Evaluating large language models for anaphylaxis detection in clinical notes

Avaliando modelos de linguagem de grande escala para detecção de anafilaxia em anotações clínicas

Evaluación de modelos de lenguaje de gran escala para la detección de anafilaxia en notas clínicas

Matheus Matos Machado[1], Joice Basílio Machado Marques[3], Fabrício A. Gualdani[4], Monica Pugliese Heleodoro dos Santos[5], Fabio Cerqueira Lario[6], Chayanne Andrade de Araujo[5], Fabiana Andrade Nunes Oliveira[5], Luis Felipe Chiaverini Ensina[6], Ricardo Marcondes Marcacini[2], Dilvan Moreira[2].

1. B.Sc., Department of Computer Science, University of São Paulo (USP), São Carlos (SP), Brazil.
2. Ph.D., Department of Computer Science, University of São Paulo (USP), São Carlos (SP), Brazil.
3. Ph.D., Department of Research, Sofya, São Paulo (SP), Brazil.
4. M.Sc., Department of Information Science, Universidade Estadual Paulista (UNESP), Marília (SP), Brazil.
5. M.Sc., Division of Allergy, Hospital Sírio-Libanês, São Paulo (SP), Brazil.
6. Ph.D., Division of Allergy, Hospital Sírio-Libanês, São Paulo (SP), Brazil.

Corresponding author: B.Sc. Matheus Matos Machado
E-mail: matheusmatos@usp.br

Resumo

Objetivo: Este estudo tem como objetivo avaliar o potencial de quatro Modelos de Linguagem de Grande Escala (LLMs) (GPT-4 Turbo, GPT-3.5 Turbo, Gemini 1.0 Pro e OpenChat 3.5) na detecção de anafilaxia em Registros Médicos Eletrônicos (EMRs). **Método:** O método empregado envolveu a análise de 150 relatórios médicos, utilizando diferentes prompts para testar a capacidade dos LLMs em identificar a anafilaxia. **Resultados:** Os resultados indicam que todos os modelos obtiveram zero falsos negativos, com destaque para o GPT-4 Turbo, que alcançou 97% de acurácia e 91% de precisão. **Conclusão:** Conclui-se que os LLMs demonstram potencial para auxiliar na identificação da anafilaxia, especialmente o GPT-4 Turbo. A pesquisa reforça a importância do design eficiente de prompts para otimizar a acurácia dos resultados.

DOI: 10.59681/2175-4411.v16.iEspecial.2024.1364

Descritores: Inteligência Artificial; Modelos de Linguagem de Grande Escala; Anafilaxia.

Abstract

Objective: This study aims to evaluate the potential of four Large Language Models (LLMs) (GPT-4 Turbo, GPT-3.5 Turbo, Gemini 1.0 Pro, and OpenChat 3.5) in detecting anaphylaxis in Electronic Medical Records (EMRs). **Method:** The method employed involved the analysis of 150 medical reports, using different prompts to test the ability of the LLMs to identify anaphylaxis. **Results:** The results indicate that all models obtained zero false negatives, with GPT-4 Turbo standing out, achieving 97% accuracy and 91% precision. **Conclusion:** It is concluded that LLMs demonstrate the potential to assist in the identification of anaphylaxis, especially GPT-4 Turbo. The research reinforces the importance of efficient prompt design to optimize the accuracy of results.

Keywords: Artificial Intelligence; Large Language Models; Anaphylaxis.

Resumen

Objetivo: Este estudio tiene como objetivo evaluar el potencial de cuatro Modelos de Lenguaje de Gran Escala (LLMs) (GPT-4 Turbo, GPT-3.5 Turbo, Gemini 1.0 Pro y OpenChat 3.5) en la detección de anafilaxia en Registros Médicos Electrónicos (EMRs). **Método:** El método empleado involucró el análisis de 150 informes médicos, utilizando diferentes prompts para probar la capacidad de los LLMs para identificar la anafilaxia. **Resultados:** Los resultados indican que todos los modelos obtuvieron cero falsos negativos, destacándose el GPT-4 Turbo, que alcanzó un 97% de precisión y un 91% de exactitud. **Conclusión:** Se concluye que los LLMs demuestran potencial para ayudar en la identificación de la anafilaxia, especialmente el GPT-4 Turbo. La investigación refuerza la importancia del diseño eficiente de prompts para optimizar la precisión de los resultados.

Descriptores: Inteligencia artificial; Modelos de Lenguaje de Gran Escala; Anafilaxia.

Introduction

Anaphylaxis is a potentially life-threatening systemic hypersensitivity reaction that can affect multiple organ systems, including the skin, respiratory, gastrointestinal, and cardiovascular systems. Typical symptoms include hives, angioedema, difficulty breathing, vomiting, diarrhea, and significant blood pressure drops, which can be fatal without prompt treatment (Simons et al., 2014; Ensina et al., 2022). The rapid onset and progression of these symptoms, usually peaking within minutes after exposure to the trigger, underscore the need for swift recognition and immediate medical intervention to prevent severe outcomes (Simons et al., 2014; Ensina et al., 2022). Common triggers include foods (such as peanuts, tree nuts, and shellfish), insect stings, medications, and latex. In some cases, the exact trigger may remain unidentified (Simons et al., 2014; Ensina et al., 2022).

Diagnosis is primarily based on clinical evaluation. The World Allergy Organization (WAO) identifies two main clinical criteria for diagnosis: (1) visible skin symptoms along with significant symptoms in at least one other organ system, and (2) known or suspected allergen exposure causing respiratory or cardiovascular issues (Cardona et al., 2020). Manually applying these criteria can be challenging, particularly in resource-limited, fast-paced clinical settings. This potential gap in documented diagnoses presents an opportunity for using automated technologies to improve anaphylaxis detection (Liu et al., 2023).

One promising technology is Large Language Models (LLMs), which have demonstrated exceptional capabilities in parsing text and generating human-like responses across various applications, including healthcare (Gao et al., 2023). These models are trained on extensive text corpora, enabling them to process and produce text with contextual accuracy (Gao et al., 2023).

In this study, we investigate the effectiveness of four LLMs—GPT-4 Turbo, GPT-3.5 Turbo, Gemini 1.0 Pro, and OpenChat 3.5—in independently identifying anaphylaxis diagnoses from medical reports in Electronic Medical Records (EMRs). A set of 150 labeled medical reports was created for testing, and the performance of six different prompts was analyzed for each LLM. All LLMs had no false negatives (sensitivity 100%), with GPT-4 Turbo showing the highest accuracy (97%) with the best prompt.

Background and Significance

The rich information contained in clinical narratives has led to the utilization of various deep learning models, such as LLMs and Natural Language Processing (NLP), for analyzing Electronic Medical Records (EMRs). Carrell et al. (2023) enhanced anaphylaxis identification by employing machine learning (ML) and NLP methodologies, using logistic regression models based on structured claims data, achieving an AUC of 58%. This provided a significant benchmark in computational anaphylaxis identification.

Kural et al. (2023) used ML to analyze claims data from a CMS database to identify anaphylaxis cases. Their method combined unsupervised and supervised learning techniques to identify features indicative of anaphylaxis in claim documents. However, traditional ML models cannot fundamentally understand the underlying narrative or context of the text. In contrast, LLMs can understand human-like text, offering potential for more nuanced and context-aware analyses.

Tu, Han, and Nenadic (2023) evaluated deep learning and LLMs in recognizing clinical domain-named entities and temporal relation extraction, achieving good precision, recall, and F1 scores of 75.67, 77.83, and 78.17, respectively.

Several studies highlight the broader application of deep learning techniques in analyzing clinical texts to identify medical conditions. Pan et al. (2023) focused on cerebrovascular disease, Lin et al. (2023) on PD-L1 expression values, Zitu et al. (2023) on adverse drug events, and Afshar et al. (2023) on opioid misuse. These studies demonstrate the potential of advanced learning structures for tasks such as Named Entity Recognition, temporal relation extraction, disease detection, and adverse drug event detection in clinical narratives, providing valuable insights for our research in anaphylaxis detection using LLMs.

Materials and Methods

To analyze the proficiency of LLMs in identifying potential anaphylaxis diagnoses, we employed an annotated dataset with 150 medical reports in Brazilian Portuguese. Due to privacy concerns, anonymized medical reports were used. This kind of dataset is difficult to find, especially for a relatively rare condition. Therefore, we asked physicians to compile 150 cases from various sources: clinical cases, scholarly articles, and the SemClinBr corpus (Oliveira et al., 2022). The SemClinBr corpus compiles 1000 anonymized clinical narratives in Brazilian Portuguese, spanning various medical fields and institutions, and is a valuable asset for clinical Natural Language Processing (NLP) research.

A trio of medical professionals with expertise in anaphylaxis meticulously vetted all 150 reports in this dataset, classifying each one as referring to or not referring to anaphylaxis. We ended up with 50 positive cases and 100 negative. Although we would have liked more cases, the difficulties of analyzing a large number limited their numbers. Vetting all 150 cases took considerable time. To compensate, the physicians tried to make these cases as representative as possible. There are more negative cases because anaphylaxis is not a common condition. Most importantly, the physicians included several challenging differential cases to test performance in borderline cases, even if they are uncommon.

The cases were organized into four categories:

1. Anonymized Anaphylaxis Incidents (32 documents): These are anaphylaxis diagnoses anonymized from patient records.
2. Scholarly Articles (33 documents): Comprising texts from academic sources, with 18 confirmed anaphylaxis diagnoses and 15 unrelated ones.
3. Differential Diagnostic Cases (35 documents): These documents showcase scenarios where diagnosing is difficult due to symptoms that mimic anaphylaxis. Of these, 10 were sourced from scholarly articles and 25 from modified anonymized patient records, all with diagnoses different from anaphylaxis.
4. SemClinBr Documents (50 documents): Consisting of 50 randomly selected narratives from the SemClinBr corpus, all exceeding 200 characters to ensure adequate informational content. None of these cases involved anaphylaxis.

We opted to employ LLMs due to their novelty and proven proficiency in understanding human-like text. In this study, we selected four distinct LLMs to evaluate: three commercial and one open source. The commercial LLMs were GPT-4 Turbo, GPT-3.5 Turbo, and Gemini 1.0 Pro. GPT-4 Turbo is designed for high-speed responses with increased comprehension, while GPT-3.5 Turbo is a less capable, lower-cost alternative. Gemini 1.0 Pro is the free version of Google's advanced AI model designed for a wide range of tasks. All commercial LLMs run online.

For an open-source local LLM, we chose OpenChat 3.5. It is a language model based on Llama-2 and fine-tuned with C-RLFT (Conditioned - Reinforcement Learning Fine-Tuning), a strategy inspired by offline reinforcement learning, delivering performance comparable with ChatGPT 3.5 even with a 7B model (Wang et al., 2024). We used a small (7B) model due to the high costs involved in running an LLM locally.

Prompts: Development and Operation

The four LLMs were tasked with analyzing the same medical texts, providing recommendations on anaphylaxis detection, identifying probable allergens, and elucidating the reasoning behind their recommendations using uniform prompts. Prompts are textual inputs that guide an LLM's responses, acting as commands that direct the model's output. The design of these prompts is crucial as it sets the context and direction for the model's responses. A well-crafted prompt ensures the model focuses on relevant information. Therefore, our study also evaluates the impact of different prompts on the performance of the four models.

In our study, we employed a structured approach to prompt engineering:

1. Initial Prompt Drafting: Preliminary prompts were crafted to guide the LLMs in analyzing medical texts and determining the presence or absence of indicators suggestive of anaphylaxis.
2. Iterative Refinement: The initial prompts were tested on a subset of the medical texts. Feedback from anaphylaxis experts and the LLMs' responses informed subsequent refinements.

In our final prompt, we provided the LLMs with explicit instructions to provide five items of information:

1. Explain their reasoning process step-by-step, indicating which WAO criteria were used and why.
2. Suggest whether the case indicates anaphylaxis (true or false).
3. Identify a probable allergen, if possible.
4. Provide a probability that the case is anaphylaxis (as a measure of their confidence).
5. Describe the reasons for their recommendation, citing passages from the medical texts.

Item 1 is returned as unconstrained text, while items 2 to 5 are returned as a JSON object. We tested variations of this final prompt to assess the impact of different formulations on the LLMs' performance. We focused on two main aspects:

1. Explicit Information: Including or excluding the WAO criteria for anaphylaxis detection in the prompt.
2. Explanation Levels: Explaining its reasoning may help the model get the correct answers. We used three levels: Full Response (all five information items), Summarized Response (only the JSON object items 2-5), and No Explanation (only items 2-4 in the JSON object).

We used six prompts to test the combination of these two aspects, assessing how different levels of response detail affect the models' ability to identify anaphylaxis. The returned fields were processed and saved as a row in a CSV table (one row per medical text).

A known limitation of LLMs is the occurrence of hallucinations, where the model generates incorrect, misleading, or nonsensical information presented as factual or logical. Our experiments primarily evaluated the LLMs' ability to recognize anaphylaxis cases accurately. If an LLM hallucinates and provides an incorrect answer, it is classified as either a false positive or a false negative, depending on the specific error made.

Experiments Setup

We utilized a Python program to interact with the APIs of the LLMs, providing the prompts to conduct our experiments directly. This approach automated the experimentation process and allowed for better control over the LLM parameters. LLMs are inherently stochastic systems, meaning their responses can vary with each execution. This variance can be managed using parameters such as temperature. A lower temperature setting leads to more deterministic responses by discarding less probable answers, while higher temperatures encourage more diverse and creative responses by selecting less likely, more unexpected answers. In all experiments, we set the temperature parameter to zero to ensure the most precise responses from the LLMs, while using defaults for any other parameters.

To conduct the experiments, the following model versions were used:

1. GPT-4 Turbo: gpt-4-0125-preview model.
2. GPT-3.5 Turbo: gpt-3.5-turbo-0125 model.
3. Gemini 1.0 Pro: gemini-pro model.
4. OpenChat 3.5: openchat_3.5 model.

We created a Google Colab notebook with the Python program to run the experiments for each configuration, comprising 6 prompts and 4 LLMs (24 configurations in total), and an input table (CSV file) with all medical documents (one per line). The program reads this table, assembles the prompt for each line, sends it to the LLM, gets the LLM response, parses the returned JSON string, and records the returned fields. For each medical text (line), it saves the fields in a corresponding column in an output table (CSV file). Finally, it compares the diagnostic suggestions of the LLMs to the physicians' diagnoses and classifies the suggestions as true and false positives and negatives, as shown in Table 1.

To account for possible variations between runs, each of the 24 configurations was run 4 times on different days. Since we used a temperature of 0, these variations were minimal (9 configurations had no differences). Where differences occurred, we averaged them when creating the confusion matrix.

Results

The variation among the four runs of each of the 24 configurations was minimal. Nine configurations had no changes, one had five, and the rest had three or fewer. On average, out of the 150 diagnostic suggestions, 148.58 remained consistent across the 24 configurations, with a coefficient of variation of 1.95%.

From the confusion matrices, we calculated precision, sensitivity (recall), specificity, accuracy, and Cohen's Kappa values for each configuration (Table 2). Cohen's Kappa was used to evaluate the agreement between the program's predictions and medical diagnostic evaluations, adjusting for the chance agreement, which is crucial in imbalanced datasets.

All models consistently identified all true anaphylaxis cases, as indicated by the consistent true positive (TP) count of 50 across all prompts, demonstrating a strong ability to recognize anaphylaxis when present. Additionally, all models had zero false negatives (FNs), indicating they did not miss any anaphylaxis cases. The false positive (FP) and true negative (TN) rates varied, with GPT-4 Turbo having fewer FPs. Gemini 1.0 Pro had slightly lower false positives compared to GPT-3.5 Turbo, except in Prompt 2. OpenChat 3.5 exhibited higher FP rates across all prompts, indicating a tendency to overdiagnose anaphylaxis.

Regarding performance metrics, GPT-4 Turbo had the highest precision across prompts, indicating higher efficiency in identifying true anaphylaxis cases. Gemini 1.0 Pro and GPT-3.5 Turbo had moderate precision, while OpenChat 3.5 had the lowest, reflecting its higher FP rates. All models showed 100% sensitivity, indicating they correctly identified all true anaphylaxis cases.

Specificity was highest for GPT-4 Turbo, indicating better performance in correctly identifying non-anaphylaxis cases. Gemini 1.0 Pro was second but not much better than GPT-3.5 Turbo, while OpenChat 3.5 had the lowest specificity. Accuracy and Kappa followed a similar pattern. For F2-Score, the results were similar, but Gemini was much closer to GPT-3.5 Turbo.

In summary, while all models are highly sensitive in detecting anaphylaxis, their precision and specificity vary, with GPT-4 Turbo outperforming the others. OpenChat had the worst results, as expected, being the smallest model.

Table 1 - Confusion matrix for each LLM prompt configuration.

Prompt	GPT-4 Turbo	GPT-3.5 Turbo	Gemini 1.0 Pro	OpenChat 3.5
1. WAO Criteria Full Response	*TP: 50 FP: 9* *FN: 0 TN: 91*	*TP: 50 FP: 16* *FN: 0 TN: 84*	*TP: 50 FP: 20* *FN: 0 TN: 80*	*TP: 50 FP: 30* *FN: 0 TN: 70*
2. WAO Criteria Summarized	*TP: 50 FP: 12* *FN: 0 TN: 88*	*TP: 50 FP: 22* *FN: 0 TN: 78*	*TP: 50 FP: 16* *FN: 0 TN: 84*	*TP: 50 FP: 37* *FN: 0 TN: 63*
3. WAO Criteria No Explanation	*TP: 50 FP: 12* *FN: 0 TN: 88*	*TP: 50 FP: 24* *FN: 0 TN: 76*	*TP: 50 FP: 20* *FN: 0 TN: 80*	*TP: 50 FP: 40* *FN: 0 TN: 60*
4. No WAO Full Response	*TP: 50 FP: 5* *FN: 0 TN: 95*	*TP: 50 FP: 23* *FN: 0 TN: 77*	*TP: 50 FP: 19* *FN: 0 TN: 81*	*TP: 50 FP: 21* *FN: 0 TN: 79*
5. No WAO Summarized	*TP: 50 FP: 8* *FN: 0 TN: 92*	*TP: 50 FP: 23* *FN: 0 TN: 77*	*TP: 50 FP: 14* *FN: 0 TN: 86*	*TP: 50 FP: 29* *FN: 0 TN: 71*
6. No WAO No Explanation	*TP: 50 FP: 9* *FN: 0 TN: 91*	*TP: 50 FP: 23* *FN: 0 TN: 77*	*TP: 50 FP: 15* *FN: 0 TN: 85*	*TP: 50 FP: 35* *FN: 0 TN: 65*

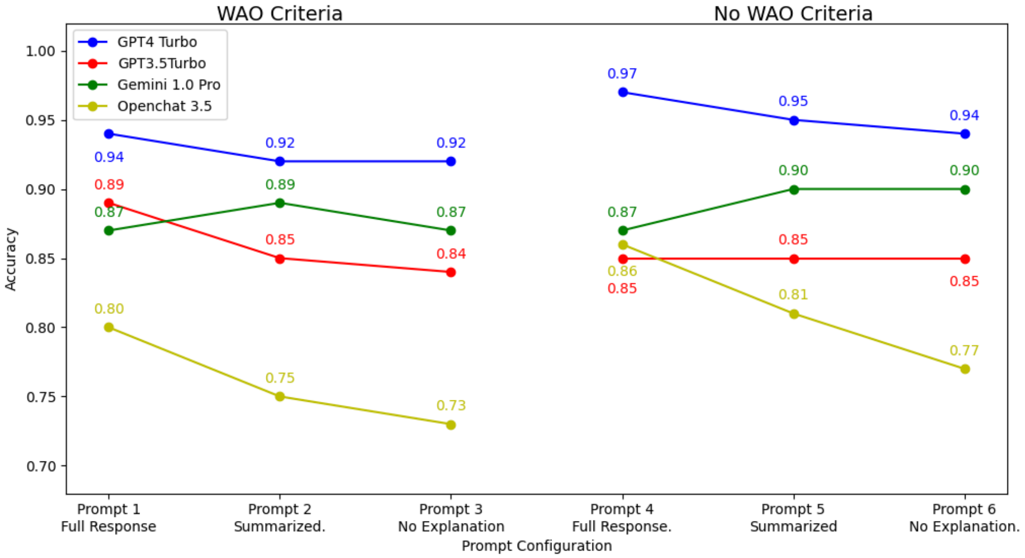
Table 2: Performance metrics across different prompt configurations for GPT-4 Turbo (c1), GPT-3.5 Turbo (c2), Gemini 1.0 Pro (c3), and OpenChat 3.5 (c4). Sensitivity (recall) is 100% for all configurations.

Prompt	Precision %				Specificity %				Accuracy %				Kappa %				F2 %			
	c1	c2	c3	c4	c1	c2	c3	c4	c1	c2	c3	c4	c1	c2	c3	c4	c1	c2	c3	c4
1	85	76	71	60	91	84	80	70	94	89	87	80	87	78	73	61	97	94	93	89
2	81	69	76	50	88	78	84	63	92	85	89	75	83	70	78	53	95	92	94	87
3	81	68	71	56	88	76	80	60	92	84	87	73	83	68	73	50	95	91	93	86
4	91	69	73	70	95	78	81	79	97	85	87	86	93	70	74	71	98	92	93	92
5	86	69	78	63	92	77	86	71	95	85	91	81	88	69	80	62	97	92	95	90
6	85	69	77	59	91	77	85	65	94	85	90	77	87	69	79	55	97	92	94	88

The accuracy analysis (Figure 1) supports our previous discussion on the confusion matrix and performance metrics. In the accuracy graph, GPT-4 Turbo consistently shows high accuracy across all prompts, highlighting its effectiveness in identifying true anaphylaxis cases. Gemini 1.0 Pro and GPT-3.5 Turbo show moderate accuracy, while OpenChat 3.5 has lower accuracy and a tendency towards overdiagnosis. This behavior is consistent with the F2-Score values (Table 2).

These graphs emphasize the importance of prompt design and model selection in achieving high accuracy and F2-Score in anaphylaxis detection. GPT-4 Turbo's robust performance across various prompts suggests it is the most reliable tool.

Figure 1 - Accuracy Evaluation of Models Across Prompts



Discussion

The experiment explored the use of LLMs for the detection of anaphylaxis in Electronic Medical Records. A key finding was the perfect sensitivity across all tested LLMs – all 50 positive cases were correctly identified, ensuring no positive patient was missed. However, precision varied significantly across the models.

All LLMs showed a tendency towards false positives when encountering anaphylaxis-related vocabulary without full-text comprehension, with OpenChat 3.5 being the most prone to this issue (60% precision). GPT-4 Turbo demonstrated superior performance with an average precision of 85% (peaking at 91%), followed by Gemini 1.0 Pro (74%) and GPT-3.5 Turbo (70%). The study highlights that high precision is crucial due to the significantly larger number of negative cases, as excessive false positives can render the results useless.

The "differential cases" designed to be particularly challenging contributed heavily to false positives across all models. For instance, in the case of Prompt 1, 25 out of the 30 false positives from OpenChat 3.5 came from differential cases. All false positives originated from these cases for GPT-4 Turbo, GPT-3.5 Turbo, and Gemini 1.0 Pro. This emphasizes the need for parameter optimization to enhance LLMs' ability to differentiate true positives from challenging negatives, thereby improving overall diagnostic accuracy in real-world medical text analysis.

Prompt Variation and Consistency

We observed that including explicit information, such as the WAO criteria, improved performance only for GPT-3.5 Turbo. However, for GPT-4 Turbo and OpenChat, there was a reduction in performance, while for Gemini, there was no significant difference (Figure 1). Regarding reasoning explanations, it was expected that more detailed explanations would improve performance or at least not reduce it. This held true for all LLMs except Gemini, which performed better with summarized prompts (Figure 1).

Prompt 4: No WAO Criteria and Full Response achieved the best overall performance. Using this prompt, GPT-4 Turbo emerged as the most precise model for this dataset, demonstrating its potential for effective medical text analysis.

Consistency tests conducted on the LLMs revealed varying degrees of stability across different prompts over consecutive days. OpenChat 3.5 and Gemini 1.0 Pro showed the least variation in their results, indicating a high level of consistency. In contrast, GPT-3.5 Turbo exhibited more significant variations, especially with Prompt 4, showing 5 changes (3.3%). The average variation for GPT-4 Turbo was 1.7%, which is a small number. The overall performance of the LLMs suggests a consistent ability to analyze medical texts.

Conclusion

The results demonstrate the potential of LLMs for automated detection of anaphylaxis in medical texts. They highlight the high sensitivity and good accuracy of LLMs in identifying true positive cases. Results showed that prompt design significantly influenced model performance.

While LLMs showed promising results in anaphylaxis detection, further research with bigger medical report sets is needed. Future research should focus on refining prompt design, using real patient data in a hospital setting, and exploring the integration of LLMs with other diagnostic tools to improve patient care and clinical decision-making.

Acknowledgments

This work was supported by the University of São Paulo, Sofya Technology, the Institute of Educational and Research (IEP), and Hospital Sírio-Libanês. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brazil (CAPES) – Finance Code 001 and by grant CEPID 2013/07375-0 and grant C4AI 2019/07665-4 (C4AI), São Paulo Research Foundation (FAPESP).

References

Simons FER, Ardusso LR, Bilò MB, Cardona V, Ebisawa M, El-Gamal YM, et al. International consensus on (ICON) anaphylaxis. World Allergy Organ J. 2014;7(1):9. doi: 10.1186/1939-4551-7-9. PMID: 24920969; PMCID: PMC4038846.

Ensina LF, Min TK, Félix MMR, de Alcântara CT, Costa C. Acute urticaria and anaphylaxis: Differences and similarities in clinical management. Front Allergy. 2022;3:840999. Available from: https://www.frontiersin.org/articles/10.3389/falgy.2022.840999.

Cardona V, Ansotegui IJ, Ebisawa M, El-Gamal Y, Fernandez Rivas M, Fineman S, et al. World Allergy Organization anaphylaxis guidance 2020. World Allergy Organ J. 2020 Oct;13(10):100472. doi: 10.1016/j.waojou.2020.100472. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1939455120303756.

Liu X, Shi Y, Zhang D, Chen M, Xu Y, Zhao J, et al. Management of immune related adverse events through electronic multidisciplinary consultation: Five years of experience from Peking Union Medical College Hospital. J Clin Oncol. 2023;41(16_suppl)
. doi: 10.1200/JCO.2023.41.16_suppl.e14712. Available from: https://doi.org/10.1200/JCO.2023.41.16_suppl.e14712.

Gao Y, Li R, Caskey J, Dligach D, Miller T, Churpek MM, et al. Leveraging a medical knowledge graph into large language models for diagnosis prediction. arXiv [cs]. 2023 Aug. Available from: http://arxiv.org/abs/2308.14321.

Carrell DS, Gruber S, Floyd JS, Bann MA, Cushing-Haugen KL, Johnson RL, et al. Improving methods of identifying anaphylaxis for medical product safety surveillance using natural language processing and machine learning. Am J Epidemiol. 2023 Feb;192(2):283-295. doi: 10.1093/aje/kwac182. Available from: https://academic.oup.com/aje/article/192/2/283/6795959.

Kural KC, Mazo I, Walderhaug M, Santana-Quintero L, Karagiannis K, Thompson EE, et al. Using machine learning to improve anaphylaxis case identification in medical claims data. JAMIA Open. 2023 Oct;6(4). doi: 10.1093/jamiaopen/ooad090. Available from: https://academic.oup.com/jamiaopen/article/doi/10.1093/jamiaopen/ooad090/7331170.

Tu H, Han L, Nenadic G. Extraction of medication and temporal relation from clinical text using neural language models. arXiv [cs]. 2023 Oct. doi: 10.48550/arXiv.2310.02229. Available from: http://arxiv.org/abs/2310.02229.

Pan J, Zhang Z, Peters SR, Vatanpour S, Walker RL, Lee S, et al. Cerebrovascular disease case identification in inpatient electronic medical record data using natural language processing. Brain Inform. 2023 Sep;10(1):22. doi: 10.1186/s40708-023-00203-w. Available from: https://doi.org/10.1186/s40708-023-00203-w.

Lin E, Zwolinski R, Wu JT, La J, Goryachev S, Huhmann L, et al. Machine learning-based natural language processing to extract PD-L1 expression levels from clinical notes. Health Inform J. 2023 Jul;29(3):14604582231198021. doi: 10.1177/14604582231198021. Available from: http://journals.sagepub.com/doi/10.1177/14604582231198021.

Zitu MM, Zhang S, Owen DH, Chiang C, Li L. Generalizability of machine learning methods in detecting adverse drug events from clinical narratives in electronic medical records. Front Pharmacol. 2023 Jul;14:1218679. doi: 10.3389/fphar.2023.1218679. Available from: https://www.frontiersin.org/articles/10.3389/fphar.2023.1218679/full.

Afshar M, Adelaine S, Resnik F, Mundt MP, Long J, Leaf M, et al. Deployment of real-time natural language processing and deep learning clinical decision support in the electronic health record: Pipeline implementation for an opioid misuse screener in hospitalized adults. JMIR Med Inform. 2023 Apr;11. doi: 10.2196/44977. Available from: https://medinform.jmir.org/2023/1/e44977.

Oliveira LES, Peters AC, Silva AMP, Gebeluca CP, Gumiel YB, Cintho LMM, et al. SemClinBr - a multi-institutional and multi-specialty semantically annotated corpus for Portuguese clinical NLP tasks. J Biomed Semantics. 2022 May;13(1):1. doi: 10.1186/s13326-022-00269-1. Available from: https://doi.org/10.1186/s13326-022-00269-1.

Wang G, Cheng S, Zhan X, Li X, Song S, Liu Y. OpenChat: Advancing open-source language models with mixed-quality data. arXiv [cs.CL]. 2024. Available from: https://arxiv.org/abs/2309.11235.