THERAPEUTIC ADVANCES in
Musculoskeletal Disease

Review

RAGing ahead in rheumatology: new language model architectures to tame artificial intelligence

Ther Adv Musculoskelet Dis
2025, Vol. 17: 1–10
DOI: 10.1177/1759720X251331529

Diego Benavent, Vincenzo Venerito and Xabier Michelena

***Abstract*:** Artificial intelligence (AI) is increasingly transforming rheumatology with research on disease detection, monitoring, and outcome prediction through the analysis of large datasets. The advent of generative models and large language models (LLMs) has expanded AI's capabilities, particularly in natural language processing (NLP) tasks such as question-answering and medical literature synthesis. While NLP has shown promise in identifying rheumatic diseases from electronic health records with high accuracy, LLMs face significant challenges, including hallucinations and a lack of domain-specific knowledge, which limit their reliability in specialized medical fields like rheumatology. Retrieval-augmented generation (RAG) emerges as a solution to these limitations by integrating LLMs with real-time access to external, domain-specific databases. RAG enhances the accuracy and relevance of AI-generated responses by retrieving pertinent information during the generation process, reducing hallucinations, and improving the trustworthiness of AI applications. This architecture allows for precise, context-aware outputs and can handle unstructured data effectively. Despite its success in other industries, the application of RAG in medicine, and specifically in rheumatology, remains underexplored. Potential applications in rheumatology include retrieving up-to-date clinical guidelines, summarizing complex patient histories from unstructured data, aiding in patient identification for clinical trials, enhancing pharmacovigilance efforts, and supporting personalized patient education. RAG also offers advantages in data privacy by enabling local data handling and reducing reliance on large, general-purpose models. Future directions involve integrating RAG with fine-tuned, smaller LLMs and exploring multimodal models that can process diverse data types. Challenges such as infrastructure costs, data privacy concerns, and the need for specialized evaluation metrics must be addressed. Nevertheless, RAG presents a promising opportunity to improve AI applications in rheumatology, offering a more precise, accountable, and sustainable approach to integrating advanced language models into clinical practice and research.

Keywords: clinical decision support, large language models, retrieval-augmented generation, rheumatology

Received: 22 October 2024; revised manuscript accepted: 13 March 2025.

Correspondence to:
Xabier Michelena
Catalan Health Service, Government of Catalonia, Gran Via de les Corts Catalanes, 587, Barcelona 08007, Spain
Digitalization for the Sustainability of the Healthcare System (DS3), Barcelona, Spain
Rheumatology Research Group, Vall d'Hebron Research Institute, Barcelona, Spain
xmichelena@gencat.cat

Diego Benavent
Department of Rheumatology, Hospital Universitari de Bellvitge, Barcelona, Spain

Vincenzo Venerito
Department of Precision and Regenerative Medicine and Ionian Area, Polyclinic Hospital, University of Bari, Bari, Italy

Introduction to large language models

Artificial intelligence (AI) is reshaping healthcare research and clinical practice by adding capabilities such as improvements in pattern recognition and decision-making support.[1] Particularly in complex fields like rheumatology, AI can aid in diagnosis, disease monitoring, and outcome prediction by analyzing large data sets and developing predictive models.[2] Besides the emergence of machine learning models, the recent emergence of generative models and their application to text has further expanded the use cases of these technologies. Generative AI offers potential in research by supporting literature synthesis, generation of ideas, or medical writing.[3]

Natural language processing (NLP), a subfield of AI, focuses on enabling machines to interpret and generate human language. In medicine, NLP can be helpful in the analysis of clinical notes or research literature, being particularly useful for assisting in documentation management and analyzing clinical notes.[4] NLP has evolved with the emergence of large language models (LLMs), which use enormous datasets to learn and produce text replies that resemble those of a human.[5,6] LLMs are sophisticated neural networks based on the so-called transformer architectures, trained to predict the next word in a sequence based on the surrounding context. These models can learn language patterns without labeled datasets since they have been pre-trained on massive volumes of text data through self-supervised learning. These models have become well known in the medical field due to their proficiency with difficult language tasks, including question-answering and medical literature synthesis.[6] Beyond conventional text-based jobs in healthcare, LLMs may also handle multimodal data, incorporating images and audio for analysis.[7]

NLP may be used to support research and clinical practice in rheumatology, as seen in recent studies. For instance, rheumatoid arthritis (RA) patients have been identified from electronic health records (EHR) using NLP in conjunction with machine learning models such as support vector machines and gradient boosting, yielding impressive accuracy (AUROC 0.98).[8] Moreover, NLP has demonstrated the potential to identify information about the patient's background and the features of their disease in patients with RA and interstitial lung disease.[9] NLP-based techniques have also been used to identify axial spondyloarthritis—with a sensitivity of 0.78 and specificity of 0.94[10]—and ANCA-associated vasculitis with a positive predictive value of 86.1%, outperforming structured ICD-10-coding.[11] While there is increasing evidence to support the use of NLP for precise identification of disease and their characteristics, the potential of this technology goes far beyond this.

With the emergence of LLMs, promising research on their rheumatology-based knowledge has been published. In a recent study, ChatGPT-4 accurately determined the most likely diagnosis in 35% of cases, closely matching the 39% ($p = 0.30$) of rheumatologists; interestingly, the model performed better in inflammatory rheumatic illness situations (71% accuracy compared to 62% accuracy for rheumatologists).[12] ChatGPT-4 was tested in the ChatSLE study against top rheumatology specialists, answering 100 patient-related queries from a website, with a mean quality score of 4.55 (95% confidence interval (CI) 4.48–4.62) for ChatGPT-4 responses compared to 4.31 (95% CI 4.23–4.39) for expert responses ($p < 0.0001$).[13] In an analysis of the Spanish Medical Training Exam (MIR), including rheumatology-related questions from 2009 to 2023, ChatGPT-4 outperformed previous LLM versions with a median clinical reasoning score of 4.67 on a 5-point Likert scale, answering 93.7% of all rheumatology-related questions correctly.[14] Yet, not all findings are equally favorable. As an example, in a cross-sectional study, although patients rated AI-generated responses comparable to physician-generated ones in comprehensiveness and readability, rheumatologists deemed these AI responses significantly less accurate.[15] This discrepancy highlights that while LLMs may be well received by patients, concerns persist among specialists, reinforcing the need for ongoing validation and careful integration of these models into rheumatology practice.

Even with their considerable potential, LLMs also present significant challenges. Despite their impressive performance, it is of note that generative models were not designed for specific use in medicine; rather, they are general-purpose models.[16,17] The lack of domain-specific knowledge in generative models is a notable limitation, particularly when dealing with highly specialized fields like rheumatology.[18] Traditional generative models, such as GPT or other language models, are trained on large-scale general datasets like Wikipedia, or publicly available books. While this provides a broad understanding of language and general knowledge, it does not equip the model with the depth of expertise required for accurate responses in niche domains.[19] As an example, in rheumatology, models trained on general data may struggle with precision in assessing immunology tests or specific disease recommendations, given their input from heterogeneous and potential non-professional sources.

Another significant problem of LLMs is the occurrence of hallucinations, which are instances in which the model produces false or inaccurate data.[16] For example, an LLM might incorrectly state that biologic therapy is the first-line treatment for early RA, while a conventional synthetic disease-modifying drug should be recommended instead. In another case, the model might inaccurately assert that antinuclear antibody testing is a reliable screening tool for all patients with arthralgia to discard suspected lupus, not considering its low specificity. These hallucinations can arise unpredictably, even when identical prompts are used, and can be the result of the model being influenced by irrelevant data.[20] The randomness of these errors is problematic in medical contexts, where clinical decisions rely on precise, evidence-based information. Addressing hallucinations is relevant to ensure that LLMs are trustworthy, reducing the risk of misinformation in patient care. Techniques to improve hallucinations can help users rely on responses while ensuring factual accuracy for clinical decision support.

Retrieval-augmented generation

Retrieval-augmented generation (RAG) has been introduced to address these problems. RAG models incorporate pertinent external information from external databases in real-time, improving LLMs and allowing models to carry out more precise in-context learning.[21] These models aim to predict the output based on the source input from a corpus, while other documents are still accessible. RAG combines the internal knowledge of LLMs with large, dynamic databases from outside sources in a synergistic way. RAG improves the language generation's credibility and accuracy, especially for activities requiring vast knowledge.[22] It also makes it possible to integrate domain-specific data and refresh knowledge continuously.

Indeed, RAG has been successfully applied across various fields outside of medicine, improving the capabilities of LLMs by integrating external knowledge sources.[22] RAG models are employed in open-domain question-answering systems in different fields. In customer service, chatbots utilize RAG to provide real-time, personalized assistance by accessing updated product information and user data; the legal sector benefits from RAG through enhanced document analysis and legal research, where models retrieve pertinent statutes and case law to support legal reasoning; in finance, RAG aids in summarizing financial reports and retrieving market trends to support investment decisions.[22] Despite its success in these areas, and the promise that RAG could be particularly useful in the healthcare industry, as clinical decision-making and mistake reduction depend on having access to continuously updated information, it remains remarkably scarce in medicine. Thus, only three studies were listed with the search "Retrieval-Augmented Generation" in PubMed over the year 2023, while there were almost 60 studies by the end of September 2024.

Exploring RAG's applications in medicine could improve domain-specific tasks, by providing up-to-date, evidence-based information. Recently, researchers evaluated GPT-4's ability to interpret guidelines from the American Society of Clinical Oncology and the European Society for Medical Oncology. They assessed GPT-4's performance in answering clinically relevant questions about the management of different types of cancer, comparing outputs with and without RAG. Eighty-four percent of the GPT-4 with RAG replies were correct, while just 57% of GPT-4 responses without RAG were accurate.[18] Novel language model-based systems such as PaperQA2 can support literature review; this model demonstrates superior performance in tasks like question answering, summarization, and contradiction detection compared to human experts, surpassing conventional literature search methods.[23]

Deep dive into RAG

The foundation of RAG lies in the ability of computers to process natural language, enabling them to search knowledge sources and retrieve relevant information. RAG typically involves five key components, including text embeddings, vector databases, information retrieval, generating responses, and prompt engineering. A summary of these components, adapted for their application in rheumatology, is explained below.

Text embeddings

Text embeddings are used to convert words, sentences, and paragraphs into numerical form. These are dense vector representations of words, sentences, or entire documents that capture semantic meaning.[24] With the emergence of LLMs, specialized embedding models have significantly enhanced the performance of these embeddings. Furthermore, domain-specific models like BioClinical_BERT[25] have greatly improved adaptability within the medical field. RAG uses these embeddings to rank documents according to similarity. Although a detailed explanation of the mathematics is beyond the scope of this review, a simple example can help clarify the concept. As shown in Figure 1, in a rheumatology

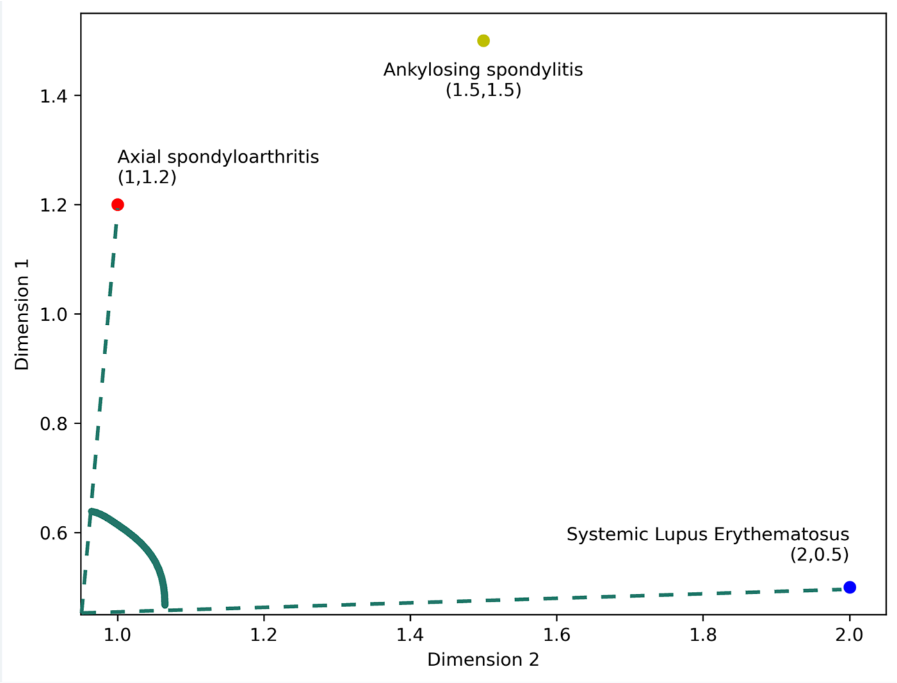

Figure 1. Vector representations of Rheumatology Keywords and similarity search. For simplicity and ease of understanding, the vector representations of rheumatological disease entities have been reduced to two dimensions. As shown in the image, the angle between "Axial spondyloarthritis" and "Systemic Lupus Erythematosus" is larger than the angle between "Axial spondyloarthritis" and "Ankylosing spondylitis," illustrating the cosine similarity technique employed during the retrieval stage.

context, embeddings for terms like "axial spondyloarthritis," "ankylosing spondylitis," and "systemic lupus erythematosus" are represented in a vector space. Since "axial spondyloarthritis" and "ankylosing spondylitis" are more closely related, the angle between their vectors is smaller, illustrating a common similarity calculation technique called cosine similarity.[26]

Vector databases

Once embeddings are generated, they need to be stored efficiently for quick retrieval. This is where vector databases come into play. Vector databases are specialized systems designed to handle high-dimensional vector data, making them ideal for storing embeddings. Vector databases can quickly return the most relevant results by organizing data in a way that prioritizes vector distance—as explained in the previous example—even in complex multidimensional spaces.[27]

In a rheumatology application, a vector database might store embeddings of clinical guidelines, research papers, and anonymized patient records. When a query is made, the database can quickly return the most relevant documents based on the similarity of their embeddings to the query embedding.

Information retrieval

The retrieval component of RAG acts as an intelligent search engine, sifting through the vector database to find the most pertinent information. This process involves three steps: (i) Query translation, in which the system converts the input query (e.g., a patient's symptoms) into the same mathematical language (vector space) used for storing documents; (ii) similarity matching, that uses mathematical techniques like cosine similarity, the system identifies documents that are conceptually close to the query; (iii) relevance ranking, where the system then orders these matches, selecting the most relevant ones for further use. This could mean automatically finding the most applicable clinical guidelines or similar patient cases based on a given patient's specific symptoms and laboratory results.

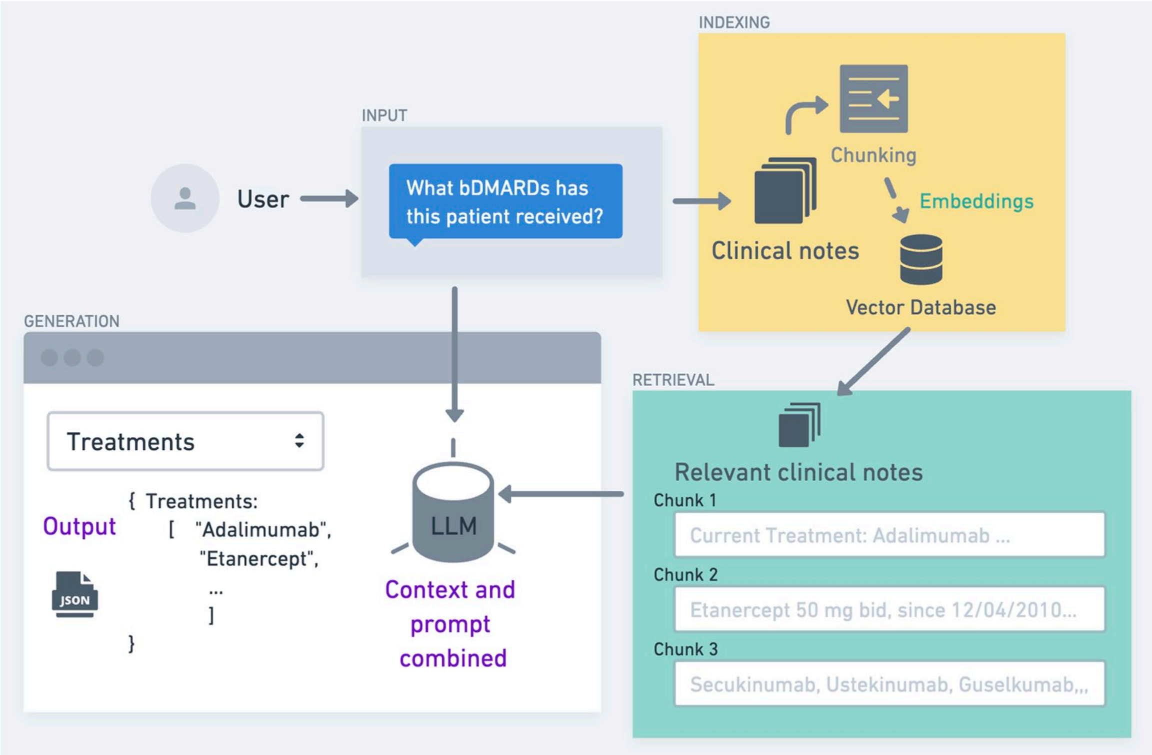

Figure 2. Outline of a RAG ecosystem to extract knowledge from clinical notes. This figure illustrates a system for extracting bDMARDs from clinical notes. In the indexing phase, clinical notes are broken into smaller chunks, transformed into embeddings, and stored in a vector database. During retrieval, the system uses similarity-based methods to find the most relevant notes in response to user input (e.g., "What bDMARDs has this patient received?"). The retrieved notes are combined with the user query and passed to an LLM to generate a structured JSON listing the treatments, which can be utilized in further applications.
bDMARDs, biologic disease-modifying anti-rheumatic drugs; JSON, JavaScript Object Notation; LLM, large language models; RAG, retrieval-augmented generation.

Generating responses

The generation component is typically a LLM core like GPT. This model takes the retrieved information along with the original query and generates a coherent response. The key here is that the LLM is "augmented" with the retrieved information, allowing it to produce more accurate and contextually relevant outputs.[22] By ranking and retrieving documents based on their similarity to the input query, these passages provide a rich, context-driven foundation for the generative model to construct coherent responses.

Prompt engineering

An often underappreciated yet vital component of RAG is prompt engineering, a process that significantly influences the system's output quality. This technique involves crafting the input to the LLM to optimize the relevance and accuracy of its generated response.[28] A well-constructed prompt typically encompasses three key elements: the original query, pertinent context derived from retrieved documents, and explicit instructions guiding the utilization of this retrieved information. In essence, prompt engineering serves as the interface between the retrieval mechanism and the generative model, ensuring that the wealth of retrieved information is effectively channeled into producing a coherent and relevant output. In the context of rheumatology, a thoughtfully engineered prompt could, for instance, guide the LLM to synthesize information from recent clinical guidelines, relevant case studies, and the specific patient data at hand, thereby facilitating the generation of tailored, evidence-based treatment recommendations. In Figure 2, we depict a RAG architecture to retrieve information from the unstructured data from the EHR in a rheumatology setting.

Opportunities, challenges, and evaluation

RAG architecture offers several notable advantages in the context of biomedical sciences and specifically in rheumatology. One of the primary benefits is a significant reduction in hallucinations, as it limits the LLM's access to external knowledge, retrieving precise information as demonstrated in studies extracting clinical data from EHR.[29] However, hallucinations can still occur, especially if the retrieval process fails to provide the correct context, which may lead to incorrect outputs during the generation phase. To address this, the Chain of Verification (CoVE) framework has been proposed as a mitigation strategy, showing a decrease in hallucinations in different tasks.[30] CoVE introduces a multi-step validation process for all retrieved information. In this approach, each referenced source is checked against an established, domain-specific database or a vetted repository. If discrepancies are detected—such as conflicting claims or unclear provenance—the system flags them for further review. CoVE bolsters the factual accuracy of the generated content and creates a structured audit trail by instituting this additional layer of verification, enabling clinicians and researchers to pinpoint the source of each piece of information. Another notable advantage in the medical field is the ability of RAG to pinpoint the exact source of information, thus removing the "black box" concern associated with LLMs, where the origins of the training data are often unclear. Moreover, the use of less potent or smaller LLM models in RAG architecture can be equally effective in reducing the costs of its use. In this framework, the primary knowledge source is the retrieval process, meaning that the model itself does not need to be as large or computationally intensive as standalone LLMs. These models can still provide accurate and contextually relevant responses, leveraging retrieval to complement their reasoning, processing, and generative capabilities. Research has demonstrated that smaller models like DistilBERT or MiniLM can perform comparably when augmented with retrieval mechanisms, leading to efficiency gains without sacrificing accuracy.[31] This flexibility allows for more efficient and scalable implementations, particularly valuable in settings where computational resources are limited.

The RAG proposal also offers the advantage of generating output in a pre-specified structured format. As illustrated in Figure 2, the output can be produced in JSON (JavaScript Object Notation), a simple, easy-to-read format that allows computers to exchange information and can be further processed by an app or the EHR. Finally, a significant advantage of RAG in a healthcare setting is the ability to handle information locally, thereby enhancing data privacy and security. RAG architecture allows the retrieval process to be conducted on private or internal databases, rather than relying on publicly trained models that may inadvertently expose sensitive data. Since the external knowledge is retrieved from secure, controlled repositories, the model's reliance on its training data is reduced, minimizing the risk of leakage of confidential information. This localized handling of data ensures that the sensitive content never leaves the protected environment. Furthermore, the use of smaller models as mentioned previously with RAG makes it feasible to deploy in secure, on-premise infrastructures, further safeguarding sensitive data. Studies have shown that local deployment of RAG for sensitive domains provides not only privacy benefits but also maintains high levels of accuracy by accessing up-to-date and domain-specific information.[22] One of the key challenges with RAG and LLM systems is the high cost of infrastructure, whether hosted on the cloud or on-premise. Cloud-based solutions incur ongoing expenses for high-performance computing resources, especially for real-time inference and retrieval, with costs scaling based on traffic and storage demands. On-premise deployments, while offering more control and potentially lower long-term costs, require significant upfront investment in servers, storage, and maintenance, which can be prohibitive for smaller organizations. Moreover, the energy consumption associated with large-scale deployments raises concerns about sustainability and environmental impact, making resource efficiency not only a financial but also an ecological priority.[32]

Nonetheless, there are some critical aspects for implementing RAG in rheumatology. One of them involves compliance with intellectual property rights and licensing agreements. Since RAG-based solutions often rely on chosen external data sources, ensuring that any clinical guidelines, journal articles, or educational materials are licensed to be used is essential to avoid infringing on copyrighted works. Local deployment of RAG systems can help mitigate certain legal and ethical risks by restricting access to in-house repositories or publicly licensed content; however, institutions should still institute regular audits and maintain robust version control of these databases to ensure that shared information respects all applicable copyright and licensing regulations.

Verification and evaluations for RAG models are limited in the literature. It is relevant to verify the reliability of external datasets by implementing a systematic approach—comprising peer-reviewed curation, version-controlled updates, and integration of reputable, domain-specific repositories. These measures help to guarantee that retrieved information remains accurate and up to date. Regular audits and defined protocols for data integration further reduce the risk of propagating outdated or erroneous information through the RAG pipeline. Concerning validation, if the aim of the RAG architecture is a specific downstream task such as question answering or fact-checking in which a gold standard is available, established metrics such as accuracy, and F1-score can be used as well as BLEU and ROUGE metrics to evaluate answer quality.[33] When it comes to more complex tasks, evaluation is a challenge that is mainly centered on retrieval quality and generation quality. Evaluation shifts to both retrieval effectiveness and generative fidelity, incorporating metrics like perplexity, BERTScore, and mean reciprocal rank to capture model confidence, semantic alignment, and ranking of relevant sources.[34] Real-world usability testing and human-in-the-loop assessments can help capture more qualitative parameters such as correctness, faithfulness, and contextual relevance, as illustrated in a recent paper extracting data from Oncology Guidelines.[18]

Future directions and rheumatology applications

To our knowledge, there is no report in the literature on the use of RAG specifically in Rheumatology. Beyond using RAG for retrieving rheumatology guidelines, especially where local guidelines are preferred, other potential applications should be considered. Clinical notes for rheumatology patients are often complex and typically stored in an unstructured format in most EHR systems. RAG could assist in efficiently searching for relevant patient information, such as identifying bDMARDs the patient has been exposed to or locating immunology test results, while always citing the original source. In this sense, RAG could help find patients with certain characteristics to meet the inclusion criteria of a clinical trial. In addition, RAG could help summarize the most important points from the unstructured data within the EHR, giving insights into longitudinal patient history. In pharmacovigilance, RAG systems could continuously monitor and analyze diverse data sources, including scientific literature, clinical trial reports, and spontaneous adverse event databases, to swiftly identify emerging safety signals or drug interactions relevant to rheumatic diseases. Furthermore, RAG could play a pivotal role in advancing precision medicine by integrating and analyzing diverse data types, including genetic profiles, biomarkers, clinical manifestations, and treatment responses. Regarding patient education, RAG could generate personalized materials by synthesizing information from authoritative sources, and tailoring content to individual patients' diagnoses, treatment regimens, and health literacy levels, thereby potentially improving treatment adherence and patient empowerment. Besides, RAG could facilitate knowledge sharing by efficiently retrieving and translating relevant studies across languages, potentially accelerating global research efforts in rheumatology. Chatbots may employ RAG to leverage advanced retrieval and summarization methods to further optimize patient triage, quickly identify potential adverse drug reactions from large-scale pharmacovigilance registries, or even integrate imaging results with genetic data for robust diagnostic support. While these applications show great potential, it is crucial to validate their clinical utility and address potential challenges, such as data privacy concerns and the need for continual updating of the knowledge base, to ensure their effective integration into rheumatology practice and research.

Several advancements in the field of NLP are being applied to the RAG ecosystem that will enhance its future use. Analyzing the various strategies is beyond the scope of this manuscript, and other studies, such as Gao et al., have dedicated entire lines of research to this topic. One strategy that is worth noting is modular RAG, an approach where different components are used to retrieve relevant information from various sources, which is then processed and combined to generate accurate and context-aware answers or summaries for complex queries.[21] In addition, knowledge graphs are gaining popularity because they improve RAG by organizing and linking complex information in a structured, interconnected way, allowing the system to retrieve more accurate, contextually relevant data.[35]

A promising trend in enhancing RAG applications in medicine is the integration of hybrid approaches, combining RAG with fine-tuning.[36] Fine-tuning an LLM involves further training on a specific dataset to refine its broad knowledge for improved performance in specialized tasks or domains. For instance, a small language model—a scaled-down version designed for efficient, task-specific performance with reduced computational resources—can be fine-tuned with highly specialized rheumatology knowledge.[37] When combined with the relevant context provided by RAG, this approach enables highly accurate and efficient responses, all while maintaining low computational demands.

The advent of multimodal models, which integrate and process data from diverse sources such as text, images, and structured information, will surely revolutionize our field.[38] These models are especially impactful in our domain, where we rely on a variety of data types, including clinical notes, lab reports, medical images, and genetic information. When used in an RAG setting, these models may significantly enhance the depth of context and the accuracy of generated responses.

Conclusion

NLP-based technologies have already shown numerous capabilities to improve rheumatology research and clinical practice. The RAG architecture provides an opportunity to adopt AI and language models in a more precise, accountable, and sustainable manner. Both patients and healthcare providers in rheumatology stand to benefit from the integration of controlled knowledge retrieval and generative capabilities, potentially leading to a more efficient, up-to-date, and transparent approach to the use of AI in medicine.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Author contributions

Diego Benavent: Conceptualization; Methodology; Validation; Visualization; Writing – original draft; Writing – review & editing.

Vincenzo Venerito: Conceptualization; Methodology; Supervision; Validation; Writing – original draft; Writing – review & editing.

Xabier Michelena: Conceptualization; Supervision; Validation; Writing – original draft; Writing – review & editing.

Acknowledgements

None.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Competing interests

Dr D.B. received grants/speaker/research support from Abbvie, Lilly, Novartis, Pfizer, and UCB. He works as a part-time advisor at Savana, a company focused on natural language processing in medicine. Dr V.V. and Dr X.M. declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Availability of data and materials

Not applicable.

ORCID iDs

Diego Benavent https://orcid.org/0000-0001-9119-5330
Xabier Michelena https://orcid.org/0000-0002-5352-919X

References

1. Yu K-H, Beam AL and Kohane IS. Artificial intelligence in healthcare. *Nat Biomed Eng* 2018; 2(10): 719–731.
2. Kothari S, Gionfrida L, Bharath AA, et al. Artificial intelligence (AI) and rheumatology: a potential partnership. *Rheumatology (Oxford)* 2019; 58(11): 1894–1895.
3. Venerito V, Bilgin E, Iannone F, et al. AI am a rheumatologist: a practical primer to large language models for rheumatologists. *Rheumatology (Oxford)* 2023; 62(10): 3256–3260.
4. Chary M, Parikh S, Manini AF, et al. A review of natural language processing in medical education. *West J Emerg Med* 2019; 20(1): 78–86.
5. Demner-Fushman D, Chapman WW and McDonald CJ. What can natural language processing do for clinical decision support? *J Biomed Inform* 2009; 42(5): 760–772.

6. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature* 2023; 620(7972): 172–180.

7. Venerito V, Puttaswamy D, Iannone F, et al. Large language models and rheumatology: a comparative evaluation. *Lancet Rheumatol* 2023; 5(10): e574–e578.

8. Maarseveen TD, Meinderink T, Reinders MJT, et al. Machine learning electronic health record identification of patients with rheumatoid arthritis: algorithm pipeline development and validation study. *JMIR Med Inform* 2020; 8(11): e23930.

9. Román Ivorra JA, Trallero-Araguas E, Lopez Lasanta M, et al. Prevalence and clinical characteristics of patients with rheumatoid arthritis with interstitial lung disease using unstructured healthcare data and machine learning. *RMD Open* 2024; 10(1): e003353.

10. Zhao SS, Hong C, Cai T, et al. Incorporating natural language processing to improve classification of axial spondyloarthritis using electronic health records. *Rheumatol Oxf Engl* 2020; 59(5): 1059–1065.

11. van Leeuwen JR, Penne EL, Rabelink T, et al. Using an artificial intelligence tool incorporating natural language processing to identify patients with a diagnosis of ANCA-associated vasculitis in electronic health records. *Comput Biol Med* 2024; 168: 107757.

12. Krusche M, Callhoff J, Knitza J, et al. Diagnostic accuracy of a large language model in rheumatology: comparison of physician and ChatGPT-4. *Rheumatol Int* 2024; 44(2): 303–306.

13. Haase I, Xiong T, Rissmann A, et al. ChatSLE: consulting ChatGPT-4 for 100 frequently asked lupus questions. *Lancet Rheumatol* 2024; 6(4): e196–e199.

14. Madrid-García A, Rosales-Rosado Z, Freites-Nuñez D, et al. Harnessing ChatGPT and GPT-4 for evaluating the rheumatology questions of the Spanish access exam to specialized medical training. *Sci Rep* 2023; 13(1): 22129.

15. Ye C, Zweck E, Ma Z, et al. Doctor versus artificial intelligence: patient and physician evaluation of large language model responses to rheumatology patient questions in a cross-sectional study. *Arthritis Rheumatol* 2024; 76(3): 479–484.

16. Shen Y, Heacock L, Elias J, et al. ChatGPT and other large language models are double-edged swords. *Radiology* 2023; 307(2): e230163.

17. Bommasani R, Hudson DA, Adeli E, et al. On the opportunities and risks of foundation models [Internet]. arXiv, https://arxiv.org/abs/2108.07258 (2022, accessed 12 March 2025).

18. Ferber D, Wiest I, Wölflein G, et al. GPT-4 for information retrieval and comparison of medical oncology guidelines. *NEJM AI* 2024; 1(6): AIcs2300235.

19. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2020; 36(4): 1234–1240.

20. Farquhar S, Kossen J, Kuhn L, et al. Detecting hallucinations in large language models using semantic entropy. *Nature* 2024; 630(8017): 625–630.

21. Gao Y, Xiong Y, Gao X, et al. Retrieval-augmented generation for large language models: a survey [Internet]. arXiv, https://arxiv.org/abs/2312.10997 (2024, accessed 12 March 2025).

22. Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks [Internet]. arXiv, https://arxiv.org/abs/2005.11401 (2021, accessed 12 March 2025).

23. Skarlinski MD, Cox S, Laurent JM, et al. Language agents achieve superhuman synthesis of scientific knowledge [Internet]. arXiv, https://arxiv.org/abs/2409.13740 (2024, accessed 12 March 2025).

24. Wang L, Yang N, Huang X, et al. Improving text embeddings with large language models [Internet]. arXiv, https://arxiv.org/abs/2401.00368 (2024, accessed 12 March 2025).

25. Leroy G, Gu Y, Pettygrove S, et al. Automated extraction of diagnostic criteria from electronic health records for autism spectrum disorders: development, evaluation, and application. *J Med Internet Res* 2018; 20(11): e10497.

26. Rahutomo F, Kitasuka T and Aritsugi M. Semantic cosine similarity. In: *Proceedings of the 7th international conference on information science and technology (ICIST)*, University of Seoul, South Korea, 26–28 August 2012, Seoul, South Korea.

27. Douze M, Guzhva A, Deng C, et al. The Faiss library [Internet]. arXiv, https://arxiv.org/abs/2401.08281 (2024, accessed 12 March 2025).

28. Venerito V, Lalwani D, Del Vescovo S, et al. Prompt engineering: the next big skill in rheumatology research. *Int J Rheum Dis* 2024; 27(5): e15157.

29. Alkhalaf M, Yu P, Yin M, et al. Applying generative AI with retrieval augmented generation to summarize and extract key clinical information from electronic health records. *J Biomed Inform* 2024; 156: 104662.

30. Dhuliawala S, Komeili M, Xu J, et al. Chain-of-verification reduces hallucination in large language models [Internet]. arXiv, https://arxiv.org/abs/2309.11495 (2023, accessed 12 March 2025).

31. Izacard G and Grave E. Distilling knowledge from reader to retriever for question answering [Internet]. arXiv, https://arxiv.org/abs/2012.04584 (2022, accessed 12 March 2025).

32. Samsi S, Zhao D, McDonald J, et al. From words to watts: benchmarking the energy costs of large language model inference [Internet]. arXiv, https://arxiv.org/abs/2310.03003 (2023, accessed 12 March 2025).

33. Shen W, Gao Y, Huang C, et al. Retrieval-generation alignment for end-to-end task-oriented dialogue system [Internet]. arXiv, https://arxiv.org/abs/2310.08877 (2023, accessed 12 March 2025).

34. Gupta M. LLM evaluation metrics explained [Internet]. Medium, https://medium.com/data-science-in-your-pocket/llm-evaluation-metrics-explained-af14f26536d2 (2024, accessed 12 March 2025).

35. Gaur M, Gunaratna K, Srinivasan V, et al. ISEEQ: information seeking question generation using dynamic meta-information retrieval and knowledge graphs. In: *Proceedings of the AAAI conference on artificial intelligence*, Association for the Advancement of Artificial Intelligence, Washington, DC, 2022, vol. 36, pp. 10672–10680.

36. Lakatos R, Pollner P, Hajdu A, et al. Investigating the performance of retrieval-augmented generation and fine-tuning for the development of AI-driven knowledge-based systems [Internet]. arXiv, https://arxiv.org/abs/2403.09727 (2024, accessed 12 March 2025).

37. Schick T and Schütze H. It's not just size that matters: small language models are also few-shot learners [Internet]. arXiv, https://arxiv.org/abs/2009.07118 (2021, accessed 12 March 2025).

38. Acosta JN, Falcone GJ, Rajpurkar P, et al. Multimodal biomedical AI. *Nat Med* 2022; 28(9): 1773–1784.

Visit Sage journals online
journals.sagepub.com/
home/tab

Sage journals