Enhancing Medication Safety with LLMs

Gabriele De Vito[1,†], Filomena Ferrucci[1] and Athanasios Angelakis[2,3,4]

[1]Università degli Studi di Salerno, Salerno, Italy
[2]Department of Epidemiology and Data Science, Amsterdam University Medical Center, Amsterdam, Netherlands
[3]Digital Health; Methodology, Amsterdam Public Health Research Institute, Amsterdam, Netherlands
[4]Data Science Center, University of Amsterdam, Amsterdam, Netherlands

Abstract

Medication safety poses a significant challenge in healthcare, with adverse reactions harming patients and increasing costs. We present two solutions based on Large Language Models (LLMs): HELIOT, a Clinical Decision Support System for managing adverse drug reactions, and an approach that uses textual-drug information to predict drug-drug interactions (DDIs). HELIOT examines patient-specific clinical narratives to provide contextual alerts, reducing alert fatigue by over 50%. Our DDI system analyzes molecular structures, organisms, and drug target genes, achieving a sensitivity of 0.978 and an accuracy of 0.919 across 13 validation datasets, with smaller LLMs (2-3 billion parameters) outperforming larger ones. Both systems demonstrate the potential of LLMs to enhance medication safety through advanced language processing and pharmaceutical data interpretation. Future work will focus on practical validation and integration into healthcare systems.

Keywords

Large Language Models, Clinical Decision Support Systems, Drug-Drug Interactions, Adverse Drug Reactions

1. Introduction

The annual burden of medication management errors in England's healthcare system is striking, with approximately 237 million recorded incidents, 28% of which pose significant clinical risks, resulting in thousands of deaths and substantial financial costs. Similarly, in the United States, medication-related mistakes and adverse drug reactions create an enormous economic impact, totaling billions of dollars annually, with drug-drug interactions accounting for 18% of the amount spent addressing adverse drug reactions (ADRs). Traditional Clinical Decision Support Systems (CDSSs) have limitations in this context [1]: they use structured databases that miss critical information often provided in unstructured clinical narratives [2, 3, 4], raising many interruptive alerts that lead to alert fatigue (with override rates of 43.7%-97%), and potentially cause critical warnings to be ignored [3, 4, 5, 6, 7, 8, 9, 10, 11].

On the other hand, traditional DDI identification relies on time-consuming experimental methods [12], while computational alternatives require complex feature engineering [13]. Large Language Models (LLMs) offer promising solutions through their ability to process unstructured data [14, 15, 16]. However, while LLMs have yielded encouraging results in various pharmaceutical applications [17, 18], their potential for ADRs identification and direct drug-drug interaction prediction remains largely unexplored. We present two LLM-based approaches: HELIOT, a CDSS for ADR management through clinical narrative analysis [19], and a text-based DDI prediction method using SMILES notation, target organisms, and gene interactions [16]. Our contributions include: (1) methods for processing clinical narratives to manage ADRs; (2) a text-based DDI prediction approach eliminating complex feature engineering; (3) comprehensive evaluations against traditional approaches; and (4) architectural frameworks for clinical integration. These approaches represent a significant advancement toward comprehensive medication safety.

The remainder of this paper is structured as follows. Section 2 presents our LLM-based CDSS for ADR management, detailing its approach, architecture, and evaluation results. Section 3 describes our

Ital-IA 2025: 5th National Conference on Artificial Intelligence, organized by CINI, June 23-24, 2025, Trieste, Italy

*Corresponding author.

gadevito@unisa.it (G. De Vito); fferrucci@unisa.it (F. Ferrucci); a.angelakis@amsterdamumc.nl (A. Angelakis)
0000-0002-1153-1566 (G. De Vito); 0000-0002-0975-8972 (F. Ferrucci); 0000-0003-1226-9560 (A. Angelakis)

text-based approach for DDI prediction, including its methodology and performance across multiple datasets. Section 4 outlines future directions and challenges for LLM applications in medication safety. Finally, section 5 concludes the paper.

2. LLM-Based CDSS

CDSSs for medication safety face the challenge of balancing accurate alerting with alert fatigue. They struggle to interpret unstructured clinical narratives, including patient-specific medication tolerances and reaction histories. HELIOT is a novel LLM-based CDSS that leverages natural language processing to analyze clinical notes and provide contextually appropriate medication safety recommendations. Its innovation lies in interpreting unstructured text about patient medication experiences and differentiating between reaction types. This section outlines HELIOT's architecture, prototype development, and performance evaluation compared to traditional CDSSs.

2.1. System Architecture and Approach

HELIOT uses Retrieval Augmentation Generation (RAG) to help physicians make medication decisions based on patients' adverse reaction histories. RAG enhances LLMs by retrieving relevant information before generating responses, improving accuracy and contextual relevance. The proposed CDSS employs parallel retrieval processes, simultaneously gathering pharmaceutical data (ingredients, contraindications, side effects) and analyzing patient clinical notes to identify problematic ingredients. During this process, the system maintains and updates longitudinal patient records of adverse reactions and tolerances across encounters, ensuring continuity of care even in facilities without integrated EHR systems. The decision support logic utilizes the persona pattern [20] to guide the LLM in embodying an expert physician who evaluates potential adverse reactions. The output provides a structured assessment with clinical classification, reaction severity categorization, and detailed analysis explaining the rationale. HELIOT also addresses regional linguistic variations by standardizing all medical terminology into English, overcoming LLMs' inherent English bias [21] and ensuring optimal correspondence with international medical ontologies. Its modular architecture allows HELIOT to function as a standalone solution and as a service integrated with existing EHR systems, making it suitable for environments ranging from advanced hospital systems to primary care settings with limited digital infrastructure. The system comprises a web application, an API application with real-time response streaming, a central controller serving as the core decision-making engine, and specialized databases. The pharmaceutical database accommodates data from various sources, with minimal formatting requirements, and allows most of the information to be stored as unstructured free text. In contrast, the clinical database efficiently processes unstructured clinical notes without imposing strict formatting rules, ensuring flexibility across different healthcare contexts.

2.2. Prototype and Knowledge Base Development

To assess the potential of the proposed CDSS, we developed a full-functional prototype. To this aim, we first created a comprehensive pharmaceutical knowledge base. Starting with the Italian Medicines Agency (AIFA) [1] database containing 106,962 approved medications, we collected essential information including drug codes, names, forms, and ATC (Anatomical Therapeutic Chemical) classifications. We then acquired 19,188 leaflet files through Farmadati [2] web services, preprocessing them to obtain detailed medication compositions and properties. To ensure clinical relevance, we created a representative subset following literature-based distribution patterns [11, 10, 9, 5], which reflects real-world prescription and adverse reaction patterns, with narcotic analgesics (65%), antibiotics (15%), NSAIDs (5%), diuretics (2%), antiplatelet agents (2%), and other medications (11%), including common excipients associated with

[1] AIFA. https://www.aifa.gov.it/en/trova-farmaco

[2] Farmadati. https://www.farmadati.it/default.aspx

hypersensitivity reactions. The preprocessing phase addressed two key challenges: extracting form-specific information from leaflets containing multiple pharmaceutical forms for drugs, and standardizing ingredients from Italian to English for international compatibility. We employed GPT-4o with specialized prompts for these tasks, with results validated by healthcare professionals. A validation process involving a clinical pharmacist and a physician confirmed the high accuracy of our preprocessing approach (Cohen's kappa = 0.95). The final knowledge base comprises two components: a drug database containing detailed pharmaceutical information, and an in-memory synonyms database for the 1,035 ingredients to ensure rapid retrieval during decision support processes. Building upon the data pipeline, we finally developed a python web application for the HELIOT CDSS. This implementation showcases how the conceptual architecture outlined in section 2.1 can be realized through specific technological solutions and integration approaches.

2.3. Evaluation and Findings

We evaluated our system using 1,000 synthetic adverse reaction cases across seven clinical classes, achieving a macro-averaged F1 score of 0.9869. The system attained perfect classification for three critical categories: Direct Active Ingredient Reactivity, Drug Class Cross-Reactivity with Documented Tolerance, and Drug Class Cross-Reactivity Without Documented Tolerance. Significantly, the system could reduce interruptive alerts by 50.2% compared to traditional systems by appropriately categorizing cases as requiring interruptive alerts (45.5%), non-interruptive alerts (14.9%), or no alerts (39.6%). The system processed each case efficiently (average 2.775 seconds) due to optimizations including efficient data retrieval, an in-memory synonyms dictionary, and parallel database operations. These results suggest LLM-based approaches can substantially improve adverse drug reaction management over rule-based systems, particularly where clinical information exists primarily in unstructured formats.

3. LLMs for Drug-Drug Interaction Prediction

Existing computational DDI prediction methods require complex feature engineering and specialized architectures, limiting their practical implementation [13, 22, 23, 24, 25]. We propose a fundamentally different paradigm: leveraging LLMs to directly interpret and reason about drug interactions using purely textual representations of molecular structures (SMILES notation), target organisms, and gene interactions. This approach eliminates the need for specialized feature engineering while potentially capturing complex interaction patterns in textual drug information through LLMs' contextual understanding capabilities. The following subsections detail our text-based approach for DDI prediction, experimental results, and implications for clinical applications.

3.1. Approach

Our approach simultaneously processes multiple drug characteristics through textual inputs combining SMILES notation (representing molecular structure), target organisms, and gene interaction information. The underlying assumption follows established literature: ”two drugs potentially interact when a drug alters the other drug's therapeutic effects through targeted genes or signaling pathways” [24], incorporating molecular structure information to capture additional interaction mechanisms. We implemented three increasingly sophisticated approaches to evaluate LLMs' capabilities for DDI prediction. First, a zero-shot approach utilized a carefully engineered prompt structure, instructing the model to analyze drug information and classify whether administration causes interaction. Second, few-shot learning incorporated balanced examples (five positive, five negative) using two selection strategies: random selection and similarity-based selection that identified contextually relevant examples through embedding similarity. Finally, fine-tuning optimized selected models for DDI prediction using Low-Rank Adaptation (LoRA) [26] with hyperparameters optimized via Optuna [27]. For data preparation, we utilized DrugBank [28] as our primary source, filtering to include only approved or experimental drugs where both drugs target at least one gene. We extracted DrugBank IDs, SMILES notation, target organisms, and binary vectors representing gene targets for each drug pair. We created a balanced dataset of 2,070,300 drug pairs (50% positive, 50% negative) for training and validation, with additional validation across 13 external datasets from diverse sources [29]. We evaluated 18 different LLMs ranging from efficient models (1.5B-3B parameters) to large proprietary models (> 250B parameters), including GPT-4 [30], Claude 3.5 [31], Gemini [32], Phi-3.5 [33], and open-weight alternatives. Performance was assessed using accuracy, precision, sensitivity, and F1-score, with particular attention to sensitivity as a critical metric for medication safety. Results were validated across the 13 external datasets to ensure generalizability, with comparative analysis against established baselines including l2-regularized logistic regression [24] and the MSDAFL deep learning model [34].

3.2. Experimental Results and Implications for Clinical Applications

Our evaluation revealed distinct performance patterns across adaptation approaches. Zero-shot results showed limited effectiveness, with even proprietary models achieving modest sensitivity (0.5413-0.5927). Few-shot learning improved performance, particularly with similarity-based selection, reaching an accuracy of 0.8376 with Claude 3.5. Fine-tuning dramatically enhanced performance, with smaller models showing remarkable improvements. Across 13 external datasets, fine-tuned Phi-3.5 showed exceptional sensitivity (average 0.978) and a high accuracy (0.919), surpassing larger models such as GPT-4 and traditional baselines. These results offer significant clinical implications. The high sensitivity of fine-tuned LLMs addresses a critical safety priority in medication management, where missing interactions pose greater risks than false positives. The superior performance of smaller LLMs enables deployment on standard hardware without specialized infrastructure, addressing accessibility and privacy concerns through local processing capabilities. These systems could support clinical decision-making throughout the medication management process, from pre-prescription screening to emergency medicine. The finding that task-specific adaptation outweighs model size suggests efficient pathways for developing focused AI tools for healthcare applications, potentially improving medication safety while maintaining operational efficiency in everyday clinical workflows.

4. Future Directions and Challenges

Our LLM-based approaches are promising. HELIOT can process unstructured clinical narratives in various healthcare settings, from hospitals to primary care practices. At the same time, our DDI prediction approach offers value for research and clinical prescribing, particularly in complex patient-specific situations. Future work will focus on validation with real clinical data, expanding capabilities to address more complex clinical scenarios, and optimizing performance for point-of-care deployment. Collaboration with healthcare institutions remains essential for the comprehensive evaluation of the impact of the proposed systems on clinical workflows and decision-making processes, ensuring approaches that improve medication safety while integrating with existing healthcare systems.

5. Conclusion

Our work demonstrates how LLMs can transform medication safety management through contextual understanding and natural language processing capabilities. By enabling systems to process unstructured clinical information and complex pharmaceutical data, we address fundamental limitations in current adverse drug reaction management and interaction prediction approaches. The performance of smaller LLMs highlights the feasibility of practical clinical deployment without extensive computational resources. These findings suggest a promising path forward for AI applications in healthcare that balance effectiveness with accessibility, potentially improving patient outcomes while seamlessly integrating into clinical workflows. Continued collaboration between technical researchers and healthcare practitioners will be essential as the field evolves for successful implementation.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] R. T. Sutton, D. Pincock, D. C. Baumgart, D. C. Sadowski, R. N. Fedorak, K. I. Kroeker, An overview of clinical decision support systems: benefits, risks, and strategies for success, NPJ digital medicine 3 (2020) 17.

[2] U. Sarkar, L. Samal, How effective are clinical decision support systems?, 2020.

[3] P.-Y. Meunier, C. Raynaud, E. Guimaraes, F. Gueyffier, L. Letrilliart, Barriers and facilitators to the use of clinical decision support systems in primary care: a mixed-methods systematic review, The Annals of Family Medicine 21 (2023) 57–69.

[4] P. L. Quan, S. Sánchez-Fernández, L. Parrado Gil, A. Calvo Alonso, J. M. Bodero Sánchez, A. Ortega Eslava, M. Luri, G. Gastaminza Lasarte, Usefulness of drug allergy alert systems: Present and future, Current Treatment Options in Allergy 10 (2023) 413–427.

[5] T. K. Colicchio, J. J. Cimino, Beyond the override: Using evidence of previous drug tolerance to suppress drug allergy alerts; a retrospective study of opioid alerts, JBI 147 (2023) 104508.

[6] B. A. Van Dort, W. Y. Zheng, V. Sundar, M. T. Baysari, Optimizing clinical decision support alerts in electronic medical records: a systematic review of reported strategies adopted by hospitals, JAMIA 28 (2021) 177–183.

[7] L. Westerbeek, K. J. Ploegmakers, G.-J. De Bruijn, A. J. Linn, J. C. van Weert, J. G. Daams, N. van der Velde, H. C. van Weert, A. Abu-Hanna, S. Medlock, Barriers and facilitators influencing medication-related cdss acceptance according to clinicians: a systematic review, International Journal of Medical Informatics 152 (2021) 104506.

[8] G. Van De Sijpe, C. Quintens, K. Walgraeve, E. Van Laer, J. Penny, G. De Vlieger, R. Schrijvers, P. De Munter, V. Foulon, M. Casteels, et al., Overall performance of a drug–drug interaction clinical decision support system: quantitative evaluation and end-user survey, BMC Medical Informatics and Decision Making 22 (2022) 48.

[9] M. Topaz, D. L. Seger, S. P. Slight, F. Goss, K. Lai, P. G. Wickner, K. Blumenthal, N. Dhopeshwarkar, F. Chang, D. W. Bates, et al., Rising drug allergy alert overrides in electronic health records: an observational retrospective study of a decade of experience, JAMIA 23 (2016) 601–608.

[10] M. Topaz, D. L. Seger, K. Lai, P. G. Wickner, F. Goss, N. Dhopeshwarkar, F. Chang, D. W. Bates, L. Zhou, High override rate for opioid drug-allergy interaction alerts: current trends and recommendations for future, in: MEDINFO 2015: eHealth-enabled Health, IOS Press, 2015, pp. 242–246.

[11] T. C. Hsieh, G. J. Kuperman, T. Jaggi, P. Hojnowski-Diaz, J. Fiskio, D. H. Williams, D. W. Bates, T. K. Gandhi, Characteristics and consequences of drug allergy alert overrides in a computerized physician order entry system, JAMIA 11 (2004) 482–491.

[12] J. E. Sager, S. Tripathy, L. S. Price, A. Nath, J. Chang, A. Stephenson-Famy, N. Isoherranen, In vitro to in vivo extrapolation of the complex drug-drug interaction of bupropion and its metabolites with cyp2d6; simultaneous reversible inhibition and cyp2d6 downregulation, Biochemical pharmacology 123 (2017) 85–96.

[13] Y. Qiu, Y. Zhang, Y. Deng, S. Liu, W. Zhang, A comprehensive review of computational methods for drug-drug interaction detection, IEEE/ACM transactions on computational biology and bioinformatics 19 (2021) 1968–1985.

[14] S. Tripathi, R. Sukumaran, T. S. Cook, Efficient healthcare with large language models: optimizing clinical workflow and enhancing patient care, JAMIA 31 (2024) 1436–1440.

[15] A. Ríos-Hoyo, N. L. Shan, A. Li, A. T. Pearson, L. Pusztai, F. M. Howard, Evaluation of large language models as a diagnostic aid for complex medical cases, Frontiers in Medicine 11 (2024) 1380148.

[16] G. De Vito, F. Ferrucci, A. Angelakis, Llms for drug-drug interaction prediction: A comprehensive comparison, arXiv preprint arXiv:2502.06890 (2025).

[17] K. T. Ahmed, M. I. Ansari, W. Zhang, Dti-lm: language model powered drug–target interaction prediction, Bioinformatics 40 (2024) btae533.

[18] D. Oniani, J. Hilsman, C. Zang, J. Wang, L. Cai, J. Zawala, Y. Wang, Emerging opportunities of using large language models for translation between drug molecules and indications, Scientific Reports 14 (2024) 10738.

[19] G. De Vito, F. Ferrucci, A. Angelakis, Design and evaluation of a cdss for drug allergy management using llms and pharmaceutical data integration, arXiv preprint arXiv:2409.16395 (2024).

[20] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, D. C. Schmidt, A prompt pattern catalog to enhance prompt engineering with chatgpt, arXiv preprint arXiv:2302.11382 (2023).

[21] C. Wendler, V. Veselovsky, G. Monea, R. West, Do llamas work in english? on the latent language of multilingual transformers, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024, pp. 15366–15394.

[22] R. Ferdousi, R. Safdari, Y. Omidi, Computational prediction of drug-drug interactions based on drugs functional similarities, JBI 70 (2017) 54–64.

[23] A. Kastrin, P. Ferk, B. Leskošek, Predicting potential drug-drug interactions on topological and semantic similarity features using statistical learning, PloS one 13 (2018) e0196865.

[24] S. Mei, K. Zhang, A machine learning framework for predicting drug–drug interactions, Scientific Reports 11 (2021) 17619.

[25] H. Yu, S. Zhao, J. Shi, Stnn-ddi: a substructure-aware tensor neural network to predict drug–drug interactions, Briefings in Bioinformatics 23 (2022) bbac209.

[26] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, arXiv preprint arXiv:2106.09685 (2021).

[27] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A next-generation hyperparameter optimization framework, in: Proceedings of the 25th ACM SIGKDD international conference, 2019, pp. 2623–2631.

[28] D. S. Wishart, Y. D. Feunang, A. C. Guo, E. J. Lo, A. Marcu, J. R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, et al., Drugbank 5.0: a major update to the drugbank database for 2018, Nucleic acids research 46 (2018) D1074–D1082.

[29] S. Ayvaz, J. Horn, O. Hassanzadeh, Q. Zhu, J. Stan, N. P. Tatonetti, S. Vilar, M. Brochhausen, M. Samwald, M. Rastegar-Mojarad, et al., Toward a complete dataset of drug–drug interaction information from publicly available sources, JBI 55 (2015) 206–217.

[30] OpenAI, Gpt-4 technical report, arXiv:2303.08774 (2023).

[31] Anthropic, Claude 3.5 sonnet (version 3.5), 2024. URL: https://claude.ai/.

[32] Google, Google gemini 1.5 pro, 2024. URL: https://ai.google.dev/.

[33] M. Abdin, J. Aneja, H. Awadalla, A. Awadallah, A. A. Awan, N. Bach, A. Bahree, A. Bakhtiari, J. Bao, H. Behl, et al., Phi-3 technical report: A highly capable language model locally on your phone, arXiv preprint arXiv:2404.14219 (2024).

[34] C. Hou, G. Duan, C. Yan, Msdafl: molecular substructure-based dual attention feature learning framework for predicting drug–drug interactions, Bioinformatics 40 (2024) btae596.

6. Online Resources

The online repositories are available on GitHub: LLMDDI, HELIOT.