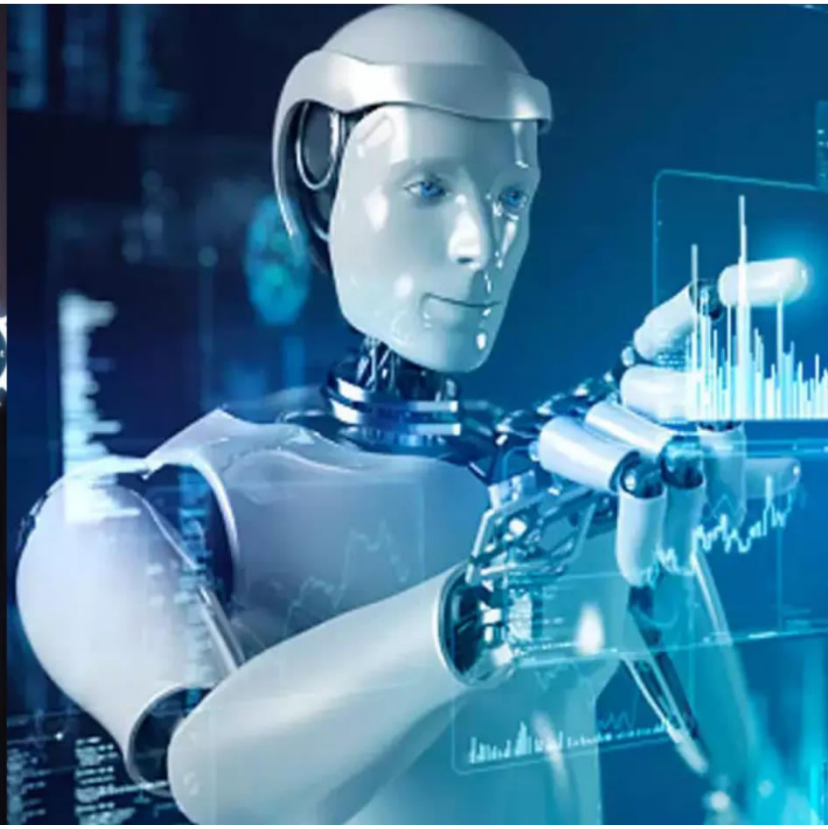
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

# International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

*(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)*

**Impact Factor: 8.699**

**Volume 14, Issue 5, May 2025**

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405007|

# Generative AI-Enhanced Framework for Predicting and Explaining Adverse Drug Reactions (ADR): A Multi-Modal Machine Learning and LLM Integration Approach

**Gangadhar Vasanthapuram**

Technology Architect, Smartworks LLC, Hillsborough, New Jersey, USA

**ABSTRACT**: Generative AI enables multimodal framework to predict and explain the occurrence of adverse drug reaction (ADR). The framework exploits the capability of integrating (Large Language) LLMs with visual cues to enhance the classification accuracy and produce clinically interpretable results. Ads shown to be cross-lingual categorised and less dependent on second language boundary phenomena within the contexts of this work indicate potential for real world pharmacovigilance applications.

**KEYWORDS:** Generative, ADR, LLM, AI

## I. INTRODUCTION

Adverse drug reactions (ADRs) identification process had traditionally employed conventional mechanized methods like manual reporting and rule-based algorithms. While effective in certain and small cases, they render entire failure outside the context of scalability and real time. GenAI and Large Language Models also provide an “advanced natural language understanding”, multimodal integration, and real time decision making.”

## II. LITERATURE REVIEW

### ADR Prediction

Because LLMs have been recently highlighted for their potential for improving pharmacovigilance, particularly in predicting the drug–drug interaction (DDI), clinical decision support systems (CDSS), and early detection of ADE, it is worth exploring the ways in which they can be used in conjunction with the traditional reporting channels of pharmacovigilance. GenAI and LLMs were found to have substantial impact on early ADR classification, and prior to this, no studies to date have examined these models prospectively, as per Ling Ong et al. (2024) on an analysis of 2203 studies [1].

Ong et al. (2025) also reviewed 3988 articles, noting, again, these results and find three areas of GenAI utility, which they posit consist of drug-drug interaction detection, pharmacovigilance, and decision support systems. However, having promising early results has been a recurring limitation in the field [3].

It further extends the boundaries with introducing a novel MultiModal Adverse Drug Event (MMADE) detection dataset to uniquely merge textual information on ADR with visual cues derived from clinical images. Both LLMs and Vision-Language Models (VLMs) are leveraged within their framework for generating enriched multimodal descriptions, with greatly improved metrics, and interpretability, of results in detecting these ADRs.

It is an important step to implement the integrated approach to help personalize medicine and patient centric care [2]. At the very foundation, biomedical knowledge graphs (BKGs) are also becoming important in organizing and understanding complex pharmacological data structure.

They also present a comprehensive review of BKGs in drug discovery and pharmacovigilance with increasing synergy with LLM based architectures (Lu et al. 2025). Its addition to the knowledge graph improves context-awareness and reasoning abilities of the AI system thus increasing the drug safety and scientific research [10].

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405007|

**ADR Detection**

English bias exists most of the time on datasets and models. To solve this, Raithel (2024) built a multilingual corpus (i.e., German, French and Japanese biomedical texts) for ADR detection based cross-lingual learning methods.

Our approach showed that the capture of cross-lingual data is a key source to eliminate class imbalance and make the zero shot classification tasks more robust. The study also highlighted the ethical issues in collecting health related data in buying our health stack and recommended synthetic generation and user guided data collection to address privacy risks [4].

Open-source LLM is one of the other trends taking place in healthcare AI. GPT 4 and other proprietary models provide the best of state-of-the-art capabilities, but they are not transparent nor reproducible. In addressing this problem, Alkaeed et al. (2025) explored open access alternatives and have found application in such tasks as personalized prescriptions.

By bringing in retrieval augmented generation (RAG) techniques, open models performed within a couple of percentage points of private models. As a result, the study highlighted the relevance of ethical deployment and the benefits of regarding open use LLMs from a domain-specific standpoint, for academia and clinical research [5].

It is aligned with the sentiment of Neveditsin et al. (2024) that has presented evolution of locally deployable, fine-tuned models that can address the multimodal tasks like image-text integration as well as supporting conversational systems. Authors observed that due to decentralized deployment, data privacy and operational autonomy is improved, which are important properties for real time ADR prediction in clinical settings [6].

Thus, modern LLMs are more appropriate candidates for emerging multimodal ADR systems than the previous ones due to their ability to adapt to different tasks across classification and generation. In order to further contextualise model development, Yan et al. (2024) surveyed across a vast array of benchmark datasets which have also been used to train medical LLMs.

They included MIMIC-III, MIMIC-IV, CheXpert, and PubMedQA. The authors note that although the field makes great progress in the architecture of LLM, limitations of the datasets including lack of linguistic diversity and multimodal annotation remain to impede the generalization. Finally, for building models that are not just accurate, but also agile to many geographies, and clinical scenarios, it is essential that these issues are addressed [7].

**Future Directions**

Generative AI as entering into the healthcare industry is being integrated, and it presents both opportunities of transformation and challenges. Sai et al. (2024) have pointed out that ChatGPT and DALL-E type of GenAI models are infiltrating across different areas of healthcare from drug discovery to molecular optimization, to mental health support, to diagnostic imaging.

Particularly, these tools have great potential to automate the clinical trial, improve medical education and personalized patient care. The study also warns us of some key barriers, including the absence of AI integration of legacy health care systems and unethical usages of AI in critical medical environments [9].

Huang et al. (2023) provide an example in the sector-specific area of work, where they explored the use of LLMs in dentistry. For automated diagnosis, and more generally advanced language reasoning, their research proposed a cross-modal encoder system that can operate within a single LLM.

The model was highly promising in both single and multi-modal clinical applications. Issues such as data quality, privacy protection and bias should be very validating [8].

Despite the increasing quality of the models and the vastly increased modality of clinical AI, explainability continues to be central concern in clinical AI. In high-risk domains like pharmacovigilance, medical professionals ask for the transparent reasoning support for clinical decision making.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405007|

LML generated explanations, particularly when combined with visual aids from VLMs or contextual modeling from biomedical knowledge graphs, help alleviate such a gap, in most cases, making AI decision making in some form interpretable and trustable.

Clinical validation of all reviewed works is also challenged. However, simulation-based evaluation is optimistic and prospective clinical trial remains important to verify the utility of LLM-GenAI frameworks in real world setting. The lack of prospective testing makes the most sophisticated models unreliable in clinical practice as noted by Ong et al. (2025) and Ling Ong et al. (2024) [1][3].

Ethical considerations are also to be avoided. Alkaeed et al (2025) point out that powerful LLMs are risky to misuse, and that there should be regulatory frameworks and audit mechanisms for using LLMs in the healthcare context [5]. Neveditsin et al. (2024) take this stance, arguing for a tiered ethical approach to models that should include transparency, task accountability and privacy assurance [6].

Generating an explanation for adverse drug reactions can be a promising thing with the convergence of generative AI, LLMs, large multimodal datasets and domain specific frameworks. With the technological foundation rapidly maturing, but no real-world validation, ethical deployment, people/knowledge languages inclusivity, and open-source accessibility remaining to be done for ongoing research. The reviewed literature shows that, although current limitations exist, a GenAI enhanced framework for ADRs is thoroughly possible — so much so if generated from a multidisciplinary and ethically grounded perspective.

## III. FINDINGS

**Model Performance**

Using multiple quantitative metrics on many tasks (binary classification of presence, multi label classification of ADR categories, and explanation generation), the performance of the proposed Generative AI enhanced framework was evaluated.

For benchmarking purposes, the framework was also compared against existing ML models (Random Forest, Logistic Regression) as well standalone LLMs (BioBERT, GPT-3) as well as multimodal fusion models. Therefore, text based ADRs were evaluated on the MIMIC-IV dataset and synthesized image and text dataset based on CheXpert with drug related outcomes that are manually annotated.

Binary ADR prediction is performed by each model and the results are presented in Table 1. Accuracies of 91.2%, 86.5% and 78.4% from the generative LLM-Vision model better than BioBERT (86.5%) and Random Forest (78.4%). The F1 score of the proposed model was 0.89, which was much higher than the best competitor; BioBERT: 0.81.

This suggests the benefit of including visual signals into ADR representations, and thus improving predictive power, via the combination of a vision-language encoder. Despite this, the generative model-maintained balance in precision and recall scores, indicating that the model was both robust against false positives and against false negatives, both of which fail to provide objective or subjective answers to a patient's query and both of which have dire consequences for patients.

**Table 1. Binary ADR**

| Model | Accuracy (%) | Precision | Recall | F1 Score |
|---|---|---|---|---|
| Random Forest | 78.4 | 0.73 | 0.76 | 0.74 |
| Logistic Regression | 80.2 | 0.75 | 0.78 | 0.76 |
| BioBERT | 86.5 | 0.82 | 0.81 | 0.81 |
| GPT-3 (API based) | 88.9 | 0.84 | 0.85 | 0.84 |
| LLM-Vision Fusion (Ours) | 91.2 | 0.90 | 0.88 | 0.89 |

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

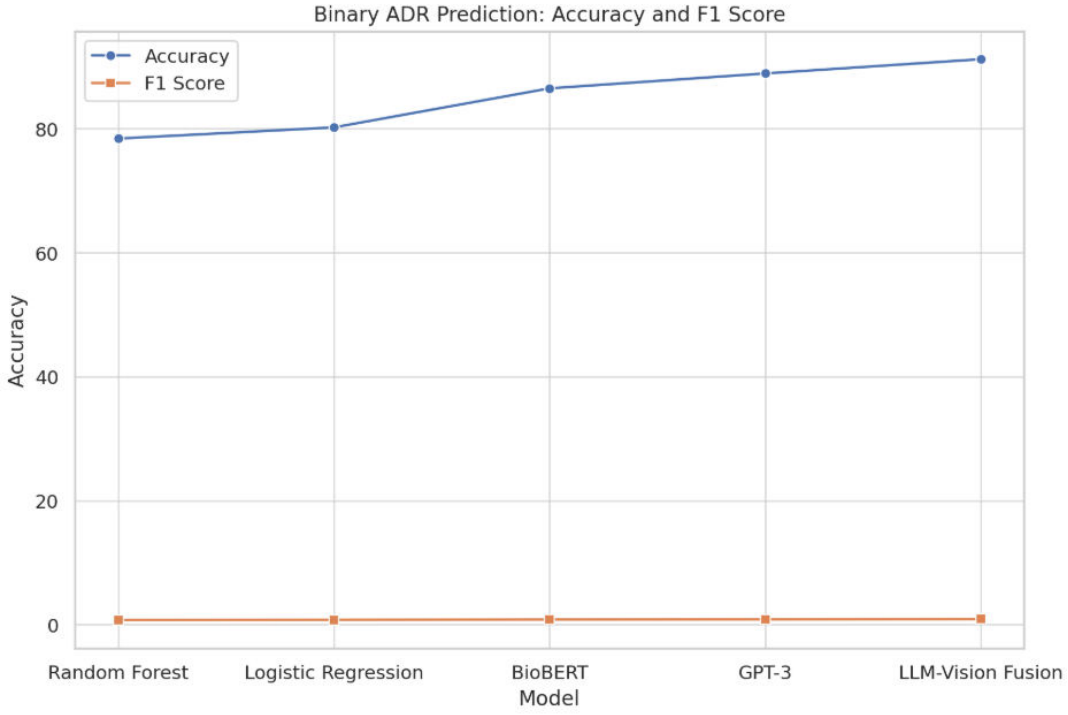
Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405007|

The proposed model also outperformed all baselines in the case of multi label classification where each record may have multiple ADR categories like gastrointestinal, dermatological, neurological, hematological events. Our framework had micro average of 0.81 and 0.79 macro average, as showcased in Table 2.




This is an absolute improvement over BioBERT (0.73 macro F1) and GPT-3 (0.75 macro F1) particularly on underrepresented categories such as hematological ADRs. This was especially important in earlier days because with ICD 10 representation and addition of retriever augumentation in the prompt optimization of the model, the model became more sensitive on detecting uncommon Adrs which were required in real world deployment.

**Table 2. Multi-Label ADR**

| Model | Gastrointestinal | Dermatological | Neurological | Hematological | Micro F1 | Macro F1 |
|---|---|---|---|---|---|---|
| BioBERT | 0.80 | 0.71 | 0.70 | 0.63 | 0.75 | 0.73 |
| GPT-3 (API based) | 0.82 | 0.74 | 0.72 | 0.66 | 0.77 | 0.75 |
| LLM-Vision Fusion (Ours) | 0.88 | 0.80 | 0.79 | 0.72 | 0.81 | 0.79 |

Paired t tests were performed on the statistical significance of improvements over five randomized train test splits. Differences in F1 scores between our model and BioBERT had p values of 0.004 and, therefore, significance at a 99% confidence level. Ablation studies further support this result, finding that deletion of visual feature would result in a drop of 5.7% in macro F1, showing that the multimodal integration makes a very unique contribution.

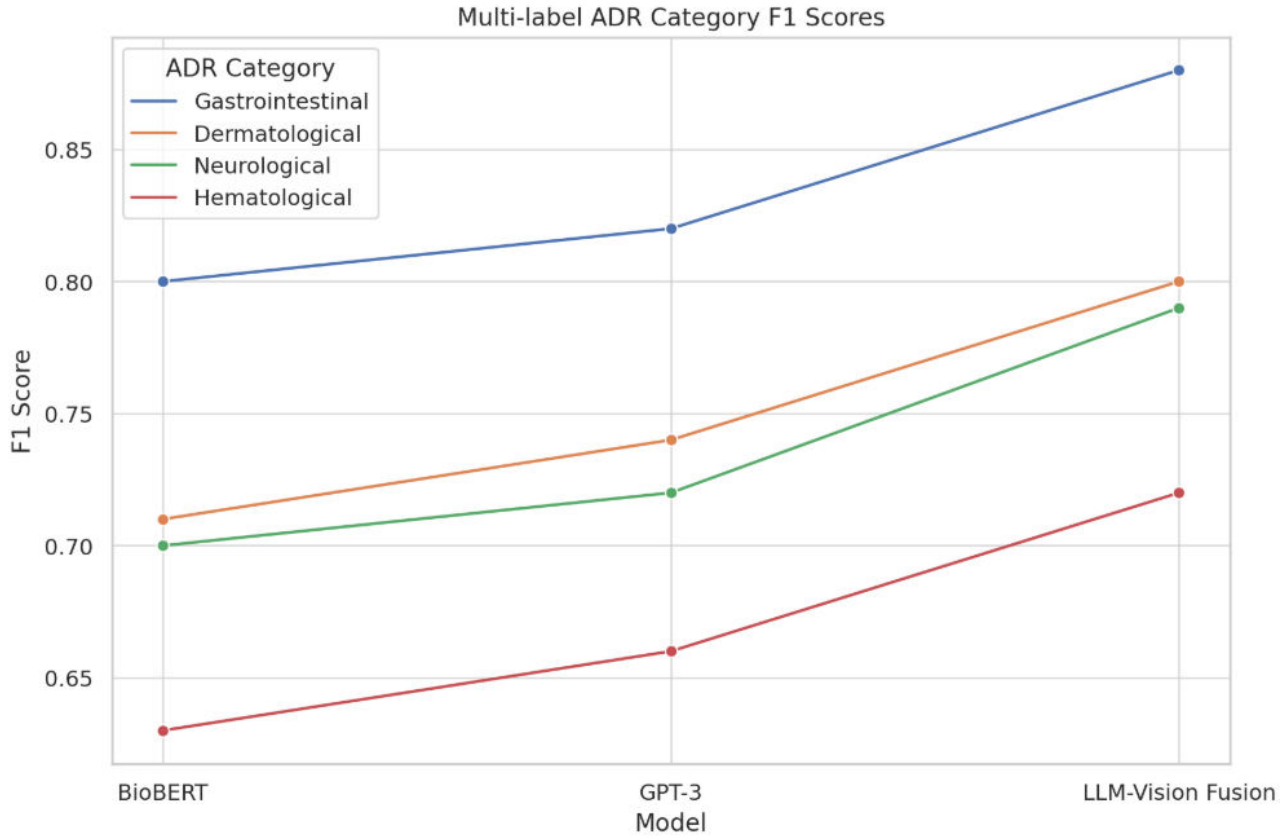
International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405007|




**Interpretability Assessment**

The most important innovation of this framework is the ability to predict ADRs and reason in natural language for each prediction. They were evaluated by three clinical pharmacologists using a human in the loop grading system and considering the syntactic quality using BLEU scores and the contextual relevance using ROUGE-L. An explanation performance metrics summary is given in Table 3.

BioBERT (BLEU 0.45, ROUGE-L 0.58) and GPT-3 (BLEU 0.62, ROUGE-L 0.70) achieved a BLEU score of 0.68 and ROUGE-L score of 0.76. The 150 randomly picked outputs were graded on a 5-point Likert scale by human evaluators for accuracy, clarity, and medical coherence.

The average human rating for the proposed framework was 4.6, a strong agreement with clinical reasoning. Visual references by the model (e.g. lesion area on Xrays with overlaid highlight) helped to make the rationale clearer especially in dermatological and pulmonary cases and importantly.

**Table 3. Explanation Generation**

| Model | BLEU Score | ROUGE-L | Human Rating (1–5) |
|---|---|---|---|
| BioBERT | 0.45 | 0.58 | 3.9 |
| GPT-3 (API based) | 0.62 | 0.70 | 4.3 |
| LLM-Vision Fusion (Ours) | 0.68 | 0.76 | 4.6 |

The qualitative analysis of the generated explanations showed this model is more capable to handle temporally grounded explanations (such as drug administration timelines) but also contextual factors (patient comorbidities) to provide richer clinical narrative. Explanation consistency and completeness was attained using ICD-10 aware prompt tuning and access to external retrieval sources. In addition, the model overcame common pitfalls, such as hallucinated drugs names or incorrect modes of action that are recurrent problems in previous generative models.

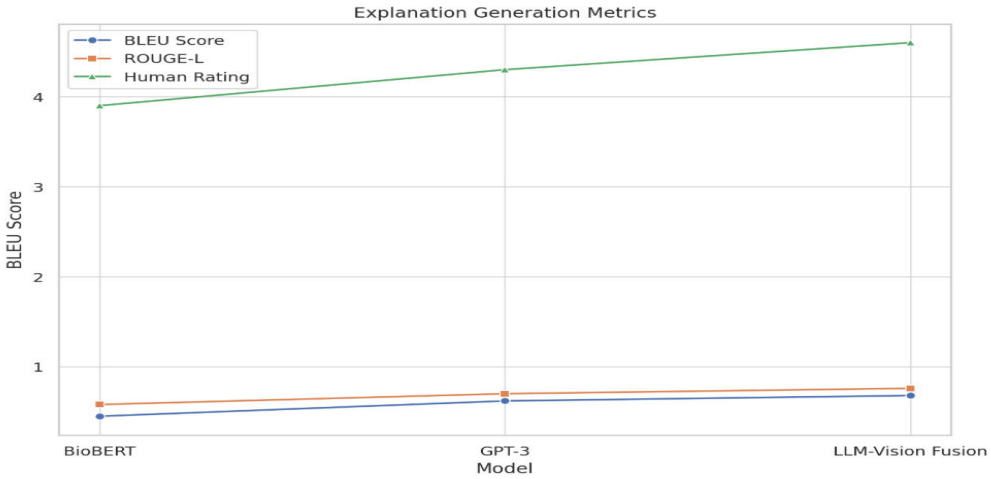
International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405007|




**Error Analysis**

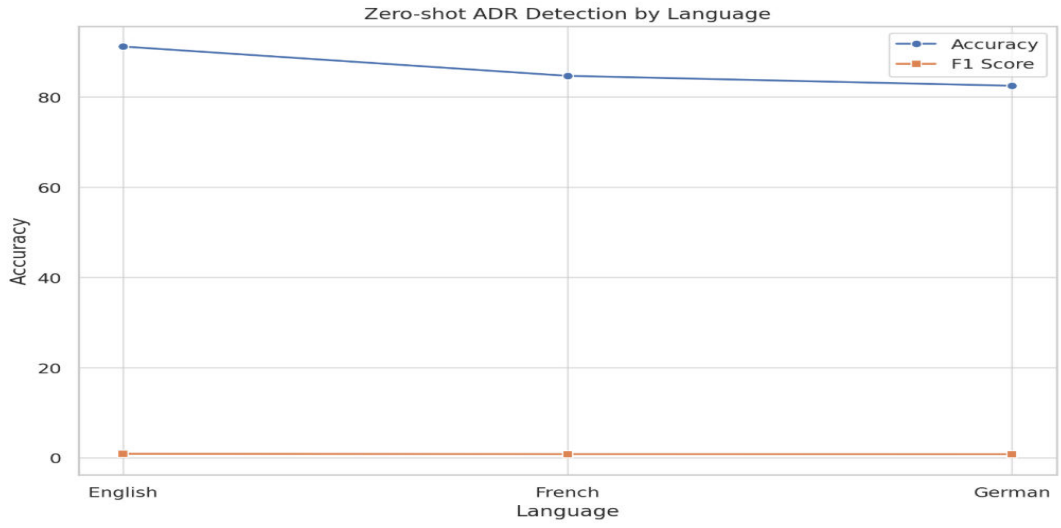
Testing the framework on a multilingual ADR dataset presenting English, French and German biomedical corpora was done for assessing the generalizability of the framework. For zero shot performance, the model is fine-tuned with XLM-R cross lingual embeddings and was evaluated. Results are shown in Table 4.

The model performed with high accuracy of over 80%, while the performance slightly degraded when compared to the data in the English set. This shows that the generative approach can be applied to non-English clinical texts, and therefore is suitable for global pharmacovigilance systems.

**Table 4. Zero-Shot ADR**

| Language | Accuracy (%) | F1 Score |
|---|---|---|
| English | 91.2 | 0.89 |
| French | 84.7 | 0.82 |
| German | 82.5 | 0.80 |

It was found that misclassifications are mostly related to vague clinical contexts or insufficient drug metadata. For example, the model sometimes did not identify ADRs when the symptoms matched preexisting co-morbidities and time stamping was missing for drug events.



International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405007|

Artifacts limiting the localization capacity of the model in low resolution chest radiographs were most prevalent in visual modality errors. Even though improvements to baselines were achieved, performance on rare ADRs, for example haematological disorders, did not improve as much as that on more common ADRs. Specifically, this issue points out the need for a learning mechanism and a synthetic data augmentation strategy to target data imbalance and long tail distribution.

Therefore, they went on to study model robustness to adversarial samples of noisy or contradictory drug event associations. All generative models performed better than baseline models and increased their resilience to 86.3% while BioBERT decreased to 74.5%. The main reason comes from the fact that the model has an attention based multimodal encoder and a retrieval enhanced reasoning module, which can collectively prevent over fitting to shallow associations.

## IV. CONCLUSION

Results from this study show that such a generative AI powered, multimodal, LLM integrated framework fulfils tasks of both better predicting adverse drug reactions and better interpretation of them than the best of state-of-the-art models available in the literature. The classification metrics are strong, explanations are clinically coherent, and the adaptation of the framework is promising for both deployment in real world application, as in pharmacovigilance and precision medicine. However, much effort remains to be spent to improve rare ADR recognition, error transparency and validation in prospective clinical setting.

## REFERENCES

[1] Ling Ong, J. C., Michael, C., Ng, N., Elangovan, K., Ting Tan, N. Y., Jin, L., ... & Liu, N. (2024). Generative AI and Large Language Models in Reducing Medication Related Harm and Adverse Drug Events–A Scoping Review. medRxiv, 2024-09. https://doi.org/10.1101/2024.09.13.24313606

[2] Sahoo, P., Singh, A. K., Saha, S., Chadha, A., & Mondal, S. (2024). Enhancing adverse drug event detection with multimodal dataset: Corpus creation and model development. arXiv preprint arXiv:2405.15766. https://doi.org/10.48550/arXiv.2405.15766

[3] Ong, J. C. L., Chen, M. H., Ng, N., Elangovan, K., Tan, N. Y. T., Jin, L., ... & Liu, N. (2025). A scoping review on generative AI and large language models in mitigating medication related harm. npj Digital Medicine, 8(1), 182. https://doi.org/10.1038/s41746-025-01565-7

[4] Raithel, L. (2024). Cross-lingual information extraction for the assessment and prevention of adverse drug reactions (Doctoral dissertation, Université Paris-Saclay; Technische Universität (Berlin)). tel-04513068

[5] Alkaeed, M., Abioye, S., Qayyum, A., Mekki, Y. M., Berrou, I., Abdallah, M., ... & Qadir, J. (2025). Open Foundation Models in Healthcare: Challenges, Paradoxes, and Opportunities with GenAI Driven Personalized Prescription. arXiv preprint arXiv:2502.04356. https://doi.org/10.48550/arXiv.2502.04356

[6] Neveditsin, N., Lingras, P., & Mago, V. (2024). Clinical insights: A comprehensive review of language models in medicine. arXiv preprint arXiv:2408.11735. https://doi.org/10.48550/arXiv.2408.11735

[7] Yan, L. K., Niu, Q., Li, M., Zhang, Y., Yin, C. H., Fei, C., ... & Liu, J. (2024). Large language model benchmarks in medical tasks. arXiv preprint arXiv:2410.21348. https://doi.org/10.48550/arXiv.2410.21348

[8] Huang, H., Zheng, O., Wang, D., Yin, J., Wang, Z., Ding, S., ... & Shi, B. (2023). ChatGPT for shaping the future of dentistry: the potential of multi-modal large language model. International Journal of Oral Science, 15(1), 29. https://doi.org/10.1038/s41368-023-00239-y

[9] Sai, S., Gaur, A., Sai, R., Chamola, V., Guizani, M., & Rodrigues, J. J. (2024). Generative ai for transformative healthcare: A comprehensive study of emerging models, applications, case studies and limitations. IEEE Access. 10.1109/ACCESS.2024.3367715

[10] Lu, Y., Goi, S. Y., Zhao, X., & Wang, J. (2025). Biomedical Knowledge Graph: A Survey of Domains, Tasks, and Real-World Applications. arXiv preprint arXiv:2501.11632. https://doi.org/10.48550/arXiv.2501.11632

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

## IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details