Entity Only vs. Inline Approaches: Evaluating LLMs for Adverse Drug Event Detection in Clinical Text (Student Abstract)

Howard Prioleau[1] **and Saurav K. Aryal**[1]

[1]EECS, Howard University, Washington, DC 20059, USA
howard.prioleau@bison.howard.edu, saurav.aryal@howard.edu

Abstract

Adverse Drug Events (ADEs) are a major healthcare issue in the United States, contributing to millions of outpatient and emergency department visits and ranking as the fourth leading cause of death. While many ADEs are identified post-market, improved detection methods are crucial for enhancing patient safety. This study explores the application of large language models (LLMs) to the n2c2 task for ADE detection, evaluating optimal prompting techniques without requiring ADE-specific training data. Results indicate that an entity-only extraction approach outperforms the inline method, offering higher precision, recall, and token efficiency. This study highlights the potential of LLMs for accurate ADE detection in clinical text, improving performance while maintaining model efficiency.

Introduction

Adverse Drug Events (ADEs) are a significant healthcare issue in the US, causing millions of outpatient and emergency visits annually, and costing billions. Despite FDA testing, many ADEs are only identified post-market, necessitating improved detection methods. ADEs also erode public trust (Cresswell et al. 2007), lead to treatment hesitancy (Balog-Way et al. 2021), and disproportionately affect under-served populations (Faasse and Petrie 2013). Automatic identification is challenging due to complex, unstructured EHRs.

ADE detection research began with the manually annotated ADE Corpus (Gurulingappa et al. 2012). The 2018 n2c2 task (Henry et al. 2020) expanded concept types and maintained original EHR formats. Modeling techniques have evolved from traditional machine learning to advanced deep learning, achieving higher F1 scores (Gurulingappa et al. 2012; Henry et al. 2020). LLMs show promise in ADE detection (Gu et al. 2023), but their application to the n2c2 task still needs to be improved, which this study aims to address.

This study compares inline entity extraction and standalone entity output for ADE detection in a Text-to-Text task. We also evaluate optimal prompting techniques and targets for LLMs to identify ADEs in the clinical text without ADE-specific training data. The research seeks to identify the most effective technique and evaluate the performance of the general-purpose language model in this task.

Methodology

Dataset & Pre-Processing

This study uses the n2c2 data set (Henry et al. 2020), which comprises 505 discharge summaries from the MIMIC-III (Medical Information Mart for Intensive Care III) database. Seven domain experts, including four physician assistant students and three nurses, annotated these documents for drugs, ADEs, strength, form, dosage, frequency, route, duration, and reason.

Due to the structured nature of medical notes, preprocessing was necessary to represent the text more naturally. This involved converting all text to lowercase, removing non-ASCII characters, and eliminating punctuation that was not a part of decimal numbers. Unnecessary special characters and a predefined set of stop words were excluded. These steps standardized the text, improving the accuracy of ADE detection in clinical narratives. Lastly, the entity labels ADE and Drug were retained, while all other entities were categorized under the label 'other'.

Prompting

Two prompting techniques will be tested: Inline and Entity-only. Inline modifies input text by inserting markers around target entity spans, integrating entity detection within the clinical text. Entity-only focuses on extracting and listing relevant entities separately, aiming to streamline the process and enhance precision, recall, and token efficiency. Top K examples were selected based on ADE and drug entity density, providing diverse representations to improve the model's generalization ability.

Modeling & Evaluation

For our modeling purposes, we exclusively utilize the Llama-3.1 8B model (AI@Meta 2024) due to its substantial context window size, robust performance, open-sourceness, and the potential for comparative analysis with larger parameter versions in future studies. We do not tweak the base generation parameters and use all the defaults of the model.

We evaluate the model by scoring their ability to do ADE and Drug Prediction lenient f1 because it is the standard for

this task. True positives are correctly predicted entities, false positives are incorrect predictions, and false negatives are missed true entities. We report Precision, Recall, and F1 for ADE and Drug classes on the test set.

Results

type	ex#	drug pr	drug re	drug f1	ade pr	ade re	ade f1
inline	0	0.457	0.285	0.351	0.036	0.222	0.062
ent_only	0	**0.575**	0.484	**0.525**	**0.067**	**0.407**	**0.115**
inline	1	0.218	0.246	0.231	0.018	0.163	0.033
ent_only	1	0.407	0.558	0.471	0.034	0.378	0.063
inline	5	0.405	0.272	0.326	0.039	0.213	0.066
ent_only	5	0.414	0.593	0.488	0.047	0.137	0.070
inline	10	0.297	0.516	0.377	0.017	0.304	0.032
ent_only	10	0.300	**0.603**	0.400	0.026	0.188	0.046

Table 1: Llama 3.1 8B Test Set Drug and ADE Precision (pr), Recall (re), and Lenient F1 scores. ex# (Example Number)

Table 1 reveals several noteworthy trends. Quite surprisingly, zero-shot entities performed the best in both drug f1 and ade f1. Although the model exhibited a high level of performance without any examples, a more intriguing observation is that its responses in the zero-shot setting closely followed the inline responses, underscoring the significance of effective prompting in further enhancing the model's performance. As well, increasing the number of examples from 0 to 10 does not consistently enhance performance. Although some metrics show improvements with more examples, others do not exhibit a clear trend, suggesting that adding more examples does not necessarily lead to better model performance. This trend is further confirmed in the plot below Figure 1.

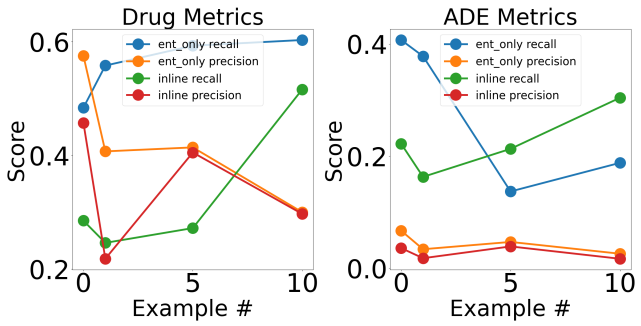

Figure 1: Precision and Recall scores for Drug and Adverse Drug Event (ADE) predictions across different example numbers for two methods: inline and entity only

As shown in Figure 1, models struggle with overprediction for both Drug and ADE entities, often repeating the same entities, which lowers precision scores. Despite duplications, recall scores remain high, indicating that the models capture many true positives but generate false positives, maintaining high recall and compromising precision.

Overall, the inline entity approach balances precision and recall, while the entity-only approach favors recall over precision. The entity-only approach outperforms the inline method across all metrics and is more token-efficient. This efficiency, focusing solely on entity extraction, likely contributes to its superior performance by reducing the complexity and potential errors in the extraction process.

Lastly, our findings suggest that entity-only extraction is more effective for Drug and ADE tasks, offering higher precision and recall with token efficiency. However, the inconsistent performance scaling with example numbers indicates that other factors may influence model effectiveness. This underscores the importance of further study and the potential for your work to advance the field.

Conclusion

In conclusion, this study showcases the impressive abilities of LLMs capabilities with and without examples of identifying drugs and ADEs within the clinical text. Although the performance leaves much to be gained through further investigation, this study is limited by using only one relatively small LLM. Relying on a single, smaller-scale model may restrict the generalizability of the findings and overlook the potential improvements and variations that could be achieved with larger or alternative language models. Future work should incorporate a broader range of language models to identify the most effective architectures to accurately detect drugs and ADEs in medical texts. Subsequently, fine-tuning these models will enhance their performance and better adapt them to the nuances of medical data. These efforts will build on the findings of this study, improving the accuracy and reliability of LLM for the analysis of medical text in the real world.

Acknowledgements

This research was, in part, funded by the National Institutes of Health (NIH) Agreement NO. 1OT2OD032581-01. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the NIH. Additional support was provided by the Institute for Socially and Culturally Relevant Human-Centered Artificial Intelligence at Howard University (GRT000370) and the Artificial Intelligence for Positive Change (AI4PC) lab.

References

AI@Meta. 2024. Llama 3.1 Model Card.

Balog-Way, D.; Evensen, D.; Löfstedt, R.; and Bouder, F. 2021. Effects of public trust on behavioural intentions in the pharmaceutical sector: data from six European countries. *Journal of Risk Research*, 24(6): 645–672.

Cresswell, K. M.; Fernando, B.; McKinstry, B.; and Sheikh, A. 2007. Adverse drug events in the elderly. *British medical bulletin*, 83(1): 259–274.

Faasse, K.; and Petrie, K. J. 2013. The nocebo effect: patient expectations and medication side effects. *Postgraduate medical journal*, 89(1055): 540–546.

Gu, Y.; Zhang, S.; Usuyama, N.; Woldesenbet, Y.; Wong, C.; Sanapathi, P.; Wei, M.; Valluri, N.; Strandberg, E.; Naumann, T.; et al. 2023. Distilling large language models for

biomedical knowledge extraction: A case study on adverse drug events. *arXiv preprint arXiv:2307.06439*.

Gurulingappa, H.; Rajput, A. M.; Roberts, A.; Fluck, J.; Hofmann-Apitius, M.; and Toldo, L. 2012. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of Biomedical Informatics*, 45(5): 885–892. Text Mining and Natural Language Processing in Pharmacogenomics.

Henry, S.; Buchan, K.; Filannino, M.; Stubbs, A.; and Uzuner, O. 2020. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*, 27(1): 3–12.