Evaluating LLMs as Pharmaceutical Care Decision Support Tools Across Multiple Case Scenarios

Vania Amanda Samor[1], Muhammad Yeza Baihaqi[2], Edmun Halawa[3], Luh Rai Maduretno Asvinigita[4], Sarah Nabila Hakim[5], and Mela Septi Rofika[6]

[1] Pharmacy Study Program, Faculty of Health Sciences, Universitas Malahayati, Lampung, Indonesia
[2] Information Sciences Division, Nara Institute of Science and Technology, Nara, Japan
[3] Department of Mechanical Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan
[4] Bhakti Widya Farma (BWF) Pharmacy, Badung, Bali, Indonesia
[5] Department of Pharmacy, Pertamina Central Hospital, Jakarta, Indonesia
[6] Department of Pharmacy, Public Health Center of Pamolokan, Sumenep, Indonesia
svaniamanda@malahayati.ac.id

Abstract. In the evolving landscape of healthcare, pharmacists face increasing challenges in providing accurate, reliable, and prompt patient care amidst growing complexity in clinical settings. The continuous advancement of diseases, pharmaceutical sciences, and treatment guidelines requires pharmacists to stay up-to-date. However, the real-world burden of non-clinical tasks often impedes this effort. Recent practice of Large Language Models (LLMs) offers promising potential to support pharmacists in their professional duties. This study aims to evaluate the capability of LLMs in assisting pharmacists with pharmaceutical care decision-making. Three pharmaceutical cases (hypertension, hyperlipidemia, and angina pectoris) and related guidelines were input into the LLM, and their responses were assessed through both subjective and objective evaluations. The results indicated that, despite our efforts, the LLM fell short of satisfactory performance in terms of accuracy and reasoning. It was evident that the LLM's outputs still required human supervision and could not be accepted without scrutiny. However, the experts agreed that the LLM would be beneficial as a reference tool and in facilitating faster decision-making. Future research will focus on improving LLM performance.

Keywords: Healthcare, Large Language Models, Pharmaceutical Care Decision-Making, Artificial Intelligence

INTRODUCTION

Clinical pharmacists face challenges daily, reflecting the evolving demands of their profession. Their roles encompass a broad spectrum of responsibilities, including clinical tasks such as performing medication review and reconciliation, adverse drug events and interactions monitoring and prevention, and providing pharmaceutical care

D. W. S. R. Wardani and S. Hadi (eds.), *Proceedings of the International Conference on Medical Science and Health (ICOMESH 2024)*, Advances in Health Sciences Research 82,
https://doi.org/10.2991/978-94-6463-604-8_24

and interactions monitoring and prevention, and providing pharmaceutical care suggestions to patients and physicians [1, 2]. In some regions, clinical pharmacists even handle non-clinical duties including administrative tasks [3, 4]. If not properly addressed these factors can lead to medication errors, caused by pharmacist-related burnout [5, 6]. Pharmaceutical care decision-making which based on their critical thinking skills were thought to be affected too [7].

Several steps had to be followed in order to make appropriate pharmaceutical care decisions, including administrative, pharmaceutical, and clinical screenings. Two information systems were developed recently for specific clinical assessment tasks. One was KALIS, an electronic database designed to detect prescribing errors and drug-drug interactions using pharmacological data, case reports, and biomolecular integration [8]. The other was PRIMA-eDS, which combined dosing support and decision support through evidence-based medicine databases to predict polypharmacy [9]. However, neither system utilized LLMs, and their functionality was limited to clinical screening.

Therefore, we utilized advance development in artificial intelligence, LLMs. The advantages of using LLMs lie within their processing and analyzing large amounts of data quickly, contextually, and provide consistent and evidence-based responses across diverse cases [10, 11]. We employed the LLM as a decision-making support tool in pharmaceutical care. By providing case studies and guidelines through prompting, the LLM could offer information for the pharmacist. In this research, we specifically focused on three case studies: hypertension, hyperlipidemia, and angina pectoris.

SUBJECT AND METHOD

Large Language Model (LLM)

Large Language Models (LLMs) are a type of artificial intelligence model designed to understand and generate human language naturally [10]. A key characteristic of LLMs is their ability to comprehend and respond to many types of questions, commands, and texts in multiple languages with high fluency and accuracy [11, 12]. This technology has become central to many modern AI applications, from dialogue agents [13], health message generator [14] and conversational recommendation systems [15].

In this research, we specifically utilized an LLM to support pharmacists in pharmaceutical care decision-making. We provided the LLM with pharmaceutical case studies and related guidelines through prompting. The case studies also featured questions that the LLM needed to answer. The LLM-generated answers were then evaluated based on both objective and subjective evaluations. In addition, our experiment was conducted using GPT-4o mini. The framework of this research is shown in Figure 1.

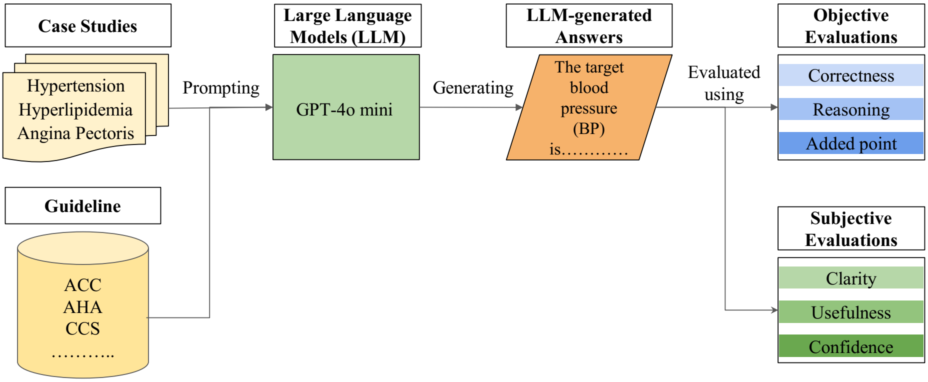

Fig 1. Framework for using LLM in pharmaceutical care decision-making

Case study selection and rationale

To evaluate the ability of LLMs to assist in pharmaceutical care decision-making, we selected three specific case studies from the PharmDia database (https://pharmdia.com/) [16]. Case 1 were selected to represent hypertension case (H), hyperlipidemia (L) was tested using case study 9, and angina pectoris (A) was assessed with case study 5. Additionally, we classified the questions into three types: therapeutic goals and targets (G), medication and therapeutic regimen (M), and follow-up and condition management (F). The combination of case study and question types abbreviations created unique question IDs. The list of questions is presented in Table 1.

Table 1. List of questions in three case studies.

Case	Type	Questions
H	G	What is the target goal for blood pressure in this patient?
	M	What are the main classes of anti-hypertensives that can be used in this case?
	M	Prepare a therapeutic regimen for this patient!
L	G	What would be the target goal for LDL-C in this patient?
	M	What is the drug of choice in this patient to treat LDL-C?
	M	What would be choice of drug in this patient, if he is intolerant to statin therapy
A	G	What is the treatment goal and strategy for this case?
	F	Suggest the best follow-up for this case!
	F	What are the conditions which worsens the symptoms of angina (in general)?

Objective evaluations on LLM-generated answer

We evaluated LLM-generated answers objectively using three metrics: correctness, reasoning and added point, based on the answer key provided by PharmDia. An answer was considered correct if it was both accurate and complete. If the answer was incomplete, it was considered incorrect. For example, if the key listed three reagents but the LLM provided only two, it was marked as wrong.

The LLM was also required to provide reasoning for each answer to gain trust. Reasoning was considered correct if it supported a correct answer with accurate rationale. An LLM could generate a correct answer with incorrect reasoning; for example, if it suggested the correct reagent but provided an incorrect reason for its use, then the reasoning metrics were counted as incorrect, even though the correctness metrics were counted as true.

At certain trials, the LLM was found to provide additional useful information, denoted as added point. For instance, if the key listed three reagents but the LLM provided more than three, these extra points were noted. However, added points were only counted if the initial answer was correct.

Expert subjective evaluations

We recruited three certified pharmacists with 1-3 years of experience in clinical and community settings for a subjective evaluation. For this evaluation, we selected several correct answers with appropriate reasoning from LLM-generated responses and asked the pharmacists to assign scores. The pharmacists assessed the quality of the LLM's responses based on three metrics: clarity (how easy the information is to understand), usefulness (how helpful the information is for making pharmaceutical care decisions), and confidence (how much the pharmacists trust the information for implementing their decisions), using a 5-point Likert scale.

RESULTS

Objective evaluation results on hypertension case

From the results, we observed on Table 2 that in addressing question HG1, all three trials yielded correct answers accompanied by precise reasoning; however, only the third trial presented an added point. For question HM2, while the first and third trials generated correct responses, only the first trial exhibited accurate reasoning, whereas the third trial provide correct reasoning. The second trial for HM2, on the other hand, was incorrect, demonstrated flawed reasoning, and lacked added point. Regarding question HM3, only the second trial provided a correct answer with accurate reasoning and included an added point.

Objective evaluation results on hyperlipidemia case

Through the investigation of hyperlipidemia case study, we noticed that all LG1 responses were accurate, well-reasoned and added point were observed in both the first and third trials, which the second trial lacked. LM2 consistently produced correct answers and reasoning across all trials but failed to provide added points. LM3 exhibited variability, with only the second trial satisfying all criteria of correctness, reasoning, and added point. Overall, LG1 showcased the most reliable performance, achieving 5 out of 6 metrics across three trials. Detailed information was shown on Table 3.

Table 2. LLM-generated answers on hypertension case evaluation results.

Question ID	Trial	Correctness	Reasoning	Added Point
HG1	1	✓	✓	✕
	2	✓	✓	✕
	3	✓	✓	✓
HM2	1	✓	✕	✓
	2	✕	✕	✕
	3	✓	✓	✕
HM3	1	✕	✕	✕
	2	✓	✓	✓
	3	✕	✕	✕

✓: Correct/Available ✕: Incorrect/Unavailable

Objective evaluation results on angina pectoris case

Table 4 indicated that evaluation done on angina pectoris, LLM consistently excelled on AG1, achieving perfect correctness, reasoning, and added point for all questions. In contrast, answer of AF2 showed mixed results: the LLM correctly answered the first question, provide well reason and added points but failed to provide so for the remaining two. Lastly, LLM unable to produce good results on all metrics for AF3.

Objective evaluation results on all case studies

We summarized LLM performance as percentage of metrics in all case studies in Table 5. For therapeutic goals and targets type of question, the two questions demonstrated exceptional performance, achieving 100% in both correctness and reasoning scores, with an added point score of 66.67%. Whereas, in the category of medication and therapeutic regimen type, which included four questions, the performance was lower, with correctness at 58.33%, reasoning at 50%, and additional points at 25%. The follow-up and condition management questions type, comprising three questions, exhibited the lowest performance, with both correctness and reasoning at 16.67%, and additional points also at 16.67%. Overall, the mean scores across all categories indicated a general correctness rate of 69.97%, reasoning 59.26% and added point of 37.04%.

Table 3. LLM-generated answers on hyperlipidemia case evaluation results.

Question ID	Trial	Correctness	Reasoning	Added Point
LG1	1	✓	✓	✓
	2	✓	✓	×
	3	✓	✓	✓
LM2	1	✓	✓	×
	2	✓	✓	×
	3	✓	✓	×
LM3	1	×	×	×
	2	✓	✓	✓
	3	×	×	×

✓: Correct/Available ×: Incorrect/Unavailable

Table 4. LLM-generated answers on angina pectoris case evaluation results.

Question ID	Trial	Correctness	Reasoning	Added Point
AG1	1	✓	✓	✓
	2	✓	✓	✓
	3	✓	✓	✓
AF2	1	✓	✓	✓
	2	×	×	×
	3	×	×	×
AF3	1	×	×	×
	2	×	×	×
	3	×	×	×

✓: Correct/Available ✕: Incorrect/Unavailable

Expert subjective evaluation

The LLM's information was rated highly for Clarity by pharmacists, achieving an average score of 4.67 with standard deviation (*SD* = 0.33), indicating it was exceptionally clear. For Usefulness, the LLM received average rating of 3.67 (*SD* = 0.67), showing variability in perceived utility. Confidence in using LLM-generated information in practice was moderately high, with a mean score of 4.22 (*SD* = 0.69).

Table 5. Objective evaluation in all case studies.

Question Type	Total Question	Percentage (%)		
		Correctness	Reasoning	Added Point
G	2	100	100	66.67
M	4	58.33	50	25
F	3	16.67	16.67	16.67
Mean		69.97	59.26	37.04

DISCUSSION

Our study investigated the capability of LLM in assisting pharmacists with pharmaceutical care decision support tools by testing the LLM with three case studies often occurring in clinical pharmacy setting such as hypertension, hyperlipidemia and angina pectoris.

LLMs achieved perfect accuracy in addressing questions about therapeutic goals and targets. For example, questions related to blood pressure target goals or triglyceride levels. This was due to the straightforward and universally agreed-upon nature of these metrics.

Furthermore, questions about medication and therapeutic regimens, such as those concerning the drug of choice obtained lower accuracy. Frequently, even when the LLM provided an answer, it introduced information not present in the key answers, or while the answers were generally correct, they included extraneous details not aligned with the key answers. It was due to the complex integration of multiple factors, including drug lists and patient conditions. The decision-making process was further complicated by the need to consider patient-specific factors, such as comorbidities and contraindications, which often required alternative therapeutic approaches.

LLM performance was notably lowest for follow-up and condition management questions. For example, questions asked for the best follow-up recommendations for patients with specific conditions. Because the questions were too abstract, the LLM tended to generate broad responses that often deviated significantly from the key answers. While the answers were not inherently incorrect, they were highly subjective and reliant on the assessor's interpretation. This lower accuracy was due to the requirement for complex, patient-specific critical thinking. The reduced accuracy was attributable to the variability in follow-up recommendations and management strategies, which differed based on patient conditions and diverse clinical guidelines.

Despite our efforts in utilizing the LLM, it did not reach satisfactory levels in both correctness and reasoning. It was evident that the LLM's output still necessitated human oversight and could not be accepted uncritically. Nevertheless, expert reviews indicated that such a system holds significant promise and utility in aiding decision-making processes.

The expert reviews of using LLM for pharmaceutical care decision-making highlighted several key points. Firstly, the information generated by the LLM was easy to understand, making it accessible to both healthcare professionals and non-professionals. Secondly, LLMs significantly sped up the decision-making process by providing quick and relevant information. With practice, LLMs helped pharmacists recall specific health conditions and care details, aiding newly graduated pharmacists in discussions with senior colleagues. However, experts agreed that LLMs were a supportive tool and should not replace professional judgment, as discussions among health practitioners were still necessary for making well-informed pharmaceutical care decisions.

Future research could focus on developing novel techniques to enhance LLM performance, such as applying different prompting strategies and fine-tuning. Additionally, experts noted that the LLM's responses were often too lengthy, suggesting that future efforts should aim to make answers more concise yet insightful. Furthermore, exploring various interface types, such as application-based, web-based, or even social robots [17], could be important for enhancing user interaction and accessibility. Lastly, incorporating considerations such as the availability of drugs in the hospital when generating the answer would be beneficial.

CONCLUSION

Our study demonstrated the potential of LLMs in supporting pharmacists with pharmaceutical care decision-making. LLMs excelled in determining standardized therapeutic goals, providing reliable augmentation for pharmacists in this area. However, their performance was still limited when addressing complex questions requiring critical thinking. Further advancements in various aspects, such as performance and application interface, are required before LLMs can be fully utilized as decision support tools in pharmaceutical care.

Disclosure of Interests. The authors declare they have no conflicts of interest.

References

[1] Kaboli, P. J., Hoth, A. B., McClimon, B. J., & Schnipper, J. L.: Clinical pharmacists and inpatient medical care. Archives of Internal Medicine, 166(9), 955–964. https://doi.org/10.1001/archinte.166.9.955 (2006).

[2] Hambisa, S., Abie, A., Nureye, D., Yimam, M.: Attitudes, Opportunities, and Challenges for Clinical Pharmacy Services in Mizan-Tepi University Teaching Hospital, Southwest Ethiopia: Health Care Providers' Perspective. Adv Pharmacol Pharm Sci, 5415290. doi: 10.1155/2020/5415290 (2020).

[3] Verawaty, V., Ramdani, M. I., Laksmitawati, D. R., & Meidiawati, C.: Analysis of pharmaceutical staffing needs in the pharmacy installation of Grha Permata Ibu Hospital Depok 2016. Jurnal Manajemen dan Pelayanan Farmasi (Journal of Management and Pharmacy Practice), 7(2), 65–74. https://doi.org/10.22146/jmpf.30124 (2017).

[4] Kong, K. K., Ong, S. C., Ooi, G. S., & Hassali, M. A.: Measuring the proportion of time spent on work activities of clinical pharmacists using work sampling technique at a public hospital in Malaysia. Pharmacy Practice, 19(3), 1–19. https://doi.org/10.18549/pharmpract.2021.3.2469 (2021).

[5] Melnyk, B. M., Hsieh, A. P., Tan, A., McAuley, J. W., Matheus, M., Larson, B., & Dopp, A. L.: The state of health, burnout, healthy behaviors, workplace wellness support, and concerns of medication errors in pharmacists during the COVID-19 pandemic. Journal of Occupational and Environmental Medicine, 65(8), 699–705. https://doi.org/10.1097/jom.0000000000002889 (2023).

[6] Prasad-Reddy, L., Kaakeh, R., & McCarthy, B. C.: Burnout among health system pharmacists: presentation, consequences, and recommendations. Hospital Pharmacy, 56(4), 195–205. https://doi.org/10.1177/0018578720910397 (2020).

[7] Mertens, J. F., Koster, E. S., Deneer, V. H. M., Bouvy, M. L., & van Gelder, T.: Factors influencing pharmacists' clinical decision making in pharmacy practice. Research in Social and Administrative Pharmacy, 19(9), 1267–1277. https://doi.org/10.1016/j.sapharm.2023.05.009 (2023).

[8] Shoshi, A., Müller, U., Shoshi, A., Ogultarhan, V., & Hofestädt, R.: KALIS - An eHealth System for Biomedical Risk Analysis of Drugs. Studies in Health Technology and Informatics, 236, 128–135. https://doi.org/10.3233/978-1-61499-759-7-128 (2017).

[9] Sönnichsen, A., Trampisch, U. S., Rieckert, A., Piccoliori, G., Vögele, A., Flamm, M., … Kunnamo, I.: Polypharmacy in chronic diseases–Reduction of Inappropriate Medication and Adverse drug events in older populations by electronic Decision Support (PRIMA-eDS): study protocol for a randomized controlled trial. Trials, 17(1). https://doi.org/10.1186/s13063-016-1177-8 (2016).

[10] Min, B., Ross, H., Elior Sulem, Veyseh, B., Thien Huu Nguyen, Sainz, O., … Roth, D.: Recent advances in natural language processing via large pre-trained language models: A survey. ACM Computing Surveys, 56(2), 1–40. https://doi.org/10.1145/3605943 (2023).

[11] Enis, M., & Hopkins, M.: From LLM to NMT: Advancing low-resource machine translation with claude. https://doi.org/10.48550/arXiv.2404.13813 (2023).

[12] Dong, X. L., Moon, S., Xu, E. Y., Malik, K., & Yu, Z.: Towards next-generation intelligent assistants leveraging LLM techniques. In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23). Association for Computing Machinery, New York, USA, 5792–5793. https://doi.org/10.1145/3580305.3599572 (2023).

[13] Baihaqi, M. Y., Contreras, A. G., Kawano, S., & Yoshino, K.: Rapport-driven virtual agent: Rapport building dialogue strategy for improving user experience at first meeting. In Proceedings of the INTERSPEECH. International Speech Communication Association, Kos, Greece, 4059-4063 (2024).

[14] Lim, S. M., & Schmälzle, R.: Artificial intelligence for health message generation: an empirical study using a large language model (LLM) and prompt engineering. Frontiers in Communication, 8, 1–15. https://doi.org/10.3389/fcomm.2023.1129082 (2023).

[15] Yang, D., Chen, F., & Fang, H.: Behavior alignment: A new perspective of evaluating LLM-based conversational recommendation systems. ArXiv (Cornell University), 2286–2290. https://doi.org/10.1145/3626772.3657924 (2024).

[16] PharmDia for PharmD.: Retrieved July 19, 2024, from PharmDia website: https://pharmdia.com/ (n.d.).

[17] Baihaqi, M. Y., & Xu, S. S.-D.: Impact of showing robot demonstration on introducing social robotics field to university students. International Journal of Humanoid Robotics, 21(2), 1-23. https://doi.org/10.1142/s0219843623500184 (2023).