Rx-LLM: a benchmarking suite to evaluate safe LLM performance for medication-related tasks

Xingmeng Zhao, PhD
Xingmeng.zhao@cuanschutz.edu
Department of Biomedical Informatics, University of Colorado Anschutz, Aurora, CO 80045, USA

Kaitlin Blotske, PharmD
Kaitlin.blotske@cuanschutz.edu
Department of Biomedical Informatics, University of Colorado Anschutz, Aurora, CO 80045, USA

Moriah Cargile
Moriah.cargile@cuanschutz.edu
University of Colorado Anschutz Medical Campus, Skaggs School of Pharmacy and Pharmaceutical Sciences. Department of Clinical Pharmacy, Aurora, CO, USA

Adeleine Tilley
adeleine.tilley@cuanschutz.edu
University of Colorado Anschutz Medical Campus, Skaggs School of Pharmacy and Pharmaceutical Sciences. Department of Clinical Pharmacy, Aurora, CO, USA

Brian Murray, PharmD
University of Colorado Anschutz Medical Campus, Skaggs School of Pharmacy and Pharmaceutical Sciences. Department of Clinical Pharmacy, Aurora, CO, USA
Brian.2.murray@cuanschutz.edu

Yanjun Gao, PhD
Yanjun.gao@cuanschutz.edu
Department of Biomedical Informatics, University of Colorado Anschutz, Aurora, CO 80045, USA

Kelli Henry, PharmD, MPA
Kelli.henry@cuanschutz.edu
Department of Biomedical Informatics, University of Colorado Anschutz, Aurora, CO 80045, USA

Susan E. Smith, PharmD, FCCM, FCCP
Susan.smith@uga.edu
University of Georgia College of Pharmacy, Department of Clinical and Administrative Pharmacy, Athens, GA, USA

Erin F. Barreto, PharmD, PhD, FCCM, FASN
Barreto.Erin@mayo.edu
Mayo Clinic, Rochester, MN, USA

Seth Bauer, PharmD, FCCM
Seth.bauer@ccf.org
Cleveland Clinic, Cleveland, OH, USA

Sunghwan Sohn, PhD
Sohn.sunghwan@mayo.edu
Mayo Clinic, Rochester, MN, USA

Tianming Liu, PhD
tianming.liu@gmail.com
Department of Computer Science, University of Georgia, Athens, GA

Andrea Sikora, PharmD, MSCR, FCCM, FCCP
Department of Biomedical Informatics, University of Colorado Anschutz, Aurora, CO 80045, USA
Andrea.sikora@cuanschutz.edu
University of Georgia College of Pharmacy, Department of Clinical and Administrative Pharmacy, Augusta, GA, USA

Conflicts of Interest: The authors have no conflicts of interest.

Funding: Funding through Agency of Healthcare Research and Quality for Dr. Sikora was provided through R21HS028485 and R01HS029009.

Abstract

Background: For large language models (LLMs) to make medication use safer, clinically relevant benchmarks capable of automated grading and measuring the performance of LLMs for medication tasks are required. The purpose of this study was to design a suite of benchmarking tests reflective of comprehensive medication management or CMM (the standard of care for medication optimization) and quantify the baseline performance of the latest LLMs.

Methods: We established six benchmarks representing critical stages of CMM: drug formulation matching, drug order (sig) generation, drug route matching, drug-drug interaction (DDI) identification, renal dose identification, and drug-indication matching. For each benchmark, we curated a clinician-annotated dataset comprising 250 standardized input-output pairs. We evaluated the clinical knowledge retrieval capabilities of three large language models, GPT-4o-mini, MedGemma-27B, and LLaMA3-70B, using a zero-shot prompting strategy. To check reliability, each model was run three times using a temperature of 0.7. Performance was assessed using precision, recall, F1-score, accuracy, and response consistency across trials.

Results: Across six benchmarks, **LLaMA3-70B** demonstrated the highest performance in four tasks: drug-formulation matching (F1, 54.03%), drug-order generation (accuracy, 88%), drug-route identification (F1, 74.31%), and drug-indication identification (accuracy, 97.6%). In the DDI task, **GPT-4o-mini** achieved the highest accuracy (70.4%). For renal dose–adjustment identification, GPT-4o-mini demonstrated the highest F1 score (83.29%). Correctness-consistency scores ranged from 8% to 97.6% across benchmarks, with no model exhibiting uniformly superior consistency.

Conclusions: Model performance varied substantially across medication-related tasks. These findings underscore the need for task-specific evaluation when deploying models for medication-focused clinical decision support.

Keywords: large language model; artificial intelligence; healthcare; medications; pharmacy

Background

Large language models (LLMs) have demonstrated potential for use as information technology (IT) tools in the space of comprehensive medication management (CMM). However, the use of LLMs for medication-related tasks requires standardized benchmarks that have the capability of being automated to avoid resource intensive clinician annotation to ensure safety and efficacy.[1-4]

Medications represent both a unique knowledge domain for LLMs and a construct far removed from diagnosis-related tasks. In disease diagnosis, the ground truth is readily available in large, open-source datasets (e.g., radiology images corresponding to a patient's cancer diagnosis). Treatment is a far less structured problem wherein potential treatments have a range of potential characteristics (e.g., highly safe and efficacious, reasonably efficacious, safe but inefficacious, efficacious but harmful, harmful and inefficacious). There are multiple "right" answers for most treatment decisions, a range of acceptable answers, and finally "wrong" answers that range from benign to deadly. The complexities of medication treatment are managed by interprofessional and interdisciplinary teams including clinical pharmacists who are charged with performing comprehensive medication management (CMM). CMM is defined as "the standard of care that ensures each patient's medications are individually assessed to determine that each medication is appropriate for the patient, effective for the medical condition, safe given the comorbidities and other medications being taken, and able to be taken by the patient as intended."[5]

Similar to drug therapy, which undergoes Food and Drug Administration (FDA) approval for safety and efficacy outcomes, LLMs require careful evaluation for their ability to provide information that is safe from harm and has the potential for benefit. However, an FDA Perspective on artificial intelligence (AI) regulation states "The sheer volume of these changes and their impact also suggests the need for industry and other external stakeholders to ramp up assessment and quality management of AI across the larger ecosystem beyond the remit of the FDA" and underscored that "all involved sectors will need to attend to AI with the care and rigor this potentially transformative technology merits.[6,7] This is particularly important given that LLMs have shown vulnerability to poisoning and hallucinations, which could result in life-threatening medication regimen recommendations.[8,9]

To date, no organization or institution has taken on the necessary task of establishing benchmarks for the safe performance of LLMs for clinical acceptance in the domain of medication tasks. The purpose of this study was to design a suite of benchmarks reflective of the CMM process to allow for the performance evaluation of LLM-based information technology (IT) tools.

Methods

Study design. A panel of 5 licensed, board-certified critical care pharmacists was convened to review the best practices for CMM in the ICU based on published consensus recommendations and guidelines.[10-12] The panel developed a visual diagram of the reasoning process for clinical decision-making within the CMM process (see **Figure 1**). Tasks were selected to capture a range of clinical reasoning skills, medication knowledge, and structured decision outputs relevant to real-world prescribing. Each task was formulated to require direct model generation or classification without access to external tools.

We conducted a structured evaluation of three large language models (LLMs), LLaMA3-70B, GPT-4o-mini, and MedGemma-27B, across six medication-related benchmarks designed to reflect common and safety-critical tasks in clinical medication management. The panel identified six benchmarks at key junctures in the CMM process representing various medication-related tasks from the major categories of CMM (safety, adherence, operations, indication, efficacy). The primary input concept and output concept are mapped in **Table 1**.

(1) **Drug formulation matching**: Ability to correctly match the generic drug name with the available dosage forms
(2) **Drug order (sig) generation:** Ability to generate a safe and appropriate complete medication order given a drug
(3) **Drug route matching:** Ability to correctly identify appropriate routes given a drug
(4) **Drug-drug interaction identification**: Ability to correctly identify drug-drug interactions among simple lists
(5) **Renal dose identification**: Ability to correctly identify if any of the available routes of the drug require a renal dose adjustment or are contraindicated in renal failure
(6) **Drug-indication matching**: Ability to correctly identify which medication (among a list of 3 medications) is being used for a specific indication

While not comprehensive for all medication tasks, these tasks were considered absolutely essential for safe performance of an LLM. Notably, the six tasks are already safely performed or checked by non-artificial intelligence information technology (IT) tools (which do not run the risk of hallucinations and have transparent data sources and coding) such that AI would ultimately need to surpass performance in these areas to supercede existing safe standards of care.

This project was reviewed and approved by the University of Colorado Institutional Review Board (COMIRB #25-1631). All methods were performed in accordance with the ethical standards of the Helsinki Declaration of 1975.

Benchmark Descriptions

The benchmarks were designed to capture core elements of medication knowledge and prescribing practice. For drug formulation and route identification, models were asked to list all clinically valid formulations and routes of administration for each medication; both tasks required multilabel predictions because many medications have multiple acceptable forms and routes. The drug-order generation task evaluated a model's ability to produce a complete prescription sig, including dose, route, and frequency, based on a medication prompt and clinical context, with outputs compared against an expert-validated reference.

The remaining tasks focused on clinical appropriateness and safety. For drug-indication identification, models judged whether a medication was indicated for a specific condition. Drug–drug interaction classification required categorizing medication pairs into established severity levels, from no interaction to contraindicated. Renal dose–adjustment identification assessed whether a medication required modification in the presence of renal impairment and whether the criteria for adjustment were met; performance for this task was evaluated using F1 scores.

Data source. A total of six datasets were curated to test the proposed benchmarks with 250 input-output pers per benchmark. These datasets are posted on Github: https://github.com/sikora07/AIChemist. Source material was taken from LexiDrug for each medication, with review by 2 board-certified pharmacists to ensure accuracy and clinical relevance.[13]

Model Evaluation Framework. The primary outcome was the ability of each LLM to correctly identify the standardized answer established from ground truth clinical annotations. This primary outcome of accuracy was measured by task-specific evaluation metrics: for multi-label route classification, we computed precision and recall from set-based true positives (correctly predicted routes), false positives (hallucinated routes), and false negatives (missing routes); for binary renal dose classification, we used standard confusion matrix metrics (true positives, false positives, true negatives, false negatives) to calculate precision, recall, F1-score, and accuracy; and for drug-pair identification tasks, we measured exact match accuracy for drug-drug interaction pairs. Additionally, we evaluated correctness consistency, which measuring the fraction of runs where predictions were correct, averaged across all instances. We also reported standard deviation across runs to quantify variability in model performance.

We evaluated three large language models representing different architectural approaches and training strategies: GPT-4o-mini (OpenAI), a general-purpose model optimized for efficiency; MedGemma-27B (Google), a domain-specific model trained on biomedical literature; and LLaMA3-70B (Meta), a large-scale general-purpose model. Model selection was designed to assess whether domain-specific training or model scale provides advantages for medication management tasks.

All models were evaluated using zero-shot prompting without in-context examples to assess inherent clinical knowledge embedded during pretraining, rather than few-shot learning capabilities. This approach isolates model knowledge from prompt engineering effects and reflects real-world deployment scenarios where examples may be unavailable. Task-specific prompts were developed through iterative refinement with clinical pharmacists to ensure clarity, minimize ambiguity, and maintain clinical authenticity while avoiding leading language that could bias responses.

To quantify model reliability and prediction consistency—critical metrics for clinical deployment—each model was evaluated three times per instance using different random seeds. This repeated evaluation approach enables assessment of both accuracy (correctness of predictions) and consistency (reproducibility across runs), which directly impact patient safety in clinical settings.[14,15] Prediction variability was quantified using standard deviation across runs, providing insight into model stability. This methodology aligns with uncertainty quantification frameworks for large language models and reflects best practices for evaluating AI trustworthiness in healthcare applications.[16]

Evaluation Metrics

We evaluated model performance using commonly used classification metrics: precision, recall, F1 score, and accuracy. These metrics were applied to the drug-formulation, drug-route, and renal dose–adjustment tasks. For all benchmarks, model predictions were compared with a reference standard created using LexiDrug, including categories for drug-drug interaction severity.[13] Two board-certified clinical pharmacists reviewed and confirmed all ground-truth labels. Additionally, the Harm Associated with Medication Error Classification (HAMEC) score was used to classify drug errors noted in the order generation benchmark and was validated by two clinical pharmacists.[17]

Classification Metrics. For tasks with single-label outputs, evaluation was based on the four parts of the confusion matrix: true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN). Precision measured how often the model's positive predictions were correct (TP / [TP + FP]). Recall measured how often the model correctly identified the true positive cases (TP / [TP + FN]). The F1 score combined precision and recall into a single number that balances both. Accuracy measured how often the model was correct overall and was calculated as:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

Multilabel Metrics. The drug-route and drug-formulation benchmarks required multilabel predictions, where a single medication could have more than one correct answer. For these tasks, we used two additional measures. Exact-match accuracy counted how often the model predicted every correct label for an item with no mistakes. Micro-averaged label accuracy looked across all labels and measured the proportion of labels the model predicted correctly out of all labels that appeared in either the predictions or the reference standard.

For all metrics, we report the mean and 95% confidence interval derived from repeated evaluations. 95% confidence interval reflects variability in model outputs across the test set and captures the stability of performance. To characterize reliability, we computed a correctness-consistency score for each model, defined as the proportion of instances for which the model provided the same correct answer across repeated prompts or structurally similar variants of the same clinical input. This quantifies robustness rather than single-run accuracy alone. All statistical analysis was conducted in Python.[18]

Results

Across the six medication-related benchmarks, model performance differed substantially by task. As shown in **Table 2**, LLaMA3-70B achieved the highest performance in four domains. In drug-formulation matching, LLaMA3-70B produced the strongest F1 score (54.03% [95% CI: 50.13-57.98]), exceeding GPT-4o-mini (53.08% [95 CI: 49-57.01]) and MedGemma-27B (40.88% [95% CI: 36.45-45.17]). In drug-order (sig) generation, LLaMA3-70B achieved the highest overall accuracy (88%) with fewer incorrect-dose, incorrect-frequency, and incomplete-instruction errors than the other models. LLaMA3-70B also outperformed the other models in drug-route identification (F1 74.31% [95% CI: 71.01-77.99]) and drug-indication identification (97.6% [95% CI: 95.6-99.2]), as reported in **Table 2**.

Performance patterns diverged in safety-critical tasks. As shown in **Table 3**, GPT-4o-mini achieved the highest overall accuracy for drug–drug interaction classification (70.4% [95% CI: 64.8-75.73]) and the highest correctness-consistency score (67.6%). Although MedGemma-27B did not lead overall, it showed the strongest performance in Category D interactions (76.45% [95% CI: 67.05-84.68]), a clinically significant subclass that typically requires therapeutic modification.[13] For renal dose-adjustment identification, GPT-4o-mini again demonstrated the highest performance (F1 83.29% [95% CI: 77.58-88]), whereas LLaMA3-70B and MedGemma-27B showed lower precision in identifying medications requiring adjustment.

Correctness-consistency scores varied widely across tasks, such as from 8% to 97.6%, reflecting differences in the stability of correct outputs across repeated prompt variants. These trends, detailed in **Table 2** and **Table 3**, show that consistency was highest in binary tasks such as drug-indication identification and lowest in multilabel tasks such as formulation and route identification. No model demonstrated uniformly superior consistency across all benchmarks.

A comparison of performance profiles across all six benchmarks is shown in **Figure 2**. LLaMA3-70B showed strong performance in structured medication-knowledge tasks requiring multilabel or compositional reasoning. GPT-4o-mini showed its greatest strength in safety-critical tasks involving drug–drug interactions and renal dose adjustments. MedGemma-27B performed competitively within selected interaction categories. These findings illustrate heterogeneous capability distributions among current LLMs and highlight the need for multi-domain evaluation when assessing medication-related clinical readiness.

Discussion

We present six benchmarking tasks, including clinician-annotated datasets and standardized prompts, available for use to test the safety and efficacy performance of LLMs for medication tasks. This evaluation marks the first time that a suite of benchmarking tasks has been specifically designed and annotated by clinicians for this use. This evaluation is an important development for the development of safe, transparent, and high-performing LLMs because each benchmark represents a discrete, real-world patient care task that must be performed with nearing perfect accuracy to avoid medication-related harm. While these tasks are not fully representative of the myriad of tasks related to the safe use of medications, this suite represents an important early indicator that an LLM has the appropriate knowledge, pre-training, and prompt engineering to be used for medication management.

Recent viewpoints have drawn parallels between education and AI benchmarking. Healthcare professions, including both medical and pharmacy schools, use a construct of competency-based medical education (CBME), in which entrustable professional activities or "EPAs" serve as observable and measurable units of practice that are teachable through modeling and feedback. These viewpoints suggest AI-CBME approach that defines competencies and milestones in terms of EPAs.[19,20] The six benchmarks are reflective of distinct tasks necessary for a broader EPA around the safe and responsible use of medication therapy.

The Three Principles is a previously proposed paradigm for safe medication use that include: (1) treatment of the underlying cause (e.g., an antibiotic for an infection), (2) provision of supportive care (e.g., an analgesic for pain secondary to the infection), and (3) avoidance of iatrogenic harm (e.g., avoidance of antibiotic that the patient has a documented drug allergy for).[21,22] While a seemingly simple paradigm that places the patient at the center of the process, medication harm remains a leading cause of morbidity and mortality.[23,24] The high rates of medication errors, ADEs, polypharmacy, and other inappropriate prescribing are a motivation for the development of data-driven tools to make medication use safer; however, indiscriminate use of the general Internet (which contains high amounts of medication misinformation) and real-world electronic health record data (which certainly contains inappropriate prescriptions) has substantial risk for developing LLMs that provide unsafe medication recommendations. Indeed, rigorous clinician validation is absolutely essential for training and benchmarking datasets in this space. A recent example was DrugGPT,[25] which included a DDI dataset without clinical validation that listed DDI pairs with directly conflicting / inaccurate information as well as outdated medications, including about anticoagulants, which are one of the most dangerous and often misused medications on the market.[26] Recent studies have shown the fragility of LLMs in the face of poisoning (also known as adversarial attacks)[9,27] and moreover even simple tasks like discriminating brand and generic drug names proved challenging.[28,29]

One strength of this analysis was robust review of LLM outputs by clinicians and measurement of harm potential. The drug formulation benchmark yielded many inconsistencies, including a variety of false positives and false negatives. Less than 20% of drugs had all correct formulations provided in each run, although this information is readily available on several drug information references, including LexiDrug and has capabilities to be programmed into EHRs.[13] For the drug order generation benchmark, the prompt specifically limited order generation to oral routes in an attempt to decrease possibilities and complexity that may be seen with other routes. However, even with one of the simplest routes for order generation, error rates exceeded 10% for each model, with a range of harm scores. Of note, MedGemma-27B recommended in all runs to give dofetilide 12.5mg by mouth twice daily. The correct dose is 125mcg twice daily, so the recommended dose was a 100-fold overdose, which would have high potential result in a life-threatening arrythmia (rated as a 4 on the HAMEC score).[13,17] Additional errors that were classified as a 3 on the HAMEC score included all models recommending an underdose of flucytosine, a medication used to treat serious fungal infections, and an overdose of nitroglycerin from each model.[13,17] The drug route benchmark revealed significant inconsistencies, including high false positives. This was present for most drugs, including drugs with relatively few routes (e.g., amitriptyline is only available as oral but was listed as IV continuous, IV intermittent, intramuscular, oral, and topical by GPT-4o-mini, intramuscular, IV continuous, and oral by MedGemma-27B, and intramuscular and oral by LLaMA3-70B).[13] The DDI benchmark yielded moderate performance from all models; however, several severe DDIs were missed, including all models not recognizing the interaction between Andexanet alfa and heparin (a category X interaction) and carbapenems and valproic acid (a category D interaction that could result in seizures). Additionally, each model often reported only a partial drug name (such as “globulin” instead of “immune globulin” and “chloride” instead of “potassium chloride”), which could increase confusion.[13] Interestingly, each model had 2-4 spelling errors where the medications within the DDIs were correct, but the LLM misspelled the medication, for example “mexelitine” instead of “mexiletine” and “cefpime” instead of “cefepime.” The renal dose adjustment benchmark yielded moderate performance as well, with interesting false positives including propranolol and rasburicase and false negatives including chlordiazepoxide and methylnaltrexone, which have clear dose adjustments in the package insert.[13] The drug indication identification benchmark yielded the highest performance from all models; however, errors here would likely have resulted in patient harm as well. For example, GPT-4o-mini and MedGemma-27B both selected daptomycin as the medication associated with treatment of methicillin-resistant *staphylococcus aureus* (MRSA) pneumonia over ceftaroline; however, daptomycin is inactivated by pulmonary surfactant in the lungs and cannot be used for MRSA pneumonia.[30] Additionally, LLaMA3-70B and MedGemma-27B selected valproic acid as the correct medication for the prevention of atrial fibrillation-related complications as opposed to argatroban, an anticoagulant. Giving valproic acid as opposed to an anticoagulant would have increased the risk of stroke.[31] GPT-4o-mini selected alteplase as the treatment for septic shock over methylene blue, potentially causing severe bleeding and undertreatment of shock.[32]

MedGemma-27B selected propranolol for thyrotoxicosis, which is correct, but propranolol was not on the answer list; metoprolol was the correct answer. This example of not following prompt instructions was also present in other benchmarks, notably the order generation benchmark where 1-7% of outputs did not follow prompt instructions. While the LLMs showed promise for completing medication-related benchmarking tasks, substantial performance improvements are needed as it relates to safety.

This study has several limitations: most notably, with just six tasks, this series is not fully reflective of the complexities of the medication management process. Thus, an LLM could conceivably have high performance on these tasks and still show safety concerns in other domains. Moreover, given that prompt engineering and other methodologies can substantially improve performance, there is a possibility that with more advanced methodology, performance would have improved. While these aspects were deemed outside of the scope of this foundational work, these are important future areas of exploration.

Conclusion

Six clinically validated medication task benchmarks were created to assess the five major categories (safety, adherence, operations, indication, efficacy) of comprehensive medication management. The tested LLMs demonstrated inconsistent performance with accuracy or F1 scores ranging widely across benchmarks and LLMs. Future applications of these benchmarks could improve confidence in the safety of LLMs for clinical use.

References

1. Chase A, Most A, Sikora A, et al. Evaluation of large language models' ability to identify clinically relevant drug-drug interactions and generate high-quality clinical pharmacotherapy recommendations. Am J Health Syst Pharm 2025. DOI: 10.1093/ajhp/zxaf168.
2. Xu S, Most A, Chase A, et al. Large language models management of complex medication regimens: a case-based evaluation. medRxiv 2024:2024.07.03.24309889. DOI: 10.1101/2024.07.03.24309889.
3. Yang H, Hu M, Most A, et al. Evaluating accuracy and reproducibility of large language model performance on critical care assessments in pharmacy education. Front Artif Intell 2024;7:1514896. DOI: 10.3389/frai.2024.1514896.
4. Zhengliang Liu ZW, Mengxuan Hu, Bokai Zhao, Lin Zhao, Tianyi Zhang, Haixing Dai, Xianyan Chen, Ye Shen, Sheng Li, Brian Murray, Tianming Liu, Andrea Sikora. PharmacyGPT: The AI Pharmacist. arixV https://arxivorg/abs/230710432 2023.
5. ASHP. Comprehensive medication management. (https://www.ashp.org/advocacy-and-issues/key-issues/other-issues/comprehensive-medication-management?loginreturnUrl=SSOCheckOnly#:~:text=Definition%20of%20CMM%3A%20The%20standard,effective%20for%20the%20medical%20condition).
6. Warraich HJ, Tazbaz T, Califf RM. FDA Perspective on the Regulation of Artificial Intelligence in Health Care and Biomedicine. Jama 2025;333(3):241–247. (In eng). DOI: 10.1001/jama.2024.21451.
7. Gilbert S, Harvey H, Melvin T, Vollebregt E, Wicks P. Large language model AI chatbots require approval as medical devices. Nat Med 2023;29(10):2396–2398. (In eng). DOI: 10.1038/s41591-023-02412-6.
8. Shrey Pandit JX, Junyuan Hong, Zhangyang Wang, Tianlong Chen, Kaidi Xu, Ying Ding. MedHallu: A Comprehensive Benchmark for Detecting Medical Hallucinations in Large Language Models. https://arxivorg/abs/250214302 2025.
9. Alber DA, Yang Z, Alyakin A, et al. Medical large language models are vulnerable to data-poisoning attacks. Nat Med 2025;31(2):618–626. DOI: 10.1038/s41591-024-03445-1.
10. Sikora A, Murray B, Henry K, et al. Consensus recommendations for the integration of critical care pharmacists on intensive care unit teams: Endorsed by the American Association of Critical-Care Nurses, American College of Clinical Pharmacy, American Society of Health-System Pharmacists, Institute for Safe Medication Practices, and Society of Critical Care Medicine. JACCP: JOURNAL OF THE AMERICAN COLLEGE OF CLINICAL PHARMACY 2025;8(9):914–930. DOI: https://doi.org/10.1002/jac5.70084.
11. Lat I, Paciullo C, Daley MJ, et al. Position Paper on Critical Care Pharmacy Services: 2020 Update. Crit Care Med 2020;48(9):e813–e834. DOI: 10.1097/CCM.0000000000004437.
12. Haas CE, Kliethermes MA, Armistead LT, et al. Comprehensive medication management: Review and recommendations for quality measures. JACCP: JOURNAL OF THE AMERICAN COLLEGE OF CLINICAL PHARMACY 2023;6(4):404–415. DOI: https://doi.org/10.1002/jac5.1767.
13. Lexidrug U. 2025.
14. Lin Z, Trivedi S, Sun J. Generating with confidence: Uncertainty quantification for black-box large language models. arXiv preprint arXiv:230519187 2023.
15. Miao N, Teh YW, Rainforth T. Selfcheck: Using llms to zero-shot check their own step-by-step reasoning. arXiv preprint arXiv:230800436 2023.

16. Kruse M, Afshar M, Khatwani S, Mayampurath A, Chen G, Gao Y. Simple Yet Effective: An Information-Theoretic Approach to Multi-LLM Uncertainty Quantification. Suzhou, China: Association for Computational Linguistics; 2025:30481–30492.
17. Gates PJ, Baysari MT, Mumford V, Raban MZ, Westbrook JI. Standardising the Classification of Harm Associated with Medication Errors: The Harm Associated with Medication Error Classification (HAMEC). Drug Saf 2019;42(8):931–939. (In eng). DOI: 10.1007/s40264-019-00823-4.
18. Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods 2020;17(3):261–272. DOI: 10.1038/s41592-019-0686-2.
19. Abdulnour RE, Gin B, Boscardin CK. Educational Strategies for Clinical Supervision of Artificial Intelligence Use. N Engl J Med 2025;393(8):786–797. DOI: 10.1056/NEJMra2503232.
20. Ilgen JS, Dhaliwal G. Educational Strategies to Prepare Trainees for Clinical Uncertainty. N Engl J Med 2025;393(16):1624–1632. DOI: 10.1056/NEJMra2408797.
21. Murray B, Sikora A, Mock JR, et al. Reverse Triggering: An Introduction to Diagnosis, Management, and Pharmacologic Implications. Front Pharmacol 2022;13:879011. (In eng). DOI: 10.3389/fphar.2022.879011.
22. Sikora A. Critical Care Pharmacists: A Focus on Horizons. Crit Care Clin 2023;39(3):503–527. (In eng). DOI: 10.1016/j.ccc.2023.01.006.
23. Hodkinson A, Tyler N, Ashcroft DM, et al. Preventable medication harm across health care settings: a systematic review and meta-analysis. BMC Medicine 2020;18(1):313. DOI: 10.1186/s12916-020-01774-9.
24. Sunshine JE, Meo N, Kassebaum NJ, Collison ML, Mokdad AH, Naghavi M. Association of Adverse Effects of Medical Treatment With Mortality in the United States: A Secondary Analysis of the Global Burden of Diseases, Injuries, and Risk Factors Study. JAMA Network Open 2019;2(1):e187041–e187041. DOI: 10.1001/jamanetworkopen.2018.7041.
25. Zhou H, Liu F, Wu J, et al. A collaborative large language model for drug analysis. Nat Biomed Eng 2025 (In eng). DOI: 10.1038/s41551-025-01471-z.
26. Herrero-Zazo M, Segura-Bedmar I, Martínez P, Declerck T. The DDI corpus: An annotated corpus with pharmacological substances and drug–drug interactions. Journal of Biomedical Informatics 2013;46(5):914–920. DOI: https://doi.org/10.1016/j.jbi.2013.07.011.
27. Han T, Nebelung S, Khader F, et al. Medical large language models are susceptible to targeted misinformation attacks. npj Digital Medicine 2024;7(1):288. DOI: 10.1038/s41746-024-01282-7.
28. Gallifant J, Chen S, Jain SK, et al. Reliability of Large Language Model Knowledge Across Brand and Generic Cancer Drug Names. JCO Clin Cancer Inform 2025;9:e2400257. (In eng). DOI: 10.1200/cci-24-00257.
29. Gallifant J, Chen S, Moreira PJF, et al. Language Models are Surprisingly Fragile to Drug Names in Biomedical Benchmarks. Miami, Florida, USA: Association for Computational Linguistics; 2024:12448–12465.
30. Silverman JA, Mortin LI, Vanpraagh AD, Li T, Alder J. Inhibition of daptomycin by pulmonary surfactant: in vitro modeling and clinical impact. J Infect Dis 2005;191(12):2149–52. (In eng). DOI: 10.1086/430352.
31. Joglar JA, Chung MK, Armbruster AL, et al. 2023 ACC/AHA/ACCP/HRS Guideline for the Diagnosis and Management of Atrial Fibrillation: A Report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. Circulation 2024;149(1):e1–e156. (In eng). DOI: 10.1161/cir.0000000000001193.

32. Reed M, Kerndt CC, Nicolas D. Alteplase. StatPearls. Treasure Island (FL): StatPearls Publishing
Copyright © 2025, StatPearls Publishing LLC.; 2025.

Tables

Table 1. Input and output concepts for the six CMM benchmarks

Benchmark	Description	Input	Output
Drug formulation matching	Tests the ability to correctly match the generic drug name with the available dosage forms	Generic drug name (e.g., amlodipine)	All FDA-approved formulations available (e.g., 2.5mg tablet, 5mg tablet, 10mg tablet, 1mg/mL oral solution, 1mg/mL oral suspension)
Drug order (sig) generation	Tests the ability to generate a safe and appropriate complete medication order given a drug name	Generic drug name (e.g., carvedilol)	One complete order for an oral route (e.g., carvedilol 12.5mg one tablet by mouth twice daily)
Drug route matching	Tests the ability to identify all safe routes (e.g. intravenous, oral, ophthalmic) available for a medication.	Generic drug name (e.g., prednisolone)	All routes available (e.g., oral, ophthalmic)
Drug-drug interaction identification	Tests the ability to correctly identify which drugs have a drug-drug interaction given a scenario of 4-6 concomitant medications	List of 4-6 drugs with complete orders (e.g., Tenecteplase 22.5 mg administered as IV bolus; Labetalol 20 mg IV given over 1 minute; Aspirin 325 mg by mouth once daily; Atorvastatin 40 mg by mouth once daily; Lorazepam 0.5 mg every 4 to 6 hours as needed for anxiety)	Two drugs with drug-drug interaction Category C, D, or X* (e.g., Tenecteplase and aspirin) among the list
Renal dose identification	Tests the ability to identify if the intravenous or oral formulation of the medication requires a renal dose adjustment or is contraindicated in renal failure	Generic drug name (e.g., vancomycin)	Renal dose adjustment required or contraindicated at certain renal function for intravenous or oral formulations (yes/no)
Drug-indication matching	Tests the ability to identify which medication, among a list of three medications, is being used for a specific indication	Generic drug name of 3 medications (e.g., azithromycin, docusate, norepinephrine) and an indication (e.g. community acquired pneumonia)	Medication in the given regimen that is being used for the specific indication (e.g. azithromycin)

CMM: comprehensive medication management, FDA: food and drug administration, IV: intravenous, kg: kilogram, mg: milligram, mL: milliliter
Drug interaction categories are defined as follows:
Category X — Avoid combination:
This drug pair should generally not be used together because the clinical risk outweighs any potential benefit.
Category D — Consider therapy modification:
The interaction is clinically relevant, and modification of therapy (dose adjustment, substitution, or precautions) should be considered.
Category C — Monitor therapy:
An interaction is present. The combination may be used but requires monitoring by a healthcare professional.
Category B — No action needed:
An interaction may exist, but no clinical intervention is required.
Category A — No known interaction:
There is no known documented interaction between the two drugs.

Table 2. Benchmarks 1-3: Drug Formulation Matching, Drug Order (sig) Generation, Drug Route Identification

	GPT-4o-mini (n=250)	**MedGemma-27B (n=250)**	**LLaMA3-70B (n=250)**
Drug Formulation Matching			
Recall	42.95% [38.56, 47.39]	32.98% [27.74, 38.22]	49.52% [44.45, 54.22]
Precision	69.56% [66.13, 72.95]	54.01% [50.65, 57.32]	59.51% [55.75, 63.13]
F1	53.08% [49, 57.01]	40.88% [36.45, 45.17]	54.03% [50.13, 57.98]
Exact Match Accuracy	19.14% [14.73, 23.83]	9.89% [6.42, 13.52]	16.48% [12.14, 21.20]
Label Match Accuracy	36.16% [32.45, 39.87]	25.72% [22.29, 29.18]	37.04% [33.45, 40.83]
Correctness Consistency Score, n (%)	39 (15.6)	20 (8)	41 (16.4)
Drug Order (sig) Generation			
Overall accuracy, n (%)	208 (83.2)	203 (81.2)	220 (88)
Type of Error, n			
Formulation Does Not Exist	16	17	8
Incorrect Dose	9	12	10
Incorrect Frequency	12	16	10
Incomplete Instructions	15	14	6
Harm Associated with Medication Error Classification (HAMEC Score), n (%)			
0	6 (2.4)	3 (1.2)	4 (1.6)
1	19 (7.6)	25 (10)	10 (4)
2	14 (5.6)	17 (6.8)	14 (5.6)
3	2 (0.8)	2 (0.8)	3 (1.2)
4	0 (0)	1 (0.4)	0 (0)
Did Not Follow Prompt Instructions, n (%)	4 (1.6)	17 (6.8)	2 (0.8)
Drug Route Identification			
Recall	77.40% [73.1, 82.04]	75.91% [71.63, 80.5]	76.81% [72.68, 81.28]
Precision	60.21% [56.39, 63.76]	63.35% [59.21, 67.69]	71.98% [67.54, 76.43]
F1	67.73% [64.37, 70.93]	69.06% [65.59, 72.86]	74.31% [71.01, 77.99]
Exact Match Accuracy	39.73% [33.87, 45.2]	49.73% [43.6, 55.87]	56.67% [50.8, 63.07]
Label Match Accuracy	51.21% [47.46, 54.96]	52.75% [48.8, 57.3]	59.13% [55.06, 63.93]
Correctness Consistency Score, n (%)	81 (32.4)	120 (48)	139 (55.6)

Values reported as mean [95% confidence interval] % of performance. Higher numbers indicate better performance.
Type of error reported as n, and some drugs may have multiple error types reported.
Recall: true positive/ (true positive + false negative). Low recall indicates many false negatives.
Precision: true positive/ (true positive + false positive). Low precision indicates many false positives
F1: 2 x (precision x recall)/(precision + recall). Low F1 indicates many false negatives, false positives, or both.
Overall accuracy defined as no errors present and followed prompt instructions for all 3 runs for the medication for each large language model.
Exact Match Accuracy = (# instances with perfect match) / (total instances (n=250))
Label Accuracy = True Positive / (True Positive + False Positive + False Negative) = correct routes / all routes that appeared
Correctness Consistency = (instances with all 3 runs correct) / (total instances)

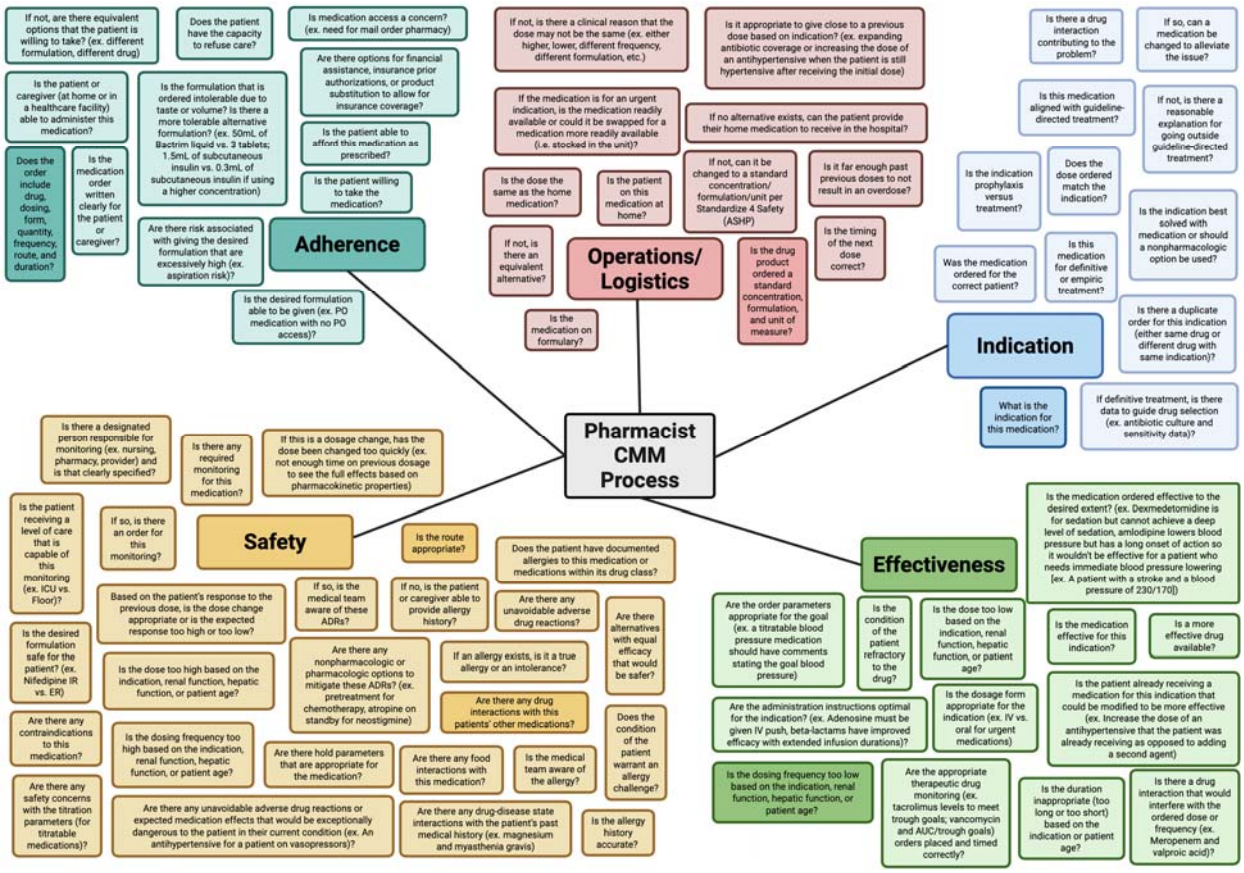
Table 3. Benchmarks 4-6: Drug-drug Interaction Identification, Renal Dose Identification, and Drug Indication Identification

	GPT-4o-mini (n=250)	medgemma27b (n=250)	llama3-70b (n=250)
Drug-drug Interaction Identification			
Accuracy	70.40% [64.8, 75.73]	66.53% [61.2, 71.87]	66.67% [61.07, 72.27]
Accuracy for Category C Interactions (n=81)	62.14% [51.98, 72.43]	56.79% [46.32, 67.05]	57.61% [47.43, 68]
Accuracy for Category D Interactions (n=92)	75.36% [66.29, 83.92]	76.45% [67.05, 84.68]	70.29% [60.67, 78.82]
Accuracy for Category X Interactions (n=77)	73.16% [63.02, 82.22]	64.94% [55.71, 75.2]	71.86% [62.14, 81.93]
Correctness Consistency Score, n (%)	169 (67.6)	155 (62)	160 (64)
Renal Dose Identification			
Recall	86.92% [80.25, 92.73]	91.59% [86.25, 96.26]	90.03% [84.17, 95.5]
Precision	79.96% [72.45, 86.8]	66.22% [58.53, 73.62]	77.27% [69.65, 84.06]
F1	83.29% [77.58, 88]	76.86% [70.78, 82.51]	83.17% [77.47, 87.95]
Correctness Consistency Score, n (%)	206 (82.4)	186 (74.4)	210 (84)
Drug Indication Identification			
Accuracy	96.53% [94.13, 98.53]	96.93% [94.66, 98.93]	97.60% [95.6, 99.2]
Correctness Consistency Score, n (%)	244 (97.6)	244 (97.6)	247 (98.8)

Values reported as mean [95% confidence interval] % of performance. Higher numbers indicate better performance
Accuracy = (True Positive + True Negative) / (True Positive + True Negative + False Positive + False Negative), this shows the fraction of correct predictions
Recall: true positive/ (true positive + false negative). Low recall indicates many false negatives.
Precision: true positive/ (true positive + false positive). Low precision indicates many false positives
F1: 2 x (precision x recall)/(precision + recall). Low F1 indicates many false negatives, false positives, or both.
Correctness Consistency = (instances with all 3 runs correct) / (total instances)
Drug interaction categories are defined as follows:
Category X — Avoid combination:
This drug pair should generally not be used together because the clinical risk outweighs any potential benefit.
Category D — Consider therapy modification:
The interaction is clinically relevant, and modification of therapy (dose adjustment, substitution, or precautions) should be considered.
Category C — Monitor therapy:
An interaction is present. The combination may be used but requires monitoring by a healthcare professional.
Category B — No action needed:
An interaction may exist, but no clinical intervention is required.
Category A — No known interaction:
There is no known documented interaction between the two drugs.

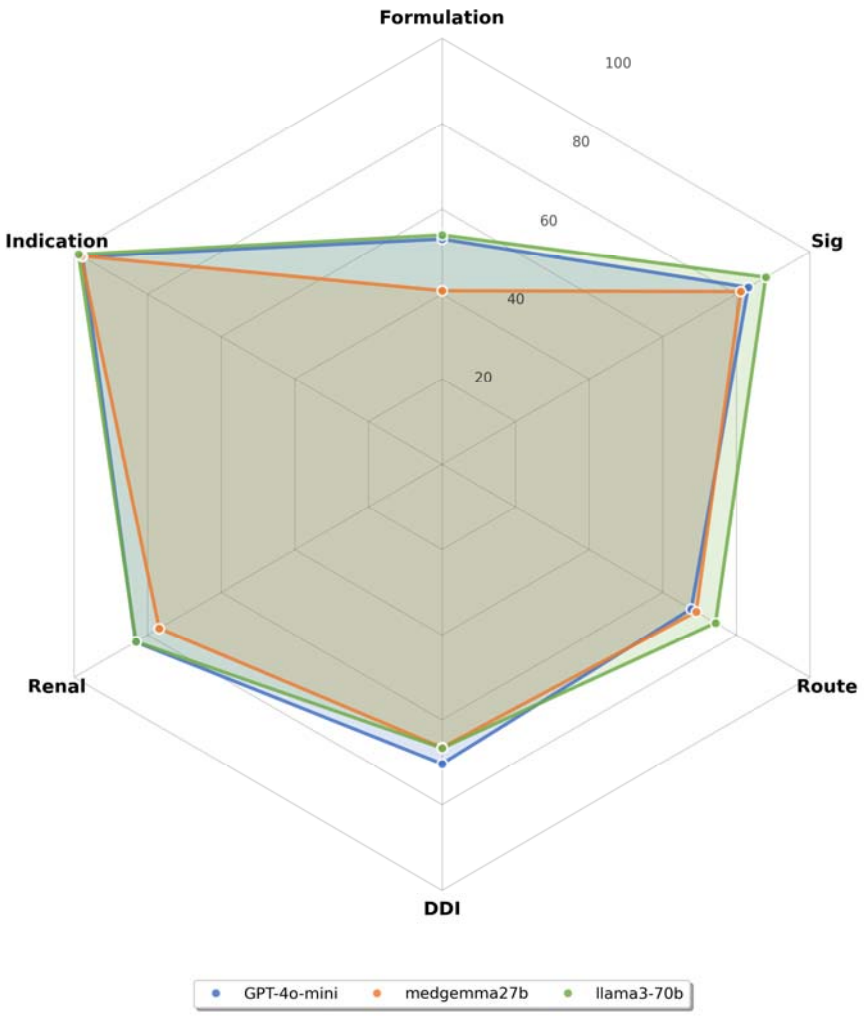
Figures

Figure 1. Comprehensive Medication Management Reasoning Process



Comprehensive medication management (CMM) is a complex reasoning process including domain specific knowledge that integrates knowledge of pharmacology, pharmaceutics, institution specific standards, disease state management, and patient specific information. This figure represents important questions asked and answered throughout that process.

Figure 2. Comparison of 3 LLMs across 6 medication benchmarks



Accuracy is reported for indication, order generation, and drug-drug interaction benchmarks. F1 is reported for renal dose adjustment, route, and formulation benchmarks.
LLM: large language model
DDI: drug-drug interaction