Right Prediction, Wrong Reasoning: Uncovering LLM Misalignment in RA Disease Diagnosis

Umakanta Maharana[1], Sarthak Verma[2], Avarna Agarwal[2], Prakashini Mruthyunjaya[2], Dwarikanath Mahapatra[3], Sakir Ahmed[2], Murari Mandal[1]
[1] RespAI Lab, KIIT Bhubaneswar,
[2] KIMS Bhubaneswar, [3] Monash University, Australia
Correspondence: murari.mandalfcs@kiit.ac.in

Abstract

Large language models (LLMs) offer a promising pre-screening tool, improving early disease detection and providing enhanced healthcare access for underprivileged communities. The early diagnosis of various diseases continues to be a significant challenge in healthcare, primarily due to the nonspecific nature of early symptoms, the shortage of expert medical practitioners, and the need for prolonged clinical evaluations, all of which can delay treatment and adversely affect patient outcomes. With impressive accuracy in prediction across a range of diseases, LLMs have the potential to revolutionize clinical pre-screening and decision-making for various medical conditions. In this work, we study the diagnostic capability of LLMs for Rheumatoid Arthritis (RA) with real world patients data. Patient data was collected alongside diagnoses from medical experts, and the performance of LLMs was evaluated in comparison to expert diagnoses for RA disease prediction. We notice an interesting pattern in disease diagnosis and find an unexpected *misalignment between prediction and explanation*. We conduct a series of multi-round analyses using different LLM agents. The best-performing model accurately predicts rheumatoid arthritis (RA) diseases approximately 95% of the time. However, when medical experts evaluated the reasoning generated by the model, they found that nearly 68% of the reasoning was incorrect. This study highlights a clear misalignment between LLMs high prediction accuracy and its flawed reasoning, raising important questions about relying on LLM explanations in clinical settings. **LLMs provide incorrect reasoning to arrive at the correct answer for RA disease diagnosis.**

1 Introduction

Early disease detection plays a crucial role in improving patient outcomes, especially in areas with limited healthcare resources. Identifying diseases at early stages can help reduce the strain on healthcare systems and improve survival rates. However, early diagnosis is still a major challenge in clinical practice due to factors such as vague early symptoms, a shortage of medical professionals, and the lengthy nature of diagnostic procedures Celermajer et al. (2012); Crosby et al. (2022); Heidari (2011). These challenges can delay treatment and negatively affect patient health. In this context, pre-screening tools that assist in the diagnostic process could offer a significant improvement by enabling faster, more efficient, and accurate disease identification.

Large Language Models (LLMs) have shown great promise as pre-screening tools in recent years, with their ability to analyze vast amounts of clinical data and assist in early diagnosis Gopeekrishnan et al. (2024); Chang et al. (2024); Cohen et al. (2024); Lucas et al. (2024). Several studies have demonstrated that LLMs can perform with impressive accuracy in predicting a range of diseases, including cancer Hein et al. (2024), cardiovascular conditions Han et al. (2023), and infectious diseases Omar et al. (2024). Their ability to

quickly process and synthesize information from patient data, such as medical records and clinical notes, positions them as a potential game-changer in healthcare, especially for underserved populations with limited access to expert medical practitioners. These models can provide preliminary insights, helping to triage patients, reduce waiting times, and enhance healthcare access in areas with a shortage of medical professionals.

However, while LLMs have shown strong *predictive* performance, their *reasoning* remains a critical area of concern. In clinical applications, the accuracy of a model's predictions is essential, but so is the transparency and trustworthiness of its reasoning. *In particular, when LLMs provide incorrect explanations or justifications for their predictions, the consequences could be severe*. A correct diagnosis with flawed reasoning may lead to misinformed decisions about treatment or further testing, undermining the model's utility in clinical settings. The importance of aligning predictions with sound reasoning cannot be overstated, especially in healthcare, where errors in reasoning can have a direct impact on patient well-being. Therefore, it is crucial to assess not only the accuracy of LLM predictions but also the reliability and correctness of the explanations that underpin these predictions.

In this study, we investigate the diagnostic capabilities of LLMs for the prediction of Rheumatoid Arthritis (RA), a common autoimmune disease with complex and often subtle early symptoms. Our work explores the alignment between LLM predictions and their reasoning for RA diagnosis using real-world patient data. We compare the performance of LLMs to expert medical diagnoses, and we uncover a critical issue: despite the high prediction accuracy (approximately 95%), the reasoning behind these predictions is often flawed. When the explanations generated by the LLMs were reviewed by medical experts, nearly 68% of the reasoning tokens were found to be incorrect. This misalignment between high prediction accuracy and flawed reasoning raises significant concerns about the reliability of LLMs in clinical decision-making.

By verifying the reasoning with input from medical experts, our study reveals an important gap in the application of LLMs for disease prediction. While LLMs can accurately predict conditions like RA, their reasoning may not align with medical expertise, limiting their effectiveness as a trustworthy clinical tool. This finding emphasizes the need for further research to improve the interpretability and reliability of LLMs in healthcare settings, ensuring that they not only provide accurate predictions but also offer sound and medically valid reasoning to support those predictions.

2 Related Works

Reasoning in LLMs. Reasoning capabilities in LLMs have garnered significant attention in recent years, as researchers explore ways to enhance their ability to solve complex tasks Jaech et al. (2024); Guo et al. (2025). Early studies focused on the emergence of reasoning skills in LLMs, particularly through in-context and few-shot learning, where models demonstrated the ability to perform basic reasoning across a range of tasks Sun et al. (2023); Kojima et al. (2022); Liu et al. (2023). Additionally, methods such as reinforcement learning (RL) and Monte Carlo Tree Search (MCTS) enable iterative reasoning over extended thought chains Li et al. (2025); Wu et al. (2024); Kumar et al. (2024), while feedback mechanisms like self-verification and error correction refine the reasoning process Guo et al. (2025); Ma et al. (2025). Efforts to improve structured and cross-lingual reasoning further broaden LLMs' applicability across languages and domains Qin et al. (2023), with hybrid feedback systems ensuring real-time corrections and enhancing reasoning quality Behrouz et al. (2024).

Rheumatoid Arthritis Diagnosis. Rheumatoid arthritis (RA) affects approximately 5 out of every 1000 individuals Aletaha & Smolen (2018) and has the potential to cause significant joint damage and disability. Early diagnosis and intervention plays a critical role in managing RA. Recent advancements in Natural Language Processing (NLP) have had significant impacts on the management and research of RA Benavent & Madrid-García (2024); Benavent et al. (2023); Venerito et al. (2023). Humbert-Droz et al. (2023)created an NLP system to extract RA-related outcomes from clinical notes, which achieved high accuracy in identifying key clinical features. Similarly, AI-based tools leveraging electronic health records (EHRs) have shown success in diagnosing RA and identifying related conditions

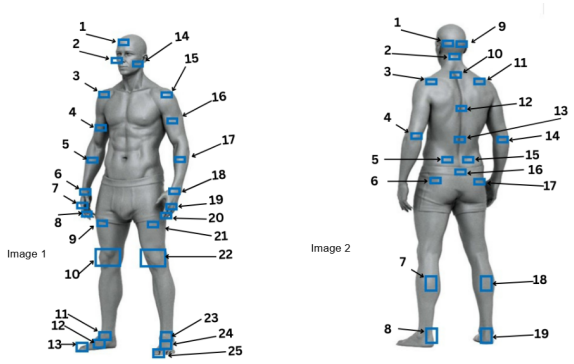

Figure 1: Diagram shown to the patient for indicating pain locations.

Category	Description
Total patients	160
Gender distribution	15% Male, 85% Female
RA diagnosis	85% RA, 15% Non-RA
Languages used	English, Odia
Data collection method	Online form

Table 1: Key statistics of the PreRAID dataset

like rheumatoid arthritis with interstitial lung disease (RA-ILD), with promising precision metrics Singhal et al. (2023); O'Neil et al. (2021); Fukae et al. (2020); Zhang et al. (2020); Bird et al. (2022); Irfan & Yaqoob (2023).

The use of LLMs, such as ChatGPT, in rheumatology is still in its early stages, with limited studies available on their effectiveness and shortcomings. Our work is the first systematic analysis of the popular LLMs (ChatGPT, Gemini, Mistral, Qwen) for diagnosing RA with real world patients data.

3 PreRAID Dataset

Data Collection and Description. We present the PreRAID (Prescreening Rheumatoid Arthritis Information Database), a structured dataset developed to assess the diagnostic capabilities of large language models (LLMs) in the context of Rheumatoid Arthritis (RA) diagnosis. The dataset comprises 160 patient records, collected via a structured online form at the Kalinga Institute of Medical Sciences (KIMS), Bhubaneswar, India (pbmh, 2025). Data gathering took place under the supervision of RA medical professionals, ensuring high-quality entries and proper patient consent throughout the process. The dataset captures a range of demographic details, including age, gender, and language (refer Table 1). Patients provided detailed descriptions of their primary complaints, with symptom onset recorded across varying time frames, ranging from days to years. Visual aids incorporated into the online form allowed patients to precisely indicate the location of pain, enhancing the accuracy of symptom reporting (refer Figure 1).

Beyond symptomatology, PreRAID includes associated symptoms such as skin rashes, fever, ocular discomfort, impact on daily activities, and medication history. The dataset also captures doctor-verified diagnoses and explanatory notes, providing a robust foundation for evaluating the reasoning behind LLM generated predictions. The dataset reflects real-world RA distribution: 85% RA cases, 15% non-RA cases, and 85% female, 15% male, a known clinical trend where RA predominantly affects women. Since we do not train an AI model but evaluate reasoning, this distribution does not impact the validity of our findings. Instead,

Patient information	Relevant fields in the online form
Demographic and Contact Details	Timestamp, Email address, First and last names, Age, Mother tongue, Gender, Mobile numbers
Unique Identifiers and Geographic Data	A unique KIMS ID for each patient Town/district information State information
Symptomatology and Disease Progression	Detailed responses on the primary problem faced by the patient Multiple entries for symptom onset (days, weeks, months, years) Comprehensive symptom checklists: pain in various body parts, early morning stiffness, joint deformities, and swelling
Visual Aids for Symptom Localization	Figure 1 for marking specific pain locations Enhances precision of symptom reporting
Additional Clinical and Lifestyle Information	Presence of other symptoms: skin rashes, fever, mouth ulcers, ocular discomfort Impact on daily activities: sleep disturbances, difficulties in rising from a chair or bed, variations in pain with physical activity or rest Use and efficacy of painkillers Previous medication history for arthritis
Follow-Up and Final Diagnosis	Self-reported prescreening data Follow-up entries comprising doctor's final diagnosis and explanatory notes

Table 2: Patient information collected through a structured online form. The details were filled by medical professionals in the presence of the patient.

PreRAID provides a controlled setting to assess how well LLMs justify their predictions, independent of dataset bias. The information collected for each patient is presented in Table 2

Dataset Preprocessing. The collected dataset is preprocessed and stored in a vector database for RAG-based analysis. ❶ *Data Structuring.* Raw patient inputs are normalized into a standardized textual format, ensuring uniform representation across all records. ❷ *Vectorization.* We embedded the structured text using GPT-4 text-embedding-3-large OpenAI (2023b), a pre-trained embedding model that captures semantic relationships within patient data. These high-dimensional vector representations support effective similarity searches and context-aware reasoning. ❸ *Storage in a Vector Database.* The resulting embeddings are stored in a vector database, forming the knowledge backbone of our framework. This setup allows retrieval of patient records, enhancing diagnostic reasoning by improving contextual awareness.

4 RA Disease Diagnosis with LLMs

4.1 Experiment Settings

We utilize the PreRAID dataset, which comprises 160 patient records, to conduct a series of experiments aimed at evaluating the diagnostic capabilities of large language models (LLMs). To examine the impact of varying knowledge base sizes on diagnostic performance, we construct knowledge bases of sizes: $\{10, 20, 30, 40, 50, 60, 70, 80, 90\}$. For each knowledge base size, we perform experiments using a range of LLMs, including GPT-3.5 Turbo, GPT-4o,

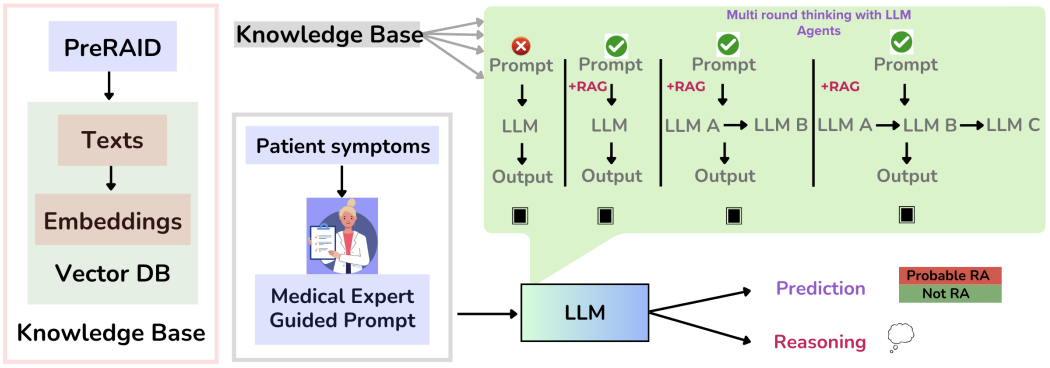

Figure 2: Overview of the framework for RA patients diagnosis. We conduct experiments with four different architectures: *LLM without knowledge base (from PreRAID), LLM+knowledge base+RAG, 2 LLM Agents+knowledge base+RAG, 3 LLM Agents+knowledge base+RAG*. We store the final diagnosis and the reasoning tokens generated by each architecture.

GPT-4o-mini, Gemini 1.5 Flash, Gemini 2.0 Flash, and QWEN 2 (Team & Google, 2023; OpenAI, 2023a; Yang et al., 2024).

For each LLM, we test four distinct architectures: ❶LLM without knowledge base ,❷ LLM + Knowledge Base (from PreRAID) + RAG (retrieval-augmented generation), ❸ 2 LLM Agents + Knowledge Base + RAG, and ❹ 3 LLM Agents + Knowledge Base + RAG. This results in a total of 24 different configurations, which are used to generate diagnostic results for RA patients.

To assess the effect of knowledge base size on performance, we incrementally expose the model to increasing amounts of historical patient data. Starting with a knowledge base of 10 records, the size is gradually increased in steps of 10, up to 90 records. For each configuration, we prompt the model to predict the diagnosis (either "Rheumatoid Arthritis" or "Not Rheumatoid Arthritis") and provide its reasoning.

4.2 Different LLM Architectures

LLM without knowledge base. In the initial phase of our experiments, we evaluate a single LLM without any access to the knowledge base, which serves as a zero-shot predictor. The prompt for this configuration directs the model to analyze patient data and output the diagnosis based solely on its internal capabilities, without any additional contextual information. This setup provides a baseline for comparing performance against models that leverage external knowledge.

Prompt → {LLM without knowledge base}

"Analyze the patient data thoroughly and then clearly state the diagnosis as 'Rheumatoid Arthritis' or 'Not Rheumatoid Arthritis'. Do not write any additional output or any patient information.\n"

LLM + knowledge base + RAG. We give access to historical diagnostic data (knowledge base) to the LLM which applies RAG and generate predictions accordingly. We use a similar prompt as before but include an instruction to incorporate the historical data.

Prompt → {LLM + Knowledge Base + RAG}

"Analyze the patient data thoroughly and then clearly state the diagnosis as 'Rheumatoid Arthritis' or 'Not Rheumatoid Arthritis'. Do not write any additional output or any patient information. \n.Use the provided data as historical diagnostic data : **"historical data"** \n"

2 LLM Agents + knowledge base + RAG. We introduce multi round thinking Tian et al. (2025) by using two LLM agents. The architecture uses a two-round reasoning approach. In the first round, the agent analyzes the patient symptoms by comparing them with the historical data, and in the second round, it re-evaluates its initial response to produce a final diagnosis.

Prompt → {2 LLM Agents + Knowledge Base + RAG}

LLM Agent A: "Extract the patient symptoms from the provided user data patient symptoms. Extract the related symptoms and corresponding diagnosis from the knowledge base **"historical data"**. Compare them and provide diagnosis."

LLM Agent B: "Analyze this comparison response of first agent and then clearly state the diagnosis as 'Rheumatoid Arthritis' or 'Not Rheumatoid Arthritis'. Do not write any additional output or any patient information. \n"

3 LLM Agents + Knowledge Base + RAG. We create 3 LLM agents to perform three rounds of reasoning. After an initial analysis and a subsequent verification of the first response, it conducts a final review before arriving at the diagnosis.

Prompt → {3 LLM Agents + Knowledge Base + RAG}

LLM Agent A: "Extract the patient symptoms from the provided user data patient symptoms. Extract the related symptoms and corresponding diagnosis from the knowledge base **"historical data"**. Compare them and provide diagnosis."

LLM Agent B: "Response from first agent Review this response is correct or incorrect and provide feedback by comparing it with the knowledge base **"historical data"**.

LLM Agent C: "Analyze this feedback feedback and then clearly state the diagnosis as 'Rheumatoid Arthritis' or 'Not Rheumatoid Arthritis'. Do not write any additional output or any patient information. \n"

The experiments are conducted across all combinations: nine different knowledge base splits (ranging from 10 to 90 records), 6 LLM models, and 4 different architectural configurations. This setup enables a comprehensive evaluation of how both the amount of historical data and the iterative reasoning process affect the diagnostic accuracy of the predictions and the quality of the generated reasoning.

4.3 Results & Discussion

Diagnostic accuracy analysis. We have tried different combinations and most of the models are actually performing similar and providing high accuracy in disease prediction. We have tested 60 patients in all possible combinations. Figure 3 shows the prediction accuracy of all the models in different frameworks. We can see the orange color bars represent the accuracies of all the models in llm without knowledge base framework. This framework shows the highest accuracy as 90% in GPT-4o and GPT-4o-mini models. And lowest in QWEN2. We can see the sky color bars represent the single agent with rag framework. It shows highest accuracy in GPT-4o model and lowest accuracy in Gemini-1.5-flash model. Gemini 2.0 flash shows 78%, GPT-4o-mini 76%, gpt3.5 turbo 85% and qwen2 78% accurate. We can see the deep blue color bars represent two agent framework with rag framework. It

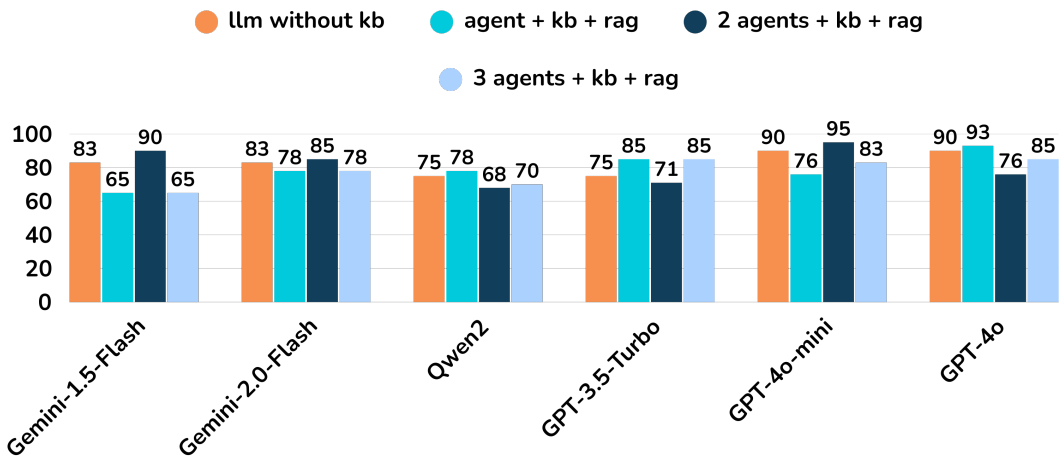

Figure 3: Accuracy of disease prediction across all the models by providing all the knowledge bases. Orange, sky blue, deep blue and light blue color represents the accuracy in llm without RAG, single agent+RAG, 2 agents+RAG, and 3 agents+RAG, respectively.

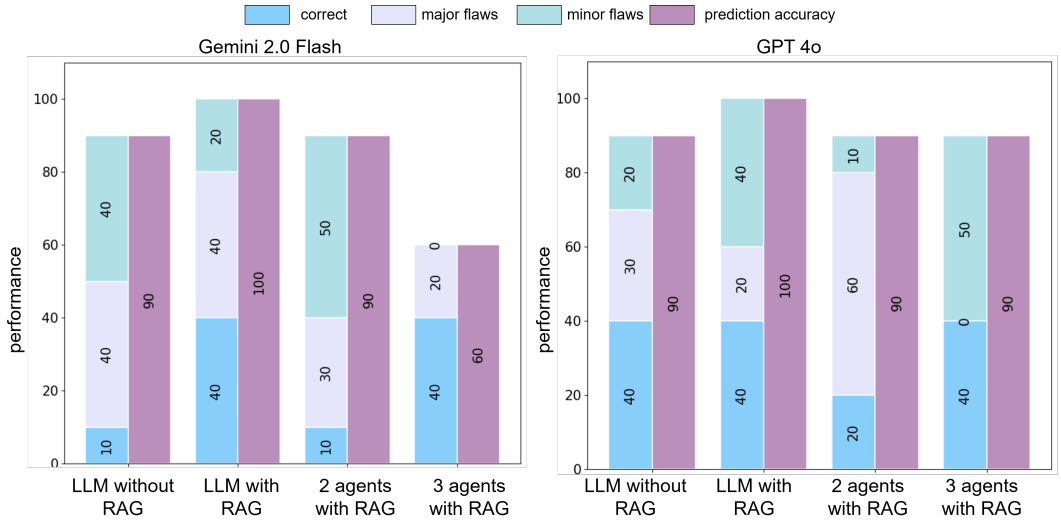

Figure 4: Comparison of diagnostic accuracy and expert-rated reasoning ratings for Gemini 2.0 Flash and GPT-4o across four architectures: LLM without RAG, LLM+RAG, 2 LLM agents+RAG, and 3 LLM agents+RAG

shows the highest accuracy in GPT4omini model and the lowest accuracy in qwen2 model. gemini1.5 flash shows 90%, gemini2.0 flash shows 85%, gpt3.5turbo models shows 71%, gpt 4o shows 76%. We can see that this framework with gpt 4o mini models shows the highest accuracy across all the combinations. We can see the light blue color represents three agents with rag framework. It shows highest accuracy in gpt 3.5 turbo and gpt 4o model and the lowest accuracy in gemini 1.5 flash model. gemini 2.0 flash shows 78%, qwen2 shows 70%, gpt4o mini shows 83%.

Reasoning validation data collected with RA experts. Our experiments reveal that while all LLM-based architectures achieve over 90% accuracy in diagnosing Rheumatoid Arthritis (RA), their underlying reasoning is not consistently aligned with expert clinical judgment. To investigate this discrepancy, we randomly selected 10 test patients: 5 diagnosed as RA and 5 as non-RA by ground truth and had domain experts evaluate the explanations

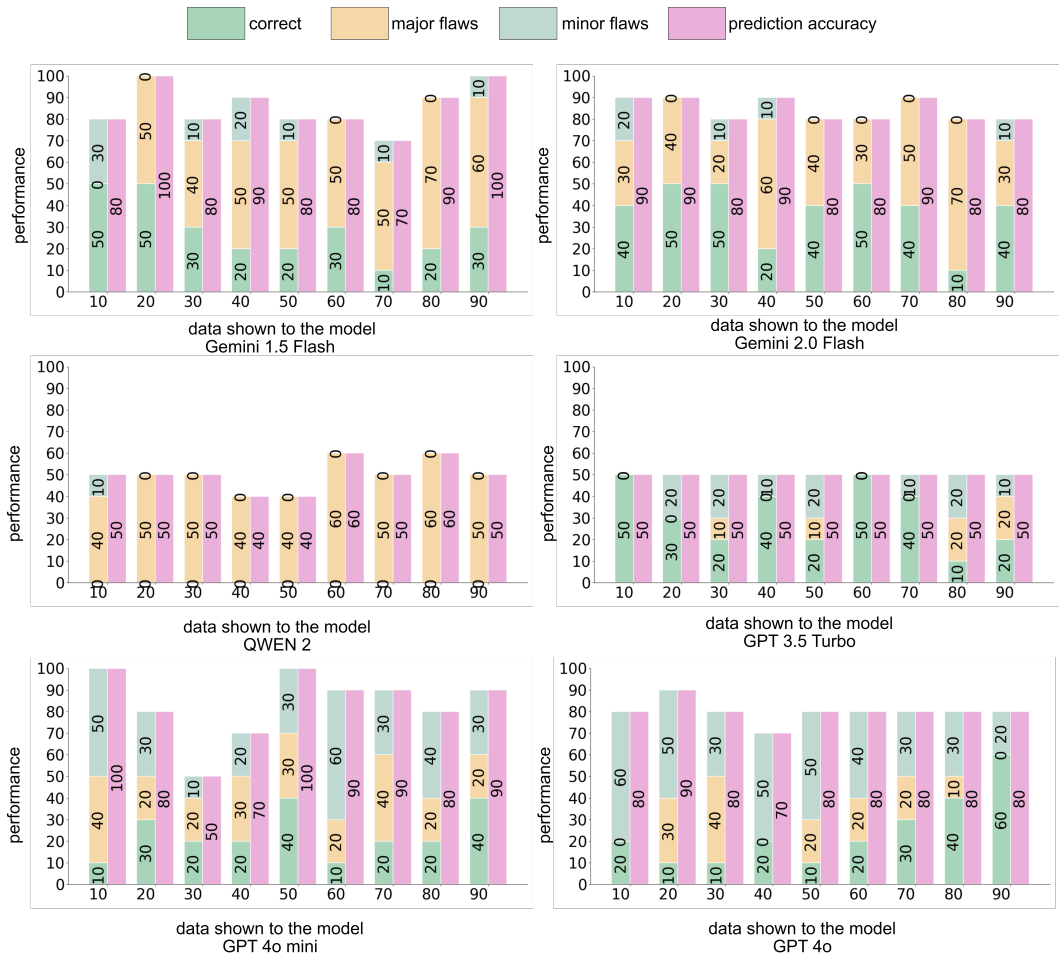

Figure 5: Reasoning performance across different LLM models in the 2 LLM agents+RAG setup. This figure compares the reasoning quality of various LLM models: GPT-3.5 Turbo, GPT-4o, GPT-4o-mini, Gemini 1.5 Flash, Gemini 2.0 Flash, and QWEN-2-7B, within the 2 LLM agents+RAG framework.

provided by the LLMs for their diagnostic decisions. The RA experts classified the LLM reasoning into three categories: "correct", "major flaw", and "minor flaw". A "correct" rating indicates that the LLM's reasoning aligns with that of human experts, while a "major flaw" suggests that the model's explanation is entirely unsound despite reaching the right diagnosis. A "minor flaw' denotes partially accurate reasoning that still contains several errors.

Contrary to the assumption that high diagnostic accuracy implies sound reasoning, our findings expose a significant misalignment: LLMs often arrive at correct predictions through flawed or clinically unsound reasoning. This **"Right Prediction, Wrong Reasoning"** phenomenon raises critical questions about the interpretability and trustworthiness of LLMs in high-stakes medical applications. We present a comparative analysis of diagnostic accuracy and reasoning quality, underscoring the need for further research into improving the explanatory mechanisms of LLMs for reliable clinical deployment.

Reasoning analysis. We conduct two types of analysis to compare the diagnostic accuracy and corresponding reasoning generated by the LLMs.

Across four distinct architectures with Gemini 2.0 Flash and GPT4o. To assess the quality of LLM reasoning in RA diagnosis, we conducted experiments using the Gemini 2.0 Flash and GPT4o models with four distinct architectures: LLM without RAG, LLM+RAG, 2 LLM agents+RAG, and 3 LLM agents+RAG. We compared the diagnostic accuracy of each architecture against the reasoning ratings provided by RA experts (see Figure 4). For Gemini 2.0 Flash, the prediction accuracies across the architectures were 90%, 100%, 90%, and 60%, while the percentage of instances rated as "correct" reasoning was 10%, 40%, 10%, and 40%, respectively. The gap in alignment is 80%, 60%, 80%, and 20%. Even when combining "correct" and "minor flaw" ratings, the gap between diagnostic accuracy and reasoning quality remained substantial: 40%, 40%, 30%, and 20%. This pronounced misalignment underscores a critical concern, particularly given the high-stakes nature of clinical applications.

Similarly, for GPT 4o, the prediction accuracy for different architectures are: 90%, 100%, 90%, 90% with corresponding "correct" reasoning as 40%, 40%, 20%, 40%, respectively. The gap in alignment is 50%, 60%, 70%, and 50%. Although, our sample size is small right now (total 10 test patients), the GPT 4o seems to perform slightly better than Gemini 2.0 Flash.This however is still far from the expected alignment between diagnosis accuracy and reasoning.

Across different LLM models in 2 LLM agents+RAG setup. We employed the 2 LLM agents+RAG architecture while varying the backend LLM models among GPT-3.5 Turbo, GPT-4o, GPT-4o-mini, Gemini 1.5 Flash, Gemini 2.0 Flash, and QWEN-2-7B (Team & Google, 2023; OpenAI, 2023a; Yang et al., 2024) (see Figure 5). We observed a consistent pattern across models i.e., all models fail in proving correct reasoning for the high predictive accuracy for RA diagnosis. Notably, QWEN-2-7B failed to provide any correct reasoning, likely due to its significantly smaller model size. In contrast, all other models, which are proprietary and considerably larger, demonstrated better reasoning performance.

4.4 Final Takeaway

These are the final takeaways of our results: **High Diagnostic Accuracy:** Across multiple architectures and models, LLMs consistently achieve over 90% accuracy in predicting Rheumatoid Arthritis (RA), demonstrating strong potential as diagnostic tools.

Misalignment in Reasoning: Despite the high prediction accuracy, there is a significant gap between the diagnostic outcomes and the quality of the generated reasoning. Many LLMs arrive at correct predictions through explanations that are either partially or completely unsound, a phenomenon we term "Right Prediction, Wrong Reasoning."

Architectural and Model Variations: In experiments with Gemini 2.0 Flash and GPT-4o across four architectures, both models show high diagnostic performance, yet the reasoning quality remains inconsistent, with notable gaps (ranging from 20% to 80%) between correct predictions and valid reasoning. When comparing different LLM models under the 2 LLM agents+RAG framework, larger proprietary models (e.g., GPT-4o, GPT-3.5 Turbo) perform better in reasoning than smaller models like QWEN-2-7B, which fails to provide any correct reasoning.

Implications for Clinical Use: The pronounced misalignment between high prediction accuracy and flawed reasoning raises concerns about the interpretability and reliability of LLMs in clinical settings. For safe and effective deployment, it is crucial to improve the explanatory mechanisms so that model reasoning aligns closely with expert clinical judgment.

Overall, while LLMs show promise in early RA diagnosis, enhancing their reasoning capabilities is essential to ensure trustworthiness and clinical utility.

5 Conclusion

This study investigates the diagnostic performance of Large Language Models (LLMs) for Rheumatoid Arthritis (RA) using real-world patient data. While LLMs demonstrated high prediction accuracy (95%) in diagnosing RA, their reasoning was often flawed, with nearly

68% of explanations deemed incorrect by medical experts. This misalignment between accurate predictions and incorrect reasoning raises concerns about the reliability of LLMs in clinical decision-making. Although LLMs can assist in disease detection, their explanations must be improved to ensure clinical validity and patient safety. Our findings underscore the need for further research to enhance the interpretability and reasoning capabilities of LLMs, ensuring that they provide not only accurate predictions but also reliable, medically sound justifications for their diagnoses.

References

Daniel Aletaha and Josef S Smolen. Diagnosis and management of rheumatoid arthritis: a review. *Jama*, 320(13):1360–1372, 2018.

Ali Behrouz, Peilin Zhong, and Vahab Mirrokni. Titans: Learning to memorize at test time. *arXiv preprint arXiv:2501.00663*, 2024.

Diego Benavent and Alfredo Madrid-García. Large language models and rheumatology: are we there yet? *Rheumatology Advances in Practice*, pp. rkae119, 2024.

Diego Benavent, Santiago Muñoz-Fernández, Isabel De la Morena, Antonio Fernández-Nebro, Judith Marín-Corral, Eva Castillo Rosa, Miren Taberna, Cristina Sanabra, Carlos Sastre, and SAVANA Research Group. Using natural language processing to explore characteristics and management of patients with axial spondyloarthritis and psoriatic arthritis treated under real-world conditions in spain: Spainet study. *Therapeutic Advances in Musculoskeletal Disease*, 15:1759720X231220818, 2023.

Alix Bird, Lauren Oakden-Rayner, Christopher McMaster, Luke A. Smith, Minyan Zeng, Mihir D. Wechalekar, Shonket Ray, Susanna Proudman, and Lyle J. Palmer. Artificial intelligence and the future of radiographic scoring in rheumatoid arthritis: a viewpoint. *Arthritis Research & Therapy*, 24(1):268, 2022. doi: 10.1186/s13075-022-02972-x.

David S Celermajer, Clara K Chow, Eloi Marijon, Nicholas M Anstey, and Kam S Woo. Cardiovascular disease in the developing world: prevalences, patterns, and the potential of early disease detection. *Journal of the American College of Cardiology*, 60(14):1207–1216, 2012.

Ying Chang, Jian-ming Yin, Jian-min Li, Chang Liu, Ling-yong Cao, and Shu-yuan Lin. Applications and future prospects of medical llms: A survey based on the m-kat conceptual framework. *Journal of Medical Systems*, 48(1):1–18, 2024.

Aaron M Cohen, Jolie Kaner, Ryan Miller, Jeffrey W Kopesky, and William Hersh. Automatically pre-screening patients for the rare disease aromatic l-amino acid decarboxylase deficiency using knowledge engineering, natural language processing, and machine learning on a large ehr population. *Journal of the American Medical Informatics Association*, 31(3):692–704, 2024.

David Crosby, Sangeeta Bhatia, Kevin M Brindle, Lisa M Coussens, Caroline Dive, Mark Emberton, Sadik Esener, Rebecca C Fitzgerald, Sanjiv S Gambhir, Peter Kuhn, et al. Early detection of cancer. *Science*, 375(6586):eaay9040, 2022.

Jun Fukae, Masato Isobe, Toshiyuki Hattori, Yuichiro Fujieda, Michihiro Kono, Nobuya Abe, Akemi Kitano, Akihiro Narita, Mihoko Henmi, Fumihiko Sakamoto, et al. Convolutional neural network for classification of two-dimensional array images generated from clinical information may support diagnosis of rheumatoid arthritis. *Scientific reports*, 10(1):5648, 2020.

Anand Gopeekrishnan, Shibbir Ahmed Arif, and Hao Liu. Accelerating patient screening for clinical trials using large language model prompting. In *2024 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, pp. 214–215. IEEE, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Changho Han, Dong Won Kim, Songsoo Kim, Seng Chan You, SungA Bae, and Dukyong Yoon. Large-language-model-based 10-year risk prediction of cardiovascular disease: insight from the uk biobank data. *medRxiv*, pp. 2023–05, 2023.

Behzad Heidari. Rheumatoid arthritis: Early diagnosis and treatment outcomes. *Caspian journal of internal medicine*, 2(1):161, 2011.

David Hein, Alana Christie, Hua Zhong, Ellen Araj, James Brugarolas, Lindsay Cowell, Payal Kapur, and Andrew Jamieson. Learning llama agents for medical record analysis and standardization. *Cancer Research*, 84(6_Supplement):7390–7390, 2024.

Marie Humbert-Droz, Zara Izadi, Gabriela Schmajuk, Milena Gianfrancesco, Matthew C Baker, Jinoos Yazdany, and Suzanne Tamang. Development of a natural language processing system for extracting rheumatoid arthritis outcomes from clinical notes using the national rheumatology informatics system for effectiveness registry. *Arthritis Care & Research*, 75(3):608–615, 2023.

Bilal Irfan and Aneela Yaqoob. Chatgpt's epoch in rheumatological diagnostics: A critical assessment in the context of sjögren's syndrome. *Cureus*, 15(10):e47754, 2023. doi: 10.7759/cureus.47754. URL https://doi.org/10.7759/cureus.47754.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, et al. Training language models to self-correct via reinforcement learning. *arXiv preprint arXiv:2409.12917*, 2024.

Peiji Li, Kai Lv, Yunfan Shao, Yichuan Ma, Linyang Li, Xiaoqing Zheng, Xipeng Qiu, and Qipeng Guo. Fastmcts: A simple sampling strategy for data synthesis. *arXiv preprint arXiv:2502.11476*, 2025.

Xin Liu, Daniel McDuff, Geza Kovacs, Isaac Galatzer-Levy, Jacob Sunshine, Jiening Zhan, Ming-Zher Poh, Shun Liao, Paolo Di Achille, and Shwetak Patel. Large language models are few-shot health learners. *arXiv preprint arXiv:2305.15525*, 2023.

Mary M Lucas, Justin Yang, Jon K Pomeroy, and Christopher C Yang. Reasoning with large language models for medical question answering. *Journal of the American Medical Informatics Association*, 31(9):1964–1975, 2024.

Ruotian Ma, Peisong Wang, Cheng Liu, Xingyan Liu, Jiaqi Chen, Bang Zhang, Xin Zhou, Nan Du, and Jia Li. Teaching llms to self-verify and self-correct via reinforcement learning. *arXiv preprint arXiv:2502.12853*, 2025.

Mahmud Omar, Dana Brin, Benjamin Glicksberg, and Eyal Klang. Utilizing natural language processing and large language models in the diagnosis and prediction of infectious diseases: A systematic review. *American Journal of Infection Control*, 2024.

OpenAI. Chatgpt, 2023a. URL https://chat.openai.com.

OpenAI. Gpt-4 text embeddings, 2023b. URL https://openai.com/index/new-embedding-models-and-api-updates/.

Liam J O'Neil, Victor Spicer, Irene Smolik, Xiaobo Meng, Rishi R Goel, Vidyanand Anaparti, John Wilkins, and Hani S El-Gabalawy. Association of a serum protein signature with rheumatoid arthritis development. *Arthritis & Rheumatology*, 73(1):78–88, 2021.

pbmh. Pradyumna Bal Memorial Hospital,KIMS,KIIT — pbmh.ac.in. https://pbmh.ac.in/, 2025. [Accessed 27-03-2025].

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. *arXiv preprint arXiv:2310.14799*, 2023.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

Jiankai Sun, Chuanyang Zheng, Enze Xie, Zhengying Liu, Ruihang Chu, Jianing Qiu, Jiaqi Xu, Mingyu Ding, Hongyang Li, Mengzhe Geng, et al. A survey of reasoning with foundation models. *arXiv preprint arXiv:2312.11562*, 2023.

Gemini Team and Google. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Xiaoyu Tian, Sitong Zhao, Haotian Wang, Shuaiting Chen, Yunjie Ji, Yiping Peng, Han Zhao, and Xiangang Li. Think twice: Enhancing llm reasoning by scaling multi-round test-time thinking. *arXiv preprint arXiv:2503.19855*, 2025.

Vincenzo Venerito, Emre Bilgin, Florenzo Iannone, and Sedat Kiraz. Ai am a rheumatologist: a practical primer to large language models for rheumatologists. *Rheumatology*, 62(10): 3256–3260, 2023.

Siwei Wu, Zhongyuan Peng, Xinrun Du, Tuney Zheng, Minghao Liu, Jialong Wu, Jiachen Ma, Yizhi Li, Jian Yang, Wangchunshu Zhou, et al. A comparative study on reasoning patterns of openai's o1 model. *arXiv preprint arXiv:2410.13639*, 2024.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024. URL https://arxiv.org/abs/2407.10671.

Yuan Zhang, Yongming Liang, Limei Feng, and Liyan Cui. Diagnostic performance of 14-3-3η and anti-carbamylated protein antibodies in rheumatoid arthritis in han population of northern china. *Clinica Chimica Acta*, 502:102–110, 2020.

A Appendix

Body Part Mapping to Text. In Table 3, we describe the textual references for the body part selected by the patients, indicating pain locations in Figure 1.

Table 3: Body part mapping to corresponding text for Figure 1

	Front Side		Back Side
1	Forehead	**1**	Suboccipital region
2	Right TMJ joint	**2**	Posterior cervical region
3	Right shoulder	**3**	Left shoulder (post)
4	Right upper arm	**4**	Left elbow
5	Right forearm	**5**	Lower back (left PSIS)
6	Right wrist	**6**	Left hip buttock
7	Right MCP (Metacarpal phalangeal joint)	**7**	Left calf
8	Right PIP (Proximal interphalangeal joint)	**8**	Left heel
9	Right thigh	**9**	Suboccipital region
10	Right knee	**10**	Upper back
11	Right ankle	**11**	Right shoulder (post)
12	Right midfoot	**12**	Middle back
13	Right toes	**13**	Lower midback
14	Left TMJ joint	**14**	Right elbow
15	Left shoulder	**15**	Right PSIS
16	Left upper arm	**16**	Low back
17	Left forearm	**17**	Right hip buttock
18	Left writst	**18**	Right calf
19	Right MCP (Metacarpal phalangeal joint)	**19**	Right heel
20	Right PIP (Proximal interphalangeal joint)		
21	Right thigh		
22	Right knee		

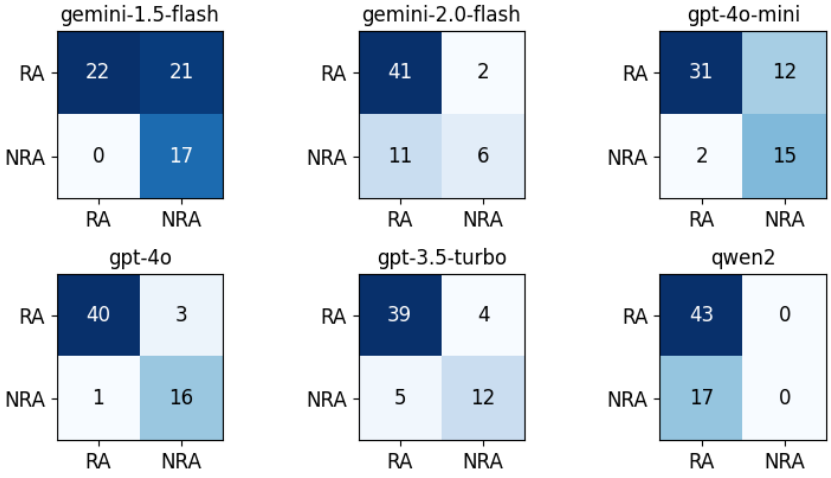

Figure 6: Confusion matrices for RA diagnosis in LLM without RAG

Confusion Matrix Analysis. We provide a comprehensive overview of the confusion matrices that assess the predictive performance of various LLM configurations in diagnosing rheumatoid arthritis (RA). These matrices capture the models' capabilities in accurately distinguishing between RA and non-RA (NRA) cases, although they do not reflect the soundness of the models' reasoning processes. In the single LLM without RAG configuration (Figure 6), high-capacity models such as GPT-4o and GPT-3.5 Turbo demonstrate better

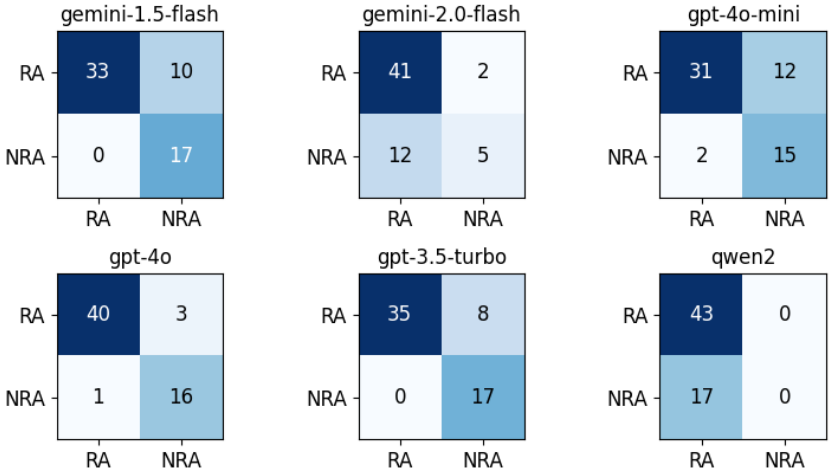

Figure 7: Confusion matrices for RA diagnosis in LLM+RAG

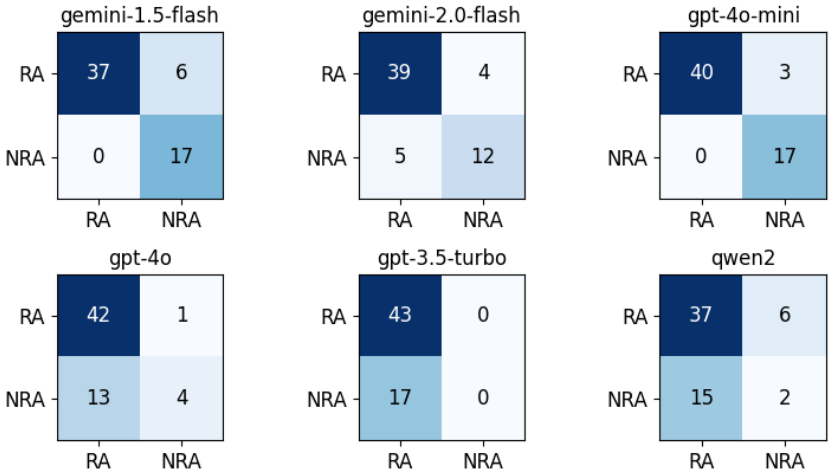

Figure 8: Confusion matrices for RA diagnosis in 2 LLM agents+RAG

performance in predicting RA cases, while smaller models like QWEN-2-7B exhibit noticeable difficulties in classification. The LLM+RAG configuration (Figure 7) shows marked improvements, with models like Gemini-2.0 Flash significantly reducing misclassifications.

The 2 LLM agent+RAG (Figure 8) further enhances diagnostic accuracy, particularly for GPT-4o and GPT-3.5 Turbo, which now register minimal false classifications. However, some models, notably QWEN-2-7B, continue to face challenges in effectively discriminating between RA and NRA cases. In the 2 LLM agent+RAG configuration (Figure 9), most models, including GPT-4o and Gemini-2.0 Flash, attain near-optimal classification performance, underscoring the benefits of multi-agent collaboration in disease prediction.

It is crucial to emphasize that these confusion matrices exclusively quantify diagnostic accuracy. They do not assess the validity or depth of the models' underlying reasoning, which may remain imperfect even when classification metrics are high, as evidenced by expert evaluations.

Examples of Flawed Reasoning Despite Correct Predictions. As we show in the results, in several instances, the LLMs produce correct diagnostic outcomes yet demonstrated problematic reasoning in their justifications. For example, a model might accurately classify

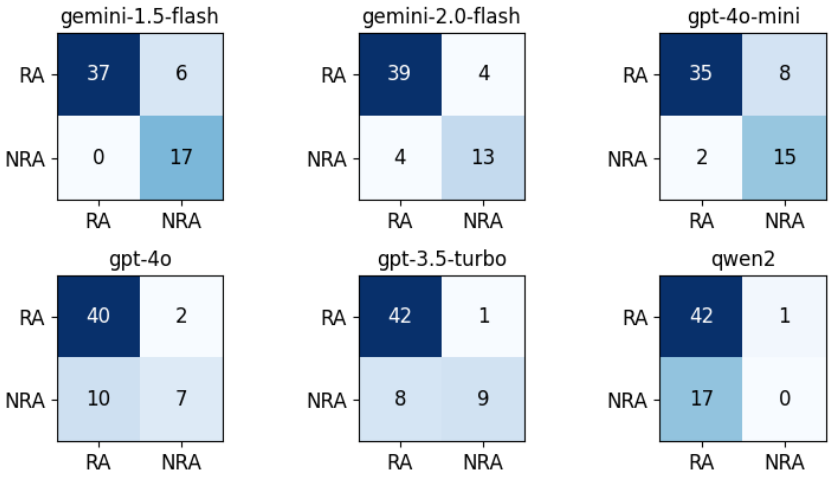

Figure 9: Confusion matrices for RA diagnosis in 3 LLM agents+RAG

a case as rheumatoid arthritis (RA) while basing its decision on peripheral or outdated clinical indicators rather than the core pathophysiological evidence. In other instances, correct answers were supported by reasoning that relied on superficial statistical correlations or misinterpreted clinical data. These examples underscore the critical need to not only improve predictive accuracy but also to enhance the depth and reliability of the underlying reasoning processes in model-based diagnostics. Detailed examples of such cases are provided below.