Journal of Biomedical Research

Prompt-guided knowledge-augmented large language model for reliable clinical medication review

Yuhao Sun, Jiayu Wen, Yuxuan Zhu, Jizhong Zhang, Silu Xu, Zhixian Liu, Nan Wu, Xin Liu, Jifu Wei

Cite this article as:
Yuhao Sun, Jiayu Wen, Yuxuan Zhu, Jizhong Zhang, Silu Xu, Zhixian Liu, Nan Wu, Xin Liu, Jifu Wei. Prompt-guided knowledge-augmented large language model for reliable clinical medication review[J]. *Journal of Biomedical Research*, 2026, 40(6): 1-5. doi: 10.7555/JBR.39.20250515

View online: https://doi.org/10.7555/JBR.39.20250515

Articles you may be interested in

Medical image translation using an edge-guided generative adversarial network with global-to-local feature fusion
Journal of Biomedical Research. 2022, 36(6): 409 https://doi.org/10.7555/JBR.36.20220037

Impaired autophagy in amyloid-beta pathology: A traditional review of recent Alzheimer's research
Journal of Biomedical Research. 2023, 37(1): 30 https://doi.org/10.7555/JBR.36.20220145

Clinical management of primary odontogenic sarcoma in the mandible: a case report after WHO nomination
Journal of Biomedical Research. 2022, 36(1): 58 https://doi.org/10.7555/JBR.35.20210133

A short review on cell-based biosensing: challenges and breakthroughs in biomedical analysis
Journal of Biomedical Research. 2021, 35(4): 255 https://doi.org/10.7555/JBR.34.20200128

Unilateral pleural effusion secondary to Takayasu arteritis: a case report and literature review
Journal of Biomedical Research. 2022, 36(2): 141 https://doi.org/10.7555/JBR.36.20210190

Hypotension prediction index together with cerebral oxygenation in guiding intraoperative hemodynamic management: a case report
Journal of Biomedical Research. 2022, 36(1): 63 https://doi.org/10.7555/JBR.36.20210164

Unproofed

Available online at www.jbr-pub.org.cn

Open Access at PubMed Central

Journal of Biomedical Research, 2026 40(6): 1–5

JBR

Letter

Prompt-guided knowledge-augmented large language model for reliable clinical medication review

Dear Editor,

Medication review (MR) is a critical process in clinical pharmacy, enabling the optimization of pharmacotherapy and improving patient safety by reducing adverse drug reactions and medication errors[1]. However, MR is a highly specialized task that requires the integration of multiple patient-specific data sources, such as medication histories, diagnoses, and laboratory results, along with the application of advanced clinical expertise and contextual judgment. Consequently, it is time-consuming and resource-intensive, limiting its routine implementation in resource-constrained healthcare settings[2].

In recent years, large language models (LLMs), as a form of generative artificial intelligence (AI), have demonstrated potential in clinical documentation, diagnostic reasoning, and medical education[3]. However, most LLMs are closed-source, which limits domain-specific customization in healthcare and raises potential concerns regarding data privacy and security. Moreover, their large model size and high computational demands further constrain practical deployment for many healthcare stakeholders. Open-source models offer greater flexibility and customization, yet their integration into hospital information systems (HIS) may face challenges related to data privacy, system compatibility, and security considerations. To optimize LLM performance for clinical pharmacy tasks, prompt engineering represents an alternative approach by guiding model outputs toward specific tasks and enabling high-quality results even with limited training data[4]. Nevertheless, LLMs may still generate inaccurate or incomplete information, a phenomenon known as hallucination. A principal approach to domain specialization is retrieval-augmented generation, which incorporates external medical knowledge to improve factual accuracy and reduce hallucinations[5].

It remains uncertain whether these technologies can match pharmacists' expertise in generating MR reports. To address this question, we developed a Knowledge-Augmented Pharmacy Agent (KAPA), a clinical-support tool that integrates a domain-specific knowledge base with an open-source LLM to assist pharmacists in generating MRs. The knowledge base was constructed using information from currently in-use drug labels, clinical guidelines, and consensus statements. Drug labels were obtained from the latest official versions, while the inclusion of guidelines and consensus statements was determined using a structured quality assessment based on the Appraisal of Guidelines for Research and Evaluation Ⅱ (AGREE Ⅱ) instrument (***Supplementary Note 1***), to ensure that only high-quality references were incorporated. The knowledge base was subsequently transformed into a vector database using an embedding model, enabling the retrieval of domain knowledge, such as drug indications, dosage ranges, and common clinical combination therapies. These retrieved data were then provided as contextual input to the LLM. The LLM used in the present study was DeepSeek-V3, chosen for its open-source availability and seamless compatibility with the composable framework.

This retrospective study was conducted using de-identified data obtained from Jiangsu Cancer Hospital. The study protocol was approved by the Institutional Review Board of Jiangsu Cancer Hospital (Approval No. KY-2025-085), with a waiver of informed consent, and was conducted in accordance with the Declaration of Helsinki. All data were de-identified prior to analysis, and data handling complied with local data protection regulations (***Supplementary Note 2***).

Received: 19 November 2025; Revised: 09 March 2026; Accepted: 12 March 2026; Available online: 30 March 2026

CLC number: R969.3, Document code: A

The authors reported no conflicts of interest.

This is an open access article under the Creative Commons Attribution (CC BY 4.0) license, which permits others to distribute, remix, adapt and build upon this work, for commercial use, provided the original work is properly cited.

© 2026 by Journal of Biomedical Research.

https://doi.org/10.7555/JBR.39.20250515

Unproofed

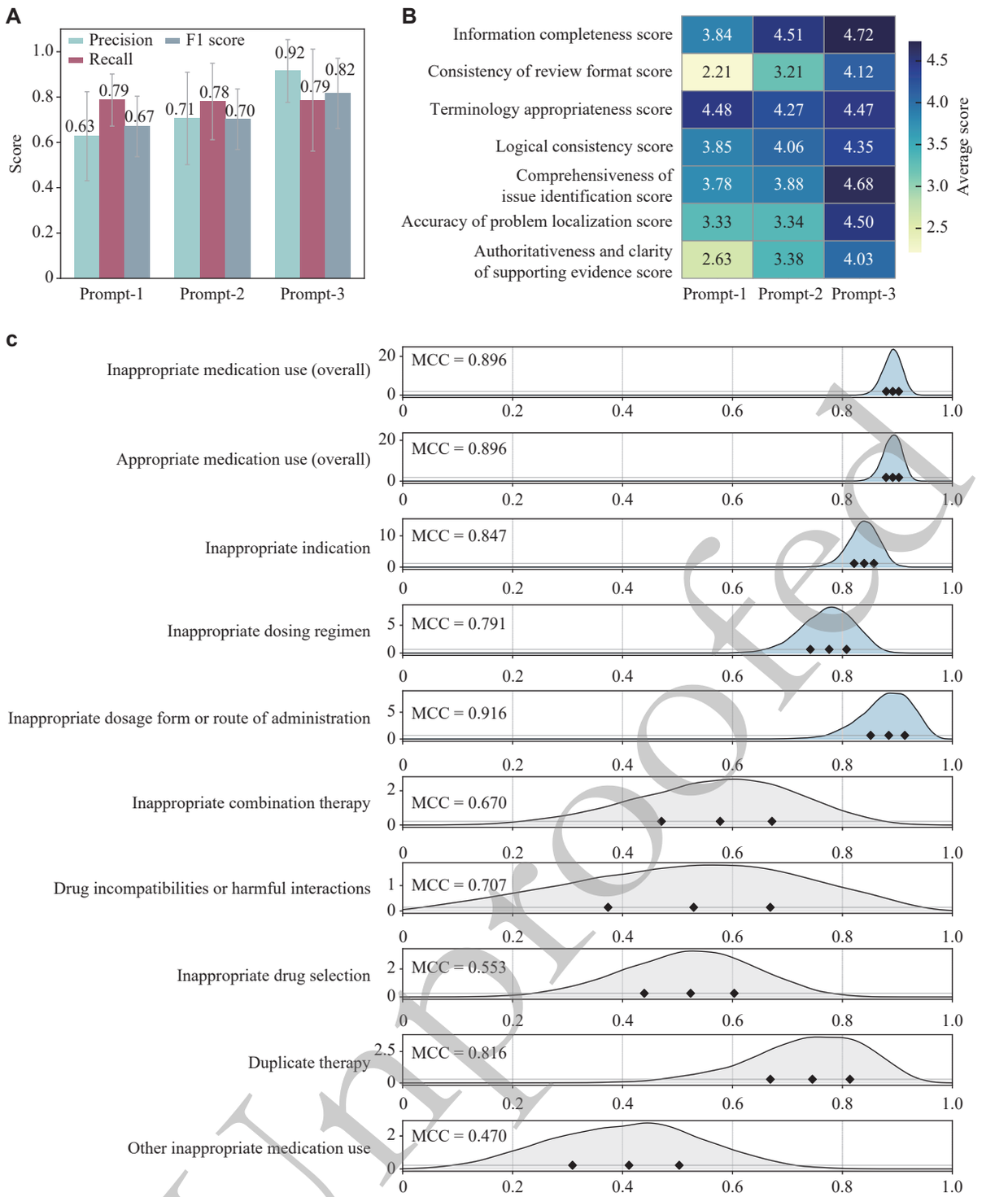
For prompt optimization, we used 154 prescription items derived from 20 patients, along with their corresponding patient records. All data were exported from the HIS by the hospital's Information Technology department. Subsequently, researchers applied Python scripts implementing regular expression-based rules to remove or transform any personally identifiable information. Two senior clinical pharmacists with experience in hospital MR practice independently generated gold-standard MRs using a predefined and standardized evaluation framework that incorporated patient demographic information, confirmed diagnoses, and complete medication orders. Medication appropriateness was assessed according to a predefined eight-category classification system for inappropriate medication use, based on drug labels, relevant clinical guidelines, and individual patient information. In cases of disagreement, discrepancies were adjudicated by a third senior pharmacist. The resulting consensus MR reports served as the reference standard for all subsequent evaluations. Three prompt designs were iteratively evaluated: Prompt-1 applied zero-shot prompting; Prompt-2 used role-based prompting with examples; and Prompt-3 implemented chain-of-thought reasoning guided by the Role-Task-Outcome-Goal-Chain-of-Thought (RTOG-CoT) framework, which structures the model's outputs according to role, task, outcome, and goal while guiding stepwise reasoning (***Supplementary Note 3***).

To assess the similarity between MRs generated by KAPA under different prompt conditions and those generated by humans, we adopted a combination of quantitative and qualitative evaluation methods. Quantitatively, Bidirectional Encoder Representations from Transformers Score (BERTScore) was used to assess semantic consistency between the generated MR reports and the gold-standard reference, yielding semantic precision, semantic recall, and semantic F1 scores that revealed performance differences across prompt designs. While BERTScore captures semantic similarity, it does not guarantee clinical correctness; therefore, we also performed a qualitative assessment. Two clinical pharmacists, who did not participate in generating the gold-standard MRs, independently evaluated the model-generated outputs in Chinese, focusing on clinical accuracy, completeness, and consistency with standard review practices. Inter-rater reliability was confirmed with a Cohen's kappa of 0.773 (***Supplementary Fig. 1***). Responses were rated using a six-point Likert-type scale ranging from 0 to 5 across seven predefined dimensions, hereafter referred to as the seven-dimensional evaluation indicators (***Supplementary Note 4***). Any discrepancies were resolved by a third senior pharmacist.

As shown in ***Fig. 1A*** and ***Supplementary Table 1***, the model using Prompt-1 achieved moderate performance in semantic similarity, with a semantic precision of 0.627 ± 0.196, a recall of 0.787 ± 0.115, and an F1 score of 0.670 ± 0.134. Although the model captured some relevant medication-related problems, its ability to extract complete prescription details and interpret therapeutic intent was limited. The reports showed inconsistent structure and occasional omission or misclassification of clinically important information. Prompt-2 substantially improved semantic precision (0.706 ± 0.204) while maintaining high recall (0.780 ± 0.169), suggesting that example-based role alignment aids the model in organizing information logically and producing more complete and consistent reports. However, challenges remained in evaluating complex multi-drug regimens and distinguishing intentional combination therapy from inappropriate duplication. Prompt-3 demonstrated the highest semantic precision and F1 score, while maintaining recall comparable to that of the other prompts. With RTOG-CoT reasoning and knowledge retrieval, semantic precision reached 0.915 ± 0.138, and the F1 increased to 0.816 ± 0.155. Beyond semantic metrics, Prompt-3 markedly outperformed other prompts in the seven-dimensional clinical evaluation (***Fig. 1B***), achieving a total score of 31.981 ± 3.109 from Rater 1 (***Supplementary Table 2***). The reports generated under this prompt were better structured, used more consistent terminology, and showed improved alignment with pharmacists' reasoning. The model's ability to identify inappropriate medication use, justify its assessments using retrieved authoritative evidence, and articulate clinically relevant recommendations showed marked improvement. These findings underscore that structured prompting combined with contextual knowledge augmentation enables LLMs to generate text that is closer in content to expert human outputs.

To assess the practical utility of LLM-driven MRs, a KAPA-based workflow using Prompt-3 (RTOG-CoT prompt) was developed to assist clinical pharmacists. The validation cohort consisted of 100 oncology inpatients (median age 65.0 years) representing a spectrum of clinical scenarios, including surgical, chemotherapeutic, and palliative interventions. The evaluation encompassed 1 762 individual prescription items, with an average density of 17.6 items per patient. Detailed patient characteristics are provided in ***Supplementary Table 3***. For each prescription, KAPA-generated reviews were audited by pharmacists,

Unproofed



Unproofed

Fig. 1 **Performance evaluation of KAPA-generated medication reviews across semantic, pharmaceutical, and medication problem identification dimensions.** A: Model evaluation metrics of MRs generated by KAPA. Bar plots show the semantic similarity assessed by Bidirectional Encoder Representations from Transformers Score (BERTScore), including precision, recall, and F1 score. B: Performance of the three prompts across the seven dimensions. C: Distribution of the Matthews correlation coefficient (MCC) across medication problem categories. Kernel density plots display the posterior distribution of MCC for each category estimated via Bayesian Beta-binomial sampling with uniform priors, with the 95% highest density interval indicated. Rhomboid markers denote quartiles, and categories with fewer than 10 true positives (TP) are highlighted in light gray, indicating limited sample sizes; results are for reference only. Abbreviation: KAPA, Knowledge-Augmented Pharmacy Agent.

whose assessments served as the gold standard. True positives, false positives, true negatives, and false negatives were computed for overall inappropriate medication use (see ***Supplementary Table 4*** for the calculation algorithm). Metrics were also calculated for appropriate medication use and specific categories of

inappropriate use. As shown in ***Supplementary Tables 5*** and ***6***, KAPA performed well in the overall assessment of prescription inappropriateness, achieving an accuracy of 0.980 for both inappropriate and appropriate use. Corresponding F1 scores were 0.907 and 0.989, and the Matthews correlation coefficient (MCC) was 0.896 for both overall judgments, indicating high reliability in overall judgment. ***Fig. 1C*** shows the detailed MCC distribution across different categories, highlighting variations in model performance among types of medication problems. The model achieved high reliability for well-defined and common inappropriate use categories, including inappropriate indication, duplicate therapy, and inappropriate dosage form or route of administration, with MCC values exceeding 0.80. In contrast, performance was lower for more complex or less frequent categories, such as inappropriate combination therapy, drug incompatibilities, or harmful interactions, reflecting potential limitations in assessing situations involving multiple medications. Additionally, some categories of inappropriate medication use had limited sample sizes due to their uneven occurrence in real-world prescriptions, which may limit the interpretability of the corresponding results. At the real-world implementation level, KAPA was incorporated into an assisted medication review workflow, and 800 cases were included to assess performance and time efficiency. These cases are separate from the 1,762 validation prescriptions, and the manual review comparator used an independent historical dataset from the same department. While maintaining high overall reliability (accuracy of 0.963, MCC of 0.858), the KAPA-assisted workflow was associated with an approximately 42% shorter medication review time than the historical manual-review workflow (***Supplementary Table 7***). Compared with prior studies reporting inferior performance of ChatGPT in clinical pharmacy settings (with mean accuracy scores of 5.23 for ChatGPT and 9.90 for clinical pharmacists, $P = 0.008\,9$)[6], our results demonstrate that a knowledge-augmented framework combined with well-designed chain-of-thought prompts can substantially enhance the ability of LLMs to generate clinically reliable and contextually grounded MR reports. RTOG-CoT engineering played a crucial role in the present study, guiding the model to articulate intermediate reasoning steps, while least-to-most prompting decomposed complex problems into manageable subtasks[7]. Additionally, by embedding authoritative references within the model workflow, KAPA may help reduce hallucinations, improve the consistency of clinical reasoning, and support standardized report formatting. Furthermore, the integration of KAPA into an MR workflow illustrated its potential to reduce pharmacists' documentation burden and enable more efficient coverage of routine MR tasks. Such advancements may reduce repetitive, labor-intensive tasks for pharmacists, allowing the time saved to be redirected toward higher-value responsibilities requiring greater expertise and communication skills, such as medication optimization, patient education, and medication therapy management[8].

As with all non-experimental studies, several limitations of the present study should be acknowledged. First, generating multiple review points increases the likelihood of correct interventions but can also produce clinically irrelevant or overly broad comments and misclassify inappropriate prescribing issues. Second, model performance is highly sensitive to both prompt design and the specific model version[9]; further customization is needed to enhance model robustness and reliability in different contexts. Third, due to limitations in the linguistic background of the research team, this study focused exclusively on Mandarin-language interactions. Fourth, regulatory frameworks for the use of LLMs remain underdeveloped, with limited and inconsistent guidance on model validation, auditability, post-market surveillance, and legal liability. Although LLM-based supporting systems may fall under existing medical device regulations in some jurisdictions, clear regulatory pathways and evidence standards are often lacking. In particular, accountability for potential harms arising from LLM-assisted medication-related recommendations remains ambiguous, especially regarding the respective responsibilities of clinical pharmacists, healthcare institutions, and model developers. Fifth, the oncology-specific cohort and limited information on disease severity may affect the extrapolation of our findings. Future multicenter studies should evaluate the KAPA framework across more diverse and complex disease states to ensure broader clinical utility. Finally, this study did not assess the impact of LLM-assisted MR on patient outcomes, such as medication safety, adverse events, or clinical effectiveness, because the MR task was performed retrospectively on existing prescriptions. Future work should consider prospective studies or clinical trials to evaluate these effects. Local deployment of a KAPA-like framework may serve as a potential solution, providing a safer and more customizable approach that allows integration with internal data sources, implementation of retrieval-augmented generation, and model fine-tuning, which together can

Uncorrected Proof

reduce hallucinations and improve reliability. However, successful integration of LLMs into clinical practice requires addressing significant technical and organizational challenges, including seamless interoperability with existing electronic medical record systems, user-friendly interface design, and the development of explainable AI methods to enhance transparency and clinician trust[10].

In conclusion, our findings suggest that knowledge-augmented, prompt-optimized LLMs can produce structured, clinically meaningful medication review reports with high agreement with pharmacist-generated reference assessments in a retrospective oncology cohort. By enabling rapid, scalable, and evidence-grounded medication review, KAPA provides a promising approach to improving pharmacy services in resource-limited clinical settings. We hope these results contribute to the ongoing discourse on the responsible and effective integration of LLMs into clinical practice.

This research was supported by the Yishan Projectof Jiangsu Cancer Hospital (Grant No. YSZD202406 to WJF) and the Qunfeng Project of Jiangsu Cancer Hospital (Grant No. DFXK202501 to WJF).

Yours sincerely,

Yuhao Sun[1,2,△], Jiayu Wen[1,2,△], Yuxuan Zhu[1,2,△], Jizhong Zhang[1,2,△], Silu Xu[1], Zhixian Liu[1], Nan Wu[1], Xin Liu[1,✉], Jifu Wei[1,3,✉]

[1]Jiangsu Cancer Hospital, the Affiliated Cancer Hospital of Nanjing Medical University, Jiangsu Institute of Cancer Research, Nanjing, Jiangsu 210009, China;
[2]School of Basic Medicine and Clinical Pharmacy, China Pharmaceutical University, Nanjing, Jiangsu 211198, China;
[3]Jiangsu Key Laboratory of Innovative Cancer Diagnosis & Therapeutics, Nanjing, Jiangsu 210009, China.

△These authors contributed equally to this work.
✉Corresponding authors: Xin Liu, E-mails: Xinliu1005@gmail.com, ORCID: 0009-0000-3555-0465; Jifu Wei, E-mail: weijifu@njmu.edu.cn, ORCID: 0000-0001-6412-3251.

Author contributions

Yuhao Sun: Investigation, Data curation, Methodology, Writing – original draft. **Jiayu Wen**: Methodology, Writing – original draft. **Yuxuan Zhu**: Visualization, Formal analysis, Writing – original draft. **Jizhong Zhang**: Data curation, Formal analysis, Writing – original draft. **Silu Xu**: Data curation, Visualization, Writing – original draft. **Zhixian Liu**: Data curation, Visualization, Writing – original draft. **Nan Wu**: Data curation, Writing – original draft. **Xin Liu**: Writing – review and editing, Supervision. **Jifu Wei**: Writing – review and editing, Funding acquisition, Supervision.

Additional information

The online version contains supplementary material available at http://www.jbr-pub.org.cn/article/doi/10.7555/JBR.39.20250515.

References

[1] Wilkes S, Zaal RJ, Abdulla A, et al. A cost-benefit analysis of hospital-wide medication reviews: A period prevalence study[J]. *Int J Clin Pharm*, 2022, 44(1): 138–145.
[2] Graabaek T, Kjeldsen LJ. Medication reviews by clinical pharmacists at hospitals lead to improved patient outcomes: A systematic review[J]. *Basic Clin Pharmacol Toxicol*, 2013, 112(6): 359–373.
[3] Nassiri K, Akhloufi MA. Recent advances in large language models for healthcare[J]. *BioMedInformatics*, 2024, 4(2): 1097–1143.
[4] Wang L, Chen X, Deng X, et al. Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs[J]. *npj Digit Med*, 2024, 7(1): 41.
[5] Ge J, Sun S, Owens J, et al. Development of a liver disease-specific large language model chat interface using retrieval-augmented generation[J]. *Hepatology*, 2024, 80(5): 1158–1168.
[6] Huang X, Estau D, Liu X, et al. Evaluating the performance of ChatGPT in clinical pharmacy: A comparative study of ChatGPT and clinical pharmacists[J]. *Br J Clin Pharmacol*, 2024, 90(1): 232–238.
[7] Yuan J, Bao P, Chen Z, et al. Advanced prompting as a catalyst: Empowering large language models in the management of gastrointestinal cancers[J]. *Innovation Med*, 2023, 1(2): 100019.
[8] Mondal H, Dash I, Mondal S, et al. A systematic mapping review on the capability of large language models in drug-drug interaction analysis[J]. *Expert Rev Clin Pharmacol*, 2025, 18(9): 683–690.
[9] Liu S, Wright AP, Patterson BL, et al. Using AI-generated suggestions from ChatGPT to optimize clinical decision support[J]. *J Am Med Inform Assoc*, 2023, 30(7): 1237–1245.
[10] Jung KH. Large language models in medicine: Clinical applications, technical challenges, and ethical considerations[J]. *Healthc Inform Res*, 2025, 31(2): 114–124.

Unproofed