Original Study

Asia Pacific Allergy

OPEN

Artificial intelligence for detecting anaphylaxis in electronic medical records

Luis Felipe Ensina[1],*, Matheus Matos Machado[2], Joice B. Machado Marques[3], Monica Pugliese H. dos Santos[1], Fábio Cerqueira Lario[4], Chayanne Andrade Araújo[1], Fabiana Andrade Nunes Oliveira[1], and Dilvan de Abreu Moreira[2]

ABSTRACT

Background: Despite established criteria, diagnosing anaphylaxis remains challenging but critical for preventing future reactions. Fast-paced clinical settings, compounded by underrecording in electronic medical records (EMRs), increase the risk of dangerous re-exposures. Leveraging artificial intelligence through automated systems such as large language models (LLMs) offers a solution.

Objective: This study aims to assess the efficacy of artificial intelligence, specifically LLMs, in autonomously identifying anaphylaxis diagnoses from EMR text to enhance patient safety and optimize care delivery.

Methods: LLMs (GPT 3.5, 4, and 4 Turbo) analyzed 969 medical texts in Brazilian Portuguese, annotated as anaphylaxis-positive (48) or negative (921) by 3 expert physicians. A primary prompt simulated a general practitioner's role in reviewing medical narratives for anaphylaxis detection, with a secondary prompt incorporating World Allergy Organization (WAO) criteria. The experiments were conducted using 3 GPT configurations. The diagnostic suggestions of the LLM were compared to the physicians' diagnoses. Precision, sensitivity (recall), specificity, and accuracy values were calculated.

Results: Using the primary prompt, GPT 4 Turbo detected anaphylaxis cases with 90.6% precision, 100% sensitivity, 99.5% specificity, 99.5% accuracy, and a Cohen kappa coefficient of 0.95. The inclusion of WAO criteria slightly improved the performance of older models (GPT 3.5 + 4 configuration). However, for GPT 4 Turbo, additional information did not enhance precision.

Conclusion: The results highlight the potential of artificial intelligence, particularly LLMs, to automate anaphylaxis diagnosis, support healthcare professionals, and improve patient safety and care.

Keywords: Anaphylaxis; artificial intelligence; diagnosis; medical records

1. Introduction

Anaphylaxis is a severe and potentially life-threatening systemic hypersensitivity reaction, typically involving multiple body systems, with an estimated lifetime prevalence ranging from 0.3% to 5.1%. Diagnosis relies primarily on clinical assessment, which requires a comprehensive patient history and identification of characteristic physical signs [1]. In 2006, Sampson et al. proposed diagnostic criteria, which were later endorsed by the European Academy of Allergy and Clinical Immunology, the Latin American Society of Allergy, Asthma, and Immunology, and the World Allergy Organization (WAO) [1–4].

However, in fast-paced clinical environments with limited resources, where time is of the essence, manual application of these diagnostic criteria can pose challenges [5]. This challenge is further compounded by the underrecording or inadequate registration of anaphylaxis diagnoses in structured formats within electronic medical records (EMRs) [6, 7]. Failure to accurately document anaphylaxis cases in EMRs could potentially lead to dangerous re-exposure to allergenic triggers and recurrent episodes of anaphylaxis, highlighting a significant gap.

Leveraging automated systems such as large language models (LLMs) presents an opportunity to address this gap by facilitating timely identification and documentation of anaphylaxis, thereby enhancing patient safety and optimizing care delivery [8]. While LLMs, including GPT variants, have demonstrated remarkable capabilities in healthcare tasks, the complexity of EMR texts may present challenges, potentially leading to diagnostic inaccuracies [9].

This study aims to explore and validate the efficacy of LLMs in autonomously identifying and recommending diagnoses of anaphylaxis from the textual data encapsulated within EMRs.

2. Methods

2.1. Data collection

In this study, medical texts served as a foundation for assessing the capabilities of LLMs in recommending anaphylaxis diagnoses. We planned the compilation process to create a representative dataset for evaluating the LLMs. The dataset, composed

[1]Department of Allergy and Clinical Immunology, Hospital Sírio-Libanês, São Paulo, Brazil, [2]Department of Computer Sciences, University of São Paulo, São Carlos, Brazil, [3]Department of Research, Sofya, São Paulo, Brazil, [4]Department of Clinical Informatics, Hospital Sírio-Libanês, São Paulo, Brazil

*Correspondence to Luis Felipe Ensina, Department of Allergy and Clinical Immunology, Hospital Sírio-Libanês, Rua Barata Ribeiro, 490 - 01308-000, São Paulo, Brazil.

Tel: +55-11-99769-0189

Email: 100alergia@gmail.com

Received: 1 July 2024; Accepted: 18 December 2024

Published online 13 January 2025

http://dx.doi.org/10.5415/apallergy.0000000000000179

of 969 medical texts in Brazilian Portuguese, was reviewed by 3 expert physicians in anaphylaxis to determine whether the text information should lead to an anaphylaxis diagnostic indication. Ethical approval was obtained from the Ethics Committee of Hospital Sírio-Libanês, with approval number 76871024.9.0000.5461.

The texts were categorized into distinct groups to facilitate a structured analysis:

(1) Real-life anaphylaxis cases (35 texts): Texts from anonymized medical records known to have a confirmed diagnosis of anaphylaxis.
(2) Case reports (29 texts): Texts published in medical journals (13 reporting anaphylaxis and 16 reporting other nonrelated conditions).
(3) Differential diagnosis cases (35 texts): These are cases particularly challenging to diagnose, as they could present clinically with symptoms similar to anaphylaxis. Ten cases were adapted from the medical literature, and 25 were from anonymized medical records, all with other diagnoses than anaphylaxis.
(4) SemClinBR cases (870 texts): Medical texts from SemClinBR, a corpus of annotated clinical narratives about various clinical conditions [10]. None of them were suspicious of anaphylaxis. We selected texts in the corpus with more than 200 characters to avoid texts with too little information.

For a better understanding, we considered the texts with confirmed anaphylaxis diagnoses as "positive" and texts with other diagnoses as "negative."

We decided to opt for a 5% rate of positive cases to mimic real-world anaphylaxis prevalence estimates, which can reach up to 5% [1]. It aims to ensure that the method maintains a low false positive rate if applied to actual medical texts, thereby preventing users from being overwhelmed with incorrect cases.

2.2. LLM selection and configuration

We opted to employ LLMs due to their novelty and proven proficiency in understanding and generating human-like text based on the input they receive. The LLMs selected for this study were OpenAi's GPT 3.5, 4, and 4 Turbo, chosen due to their originality, widespread use, and demonstrated ability to generate coherent and contextually relevant text. GPT 4 surpasses models specifically fine-tuned on medical knowledge (ie, Med-PaLM) in competency exams [11]. They have been utilized in various applications, including content creation tools, conversational agents, and as a tool to assist in more complex decision-making processes. In our study, we used prompts to assign GPT the task of analyzing medical texts to recommend diagnoses of anaphylaxis, identify likely allergens, and explain the basis for its conclusions using prompts.

Prompt engineering methodologies often involve the iterative development and testing of prompts, ensuring they elicit the desired responses or actions from the LLM [12]. In our study, that involves several steps:

(1) Initial prompt drafting: Preliminary prompts were formulated to encapsulate the criteria. These prompts were structured to guide the LLMs in analyzing medical texts and determining the presence or absence of suggestive indicators of anaphylaxis.
(2) Iterative refinement: The initial prompts were tested on a subset of the medical texts. Based on the LLMs' responses and feedback from the experts in anaphylaxis, the prompts underwent refinement iterations to enhance clarity and specificity.

The finalized prompt was employed to guide the LLMs in analyzing the entire dataset of medical texts. In it, we explicitly instructed the LLM to:

(1) suggest an anaphylaxis diagnosis (true or false);
(2) name a probable allergen;
(3) describe the reasons for their recommendation (citing passages from the medical texts);
(4) provide the probability of anaphylaxis (as a measure of their confidence); and
(5) explain its reasoning process step-by-step, indicating which criteria it used and why.

We also instructed the LLM to return an object with items 1 to 4. The program read the fields returned by the LLM and saved them as a table line. Ultimately, we got a table in a CSV file, with one line evaluating each medical text data.

One known drawback of LLMs is hallucinations. They refer to instances where the model generates incorrect, misleading, or nonsensical information presented as factual or logical [13]. We asked the model to link its recommendation reasons to text passages as a form to "ground" the model and avoid or reduce hallucinations [14]. Our experiments only consider whether or not the LLM recognizes an anaphylaxis case. If the LLM hallucinates and returns a wrong answer, it will be counted as a false positive or negative.

2.3. Experiments setup

We used a Python program to call the LLM's application program interfaces, with the prompts, to run our experiments directly. Besides allowing experiment automation, which is indispensable as we have almost 1,000 texts, it also gives us much better control of the LLM parameters. LLMs are inherently stochastic systems, so their answer can vary from one call to another [15]. To some extent, this variance is controlled using LLM parameters like temperature. The lower the temperature, the more precise the answers, as less probable answers are discarded. High temperatures generate more creative solutions as less likely, more unexpected answers are chosen. In all experiments, we used the temperature equal to zero. All other parameters were left with their default values.

The experiments were conducted using 3 GPT configurations (Fig. 1). Configuration 2 uses the results from configuration 1 and confirms the positive cases using GPT 4. This choice was due to the high price of using GPT 4 in all texts. We created a Google Colab notebook with the Python program to run the experiments for each configuration and an input table (CSV file) with all medical documents (one per line). The prompt's text had a placeholder string to be substituted by the processed medical text. The program reads this table, and, for each line, it assembles the prompt, using the prompt string, sends it to the LLM, gets the LLM response, parses the JSON string returned, and records the 4 fields returned by the LLM. For each medical text (line), it saves the 4 fields in a corresponding column in an output table (CSV file).

After that, we incorporated the prompt text describing the WAO clinical criteria for anaphylaxis and ran the experiments for the 3 configurations [1]. Its results served to gauge the influence of explicitly stating the criteria on the model's performance.

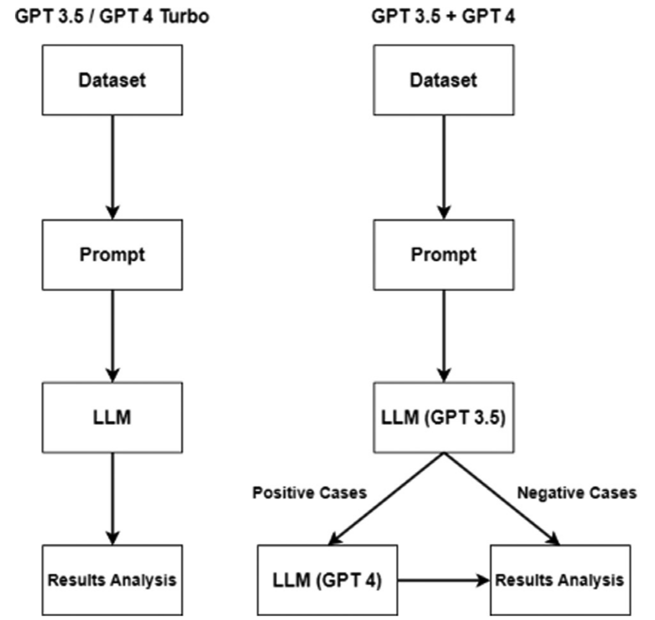

Figure 1. Flowchart depicting the 3 experimental configurations. The left section shows the standalone GPT 3.5 and GPT 4 Turbo configurations. The right section shows the hybrid GPT 3.5 and GPT 4 configuration with case-based branching.

Finally, we copied the 6 output tables to a spreadsheet. The 3 specialists reviewed all texts and classified all cases as anaphylaxis or not before these experiments. We compared the diagnostic suggestions of the LLM to the physicians' diagnoses. Based on this comparison, the spreadsheet classified the suggestions as true and false positives and negatives to calculate precision, sensitivity (recall), specificity, and accuracy values.

In addition to these values, Cohen kappa was used to evaluate the agreement between the program's predictions and medical diagnostic evaluations [16]. Kappa adjusts for the agreement that could happen by chance, which is a crucial factor in imbalanced datasets. A high agreement rate could occur by chance in such datasets if a rater and a predictive model consistently predict the majority class. Unlike accuracy, which can be misleading in such datasets, Kappa factors in the prevalence of each class considered the probability of random agreement. Based on the value of Cohen kappa, the agreement is interpreted as none (<0), slight (0–0.2), fair (0.2–0.4), moderate (0.4–0.6), substantial (0.6–0.8), and almost perfect (0.8–1.0).

3. Results

Table 1 shows the experiment's results. The best results came using configuration 3, the GPT 4 Turbo. The values with and without the WAO criteria were very close, with a slight advantage for the prompt without them with 90.6% precision, 100% sensitivity, 99.5% specificity, 99.5% accuracy, and 0.95 kappa's value (almost perfect agreement). The WAO criteria did improve the older models' performance, as for GPT 3.5 + 4 configuration, precision went from 81.0% to 84.2% and kappa from 0.88 to 0.9, and for the GPT 3.5, configuration had even more improvements, with precision going from 44.9% to 60.8% and kappa from 0.59 to 0.74.

Figure 2 shows how precision improves as the GPT models evolve, and Figure 3 shows the same for Kappa. As the models got smarter, the advantage of adding extra knowledge (the WAO criteria) to the prompt was reduced. It reached the point that GPT 4 Turbo precision did not improve with the additional information.

3.1. Bias

We considered that the 5% positive to 95% negative anaphylaxis cases rate in our 969 texts corpus approximates the real-world distribution. Despite this, we also made an analysis removing the 870 texts from SemClinBR (Table 2). Most of these texts should be easier to classify as they do not discuss anything related to anaphylaxis and could introduce a bias (some presented other allergy cases but not anaphylaxis). In the best configuration results, using GPT 4 Turbo, precision and sensitivity remained the same. Specificity went from 99.5% to 90.2%, and accuracy from 99.5% to 95%, still outstanding results. Kappa went down from 0.95 to 0.90, which still is an almost perfect agreement. The SemClinBR texts had a small effect on the overall results.

3.2. Consistency

Before running the experiments, we tested GPT 3.5 consistency. With this model, we analyzed 927 medical texts on 3 different days using a prompt incorporating the WAO criteria and returning the same fields. We compared the results from each day, revealing the following differences:

Table 1.

Experiments' confusion matrix and performance indicators for each GPT model combination with and without the WAO criteria

GPT model	Confusion matrix	Precision	Sensitivity	Specificity	Accuracy	Kappa agreement
4 Turbo	TP: 48 FP: 5 TN: 916 FN: 0	90.6%	100%	99.5%	99.5%	0.95 almost perfect
4 Turbo W/criteria	TP: 48 FP: 6 TN: 915 FN: 0	88.9%	100%	99.3%	99.4%	0.94 almost perfect
3.5 + 4	TP: 47 FP: 11 TN: 910 FN: 1	81.0%	97.9%	98.8%	98.8%	0.88 almost perfect
3.5 + 4 W/criteria	TP: 48 FP: 9 TN: 912 FN: 0	84.2%	100%	99.0%	99.1%	0.90 almost perfect
3.5	TP: 48 FP: 59 TN: 862 FN: 0	44.9%	100%	93.6%	93.9%	0.59 moderate
3.5 W/criteria	TP: 48 FP: 31 TN: 890 FN: 0	60.8%	100%	96.6%	96.8%	0.74 substantial

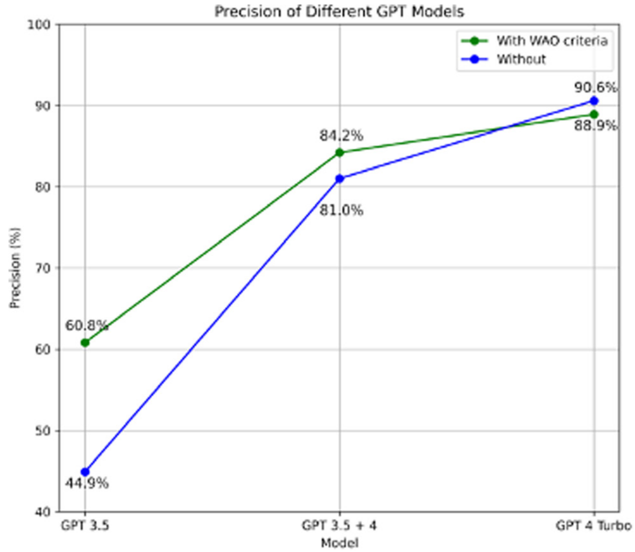

Figure 2. Evolution of precision in GPT models: This graph displays the diminishing returns of including WAO criteria as the models progress from GPT 3.5 to GPT 4 Turbo. The latest model maintains high precision without the extra information. WAO, World Allergy Organization.

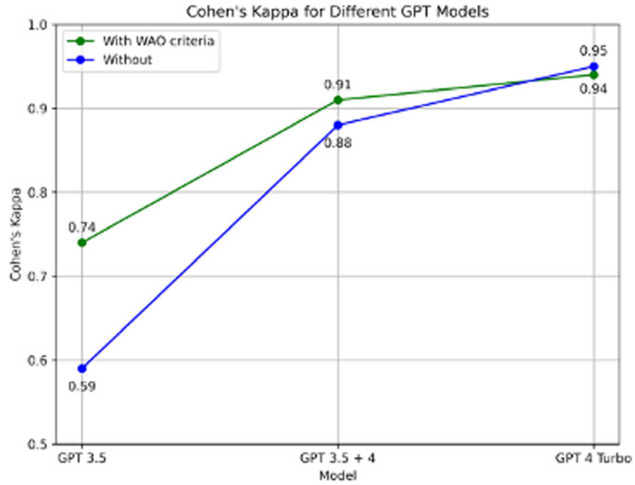

Figure 3. Advances in agreement, using Cohen kappa, for GPT models: This graph shows a consistent upward trend in Cohen kappa values from GPT 3.5 to GPT 4 Turbo, demonstrating the reduced impact of WAO criteria. WAO, World Allergy Organization.

(1) Textual differences: Of the 927 medical texts, 646 cases (70%) showed some textual variation.
(2) Probability variation: In 26 cases (2.8%), the probability values, which indicate the model's confidence in its recommendation, differed.
(3) Diagnoses: Most notably, 6 cases (0.65%) presented different diagnoses on at least 2 different days. This small percentage indicates a high consistency level in the primary diagnosis recommendation task.

We opted to run this test just with GPT 3.5 because of the higher costs of using GPT 4 and the fact that, if the GPT 3.5 variance was acceptable, the newer models' variance should be even better.

After running the experiments, we again tested the consistency of the LLM results. This time, we reran the program on a different date for all configurations. Table 3 shows each run's confusion matrix and changes. The changes were minimal (average of 1.7); consequently, the precision, sensitivity, etc. values changed very little. The changes are primarily in texts where the model is unsure (low probabilities). Most of the differences were minor textual variations in the sentences generated by the models. They did not impact the diagnostic recommendation but underscored the LLM's stochastic behavior.

4. Discussion

The results from our experiments provide a compelling insight into the potential and limitations of utilizing LLMs, specifically GPT models, in the realm of automated diagnosis from medical texts. Several key findings merit further discussion.

The GPT 4 Turbo LLM exhibited impressive performance in diagnosing anaphylaxis, with a precision of 90.6%, a sensitivity of 100%, and a specificity of 99.5%. High sensitivity is crucial here, ensuring the model effectively identifies most cases, thereby minimizing false negatives. On the other hand, high specificity means the model accurately rules out anaphylaxis in those who do not have it, reducing false positives. Too many false positives could lead to diagnostic fatigue and potential oversight of genuine cases. This balance is essential in medical diagnostics, where accurately detecting a condition without overdiagnosing is vital.

These findings align with recent studies highlighting the efficacy of LLMs in medical data extraction [17].

For earlier models (GPT 3.5 and 4), there was a drop in performance metrics when the WAO criteria text was excluded from the prompt. It underscores the importance of effective prompt engineering. This resonates with the broader understanding that while LLMs possess vast knowledge, guiding them with precise instructions can significantly enhance their performance [18].

Carrell et al. [19] research aligns closely with our focus on anaphylaxis identification; they enhanced the identification of anaphylaxis events by employing machine learning (ML) and natural language processing methodologies. Their approach utilized logistic regression models, leveraging structured claims data, and achieved a cross-validated area under the curve of 0.58. This work provides a significant benchmark in anaphylaxis identification through computational means, particularly in utilizing structured data and ML models.

They predominantly relied on structured claims data for their ML models and employed logistic regression models, fundamentally statistical models that predict the probability of a binary outcome. We focus on analyzing unstructured medical texts within EMRs using LLMs, which can potentially harness richer, more detailed patient information that might be absent or not readily accessible in structured data. We achieved a notably higher precision (86%), sensitivity (100%), and specificity (99.13%), potentially indicating a robust predictive capability to identify and recommend anaphylaxis diagnoses accurately.

Kural et al. [20] utilized ML to analyze claims data from a Content Management System database to identify anaphylaxis cases. It uses a combination of unsupervised and supervised learning techniques to identify specific words or features in claim documents indicative of anaphylaxis rather than attempt to understand the text's context or narrative content. Such a method, typical of many traditional ML models, while effective in specific contexts, cannot fundamentally understand or interpret the underlying narrative or context of the text. LLMs can understand human-like text, potentially providing more nuanced and context-aware analyses.

Table 2.

Experiment values without SemClinBr data: confusion matrix and performance indicators for each GPT model combination with and without the WAO criteria

GPT model	Confusion matrix	Precision	Sensitivity	Specificity	Accuracy	Kappa agreement
4 Turbo	TP: 48 FP: 5 TN: 46 FN: 0	90.6%	100%	90.2%	95.0%	0.90 almost perfect
4 Turbo W/criteria	TP: 48 FP: 5 TN: 46 FN: 0	90.6%	100%	90.2%	95.0%	0.90 almost perfect
3.5 + 4	TP: 47 TP: 42 FP: 9 FP: 1	83.9%	97.9%	82.4%	89.9%	0.80 substantial
3.5 + 4 W/criteria	TP: 48 FP: 9 TN: 42 FN: 0	84.2%	100%	82.4%	90.9%	0.82 almost perfect
3.5	TP: 48 TN: 32 FP: 19 FN: 0	71.6%	100%	62.7%	80.8%	0.62 substantial
3.5 W/criteria	TP: 48 FP: 9 TN: 42 FN: 0	71.6%	100%	62.7%	80.8%	0.62 substantial

Table 3.

Results for 2 program runs for 4 model configurations on different days

	Confusion matrix		
GPT model	First run	Second run	Changes
4 Turbo	TP: 48 TN: 916 FP: 5 FN: 0	TP: 48 TN: 917 FP: 4 FN: 0	1
4 Turbo W/criteria	TP: 48 TN: 915 FP: 6 FN: 0	TP: 48 TN: 912 FP: 9 FN: 0	3
3.5 + 4	TP: 47 TN: 910 FP: 11 FN: 1	TP: 48 TN: 909 FP: 12 FN: 0	2
3.5 + 4 W/criteria	TP: 48 TN: 912 FP: 9 FN: 0	TP: 48 TN: 913 FP: 8 FN: 0	1
3.5	TP: 48 TN: 862 FP: 59 FN: 0	TP: 48 TN: 862 FP: 59 FN: 0	0
3.5 W/criteria	TP: 48 TN: 890 FP: 31 FN: 0	TP: 48 TN: 887 FP: 34 FN: 0	3

There were not many changes.

Our LLM-based solution grasps not just the explicit mentions of anaphylaxis (or related words) but also the nuanced context in which these mentions occur. So, they are more flexible and can be used in a more extensive range of medical texts.

Our approach explicitly analyses medical texts and applies specific clinical criteria in Brazilian Portuguese. That may provide a culturally and linguistically relevant tool for anaphylaxis identification in Portuguese-speaking regions.

While our results are promising, there are some limitations to consider. First, the study used a specific set of medical texts in Brazilian Portuguese, which may be limited to particular clinical narratives. Second, the differential cases experts intentionally crafted to be particularly challenging allowed for more challenging experiments but might not reflect the typical anaphylaxis presentations. Furthermore, while the LLMs demonstrated high accuracy, they are not infallible and still require human oversight.

The findings of this study have important implications for clinical practice. The application of LLMs in the automatic recommendation of anaphylaxis diagnoses from medical texts has showcased significant potential, with results indicating high precision, sensitivity, and specificity. The ability of LLMs to accurately identify anaphylaxis in texts can ensure that structured data in EMRs match the information in text format. That can streamline clinical workflows by reducing the manual labor involved in case identification and documentation.

Also, we anticipate contributing to the broader understanding of the role and utility that LLMs can have in clinical decision support and patient safety.

In future research, there is ample scope to refine this methodology for identifying a more comprehensive range of medical conditions and harnessing the capabilities of LLMs for EMR optimization [15]. The amalgamation of LLMs with Clinical Decision Support Systems has the potential to autonomously extract pertinent medical data from textual sources and generate lucid explanations for clinical decisions [21]. This integration can further amplify the efficacy of real-time clinical recommendations, thereby elevating the standard of patient care. Finally, capitalizing on the multilingual proficiency of LLMs can pave the way for more globally accessible and universally applicable healthcare solutions.

Acknowledgments

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).

Conflicts of interest

The authors declare no conflicts of interest.

Author contributions

DAM, FCL, and LFE proposed hypothesis, conceived, and designed the study; FCL, LFE, JBMM, MPHS, FANO, and CAA reviewed clinical notes and created the dataset; DAM and MMM developed the program; DAM, MMM, FCL, MPHS, CAA, FANO, and LFE conducted data analysis and interpreted the results; LFE and DAM drafted the article. All the authors reviewed the text and approved the final version of the manuscript.

References

1. Cardona V, Ansotegui IJ, Ebisawa M, El-Gamal Y, Rivas MF, Fineman S, Geller M, Gonzalez-Estrada A, Greenberger PA, Sanchez Borges M, Senna G, Sheikh A, Tanno LK, Thong BY, Turner PJ, Worm M. World Allergy Organization Anaphylaxis Guidance 2020. World Allergy Organ J 2020;13:100472.
2. Sampson HA, Muñoz-Furlong A, Campbell RL, Adkinson Jr NF, Bock SA, Branum A, Brown SG, Camargo CA Jr, Cydulka R, Galli SJ, Gidudu J, Gruchalla RS, Harlor AD Jr, Hepner DL, Lewis LM, Lieberman PL, Metcalfe DD, O'Connor R, Muraro A, Rudman A, Schmitt C,

Scherrer D, Simons FE, Thomas S, Wood JP, Decker WW. Second symposium on the definition and management of anaphylaxis: summary report—Second National Institute of Allergy and Infectious Disease/Food Allergy and Anaphylaxis Network Symposium. Ann Emerg Med. 2006;47(4):373-80.

3. Muraro A, Worm M, Alviani C, Cardona V, DunnGalvin A, Garvey LH, Riggioni C, de Silva D, Angier E, Arasi S, Bellou A, Beyer K, Bijlhout D, Bilò MB, Bindslev-Jensen C, Brockow K, Fernandez-Rivas M, Halken S, Jensen B, Khaleva E, Michaelis LJ, Oude Elberink HNG, Regent L, Sanchez A, Vlieg-Boerstra BJ, Roberts G; European Academy of Allergy and Clinical Immunology, Food Allergy, Anaphylaxis Guidelines Group. EAACI guidelines: anaphylaxis (2021 update). Allergy 2022;77:357-77.
4. Cardona V, Álvarez-Perea A, Ansotegui-Zubeldia IJ, Arias-Cruz A, Ivancevich JC, Gonzales-Dias SN, Latour-Staffeld P, Sánchez-Borges M, Serrano C, Solé D, Tanno L, Cabañes-Higuero N, Chivato T, De la Hoz B, Fernández-Rivas M, Gangoiti I, Guardia-Martínez P, Herranz-Sanz MÁ, Juliá-Benito JC, Lobera-Labairu T, Praena-Crespo M, Prieto-Romo JI, Sánchez-Salguero C, Sánchez-González JI, Uixera-Marzal S, Vega A, Villarroel P, Jares E. Guía de Actuación en Anafilaxia en Latinoamérica. Galaxia-Latam [Clinical practice guide for anaphylaxis in Latin America (galaxia-latam)]. Rev Alerg Mex 2019; 66(Suppl 2):1-39.
5. Liu X, Shi Y, Zhang D, Chen M, Xu Y, Zhao J, Zhong W, Wang M. 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):e14712.
6. de Sordi D, Kappen S, Otto-Sobotka F, Kulschewski A, Weyland A, Gutierrez L, Fortuny J, Reinold J, Schink T, Timmer A. Validity of hospital ICD-10-GM codes to identify anaphylaxis. Pharmacoepidemiol Drug Saf 2021;30:1643-52.
7. Mukherjee M, Wyatt JC, Simpson CR, Sheikh A. Usage of allergy codes in primary care electronic health records: a national evaluation in Scotland. Allergy 2016;71:1594-602.
8. Jimenez-Rodriguez TW, Garcia-Neuer M, Alenazy LA, Castells M. Anaphylaxis in the 21st century: phenotypes, endotypes, and biomarkers. J Asthma Allergy 2018;11:121-42.
9. Gao Y, Li R, Caskey J, Diglach D, Miller T, Churpek MM, et al. Leveraging a medical knowledge graph into large language models for diagnosis prediction. arXiv 2023;2308.14321.
10. Oliveira LESE, Peters AC, da Silva AMP, Gebeluca CP, Gumiel YB, Cintho LMM, Carvalho DR, Al Hasan S, Moro CMC. SemClinBr - a multi-institutional and multi-specialty semantically annotated corpus for Portuguese clinical NLP tasks. J Biomed Semantics 2022; 13:13.
11. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. arXiv 2023;2303.13375v2.
12. Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. J Med Internet Res 2023;25:e50638.
13. Brameier DT, Alnasser AA, Carnino JM, Bhashyam AR, von Keudell AG, Weaver MJ. Artificial intelligence in orthopaedic surgery: can a large language model "write" a believable orthopaedic journal article? J Bone Joint Surg Am 2023;105:1388-92.
14. Wornow M, Xu Y, Thapa R, Patel B, Steinberg E, Fleming S, Pfeffer MA, Fries J, Shah NH. The shaky foundations of large language models and foundation models for electronic health records. NPJ Digit Med 2023;6:135.
15. Nashwan AJ, AbuJaber AA. Harnessing the power of large language models (LLMs) for electronic health records (EHRs) optimization. Cureus 2023;15:e42634.
16. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22:276-82.
17. Agrawal M, Hegselmann S, Lang H, Kim Y, Sontag D. Large language models are few-shot clinical information extractors. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. https://aclanthology.org/2022.emnlp-main.130.pdf. Accessed June 10, 2024.
18. Clavié B, Ciceu A, Naylor F, Soulié G, Brightwell T. Large language models in the workplace: a case study on prompt engineering for job type classification. In: Métais E, Meziane F, Sugumaran V, Manning W, Reiff-Marganiec S, eds. Natural Language Processing and Information Systems. Cham: Springer Nature; 2023.
19. Carrell DS, Gruber S, Floyd JS, Bann MA, Cushing-Haugen KL, Johnson RL, Graham V, Cronkite DJ, Hazlehurst BL, Felcher AH, Bejan CA, Kennedy A, Shinde MU, Karami S, Ma Y, Stojanovic D, Zhao Y, Ball R, Nelson JC. Improving methods of identifying anaphylaxis for medical product safety surveillance using natural language processing and machine learning. Am J Epidemiol 2023;192:283-95.
20. Kural KC, Mazo I, Walderhaug M, Santana-Quintero L, Karagiannis K, Thompson EE, Kelman JA, Goud R. Using machine learning to improve anaphylaxis case identification in medical claims data. JAMIA Open 2023;6:ooad090.
21. Yang R, Tan TF, Lu W, Thirunavukarasu AJ, Ting DSW, Liu N. Large language models in health care: development, applications, and challenges. Health Care Sci 2023;2:255-63.