizing few-shot prompting, and conducting case studies.

4.1 Overall Comparison of Identification Performance

In this section, we use the identification of five biomedical relationships as examples to evaluate both the accuracy and completeness of the LLM KB. A suite of LLMs was utilized for this purpose. The performance results for the five tasks are detailed in Fig. 2, with the following observations.

First, notable disparities in identification performance among different LLMs were observed for the same tasks. For instance, in the DGA, CPI, and DTI tasks, GPT-4 and GPT-3.5-turbo demonstrated superior performance, whereas ChatGLM-std and LLaMA2-13B lagged. Conversely, in DDA task, GPT-4 and ChatGLM-std excelled, while GPT-3.5-turbo and LLaMA2-13B showed subpar performance. In summary, no single model consistently outperformed the others across all tasks, nor did any model uniformly exhibit the lowest performance.

Second, the performance of a given LLM varied significantly across different tasks. For instance, GPT-4 achieved its best performance on the DGA task with a JI of 0.3059 and an F1-score of 0.4510. In contrast, its poorest performance was on the CPI task, with a JI of 0.0254 and an F1-score of 0.0612. Regarding ChatGLM-std, it performed optimally on the HEA task. However, it showed poor results on the DGA (JI = 0.0324, F1 = 0.1774), CPI, and DDA task.

Third, GPT-4 outperforms other models in all English tasks except for the Chinese HEA task, where ChatGLM-std, tailored to a Chinese corpus, performs better. The performance of GPT-3.5-turbo is lower than that of GPT-4 but superior to the other LLMs. While ChatGLM-std excels in Chinese HEA, it lags in English tasks. LLaMA2-13B, with a training corpus significantly smaller than that of GPTs, exhibits notably inferior performance across most tasks.

Lastly, in the only Chinese task of HEA, ChatGLM-std significantly outperforms GPTs and LLaMA2-13B, which are primarily trained on English corpora. This highlights the significant influence of linguistic bias in a training corpus on the identification performance of LLMs. ChatGLM-std is the only model primarily trained on a Chinese language corpus among all the LLMs evaluated.

Overall, a significant disparity is observed in the composite performance of LLMs across various tasks. GPT-4 leads in most areas, with GPT-3.5-turbo closely following. LLaMA2-13B shows potential for improvement, especially in medical tasks, while ChatGLM-std excels in Chinese HEA task. Through this experiment, it is evident that different LLMs exhibit varying performance levels in biomedical relationship tasks.

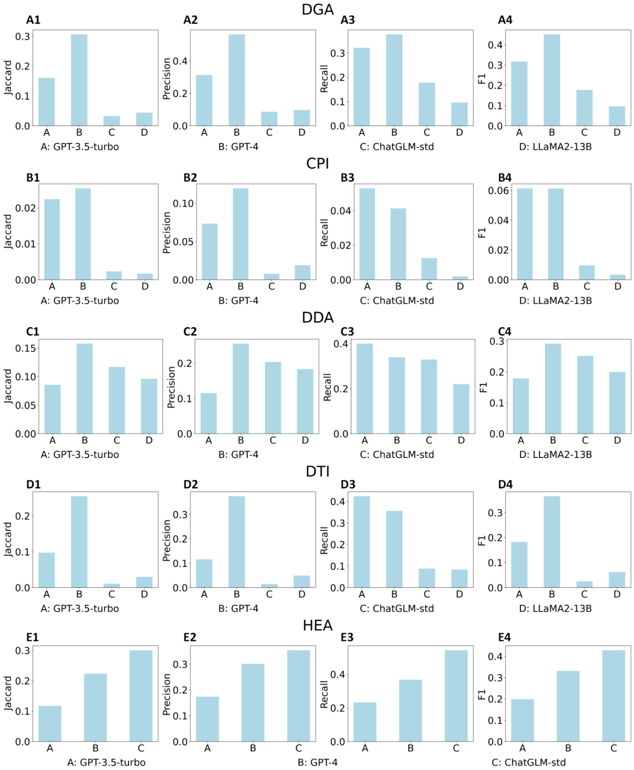


Fig. 2. Comprehensive performance evaluation of four LLMs across five biomedical relation discovery tasks using Jaccard index, precision, recall, and F1 score as metrics. Models compared are A: GPT-3.5-turbo, B: GPT-4, C: ChatGLM-std, and D: LLaMA2-13B. **(A1–A4)** Evaluation performance of DGA tasks; **(B1–B4)** Evaluation performance of CPI tasks; **(C1–C4)** Evaluation performance of DDA tasks; **(D1–D4)** Evaluation performance of DTI tasks; and **(E1–E4)** Evaluation performance of HEA tasks.

4.2 Ablation Experiment

To evaluate the effects of identity setting, output formatting, and terminology matching components within prompt design on the ability of LLMs, we conducted a series of ablation studies. These studies involved the independent removal of each component from the prompt to observe and compare the performance of different LLMs in the DDA task. The results of these experiments are presented in Table 1, with the ensuing observations.

The identity setting has a significant positive effect on the identification performance of the GPT-4 model in the DDA task. After removing this component, its performance has declined. However, GPT-3.5-turbo and LLaMA2-13B improved performance after removing this component. This variation suggests that the identity setting plays a different role across LLMs; it can significantly

Table 1. Results of the ablation study demonstrate the impact of specific prompt components on the performance of LLMs for the DDA task. The ablation components include identity setting, I/O formatting, and term matching.

LLMs	Component of LLMs	Jaccard	Precision	Recall	F1
GPT-4	Full Prompt	0.1578	0.2557	0.3390	0.2915
	w/o Identity Setting	0.1562	0.2560	0.3230	0.2856
	w/o I/O Formatting	**0.1651**	**0.2682**	**0.3526**	**0.3047**
	w/o Term Matching	0.1114	0.1812	0.2595	0.2134
GPT-3.5-turbo	Full Prompt	0.0856	0.1148	**0.4000**	0.1784
	w/o Identity Setting	0.1061	**0.1522**	0.3747	**0.2165**
	w/o I/O Formatting	**0.1076**	0.1471	0.3959	0.2145
	w/o Term Matching	0.0610	0.0820	0.2975	0.1286
LLaMA2-13B	Full Prompt	0.0960	0.1833	0.2198	0.1999
	w/o Identity Setting	**0.1066**	**0.1862**	0.2661	**0.2191**
	w/o I/O Formatting	0.0835	0.1256	**0.2707**	0.1715
	w/o Term Matching	0.0564	0.1095	0.1352	0.1210

enhance GPT-4's ability to recognize biomedical knowledge while potentially limiting the flexibility and adaptability of GPT-3.5-turbo and LLaMA2-13B in processing the task.

Upon removing the output formatting component, a decrease in LLaMA2-13B's performance was observed, while GPT-4 and GPT-3.5-turbo improved their performances. This indicates that output formatting benefits the model performance of LLaMA2-13B but may have a negative impact on GPT-4 and GPT-3.5-turbo. The removal of the terminology matching component resulted in a noticeable decline in the identification capabilities of all three models, with GPT-4 experiencing the most significant drop in performance. This underscores the positive impact of terminology matching on the overall effectiveness of LLMs, highlighting the critical role of accurate terminology matching in the performance of biomedical identification tasks by LLMs.

Overall, the identity setting, output formatting, and terminology matching, three prompt components, are specific to LLMs, and their impacts on the ability of different LLMs to handle biomedical identification tasks vary. This reveals that the design of prompts for LLMs should not adopt a one-size-fits-all approach but should be personalized for different LLMs and tasks.

4.3 Analysis of Few-Shot Prompt Optimization

In order to investigate the impact of FSP on the efficacy of LLMs in performing identification tasks, we experimented using GPT-3.5-turbo. We implemented FSP optimization to evaluate the performance of various FSP strategies on different identification tasks. Specifically, we selected three tasks for this exploration:

DGA, CPI, and DTI, to evaluate the FSP approaches of zero-shot, one-shot, and two-shot for comparison. The results presented in Table 2 delineate the performance results of the GPT-3.5-turbo model on these medical tasks post the application of FSP techniques. From this investigation, we distilled two main observations.

Table 2. Comparative analysis of GPT-3.5-turbo performance employing FSP strategies across different biomedical tasks, quantified by JI, precision, recall, and F1 score.

Evaluation Tasks	Few-shot	Jaccard	Precision	Recall	F1
DGA	zero-shot	**0.1600**	**0.3127**	0.3213	0.3169
	one-shot	0.1485	0.2722	0.4081	**0.3266**
	two-shot	0.1325	0.2385	**0.4616**	0.3145
CPI	zero-shot	0.0224	0.0732	0.0528	0.0613
	one-shot	**0.0725**	**0.2063**	**0.1366**	**0.1644**
	two-shot	0.0500	0.1484	0.0944	0.1154
DTI	zero-shot	0.0974	0.1159	0.4238	0.1821
	one-shot	**0.1617**	**0.2225**	0.3618	**0.2755**
	two-shot	0.1510	0.1971	**0.4502**	0.2741

First, implementing FSP significantly improved GPT-3.5-turbo's identification capabilities in the CPI and DTI tasks, underscoring the technique's beneficial impact. Furthermore, it was observed that employing one-shot prompting resulted in a more significant improvement in the performance of GPT-3.5-turbo compared to two-shot prompting. This indicates that accumulating too much prior knowledge does not always lead to better performance. Often, utilizing more concise information can lead to better results.

Second, the effectiveness of FSP in enhancing the identification performance of GPT-3.5-turbo is not consistently applicable to all tasks. Notably, it negatively affected the DGA task, diminishing the model's identification capabilities. This adverse outcome may stem from the prior knowledge imparting some level of misdirection to LLMs, complicating the acquisition of accurate results. This phenomenon highlights that the effectiveness of FSP techniques in enhancing LLMs' response quality is not a certainty. Their success largely depends on the specific tasks the LLMs are performing.

4.4 Case Analysis

To compare the difference of LLMs in biomedical identification tasks, we selected two representative diseases, i.e., Basal Ganglia Calcification, Idiopathic 1, and Generalized Junctional Epidermolysis Bullosa, Non-Herlitz Type, and evaluated the performance of four LLMs in the DGA task.

The results of the first case showed GPT-4 found five genes, four of which are known genes recorded in KBs, while GPT-3.5-turbo identified three known genes (Table 3). In contrast, ChatGLM-std and LLaMA2-13B failed to identify any known genes, indicating limitations in specific gene identification. In the second case, both GPT-3.5-turbo and GPT-4 accurately identified all known genes. ChatGLM-std, despite identifying a broader set of genes, included only a subset of known genes, suggesting potential issues with overgeneralization. LLaMA2-13B's outputs did not correctly identify any genes.

Above cases highlight the diverse efficacy of LLMs in biomedical tasks, showcasing the strengths of GPT-4 and GPT-3.5-turbo in accurately identifying disease genes. In contrast, the limitations observed in ChatGLM-std and LLaMA2-13B emphasize the need for a cautious evaluation of their outputs. These case studies highlight the performance differences among LLMs and provide insights into LLMs for handling biomedical tasks.

Table 3. Comparison of Cases under Different LLMs in DGA Task

Disease	LLMs	Output
Basal Ganglia Calcification, Idiopathic, 1	Known Genes	ISG15, SLC20A2, RAB39B, TYROBP, PDGFRB, THAP1, XPR1, PDGFB, IBGC1, IBGC2, SNTG1
	GPT-3.5-turbo	PANK2, **SLC20A2**, **PDGFRB**, **XPR1**, MYORG
	GPT-4	**SLC20A2**, **PDGFB**, **PDGFRB**, **XPR1**, MYORG
	ChatGLM-std	FGF23, GBA, GHR, HSD17B4, SLC34A1, SLC34A2, SLC34A3, SLC34A4, SLC34A5, SLC34A6, SLC34A7
	LLaMA2-13B	STK11, PTEN, TCF4, AKT1, P53
Generalized Junctional Epidermolysis Bullosa, Non-Herlitz Type	Known Genes	LAMA3, LAMB3, LAMC2, COL17A1
	GPT-3.5-turbo	**LAMA3**, **LAMB3**, **LAMC2**, **COL17A1**, ITGB4, PLEC1
	GPT-4	**LAMA3**, **LAMB3**, **LAMC2**, **COL17A1**
	ChatGLM-std	**COL17A1**, COL17A2, ITGB4, ITGB5, **LAMA3**, **LAMB3**, **LAMC2**, LAT1, LAT2, TGM1, TGM5
	LLaMA2-13B	COL5A1, COL5A2, CYP26B1, KRT5, KRT14, P53

5 Discussion

Advances in LLM technology are rapid. Benchmarking their problem-solving capabilities across various fields, especially in biomedicine, has become a key area of research and interest. In this study, we have designed five evaluation tasks to investigate the biomedical knowledge embedded in LLMs. To benchmark the performance of LLMs in identifying biomedical knowledge and assessing the quality of biomedical knowledge implicit in LLMs, we conducted comprehensive experiments. Our experiments with LLMs have led to several intriguing findings. First, performance comparisons show that GPT-4 leads in LLM technology. It performs better on most tasks than other large models. Notably, ChatGLM-std achieves the best results in the Chinese HEA task, distinguishing itself from other LLMs. Second, the effectiveness of GPT-4 in identification tasks significantly depends on the nature of those tasks. It shows remarkable results in the DGA task, whereas its performance dips in the CPI task. Third, our ablation study highlights the significance of prompt design in influencing LLM performance. Prompt design elements are crucial. They greatly affect which LLMs are chosen and how well they perform in various identification tasks. These elements are valid on some LLMs but not on others, and they work on some identification tasks but not on others.

There is some work to be done in the future. First, biomedical relationships vary widely. However, this study's evaluations have a limited scope, covering only a small portion of medical knowledge. Our evaluation lacks intricate relationships. These include links between symptoms and genes or drug interactions, crucial for enhancing our understanding and using LLMs effectively in medicine [29]. Second, the rapid evolution of LLM technology has led to the emergence of numerous new models (e.g., Alpaca [33] and LaMDA [35]). The LLMs examined in this research do not offer a comprehensive overview. In future studies, we aim to broaden our analysis. We plan to include more LLMs to enhance the evaluation of the biomedical knowledge these models possess. Third, prompt design and the few-shot prompting technique have enhanced task performance for certain LLMs. However, instances of negative effects are still present. Furthermore, our substring matching and similarity matching strategies may not handle biological entities with very different synonymous names effectively. For example, the drug "Tafluprost" and its synonymous "Zioptan" are semantically identical but differ significantly in string representation, posing challenges for our current methods. Future efforts will involve developing more refined and specific strategies for prompt optimization and incorporating medical semantic databases such as UMLS [4] and MeSH [19] to improve matching accuracy. Overall, our future work will focus on expanding the scope of evaluation tasks, evaluating newly developed LLMs, and refining the design of prompts. This comprehensive approach aims to enhance the evaluation of the quality of biomedical knowledge in LLMs and improve their capacity to address complex biomedical challenges [1].

Funding Information. This work is partially supported by the National Key Research and Development Program (Nos. 2023YFC3502604 and 2021YFC1712901), the National Natural Science Foundation of China (Nos. 82174533, 82374302, 82204941, and U23B2062), the Natural Science Foundation of Beijing (No. L232033), Beijing-Tianjin-Hebei Basic Research Cooperation Project (No. 22JCZXJC00070) and Key R&D Program Project of Ningxia Hui Autonomous Region (2022BEG02036).

References

1. Abd-Alrazaq, A., et al.: Large language models in medical education: opportunities, challenges, and future directions. JMIR Med. Educ. **9**(1), e48291 (2023)
2. Agrawal, G., Kumarage, T., Alghamdi, Z., Liu, H.: Can knowledge graphs reduce hallucinations in LLMs?: a survey. arXiv arXiv:2311.07914 (2024)
3. Bang, Y., et al.: A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. arXiv arXiv:2302.04023 (2023)
4. Bodenreider, O.: The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. **32**(Suppl–1), D267–D270 (2004)
5. Brown, T., et al.: Language models are few-shot learners. Adv. Neural. Inf. Process. Syst. **33**, 1877–1901 (2020)
6. Chiang, W.L., et al.: Vicuna: an open-source chatbot impressing GPT-4 with 90%* ChatGPT quality, March 2023
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota, June 2019, pp. 4171–4186. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/N19-1423
8. Du, Z., et al.: GLM: general language model pretraining with autoregressive blank infilling. arXiv arXiv:2103.10360 (2022)
9. Fei, Z., et al.: LawBench: benchmarking legal knowledge of large language models. arXiv arXiv:2309.16289 (2023)
10. Guan, X., et al.: Mitigating large language model hallucinations via autonomous knowledge graph-based retrofitting. arXiv arXiv:2311.13314 (2023)
11. He, Z., et al.: MedEval: a multi-level, multi-task, and multi-domain medical benchmark for language model evaluation. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, December 2023, pp. 8725–8744. Association for Computational Linguistics (2023). https://doi.org/10.18653/v1/2023.emnlp-main.540
12. Huang, L., et al.: A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. arXiv arXiv:2311.05232 (2023)
13. Ji, Z., et al.: Survey of hallucination in natural language generation. ACM Comput. Surv. **55**(12), 1–38 (2023)
14. Jiang, Z., Xu, F.F., Araki, J., Neubig, G.: How can we know what language models know? Trans. Assoc. Computat. Linguist. **8**, 423–438 (2020)
15. Jin, H., Chen, S., Wu, M., Zhu, K.Q.: PsyEval: a comprehensive large language model evaluation benchmark for mental health. arXiv arXiv:2311.09189 (2023)
16. Kassner, N., Schütze, H.: Negated and misprimed probes for pretrained language models: birds can talk, but cannot fly. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 7811–7818. Association for Computational Linguistics, Online, July 2020. https://doi.org/10.18653/v1/2020.acl-main.698

17. Kuhn, M., Letunic, I., Jensen, L.J., Bork, P.: The SIDER database of drugs and side effects. Nucleic Acids Res. **44**(D1), D1075–D1079 (2016)
18. Liang, P., et al.: Holistic evaluation of language models. arXiv preprint arXiv:2211.09110 (2022)
19. Lipscomb, C.E.: Medical subject headings (MESH). Bull. Med. Libr. Assoc. **88**(3), 265 (2000)
20. Liu, C.C., et al.: DiseaseConnect: a comprehensive web server for mechanism-based disease-disease connections. Nucleic Acids Res. **42**(W1), W137–W146 (2014)
21. Luo, J., Li, T., Wu, D., Jenkin, M., Liu, S., Dudek, G.: Hallucination detection and hallucination mitigation: an investigation. arXiv arXiv:2401.08358 (2024)
22. Luo, L., Vu, T., Phung, D., Haf, R.: Systematic assessment of factual knowledge in large language models. In: Findings of the Association for Computational Linguistics, EMNLP 2023, pp. 13272–13286. Association for Computational Linguistics, Singapore, December 2023. https://doi.org/10.18653/v1/2023.findings-emnlp.885
23. Omar, R., Mangukiya, O., Kalnis, P., Mansour, E.: ChatGPT versus traditional question answering for knowledge graphs: current status and future directions towards knowledge graph chatbots. arXiv arXiv:2302.06466 (2023)
24. OpenAI: GPT-4 technical report. arXiv arXiv:2303.08774 (2024)
25. Petroni, F., et al.: Language models as knowledge bases? In: Inui, K., Jiang, J., Ng, V., Wan, X. (eds.) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 2463–2473. Association for Computational Linguistics, Hong Kong, China, November 2019. https://doi.org/10.18653/v1/D19-1250
26. Piñero, J., et al.: DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res. **45**(D1), D833–D839 (2016)
27. Rappaport, N., et al.: MalaCards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search. Nucleic Acids Res. **45**(D1), D877–D887 (2017)
28. Roberts, A., Raffel, C., Shazeer, N.: How much knowledge can you pack into the parameters of a language model? arXiv arXiv:2002.08910 (2020)
29. Schaefer, M., et al.: Large language models are universal biomedical simulators. bioRxiv (2023). https://doi.org/10.1101/2023.06.16.545235
30. Singhal, K., et al.: Large language models encode clinical knowledge. nature **620**(7972), 172–180 (2023)
31. Srivastava, A., et al.: Beyond the imitation game: quantifying and extrapolating the capabilities of language models. arXiv preprint arXiv:2206.04615 (2022)
32. Talmor, A., Elazar, Y., Goldberg, Y., Berant, J.: oLMpics-on what language model pre-training captures. Trans. Assoc. Comput. Linguist. **8**, 743–758 (2020)
33. Taori, R., et al.: Alpaca: a strong, replicable instruction-following model. Stanford Center Res. Found. Models **3**(6), 7 (2023)
34. Thirunavukarasu, A.J., Ting, D.S.J., Elangovan, K., Gutierrez, L., Tan, T.F., Ting, D.S.W.: Large language models in medicine. Nat. Med. **29**(8), 1930–1940 (2023)
35. Thoppilan, R., et al.: LaMDA: language models for dialog applications. arXiv preprint arXiv:2201.08239 (2022)
36. Tian, S., et al.: Opportunities and challenges for ChatGPT and large language models in biomedicine and health. Brief. Bioinform. **25**(1), bbad493 (2024)
37. Touvron, H., et al.: LLaMA: open and efficient foundation language models. arXiv arXiv:2302.13971 (2023)

38. Vaswani, A., et al.: Attention is all you need. Adv. Neural. Inf. Process. Syst. **30**, 5998–6008 (2017)
39. Wang, C., et al.: Evaluating Open-QA evaluation. Adv. Neural Inf. Process. Syst. **36** (2024)
40. Wei, J., et al.: Chain-of-thought prompting elicits reasoning in large language models. Adv. Neural. Inf. Process. Syst. **35**, 24824–24837 (2022)
41. Wu, Y., et al.: SymMap: an integrative database of traditional Chinese medicine enhanced by symptom mapping. Nucleic Acids Res. **47**(D1), D1110–D1117 (2019)
42. Xie, J., Zhang, K., Chen, J., Lou, R., Su, Y.: Adaptive chameleon or stubborn sloth: unraveling the behavior of large language models in knowledge conflicts. arXiv preprint arXiv:2305.13300 (2023)
43. Yang, K., et al.: HerGePred: heterogeneous network embedding representation for disease gene prediction. IEEE J. Biomed. Health Inform. **23**(4), 1805–1815 (2018)
44. Yang, K., et al.: PDGNet: predicting disease genes using a deep neural network with multi-view features. IEEE/ACM Trans. Comput. Biol. Bioinf. **19**(1), 575–584 (2020)
45. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V.: XLNet: generalized autoregressive pretraining for language understanding. In: Wallach, H., et al. (eds.) Advances in Neural Information Processing Systems, vol. 32. Curran Associates, Inc. (2019)
46. Zhao, W.X., et al.: A survey of large language models. arXiv arXiv:2303.18223 (2023)
47. Zhu, K., et al.: PromptBench: towards evaluating the robustness of large language models on adversarial prompts. arXiv preprint arXiv:2306.04528 (2023)